A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data

AUTHORS

Hamel Patel [1,2], Richard J.B Dobson [1,2,3,4,5], Stephen J Newhouse [1,2,3,4,5]

Affiliations

1. Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
2. NIHR BioResource Centre Maudsley, NIHR Maudsley Biomedical Research Centre (BRC) at South London and Maudsley NHS Foundation Trust (SLaM) & Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London.
3. Health Data Research UK London, University College London, 222 Euston Road, London NW1 2DA, UK
4. Institute of Health Informatics, University College London, 222 Euston Road, London NW1 2DA
5. The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, 222 Euston Road, London NW1 2DA, UK

ABSTRACT

Background

Microarray technologies have identified imbalances in the expression of specific genes and biological pathways in Alzheimer's disease (AD) brains. However, there is a lack of reproducibility across individual AD studies, and many related neurogenerative and mental health disorders exhibit similar perturbations. We are yet to identify robust transcriptomic changes specific to AD brains. This study meta-analysed publicly available brain-related disorders along with healthy cognitive individuals to decipher common and AD-specific brain transcriptomic changes.

Methods and Results

Twenty-two AD, eight Schizophrenia, five Bipolar Disorder, four Huntington's disease, two Major Depressive Disorder and one Parkinson's disease dataset totalling 2667 samples and mapping to four different brain regions (Temporal lobe, Frontal lobe, Parietal lobe and Cerebellum) were analysed. Differential expression analysis was performed independently in each dataset, followed by meta-analysis using a combining p-value method known as Adaptively Weighted with One-sided Correction. This identified 323, 435, 1023 and 828 differentially expressed genes specific to the AD temporal lobe, frontal lobe, parietal lobe and cerebellum brain regions respectively. Seven of these genes were consistently perturbed across all AD brain regions and are considered disease-specific, while twenty-two genes are perturbed specifically in AD brain regions affected by both plaques and tangles, suggesting involvement in AD neuropathology. Biological pathways involved in the "metabolism of proteins" and viral components were significantly enriched across AD brains.

Conclusion

We identify specific transcriptomic changes in AD brains which could make a significant contribution towards the understanding of AD disease mechanisms and provides new therapeutic targets.

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia affecting over 44 million individuals worldwide, and numbers are expected to triple by 2050 [1]. The hallmark of the disease is characterised by the abnormal brain accumulation of amyloid-β (Aβ) protein and hyperphosphorylated tau filaments, which forms structures known as plaques and tangles respectively. The accumulation of these proteins contributes to the loss of connections between neurone synapses, leading to the loss of brain tissue and disruption of normal cognitive functions.

As AD progresses, the spread of plaques and tangles in the brain usually occurs in a predictable pattern and can begin up to 18 years prior to the onset of clinical symptoms [2]. In the earliest stages of the disease, plaques and tangles form in areas of the brain primarily involved in learning and memory, specifically the hippocampus and entorhinal cortex, both situated in the temporal lobe (TL) region [3]. Next, the frontal lobe (FL), a region involved in voluntary movement, is affected, followed by the parietal lobe (PL), a region involved in processing reading and writing. In the later stage of the disease, the occipital lobe, a region involved in processing information from the eyes, can become affected, followed by the cerebellum (CB), a region which receives information from the sensory systems and the spinal cord to regulates motor movement. Nerve cell death, tissue loss and atrophy occur throughout the brain as AD progresses, leading to the manifestation of clinical symptoms associated with loss of normal brain function. However, not all brain regions are neuropathologically affected in the same manner. The CB, which only accounts for 10% of the brain but contains over 50% of the brains total neurones, is often neglected in AD research because it is generally considered to be partially spared from the disease as plaques are only occasionally seen but tangles are generally not reported [4] [5].

The histopathological spread of the disease is well documented, and with the advent of high throughput genomics approaches, we are now able to study the transcriptomic and biological pathways disrupted in AD brains. Microarrays can simultaneously examine thousands of genes,

providing an opportunity to identify imbalances in the expression of specific genes and biological pathways. However, microarray reproducibility has always been questionable, with replication of differentially expressed genes (DEG's) very poor [6]. For example, two independent microarray transcriptomic studies performed differential expression analysis in the hippocampus of AD brains. The first study by Miller et al. identified 600 DEG's [7], and a similar study by Hokama et al. identified 1071 DEG's [8]. An overlap of 105 DEG's exist between the two studies; however, after accounting for multiple testing, no gene was replicated between the two studies. The Miller study consisted of 7 AD and 10 control subjects expression profiled on the Affymetrix platform while the Hakoma study consisted of 31 AD and 32 control subjects expression profiled on the Illumina platform. Replication between the Illumina and Affymetrix platform has been shown to be generally very high [9]; therefore, the lack of replication between the two studies is probably down to a range of other factors including low statistical power, sampling bias and disease heterogeneity.

Unlike DEG's, replication of the molecular changes at a pathway level are more consistent and have provided insights into the biological processes disturbed in AD. Numerous studies have consistently highlighted disruptions in immune response [10] [11] [12] [13], protein transcription/translation [10] [11] [14] [15] [16] [17], calcium signalling [10] [18] [19], MAPK signalling [16] [7], various metabolism pathways such as carbohydrates [16], lipids [16] [20], glucose [21] [22] [17], and iron [11] [23], chemical synapse [18] [7] [19] and neurotransmitter pathways [11] [18] [19]. However, many of these pathways have also been suggested to be disrupted in other brain-related disorders. For example, disruptions in calcium signalling, MAPK, chemical synapse and various neurotransmitter pathways have also been implicated in Parkinsons's Disease (PD) [24] [25]. In addition, glucose metabolism, protein translation, and various neurotransmission pathways have also been suggested to be disrupted in Bipolar Disorder (BD) [26] [27] [28] [29]. Although the biological disruptions involved in AD are steadily being identified, many other neurodegenerative and mental disorders are

showing similar perturbations. We are yet to identify robust transcriptomic changes specific to AD brains.

In this study, we combined publicly available microarray gene expression data generated from AD human brain tissue and matched cognitively healthy controls to conduct the most extensive AD transcriptomic microarray meta-analyses known to date. We generate AD expression profiles across the temporal lobe, frontal lobe, parietal lobe and cerebellum brain regions. We further refine each expression profile by removing perturbations seen in other neurodegenerative and mental disorders (PD, BD, Schizophrenia [SCZ], Major Depressive Disorder [MDD] and Huntington's Disease [HD]) to decipher specific transcriptomic changes occurring in human AD brains. These AD-specific brain changes may provide new insight and a better understanding of the disease mechanism, which in turn could provide new therapeutic targets for preventing and curing AD.

MATERIALS AND METHODS

Selection of publicly available microarray studies

Publicly available microarray gene expression data was sourced from the Accelerating Medicines Partnership-Alzheimer's Disease AMP-AD (doi:10.7303/syn2580853, doi:10.1038/ng.305, doi:10.1371/journal.pgen.1002707, doi:10.1038/ng.305, doi:10.1038/sdata.2016.89, doi:10.1038/sdata.2018.185) and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) in June 2016. For a study to be selected for inclusion, the data had to (1) be generated from a neurogenerative or mental health disorder, (2) be sampled from human brain tissue, (3) have gene expression measured on either the Affymetrix or Illumina microarray platform, (4) contain both diseased and suitably matched healthy controls in the same experimental batch and (5) contain at least 10 samples from both the diseased and control group.

Microarray gene expression data pre-processing

Data analysis was performed in RStudio (version 0.99.467) using R (version 3.2.2). All data analysis scripts used in this study are available at https://doi.org/10.5281/zenodo.823256. In brief, raw Affymetrix microarray gene expression data was "mas5" background corrected using R package "affy" (version 1.42.3) and raw Illumina microarray gene expression data Maximum Likelihood Estimation (MLE) background corrected using R package "MBCB" (version 1.18.0). Studies with samples extracted from multiple tissues were separated into tissue-specific matrices, log2 transformed and then Robust Spline Normalised (RSN) using R package "lumi" (version 2.16.0).

BRAAK staging is a measure of AD pathology and ranges from I-VI. In general, stages I-II, III-IV and V-VI represent the "low likelihood of AD", "probable AD" and "definite AD" respectively [30]. To maintain homogeneity within the sample groups and to be able to infer pathological related genetic changes, if BRAAK staging was available, clinical AD samples with BRAAK scores ≤ 3 or clinical control samples with BRAAK scores ≥ 3 were removed from further analysis.

Gender was predicted using the R package "massiR" (version 1.0.1) and used to subset the data into four groups based on diagnosis (case/control) and gender (male/female). Next, probes below the 90th percentile of the log2 expression scale in over 80% of samples were deemed "not reliably detected" and were excluded from further analysis to eliminate noise [31] and increase power [32].

Publicly available data is often accompanied by a lack of sample processing information, making it impossible to adjust for known systematic errors introduced when samples are processed in multiple batches, a term often known as "batch effects". To account for both known and latent variation, batch effects were estimated and removed using the Principal Component Analysis (PCA) and Surrogate Variable Analysis (SVA) using the R package "sva" (version 3.10.0). Gender and diagnosis information were used as covariates in sva when correcting for batch effects. Outlying samples were iteratively identified and removed from each gender and diagnosis group using fundamental network concepts described in [33]. Platform-specific probe ID's were converted to Entrez Gene ID's

using the BeadArray corresponding R annotation files ("hgu133plus2.db", "hgu133a.db", "hgu133b.db", "hugene10sttranscriptcluster.db", "illuminaHumanv4.db", "illuminaHumanv3.db") and differential expression analysis was performed using R package "limma" (version 3.20.9).

Finally, study compatibility analysis was investigated through the R package "MetaOmics" (version 0.1.13). This package uses differentially expressed genes (DEGs), co-expression and enriched biological pathways analysis to generate six quantified measures that are used to generate a PCA plot. The direction of each Quality Control (QC) measure is juxtaposed on top of the two-dimensional PC subspace using arrows. Datasets in the negative region of the arrows were classed as outliers [34] and were removed from further analysis.

Meta-analysis

Datasets were grouped by the primary cerebral cortex lobes (TL, FL, PL) and the CB. Meta-analysis was performed using a "combining p-values" method known as "Adaptively Weighted with One-sided Correction" (AW.OC), implemented through the R package "MetaDE" (version 1.0.5)[34]. This method was chosen as it permits missing information across datasets which are introduced by combining data generated from different microarray platforms and expression chips. This avoids the need to subset individual datasets to common probes, which essentially allows for the maximum number of genes to be analysed. Furthermore, the method provides additional information on which dataset is contributing towards the meta-analysis p-value, and has been shown to be amongst the best performing meta-analysis methods for combining p-values for biological associations [35]. A gene was deemed significantly differentially expressed (DE) if the FDR adjusted meta p-value ≤ 0.05.

If a gene was significantly differentially expressed (DE) according to the meta-analysis, but at least one contributing datasets (according to AW.OC weights) had directional logFC discrepancies (i.e. up-regulated in one dataset and down-regulated in another dataset), the gene was deemed to be discordant and was excluded from further analysis. This ensured we only captured robust, and consistently reproducible expression signatures.

The meta-analysis method does not provide an overall directional change for each gene; therefore, the standard error (SE) was calculated from the DE logFC values of each gene across the AW assigned significant datasets and used for standard meta-summary estimate analysis using the R package "rmeta" (version 2.16). This served as the meta logFC value in downstream analysis.

Generation of disease-specific meta-analysis expression profiles

Meta-analysis was performed across all AD datasets, followed by a separate meta-analysis across the non-AD disorder datasets. Using these meta-analysis results we generated three expression profiles; (1) "AD expression profile", (2) "AD-specific expression profile" and (3) "common mental disorder expression profile".

The first expression profile, "AD expression profile", is a direct result of the meta-analysis performed on AD studies, which represents the changes typically observed from an AD and cognitively healthy control study design. The second expression profile, deemed as the "AD-specific expression profile", is produced by subtracting significantly DEG's found in the non-AD meta-analysis results from the "AD expression profile". This profile represents transcriptomic changes specifically observed in AD and not in any other neurogenerative or mental health disorder used in this study. The third expression profile, deemed as the "common mental disorder expression profile", represents genes which are significantly DE in all disorders used in this study, including AD.

Functional and gene set enrichment analysis

Gene set enrichment analysis (GSEA), and Gene Ontology (GO) analysis was conducted using an Over-Representation Analysis (ORA) implemented through the ConsensusPathDB web platform (version 32) [36] in May 2017. ConsensusPathDB incorporates numerous well-known biological pathway databases including BioCarta, KEGG, Reactome and Wikipathways. The platform performs a hypergeometric test while integrating a background gene list, which in this case is a list of all the genes that pass quality control in this study, compiles results from each database and corrects for

multiple testing using the false discovery rate (FDR) [36]. A minimum overlap of the query signature and database was set to 2, and a result was deemed significant if the q-value was ≤ 0.05.

Network analysis

Protein-protein interaction (PPI) networks were created by uploading the meta-analysis DEG lists (referred to as seeds in network analysis) along with their meta logFC expression values to NetworkAnalyst's web-based platform http://www.networkanalyst.ca/faces/home.xhtml in June 2017. The "Zero-order Network" option was incorporated to allow only seed proteins directly interacting with each other, preventing the well-known "Hairball effect" and allowing for better visualisation and interpretation [37]. Sub-modules with a p-value ≤ 0.05 (based on the "InfoMap" algorithm [38]) were considered significant key hubs, and the gene with the most connections within this hub was regarded as the key hub gene.

RESULTS

The AD microarray datasets

We Identified and acquired nine publicly available AD studies from ArrayExpress and AMP-AD, of which seven studies contained samples extracted from differing regions of the brain. The basic characteristics of each study and dataset are provided in **Table 1**. Separating the nine studies by brain regions resulted in 46 datasets. Here a "dataset" is defined by brain region and study origin. For example, ArrayExpress study E-GEOD-36980 consists of diseased and healthy samples extracted from three different tissues (temporal cortex, hippocampus and frontal cortex). All samples originating from the same tissue were classified as one dataset; therefore, study E-GEOD-36980 generated three datasets, representing the three different tissues.

The 46 AD datasets contained both AD samples and healthy controls, and were assayed using seven different expression chips over two different microarray platforms (Affymetrix and Illumina) and

Table 1: Characteristics of individual AD studies processed in this meta-analysis

Data repository	Accession details (Publication)	Microarray platform	Expression chip	Tissue source (as stated in the original study publication)	Meta-Analysis brain region mapping	Number of samples after QC (AD/Control)
ArrayExpress	E-GEOD-118553	Illumina	HumanHT-12 v4	Entorhinal Cortex	Temporal Lobe	35/21
				Cerebellum	Cerebellum	38/19
				Frontal Cortex	Frontal Lobe	38/22
				Temporal Cortex	Temporal Lobe	51/29
ArrayExpress	E-GEOD-48350 ([60])	Affymetrix	Human Genome U133 Plus 2.0	Entorhinal Cortex	Temporal Lobe	11/38
				Hippocampus	Temporal Lobe	15/41
				Postcentral Gyrus	Parietal Lobe	19/33
				Superior Frontal Gyrus	Frontal Lobe	17/38
ArrayExpress	E-GEOD-29378 ([7])	Illumina	HumanHT-12 v3	Hippocampus CA1	Temporal Lobe	16/16
				Hippocampus CA3	Temporal Lobe	15/16
ArrayExpress	E-GEOD-36980 ([8])	Affymetrix	Human Gene 1.0 ST	Frontal Cortex	Frontal Lobe	14/17
				Hippocampus	Temporal Lobe	7/10
				Temporal Cortex	Temporal Lobe	10/19
ArrayExpress	E-GEOD-28146 ([19])	Affymetrix	Human Genome U133 Plus 2.0	Hippocampus CA1	Temporal Lobe	15/8
ArrayExpress	E-GEOD-1297 ([61])	Affymetrix	Human Genome U133A	Hippocampus	Temporal Lobe	19/9
ArrayExpress	E-GEOD-5281 ([21])	Affymetrix	Human Genome U133 Plus 2.0	Entorhinal Cortex	Temporal Lobe	10/13
				Hippocampus CA1	Temporal Lobe	10/13
				Medial Temporal Gyrus	Temporal Lobe	16/12

				Posterior Cingulate	Parietal Lobe	9/13
				Superior Frontal Gyrus	Frontal Lobe	23/11
AMP	syn3157225 ([62])	Illumina	Whole-Genome DASL HT	Temporal Cortex	Temporal Lobe	189/186
				Cerebellum	Cerebellum	169/171
AMP	syn4552659 ([63])	Affymetrix	Human Genome U133A	Frontal Pole	Frontal Lobe	25/7
				Precentral Gyrus	Frontal Lobe	20/3
				Inferior Frontal Gyrus	Frontal Lobe	19/4
				Dorsolateral Prefrontal Cortex	Frontal Lobe	19/8
				Superior Parietal Lobule	Parietal Lobe	11/5
				Prefrontal Cortex	Frontal Lobe	23/4
				Parahippocampal Gyrus	Temporal Lobe	18/7
				Hippocampus	Temporal Lobe	20/5
				Inferior Temporal Gyrus	Temporal Lobe	20/6
				Middle Temporal Gyrus	Temporal Lobe	15/7
				Superior Temporal Gyrus	Temporal Lobe	15/8
				Temporal Pole	Temporal Lobe	25/6
AMP	syn4552659 ([63])	Affymetrix	Human Genome U133B	Frontal Pole	Frontal Lobe	26/7
				Precentral Gyrus	Frontal Lobe	18/3
				Inferior Frontal Gyrus	Frontal Lobe	21/5
				Dorsolateral Prefrontal Cortex	Frontal Lobe	20/8
				Superior Parietal Lobule	Parietal Lobe	16/5
				Prefrontal Cortex	Frontal Lobe	23/4

				Parahippocampal Gyrus	Temporal Lobe	19/7
				Hippocampus	Temporal Lobe	22/5
				Inferior Temporal Gyrus	Temporal Lobe	21/7
				Middle Temporal Gyrus	Temporal Lobe	23/7
				Superior Temporal Gyrus	Temporal Lobe	23/8
				Temporal Pole	Temporal Lobe	24/6

Nine publicly available AD studies were identified and acquired for this study. Separating the studies by tissue resulted in 46 datasets, each containing AD and healthy control samples. The brain tissue in each of the 46 datasets was mapped to their corresponding cerebral cortex (temporal lobe, frontal lobe or parietal lobe) or the cerebellum.

consisted of a total 2718 samples before QC. Briefly, the MetaOmics analysis identified study syn4552659 as an outlier and was therefore removed from further analysis (see supplementary text 1), resulting in 1501 samples (746 AD, 755 controls) in the remaining 22 datasets after QC

Summary of the AD meta-analysis DEG counts

The AD meta-analysis was performed on the 22 AD datasets and independently identified differentially expressed genes within the TL, FL, PL and CB brain regions. A summary of the number of datasets in each brain region and the number of significant DEG's identified is provided in **Table 2**

The complete DE results are provided in Supplementary Table 1.

Table 2: Summary of AD study meta-analysis DEG's

Brain region	Number of datasets	Number of samples (case/control)	AW.OC Significant DEGs (FDR adjusted p ≤ 0.05)
Temporal lobe	14	850 (419/431)	323
Frontal lobe	4	180 (92/88)	460
Parietal lobe	2	74 (28/46)	1736
Cerebellum	2	397 (207/190)	867

Twenty-two AD datasets containing a total of 1501 samples remained in this study after QC. The case/control numbers represent the total number of AD/healthy controls subjects across all datasets within a particular brain region. The number of significant genes was identified through a combining p-value method known as Adaptively Weighted with One-sided Correction (AW.OC).

The non-AD disorder microarray datasets

Nine non-AD studies were identified and acquired, of which four studies consisted of samples generated from multiple disorders and brain regions. Separating the studies by disease and tissue equated to 21 datasets consisting of 8 SCZ, 6 BD, 4 HD, 2 MDD and 1 PD dataset with a total of 1166 samples after QC. The demographics of the non-AD datasets is provided in **Table 3**

Table 3: Characteristics of individual non-AD studies included in this meta-analysis

Data repository	ArrayExpress Accession details (Publication)	Microarray Platform	Expression chip	Disorder	Sample source (as stated in the original study publication)	Mapping to brain region	Number of samples after QC (AD/Control)
ArrayExpress	E-GEOD-12649 (**[43]**)	Affymetrix	Human Genome U133A	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	33/34
				Schizophrenia	Prefrontal Cortex	Frontal Lobe	33/32
ArrayExpress	E-GEOD-17612 (**[44]**)	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Prefrontal Cortex	Frontal Lobe	27/22
ArrayExpress	E-GEOD-20168 (**[45]**)	Affymetrix	Human Genome U133A	Parkinson's Disease	Prefrontal Cortex	Frontal Lobe	14/16
ArrayExpress	E-GEOD-21138 (**[46]**)	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Prefrontal Cortex	Frontal Lobe	25/28
ArrayExpress	E-GEOD-21935 (**[47]**)	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Temporal Cortex	Temporal Lobe	22/19
ArrayExpress	E-GEOD-35978 (**[48]**)	Affymetrix	Human Gene 1.0 ST	Bipolar Disorder	Cerebellum	Cerebellum	32/46
				Schizophrenia	Cerebellum	Cerebellum	43/46
				Bipolar Disorder	Parietal Lobe	Parietal Lobe	40/45
				Schizophrenia	Parietal Lobe	Parietal Lobe	51/36
ArrayExpress	E-GEOD-3790 (**[49]**)	Affymetrix	Human Genome U133A	Huntingdon's Disease	Frontal Lobe	Frontal Lobe	36/27
				Huntingdon's Disease	Cerebellum	Cerebellum	38/27
				Huntingdon's Disease	Cerebellum	Cerebellum	38/27

			Human Genome U133B	Huntingdon's Disease	Frontal Lobe	Frontal Lobe	37/29
ArrayExpress	E-GEOD-5388 (**[50]**)	Affymetrix	Human Genome U133A	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	30/29
ArrayExpress	E-GEOD-53987 (**[51]**)	Affymetrix	Human Genome U133 Plus 2.0	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	17/19
				Major Depressive Disorder	Prefrontal Cortex	Frontal Lobe	16/18
				Schizophrenia	Prefrontal Cortex	Frontal Lobe	14/19
				Bipolar Disorder	Hippocampus	Temporal Lobe	18/17
				Major Depressive Disorder	Hippocampus	Temporal Lobe	16/17
				Schizophrenia	Hippocampus	Temporal Lobe	15/18

Nine publicly available non-AD studies were identified and acquired. Separating the studies by tissue resulted in 21 datasets. Each dataset contained both diseased and complimentary healthy controls. The brain tissue in each of the 21 datasets was mapped to their corresponding cerebral cortex (temporal lobe, frontal lobe or parietal lobe) or the cerebellum.

Summary of non-AD brain disorder meta-analyses DEG counts

A second meta-analysis was performed on all non-AD disorders, and similarly to the AD meta-analysis, datasets were grouped into the TL, FL, PL and CB brain regions. An overview of the non-AD meta-analysis results are provided in **Table 4**, and a complete list of DEG's is provided in Supplementary Table 2. SCZ and BD were the only disorders with expression data available across all four brain regions and the frontal lobe brain region was the only region with expression data available from all non-AD disorders identified in this study.

Table 4: Summary of non-AD study meta-analysis DEG's

Brain region	Number of BD datasets (case/control)	Number of Schizophrenia datasets (case/control)	Number of HD datasets (case/control)	Number of MDD datasets (case/control)	Number of PD datasets (case/control)	Total number of datasets (case/control)	AW.OC Significant DEGs (FDR adjusted p ≤ 0.05)
Temporal lobe	1 (18/17)	2 (37/37)	**0**	1 (16/17)	**0**	4 (71/71)	51
Frontal lobe	3 (80/82)	4 (99/101)	2 (73/56)	1 (16/18)	1 (14/16)	11 (282/273)	149
Parietal lobe	1 (40/45)	1 (51/36)	**0**	**0**	**0**	2 (91/81)	2611
Cerebellum	1 (32/46)	1 (43/46)	2 (76/54)	**0**	**0**	4 (151/146)	177

The table illustrates the non-AD dataset and sample distribution across the four brain regions. Disease abbreviations are as follows: BD=Bipolar Disease, HD= Huntington's Disease, MDD=Major Depressive Disorder and PD=Parkinson's Disease. The case/control numbers represent the total number of diseased and healthy control subjects within a disease group and brain region. For instance, "3 (80/82)" for BD datasets in the Frontal lobe region indicates three BD datasets with a combined total of 80 BD and 82 complimentary healthy control subjects. The number of significant DEG's was identified through a combining p-value method known as Adaptively Weighted with One-sided Correction (AW.OC).

The meta-analysis expression profiles

As described in the methods, three primary expression signatures were derived from the meta-analyses for each of the four brain regions: - 1) "AD expression profile", 2) "AD-specific expression profile" and 3) "common mental disorder expression profile". The numbers of significant DEG's in each of the three expression signatures are provided in **Table 5**.

Table 5: Summary of DEGs in each expression signature and brain region

Expression Profile	Cerebellum	Frontal lobe	Parietal lobe	Temporal lobe	Total (unique)
AD	867	460	1736	323	2494
Non-AD	177	149	2611	51	2809
AD-specific	828	435	1023	323	1994
Common	39	25	713	0	755
Total (unique)	1005	584	3642	374	-

The "AD" expression profile represents genes identified as DE in the AD vs control meta-analysis. The "non-AD" expression profile represents genes identified as DE in the non-AD meta-analysis. The "AD-specific" expression profile is a list of genes DE in AD and no other disorder, and the "common" expression profile is a list of genes DE in all mental disorder used in this study. Each expression profile is brain region specific. The "Total (unique)" represents a unique list of the total number of genes identified as significantly DE across brain regions or expression profiles.

The DEG's from the "AD expression profile" in the TL brain region were not significantly DE in any other disorder included in this study. Hence, the "AD expression profile" and the "AD-specific expression profile" contained the same 323 genes for the TL brain region. The "AD-specific expression profile" for all four brain regions is provided in Supplementary Table 3.

The "common mental disorder expression profile" within the four brain regions consisted of very little or no DEG's (except for the parietal lobe); hence, the downstream analysis did not yield any statistically significant results of biological relevance. We find little robust evidence of shared biology based on this data analysis and therefore, exclude all results generated from the "common mental disorder expression profile" from this paper; however, we provide the complete list of significantly DEG's within this profile in Supplementary Table 4.

Common differentially expressed genes across multiple brain regions in AD

AD is known to affect all brain regions through the course of the disease, although not to the same degree, similar transcriptomic changes across all brain regions were deemed disease-specific, while perturbations in a single brain region were considered to be tissue-specific. We were particularly

interested in disease-specific transcriptomic changes and therefore decided to focus on genes that were found to be consistently DE across multiple brain regions.

Meta-analysis of the AD datasets identified a total of 2494 unique genes as significantly DE. The distribution of these genes across the four brain regions is shown in Figure 1. Forty-two genes were found to be perturbed across all four brain regions and can be grouped into three sets (Figure 2).

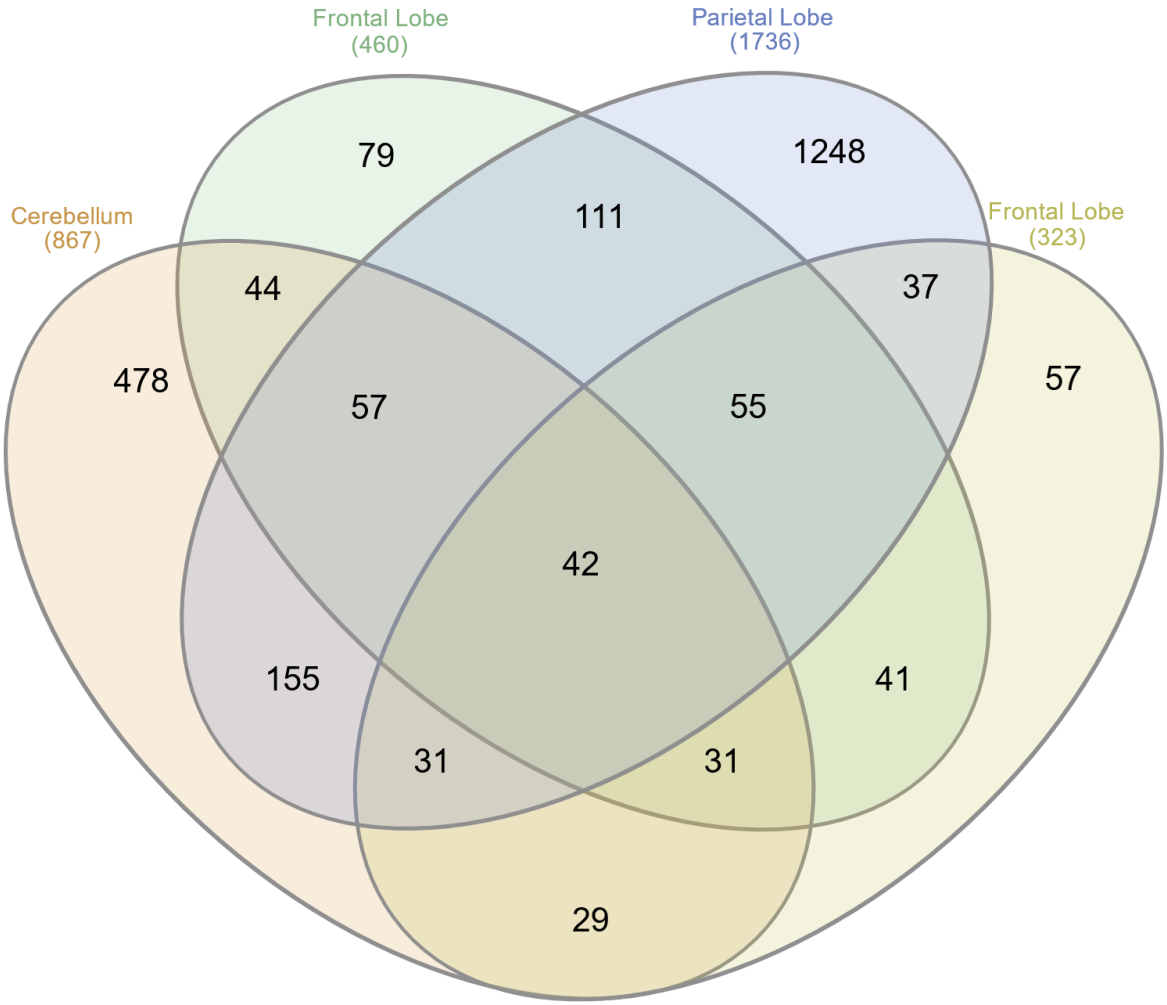

Figure 1: Overlap of DEG's in the AD expression profile across brain regions. Forty-two genes were observed to be significantly differentially expressed across all four AD brain regions.

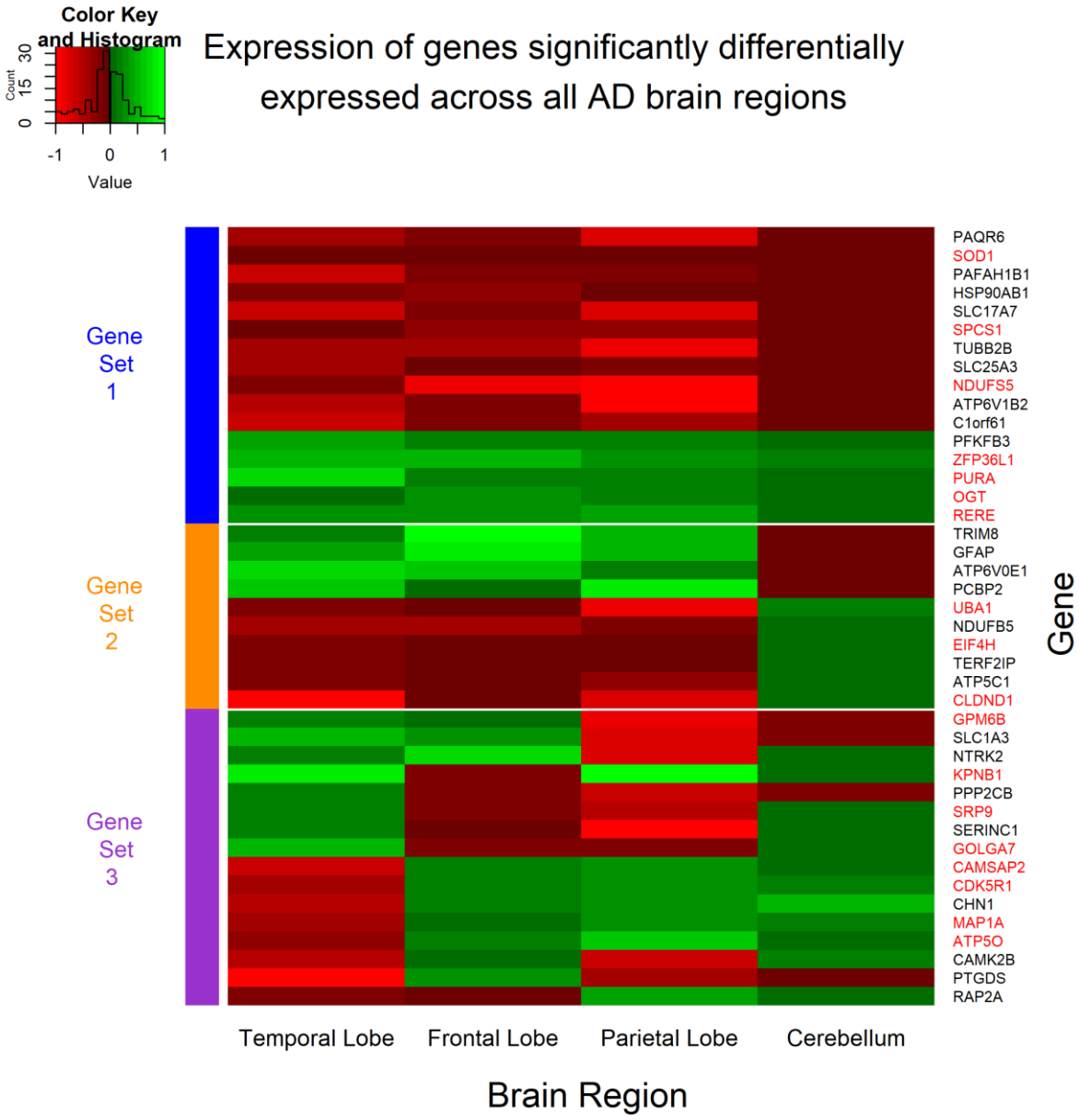

Figure 2: Expression pattern of genes significantly differentially expressed across all four AD brain regions. The expression values for each gene was obtained from the meta-summary calculations. Red cells represent down-regulated genes, and green cells represent up-regulated genes. Forty-two genes were observed to be significantly perturbed across all four AD brain regions and can be grouped into three "sets". Gene set 1 represents genes which are perturbed consistently in the same direction across all AD brain regions and can be considered disease-specific. Gene set 2 represents genes consistent in expression in the TL, FL and PL brain regions, but reversed in the CB brain region; a region often referred to be free from AD pathology. Finally, Gene set 3 represents genes which are significant DE across all four brain regions, however, directional change is not consistent across the brain regions and may represent tissue-specific genes or even false positive. The gene names highlighted in red are genes perturbed in AD and not in any other disorder used in this study and are deemed "AD-specific".

The first group (Gene set 1) are expressed consistently in the same direction across all four brain regions and can be regarded as disease-specific. The second group (Gene set 2) are expressed in the same direction in the TL, FL and PL, but expression is reversed in the CB brain region, a region suggested to be spared from AD pathology [4] [5]. This expression pattern suggests these genes may be involved in AD pathology. Finally, the third group (Gene set 3) are inconsistently expressed across the four brain regions are most likely tissue-specific or even false-positives.

From the forty-two genes significantly differentially expressed across all brain regions, seven genes were DE in the same direction and belong to the "AD-specific expression profile", that is, these seven genes (down-regulated **NDUFS5**, **SOD1**, **SPCS1** and up-regulated **OGT**, **PURA**, **RERE**, **ZFP36L1**) were consistently perturbed in all AD brain regions and not in any other brain region of any other mental disorder used in this study and can be considered unique to AD brains. The expression of these seven genes across AD brains is shown in Figure 3.

Differentially expressed genes in brain regions affected by AD histopathology

In AD, the TL, FL and PL are known to be affected by both plaques and tangles, while the CB brain region is rarely reported to be affected. In addition to identifying genes DE across all brain regions and reversed in the CB brain region, we were also interested in genes perturbed in the TL, FL and PL and not the CB. These genes may also play a role in general AD histopathology and could be new therapeutic targets in preventing or curing AD.

Fifty-five genes were found to be significantly DE in TL, FL and PL but not the CB, of which seventeen were expressed in the same direction and were not DE in the other brain disorders used in this study. Nine of the seventeen genes (**ATP1A1**, **ATP2A2**, **ATP6V1E1**, **GHITM**, **LDHA**, **NDRG4**, **NSF**, **RAB6A** and **RTN3**) were consistently up-regulated, and eight genes (**FDFT1**, **MT2A**, **NACC2**, **NFIB**, **RTN1**, **SH3GL2**, **SRRM2** and **WAC**) were consistently down-regulated in AD.

Furthermore, from the forty-two genes identified as significantly DE across all four AD brain regions, ten genes were in consensus in their expression across the TL, FL and PL brain region but expression

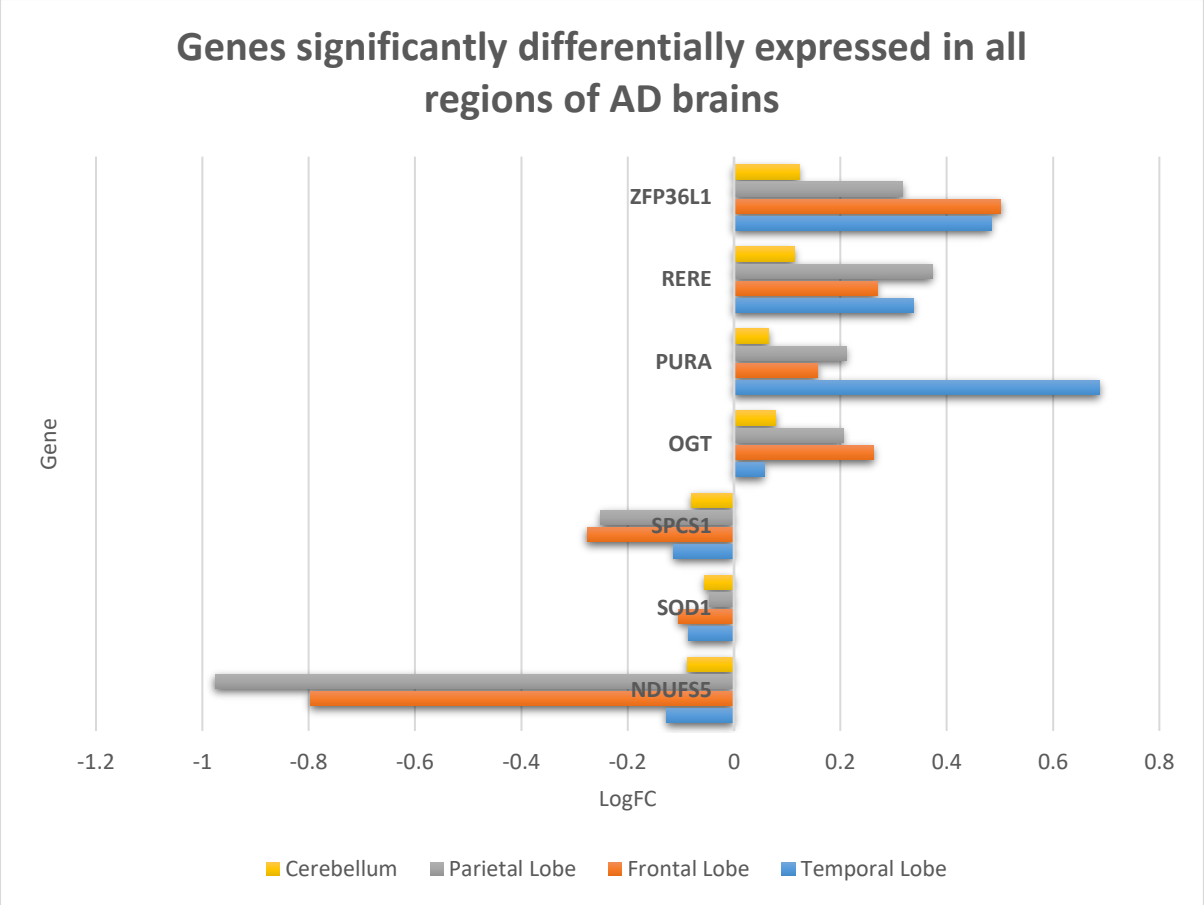

Figure 3: Seven genes consistently significantly differentially expressed in the same direction in all regions of AD brains but not in Schizophrenia, Bipolar Disorder, Huntington's disease, Major Depressive Disorder or Parkinson's disease brains. These seven genes can be assumed to be unique to AD brains and may play an important role in disease mechanisms.

is reversed in the cerebellum. Only 3 of these genes (**UBA1**, **EIF4H** and **CLDND1**) belong to the "AD-specific expression profile", and all three genes were significantly down-regulated in the TL, FL and PL, but significantly up-regulated in the CB brain region (see Gene set 2 in Figure 2).

"AD Expression Profile" functional gene set enrichment and GO analysis

Gene set enrichment analysis of the "AD expression profile" identified 205, 197, 98, and 45 biological pathways significantly enriched in the TL, FL, PL and CB brain regions respectively (Supplementary Table 5). There were ten pathways significantly enriched in all four brain regions, of which eight are involved in the "**metabolism of protein**" (specifically the translation process, the most significant being in CB brain region with a q-value=1.11e-7), one involved in "**adenosine ribonucleotides de**

novo biosynthesis" (TL q-value = 0.007, FL q-value = 7.56e-5, PL q-value = 0.04, CB q-value = 0.03) and one involved in the "**digestive system**" (TL q-value = 0.02, FL q-value = 0.02, PL q-value = 0.01, CB p-value = 0.02).

When excluding the CB brain region, 42 pathways were significantly enriched in the remaining three brain regions, of which five pathways obtained an FDR adjusted significance p-value of ≤ 0.01. The five pathways are "**Alzheimer's disease**" (TL q-value = 6.53e-4, FL q-value = 0.02, PL q-value = 0.01), "**Electron Transport Chain**" (TL q-value = 0.006, FL q-value = 2.95e-5, PL q-value = 3.69e-5), "**Oxidative phosphorylation**" (TL q-value =1.77e-4, FL q-value = 4.99e-8, PL q-value = 4.18e-05), "**Parkinson's disease**" (TL q-value =8.57e-4, FL q-value = 1.59e-6, PL q-value = 1.77e-6) and "**Synaptic vesicle cycle**" (TL q-value = 5.19e-4, FL q-value = 3.82e-7 , PL q-value = 2.03e-4).

The biological GO analysis identified 384, 417, 216, and 72 biological components as significantly enriched in the TL, FL, PL and CB brain region respectively (Supplementary Table 6). There were 36 pathways significantly enriched across all four brain regions at a p-value threshold of ≤ 0.05 and nine at an FDR adjusted significant p-value threshold of ≤ 0.01. These nine processes are "**cellular component biogenesis**" (TL q-value =1.38e-4, FL q-value = 0.002, PL q-value = 5.86e-4 , CB q-value = 0.006), "**cellular component organization**" (TL q-value = 1.96e-8, FL q-value = 1.04e-8, PL q-value = 3.35e-5 , CB q-value = 0.004), "**interspecies interaction between organisms**" (TL q-value = 1.13e-41.85e-4, FL q-value = 8.73e-5, PL q-value = 5.59e-5 , CB q-value = 0.002), "**multi-organism cellular process**" (TL q-value = , FL q-value = 4.72e-5, PL q-value = 8.04e-5, CB q-value = 0.002), "**nervous system development**" (TL q-value = 1.64e-7, FL q-value = 5.90e-14, PL q-value = 3.82e-8, CB q-value = 0.01), "**organonitrogen compound metabolic process**" (TL q-value = 0.002, FL q-value = 1.56e-5, PL q-value = 1.02e-5, CB q-value = 0.002), "**symbiosis, encompassing, mutualism through parasitism**" (TL q-value = 4.04e-4, FL q-value = 1.92e-4, PL q-value = 3.18e-4, CB q-value = 0.004), "**translational initiation**" (TL q-value = 0.007, FL q-value = 0.006, PL q-value = 2.41e-4, CB q-value = 5.24e-6), and "**viral process**" (TL q-value = 2.82e-4, FL q-value = 1.17e-4, PL q-value = 3.18e-4, CB q-value = 0.002).

Excluding the CB brain region resulted in 84 common biological components being significantly enriched across the remaining three brain regions.

"AD-Specific Expression Profile" functional gene set enrichment and GO analysis

Analysis of the "AD-specific expression profile" identified 205, 196, 40 and 42 pathways as significantly enriched in the TL, FL, PL and CB brain region respectively in the GSEA analysis (Supplementary Table 7). The analysis identified six significantly enriched pathways across all four brain regions, and all are involved in "**metabolism of protein"** (specifically the translation process, with the most significant pathway being in the PL brain region with a q-value = 8.92e-7). The same six pathways were identified when the CB region was excluded.

The GO analysis identified 384, 344, 36 and 72 significantly enriched biological components for the TL, FL, PL and CB brain region respectively. Only four common biological components were significantly enriched across all four brain regions, and all are indicative of interspecies interactions including viral. Excluding the CB identifies only "**neural nucleus development**" (TL q-value = 5.35e-5, FL q-value = 0.007, PL q-value = 0.003) as an additional component being enriched. The complete biological GO analysis results are provided in Supplementary Table 8.

Network analysis hub gene identification

PPI networks were generated for each expression profile and in each of the four brain regions (TL, FL, PL and CB) to identify genes whose protein product interacts with other protein products from the same expression profile. Genes with more interactions than expected are referred to as hub genes and may be of biological significance.

Temporal lobe hub genes

PPI network analysis was performed on the expression profiles of TL brain region to identify key hub genes. The "AD expression profile" and the "AD-specific expression profile" both consisted of the same 323 DEG's which represented 282 seed proteins with 716 edges (interactions between

proteins). Two significant key hub genes were identified; the down-regulated Polyubiquitin-C (**UBC**, p-value = 1.57e-30) and the up-regulated Small Ubiquitin-related Modifier 2 (**SUMO2**, p-value = 3.7e-4).

Frontal Lobe hub genes

The FL "AD expression profile" consisted of 460 DEG which represented 272 seed proteins and 620 edges. Two significant key hub genes were identified; up-regulated Amyloid Precursor Protein (**APP**, p-value = 1.98e-08) and down-regulated Heat Shock Protein 90-alpha (**HSP90AA1**, p-value = 0.003). Using the "AD-specific expression profile" identified the same two key hub genes, with **APP** reaching a significant p-value of 2.11e-09.

Parietal Lobe hub genes

The PL "AD expression profile" consisted of 1736 DEG which represented 1437 seed proteins and 5720 edges. Similar to the TL and FL, two significant key hub genes were identified; down-regulated Cullin-3 (**CUL3**, p-value = 1.84e-10) and down-regulated **UBC** (p-value = 1.84e-10). Using the "AD-specific expression profile" (1023 DEGs, 810 seed proteins and 2351 edges) identified **UBC** as the only key hub gene, with a more significant p-value of 1.84e-10. The **CUL3** gene is no longer a significant key hub gene in the network.

Cerebellum hub genes

The CB "AD expression profile" consisted of 867 DEG's which represented 548 seed proteins and 1419 edges. Four significant key hub genes were identified; up-regulated **APP** (p-value =4.24e-26), down-regulated Ribosomal Protein 2 (**RPS2**, p-value = 4.24e-26), down-regulated **SUMO2** (p-value = 4e-05), and up-regulated Glycyl-TRNA Synthetase (**GARS**, p-value = 0.0207). Using the "AD-specific expression profile" for the same brain region identified **APP** (p-value = 3.44e-26), **RPS2** (p value= 6.61e-06), and **SUMO2** (p-value = 3.78e-06) as the key hub genes only. The **GARS** gene is no longer a key hub gene in the network.

DISCUSSION

In this study, we acquired eighteen publicly available microarray gene expression studies covering six mental health disorders; AD, BD, HD, MDD, PD and SCZ. Data was generated on seven different expression BeadArrays and across two different microarray technologies (Affymetrix and Illumina). The eighteen studies consisted of 3984 samples extracted from 22 unique brain regions which equated to 67 unique datasets when separating by disorder and tissue. However, due to study and sample outlier analysis, only 43 datasets (22 AD, 6 BD, 4 HD, 2 MDD, 1 PD and 8 SCZ) totalling 2,667 samples passed QC. We grouped the AD datasets by tissue, into the TL, FL, PL and CB brain regions to perform the largest microarray AD meta-analysis known to date to our knowledge, which identified 323, 460, 1736 and 867 significant DEG's respectively. Furthermore, we incorporated transcriptomic information from other mental disorders to subset the initial findings to 323, 435, 1023, and 828 significant DEG's that were specifically perturbed in the TL, FL, PL and CB brain regions respectively of AD subjects.

Genes specifically perturbed across AD brain regions

Seven genes (down-regulated **NDUFS5**, **SOD1**, **SPCS1** and up-regulated **OGT**, **PURA**, **RERE**, **ZFP36L1**) were DE in AD brains and not DE in the other disorders used in this study. We deemed these seven protein-coding genes as "AD-specific". Three of these genes (**NDUFS5**, **SOD1** and **OGT)** have been previously associated with AD. Down-regulated NADH Dehydrogenase Ubiquinone Fe-S Protein 5 (**NDUFS5**) gene is part of the human mitochondrial respiratory chain complex; a process suggested to be disrupted in AD in multiple studies [38] [39]. A study investigating blood-based AD biomarkers identified 13 genes, including **NDUFS5**, which was capable of predicting AD with 66% accuracy (67% sensitivity and 75% specificity) in an independent cohort of 118 AD and 118 control subjects [41]. The perturbation in **NDUFS5** expression in the blood and brains of AD subjects suggests this gene may have potential as an AD biomarker and warrants further investigation.

Down-regulated Superoxide Dismutase 1 (**SOD1**) gene encodes for copper and zinc ion binding proteins which contribute to the destruction of free superoxide radicals in the body and is also involved in the function of motor neurons [provided by RefSeq, Jul 2008]. Mutations in this gene have been heavily implicated as causes of familial amyotrophic lateral sclerosis (**ALS**) [42] and have also been associated with AD risk [43]. A recent study discovered **SOD1** deficiency in an amyloid precursor protein-overexpressing mouse model accelerated Aβ oligomerisation and also caused Tau phosphorylation [44]. They also stated **SOD1** isozymes were significantly decreased in human AD patients, and we can now confirm **SOD1** is significantly under-expressed at the mRNA level in human AD brains as well.

The up-regulated O-Linked N-Acetyl Glucosamine Transferase (**OGT**) gene encodes for a glycosyltransferase that links N-acetylglucosamine to serine and threonine residues (O-GlcNAc). O-GlcNAcylation is the post-translational modification of O-GlcNAc and occurs on both neuronal tau and APP. Increased brain O-GlcNAcylation has been observed to protect against tau and amyloid-β peptide toxicity [45]. A mouse study has demonstrated a deletion of the encoding **OGT** gene causes an increase in tau phosphorylation [46]. In this study, we observe a significant increase in **OGT** gene expression throughout human AD brains, including the cerebellum where tangles are rarely reported, suggesting **OGT** gene is most likely not solely responsible for the formation of tangles.

OGT and O-GlcNAcase (OGA) enzymes facilitate O-GlcNAc cycling, and levels of GlcNAc have also been observed to be increased in the parietal lobe of AD brains [47]. Appropriately, OGA inhibitors have been tested for treating AD with promising preliminary results [48], prompting further investigation into targeting OGT for AD treatment.

Genes involved in AD histopathology

The CB brain region is known to be free from tau pathology and occasionally free from plaques as well. We exploited the CB brain region as a secondary control to identify seventeen genes (**ATP1A1, ATP2A2, ATP6V1E1, GHITM, LDHA, NDRG4, NSF, RAB6A, RTN3, FDFT1**, **MT2A**, **NACC2**, **NFIB**, **RTN1**,

SH3GL2, **SRRM2** and **WAC**) DE specifically in TL, FL and PL and not the CB brain region of AD subjects. From these genes, **LDHA**, **RAB6A** and **RTN3** have been previously associated with AD. Lactate Dehydrogenase A (**LDHA**) gene encodes for a protein that catalyses the conversion of L-lactate and NAD to pyruvate and NADH in the final step of anaerobic glycolysis [provided by RefSeq, Sep 2008], a process previously suggested to be disrupted in AD brains [49]. **RAB6A** regulates the intracellular processing of the amyloid precursor protein (APP) [50], and Reticulon 3 (**RTN3**) encodes for a protein that interacts and modulates the activity of beta-amyloid converting enzyme 1 (**BACE1**) and the production of amyloid-beta.

We identified an additional three AD-specific genes (**UBA1**, **EIF4H** and **CLDND1**) which were significant DE in all four brain regions. However, the genes were down-regulated in the TL FL and PL but up-regulated in the CB brain region. Ubiquitin-Like Modifier Activating Enzyme 1 (**UBA1**) encodes for a protein that catalyses the first step in ubiquitin conjugation to mark cellular proteins for degradation. Eukaryotic Translation Initiation Factor 4H (**EIF4H**) encodes for a translation initiation factors, which functions to stimulate the initiation of protein synthesis at the level of mRNA utilisation and Claudin Domain Containing 1 (**CLDND1**) is a transmembrane protein of tight junctions found on endothelial cells [51]. **As the cerebellum is the only brain region spared from tangle formation and occasionally from plaque, we suggest these 20 genes could potentially be associated with AD histopathology.**

Translation of proteins perturbed specifically in AD brains

Functional gene set enrichment analysis of the "AD expression profile" revealed more pathways were significantly perturbed in the TL, followed by the FL, PL and CB, which is the general route AD pathology is known to spread through the brain. We originally observed ten biological pathways being enriched across all AD brain regions, which included biological pathways likely to be irrelevant such as the "**digestive system**". However, when incorporating transcriptomic information from non-AD disorders, we were able to refine the AD expression signature to specific genes perturbed in AD

only. This resulted in the enrichment of pathways only involved in the "**metabolism of proteins**", specifically the translation process which has been previously suggested in be associated with AD on numerous occasions [10] [11] [14] [15] [16] [17]. **We now suggest this may be a biological process specifically disrupted in AD brains, and not BD, HD, MDD, PD or SCZ brains.**

Previous biological perturbations observed in AD are only associated Temporal Lobe brain region.

Previous AD studies have consistently suggested the immune response [10] [11] [12] [13], protein transcription/translation regulation [10] [11] [14] [15] [16] [17], calcium signalling [10] [18] [19], MAPK signalling [16] [7], chemical synapse [18] [7] [19], neurotransmitter pathways [11] [18] [19] and various metabolism pathways [16] [20] [21] [22] [17][11] [23] are disrupted in AD. We observe the same pathways enriched in our meta-analysis; however, only in the TL brain region, a brain region often heavily investigated in AD. Except for "**metabolism of proteins**", we did not observe any of these pathways significantly enriched across all of the four brain regions, suggesting these pathways observed to be perturbed in previous studies may be tissue-specific rather than disease-specific.

Interspecies interactions possibly involved in AD

Gene Ontology analysis on the "AD expression profile" identified nine different biological components enriched across all four brain regions. However, when we remove genes perturbed in other mental disorders, we only observe four biological components as significantly enriched, and all four were indicative of interspecies interactions. AD brains have a prominent inflammatory component which is characteristic of infection, and many microbes have been implicated in AD, notably herpes simplex virus type 1 (HSV1), Chlamydia pneumonia, and several types of spirochaete [52]. Furthermore, Aβ has been suggested to be an antimicrobial peptide and has been shown to protect against fungal and bacterial infections [53]. Thus, the accumulation of Aβ may be part of the

brains defence mechanism against infections. Although a controversial theory, we also observe a viral component in AD brains, and as a result of this meta-analysis, further suggest this maybe AD-specific and warrants further investigation.

Network analysis identifies AD-specific APP UBC and SUMO2 hub genes

Network analysis identified five (**APP, HSP90AA1, UBC, SUMO2** and **RPS2**) significant hub genes specific to AD brain regions. **APP, UBC** and **SUMO2** gene appear as hub genes in multiple brain regions. The **APP** gene encodes for a cell surface receptor transmembrane amyloid precursor protein (APP) that is cleaved by secretases to form a number of peptides. Some of these peptides are secreted and can bind to the acetyltransferase complex APBB1/TIP60 to promote transcriptional activation, while others form the protein basis of the amyloid plaques in AD brains. In addition, two of the peptides are antimicrobial peptides, having been shown to have bacteriocidal and antifungal activities [provided by RefSeq, Aug 2014]. Changes in APP functions have been suggested to play an essential role in the lack of AB clearance, ultimately leading to the formation of plaques [54].

UBC (ubiquitin-C) gene encodes for a Polyubiquitin-C protein which is part of the ubiquitin-proteasome system (UPS), the major intracellular protein quality control system in eukaryotic cells. UPS has an immense impact on the amyloidogenic pathway of APP processing that generates Abeta [55]. A recent GWAS study identified **UBC** as a novel LOAD gene, and through network analysis also identified **UBC** as a key hub gene. The study validated their findings in a **UBC** C. elegans model to discover **UBC** knockout accelerated age-related AB toxicity [56]. We also observe the **UBC** gene being down-regulated and as a key hub gene in multiple regions of human AD brains, further providing evidence of its key role in AD.

Small Ubiquitin-Like Modifier 2 (**SUMO2**) gene encodes for a protein that binds to target proteins as part of a post-translational modification system, a process referred to as SUMOylation [57]. However, unlike ubiquitin, which targets proteins for degradation, this protein is involved in a variety of cellular processes, such as nuclear transport, transcriptional regulation, apoptosis, and protein

stability [provided by RefSeq, Jul 2008]. Early studies have indicated that the **SUMO** system is likely altered with AD-type pathology, which may impact Aβ levels and tau aggregation [57]. Genetic studies have supported this theory with a GWAS study linking SUMO-related genes to LOAD [58], with further studies showing that the two natively unfolded proteins, tau and α-synuclein, are sumoylated in vitro [59]. We identified **SUMO2** as a significant key hub gene in both the human TL and CB brain region. However, what makes this discovery interesting is that **SUMO2** is significantly up-regulated in the TL, a region where both plaques and tangles can be observed, but significantly down-regulated in the CB, where only plaques have been occasionally observed, but tangles never reported. The up-regulation of **SUMO2** gene may play a vital role in the formation of tangles, and further investigation into this gene is warranted.

Limitations

Although this study presents novel insights to AD-specific transcriptomic changes in the human brain, limitations to this study must be addressed. Firstly, we meta-analysed a total of 22 AD and 21 non-AD datasets, and many of these datasets lacked necessary experimental processing or basic phenotypic information such as technical batches, RNA integrity numbers (RIN), age, gender, or ethnicity, all of which can have confounding effects. To address this, we incorporated recommended best practices to estimate and correct for both known and hidden batch effects using SVA and COMBAT to ensure data is comparable between experiments and studies. However, this does not guarantee that all technical variation is completely removed.

Secondly, the terminology used to label brain tissue varied across studies, with some reporting a broad region such as the “hippocampus” used in study E-GEOD-48350, while others were very specific to the tissue layer, such as “hippocampus CA3” in study E-GEOD-29378. We, therefore, decided to map all brain tissue as mentioned in each dataset publication to their hierarchical cerebral cortex lobe (TL, FL and PL) and the CB. The mapping procedure was completed using

publicly available literature defined knowledge, and we assume tissues within these brain regions are relatively comparable to infer AD-associated histopathological changes.

This study relied on publicly available transcriptomic data, and as previous research has heavily investigated brain regions known to be at the forefront of disease manifestation, this led to unbalanced datasets per brain region in both the AD and non-AD meta-analysis. Subsequently, the AD meta-analysis consisted of 14, 4, 2, and 2 datasets for the TL, FL, PL and CB brain regions respectively, with the PL brain region consisting of only 74 samples (28 AD and 46 controls) in total. In addition, the non-AD meta-analysis lacked expression signatures form all non-AD diseases across all brain regions (except for FL). Nevertheless, the brain regions most affected by each disorder was captured in this study, suggesting we most likely were able to capture key brain transcriptomic changes relating to each disorder. Furthermore, as AD is known to affect all brain regions, albeit not to the same extent, we focus on transcriptomic changes observed across all brain regions that are also not observed in any brain region of the non-AD subjects, ensuring we capture transcriptomic signatures unique to AD brains.

Finally, we assume the non-AD datasets are comparable through meta-analysis, and by identifying common expression signatures that are not associated with individual disease mechanisms may represent false positives or even a general signature for “brain disorder”. Removing this signature from the AD meta-analysis expression profile may result in transcriptomic changes specific to AD brains, revealing more relevant changes to the underlying disease mechanism rather than general mental diseases. Under this assumption, we observe more relevant and refined biological enrichment results. For example, we originally observed ten biological pathways enriched across all AD brain regions, including biological pathways such as the “**digestive system**”. However, by refining the AD expression signature by removing genes perturbed in other related mental-disorders, only pathways involved in the “**metabolism of proteins**” remain, which has been previously suggested in be associated with AD on numerous occasions [10] [11] [14] [15] [16] [17]. This observation provides

strong evidence of our assumption of incorporating non-AD diseases in this study to infer AD-specific changes as valid.

Conclusion

We present the most extensive human AD brain microarray transcriptomic meta-analysis study to date, incorporating, brain regions both affected and partially spared by AD pathology, and utilise related mental disorders to infer AD-specific brain changes. This led to the identification of seven genes specifically perturbed across all AD brain regions and are considered disease-specific, twenty genes specifically perturbed in AD brains which could play a role in AD neuropathology, and the refinement of GSEA and GO analysis results to identify specific biological pathways and components specific to AD. These AD-specific changes may provide new insights into the disease mechanisms, thus making a significant contribution towards understanding the disease and provides new therapeutic targets for the prevention and cure of AD.

References

[1] M. Prince, Albanese Emiliano, and Prina Matthew, "World Alzheimer Report 2014 Dementia and Risk Reduction," *Alzheimer's Dis. Int.*, 2014.

[2] K. B. Rajan, R. S. Wilson, J. Weuve, L. L. Barnes, and D. A. Evans, "Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia," *Neurology*, vol. 85, no. 10, pp. 898–904, 2015.

[3] A. Serrano-Pozo, M. P. Frosch, E. Masliah, and B. T. Hyman, "Neuropathological alterations in Alzheimer disease.," *Cold Spring Harb. Perspect. Med.*, vol. 1, no. 1, pp. 1–23, 2011.

[4] a. Convit, J. De Asis, M. J. De Leon, C. Y. Tarshish, S. De Santi, and H. Rusinek, "Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease," *Neurobiol. Aging*, vol. 21, no. 1, pp. 19–26, 2000.

[5] H. I. L. Jacobs *et al.*, "The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline," *Brain*, no. January, 2017.

[6] M. Zhang *et al.*, "Apparently low reproducibility of true differential expression discoveries in microarray studies," *Bioinformatics*, vol. 24, no. 18, pp. 2057–2063, 2008.

[7] J. A. Miller, R. L. Woltjer, J. M. Goodenbour, S. Horvath, and D. H. Geschwind, "Genes and pathways underlying regional and cell type changes in Alzheimer's disease," *Genome Med.*, vol. 5, no. 5, p. 48, 2013.

[8] M. Hokama *et al.*, "Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study.," *Cereb. Cortex*, vol. 24, no. 9, pp. 2476–88, Sep. 2014.

[9] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Res.*, vol. 33, no. 18, pp. 5914–5923, 2005.

[10] S. Sekar *et al.*, "Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes," *Neurobiol. Aging*, vol. 36, no. 2, pp. 583–591, 2015.

[11] Y. Li *et al.*, "Analysis of hippocampal gene expression profile of Alzheimer ' s disease model rats using genome chip bioinformatics * ☆" vol. 7, no. 5, pp. 332–340, 2012.

[12] J. C. Lambert *et al.*, "Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis," *J. Alzheimer's Dis.*, vol. 20, no. 4, pp. 1107–1118, 2010.

[13] J. Chen, C. Xie, Y. Zhao, Z. Li, and P. Xu, "Gene expression analysis reveals the dysregulation of immune and metabolic pathways in Alzheimer ' s disease," vol. 7, no. Table 1, pp. 1–6, 2016.

[14] T. Liu *et al.*, "Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform.," *Sci. Rep.*, vol. 3, p. 3467, 2013.

[15] X. Li, J. Long, T. He, R. Belshaw, and J. Scott, "Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease," *Sci. Rep.*, vol. 5, no. 1, p. 12393, 2015.

[16] N. Puthiyedth, C. Riveros, R. Berretta, and P. Moscato, "Identification of differentially expressed genes through integrated study of Alzheimer's disease affected brain regions," *PLoS One*, vol. 11, no. 4, pp. 1–29, 2016.

[17] J. A. Godoy, J. A. Rios, J. M. Zolezzi, N. Braidy, and N. C. Inestrosa, "Signaling pathway cross

talk in Alzheimer's disease," *Cell Commun. Signal.*, vol. 12, no. 1, p. 23, 2014.

[18] V. K. Ramanan *et al.*, "Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks," vol. 6, no. 4, pp. 634–648, 2013.

[19] E. M. Blalock, H. M. Buechel, J. Popovic, J. W. Geddes, and P. W. Landfield, "Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease," *J. Chem. Neuroanat.*, vol. 42, no. 2, pp. 118–126, 2011.

[20] G. Di Paolo and T. Kim, "Linking Lipids to Alzheimer ' s Disease : Cholesterol and Beyond," *Aging (Albany. NY).*, vol. 12, no. 5, pp. 284–296, 2012.

[21] W. S. Liang *et al.*, "Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 11, pp. 4441–6, 2008.

[22] K. Ishii *et al.*, "Reduction of cerebellar glucose metabolism in advanced Alzheimer's disease.," *J. Nucl. Med.*, vol. 38, no. 6, pp. 925–928, 1997.

[23] S. Oshiro, M. S. Morioka, and M. Kikuchi, "Dysregulation of iron metabolism in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis," *Adv. Pharmacol. Sci.*, vol. 2011, 2011.

[24] Y. J. K. Edwards *et al.*, "Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach," *PLoS One*, vol. 6, no. 2, 2011.

[25] S. Chandrasekaran and D. Bonchev, "a Network View on Parkinson'S Disease," *Comput. Struct. Biotechnol. J.*, vol. 7, no. 8, p. e201304004, 2013.

[26] C. L. Clelland, L. L. Read, L. J. Panek, R. H. Nadrich, C. Bancroft, and D. James, "Utilization of Never-Medicated Bipolar Disorder Patients towards Development and Validation of a Peripheral Biomarker Profile," vol. 8, no. 6, pp. 1–11, 2013.

[27] P. ohn I. Nurnberger Jr, MD *et al.*, "Identification of Pathways for Bipolar Disorder A Meta-analysis," *Curr. Drug Targets*, vol. 16, no. 7, pp. 700–710, 2015.

[28] H. Chen, N. Wang, X. Zhao, C. A. Ross, K. S. O'Shea, and M. G. Mcinnis, "Gene expression alterations in bipolar disorder postmortem brains," *Bipolar Disord.*, vol. 15, no. 2, pp. 177–187, 2013.

[29] N. Khanzada, M. Butler, and A. Manzardo, "GeneAnalytics Pathway Analysis and Genetic Overlap among Autism Spectrum Disorder, Bipolar Disorder and Schizophrenia," *Int. J. Mol. Sci.*, vol. 18, no. 3, p. 527, 2017.

[30] William R. Markesbery, "Neuropathologic Alterations in Mild Cognitive Impairment: A Review," *J Alzheimers Dis*, vol. 19, no. 1, pp. 221–228, 2010.

[31] C. Lazar *et al.*, "Batch effect removal methods for microarray gene expression data integration: A survey," *Brief. Bioinform.*, vol. 14, pp. 469–490, 2013.

[32] A. J. Hackstadt and A. M. Hess, "Filtering for increased power for microarray data analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 11, 2009.

[33] M. C. Oldham, P. Langfelder, and S. Horvath, "Network methods for describing sample relationships in genomic datasets: application to Huntington's disease.," *BMC Syst. Biol.*, vol. 6, no. 1, p. 63, Jan. 2012.

[34] D. D. Kang, E. Sibille, N. Kaminski, and G. C. Tseng, "MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis," *Nucleic Acids Res.*, vol. 40, no. 2, pp. 1–14, 2012.

[35] L.-C. Chang, H.-M. Lin, E. Sibille, and G. C. Tseng, "Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline," *BMC Bioinformatics*, vol. 14, no. 1, p. 368, 2013.

[36] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB - A database for integrating human functional interaction networks," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 623–628, 2009.

[37] J. Xia, M. J. Benner, and R. E. W. Hancock, "NetworkAnalyst - Integrative approaches for protein-protein interaction network analysis and visual exploration," *Nucleic Acids Res.*, vol. 42, no. W1, pp. 167–174, 2014.

[38] N. Zaki and A. Mora, "A comparative analysis of computational approaches and algorithms for protein subcomplex identification," *Sci. Rep.*, vol. 4, no. 1, p. 4262, 2015.

[39] E. Bonilla, K. Tanji, M. Hirano, T. H. Vu, S. Dimauro, and E. A. Schon, "Mitochondrial involvement in Alzheimer ' s disease," vol. 1410, 1999.

[40] J. Hroudová and N. Singh, "Mitochondrial Dysfunctions in Neurodegenerative Diseases : Relevance to Alzheimer ' s Disease," vol. 2014, 2014.

[41] N. Voyle *et al.*, "A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis," *J. Alzheimer's Dis.*, vol. 49, no. 3, pp. 659–669, 2016.

[42] D. P. Daniel R. Rosen, Teepu Siddique, "Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis," *Nature*, vol. 362, 1993.

[43] K. Spisak, A. Klimkowicz-Mrowiec, J. Pera, T. Dziedzic, A. Golenia, and A. Slowik, "rs2070424 of the SOD1 gene is associated with risk of alzheimer's disease," *Neurol. Neurochir. Pol.*, vol. 48, no. 5, pp. 342–345, 2015.

[44] K. Murakami *et al.*, "SOD1 (Copper / Zinc Superoxide Dismutase) Deficiency Drives Amyloid Protein Oligomerization and Memory Loss in Mouse Model of Alzheimer Disease * □," vol. 286, no. 52, pp. 44557–44568, 2011.

[45] Y. Zhu, X. Shan, S. A. Yuzwa, and D. J. Vocadlo, "The emerging link between O-GlcNAc and Alzheimer disease," *J. Biol. Chem.*, vol. 289, no. 50, pp. 34472–34481, 2014.

[46] N. O'Donnell, N. E. Zachara, G. W. Hart, and J. D. Marth, "Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability Ogt -Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability," *Mol Cell Biol*, vol. 24, no. 4, pp. 1680–1690, 2004.

[47] S. Förster, A. S. Welleford, J. C. Triplett, R. Sultana, B. Schmitz, and D. A. Butterfield, "Increased O-GlcNAc levels correlate with decreased O-GlcNAcase levels in Alzheimer disease brain," *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1842, no. 9, pp. 1333–1339, Sep. 2014.

[48] S. A. Yuzwa *et al.*, "Pharmacological inhibition of O-GlcNAcase (OGA) prevents cognitive decline and amyloid plaque formation in bigenic tau/APP mutant mice," *Mol. Neurodegener.*, vol. 9, p. 42, 2014.

[49] C. J. Valvona, H. L. Fillmore, P. B. Nunn, and G. J. Pilkington, "The Regulation and Function of Lactate Dehydrogenase A : Therapeutic Potential in Brain Tumor," 2015.

[50] I. Teber, F. Nagano, K. Bilbilis, and A. Barnekow, “Rab6 interacts with the mint3 adaptor protein,” vol. 386, no. July, pp. 671–677, 2005.

[51] G. Krause, L. Winkler, S. L. Mueller, R. F. Haseloff, J. Piontek, and I. E. Blasig, “Structure and function of claudins,” vol. 1778, pp. 631–645, 2008.

[52] A. J. Sethi, R. M. Wikramanayake, R. C. Angerer, R. C. Range, and L. M. Angerer, “Microbes and Alzheimer’s Disease,” vol. 335, no. 6068, pp. 590–593, 2016.

[53] D. K. V. Kumar *et al.*, “Amyloid- peptide protects against microbial infection in mouse and worm models of Alzheimers disease,” *Sci. Transl. Med.*, vol. 8, no. 340, p. 340ra72-340ra72, 2016.

[54] R. J. O’Brien and P. C. Wong, “Amyloid Precursor Protein Processing and Alzheimer’s Disease,” *Annu. Rev. Neurosci.*, vol. 34, no. 1, pp. 185–204, Jul. 2011.

[55] L. Hong, H.-C. Huang, and Z.-F. Jiang, “Relationship between amyloid-beta and the ubiquitin–proteasome system in Alzheimer’s disease,” *Neurol. Res.*, vol. 36, no. 3, pp. 276–282, 2014.

[56] S. Mukherjee *et al.*, “Systems biology approach to late-onset Alzheimer’s disease genome-wide association study identifies novel candidate genes validated using brain expression data and Caenorhabditis elegans experiments,” *Alzheimer’s Dement.*, no. February, pp. 1–10, 2017.

[57] L. Lee, M. Sakurai, S. Matsuzaki, O. Arancio, and P. Fraser, “SUMO and alzheimer’s disease,” *NeuroMolecular Med.*, vol. 15, no. 4, pp. 720–736, 2013.

[58] A. Grupe *et al.*, “Evidence for novel susceptibility genes for late-onset Alzheimer’s disease from a genome-wide association study of putative functional variants,” *Hum. Mol. Genet.*, vol. 16, no. 8, pp. 865–873, 2007.

[59] V. Dorval and P. E. Fraser, “Small ubiquitin-like modifier (SUMO) modification of natively unfolded proteins tau and α-synuclein,” *J. Biol. Chem.*, vol. 281, no. 15, pp. 9919–9924, 2006.

[60] N. C. Berchtold, P. D. Coleman, D. H. Cribbs, J. Rogers, D. L. Gillen, and C. W. Cotman, “Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer’s disease,” *Neurobiol. Aging*, vol. 34, no. 6, pp. 1653–1661, 2013.

[61] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, “Incipient Alzheimer’s disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 7, pp. 2173–2178, 2004.

[62] F. Zou *et al.*, “Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants,” *PLoS Genet.*, vol. 8, no. 6, 2012.

[63] M. Wang *et al.*, “Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer’s disease,” *Genome Med.*, vol. 8, no. 1, pp. 1–21, 2016.