Identification and Analysis of Transcriptomic Contributors to the Development and Onset of Alzheimer's Disease

Abstract

In the last decade, research into the genetics of Alzheimer's Disease has predominantly focused on the Amyloid Cascade Hypothesis. This study explored the genomic transcripts of 9 AD patients and 8 control patients, determining and comparing the differential expression of genes between the two groups to draw conclusions about other genes which have an impact on the development and onset of AD. The study found that genes relating to Interferon signaling and regulation, especially STAT1, STAT2, and HLA-DQB, are most significantly differentially expressed in AD patients, to a greater degree than the Amyloid Precursor Protein (APP). However, limitations regarding access to sample count means that further research is required to confirm these findings.

Significance of the Research

Alzheimer's disease (henceforth referred to as "AD") is a neurodegenerative disease which is responsible for over 60% of dementia cases in the developed world. From 2000 to 2013, every major fatal disease group (cancers, cardiovascular diseases, viral infections, etc.) experienced a significant decline in case numbers throughout the United States and Canada, with the sole exception of neurodegenerative diseases which, in the case of AD, experienced a 71% increase in case numbers over the 13-year span (Alzheimer's Foundation, 2016). AD has become an epidemic, and at a time where human life expectancies have nearly doubled from 100 years ago, it has become critical to maintain the quality of life for the aging population. As AD is the primary cause of years of life lost for individuals above 60 in developed nations (Nichols et al., 2019), research into the causes and potential treatments is a critical step in eliminating the harm caused by AD.

Literature Review

The past two decades of AD research has primarily focused on the "Amyloid Cascade Hypothesis", which has hypothesized that the build-up of Amyloid Beta plaques in the brain as the mechanism responsible for AD onset. The development and history of this line of inquiry has been explored by Tanzi and Bertram (Tanzi & Bertram, 2005), and includes the identification of the links

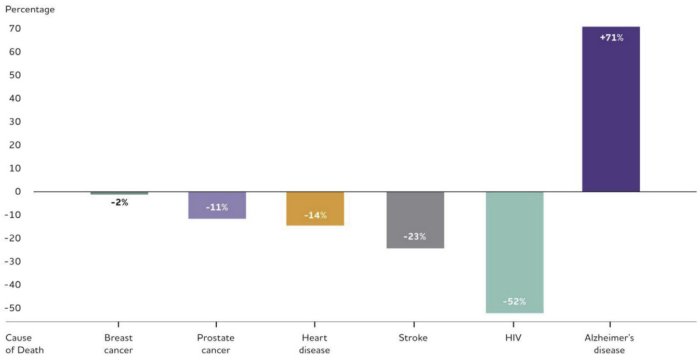

Figure 1: While most diseases have experienced a decline in mortality in the last 2 decades, Alzheimer's has become more prevalent than ever. (Alzheimer's Foundation, 2016)

between Down Syndrome (DS) and AD. The paper explores other investigations into the "Amyloid Cascade Hypothesis", identifying developments in AD research from 1985 to 2005, as well as areas of potential future investigation, such as the impact of bioinformatics and its implications for neurodegenerative disease research. This paper is significant as it produces a direct explanation for the onset of AD, as well as opening a new area of research for mitigation and treatment. [redacted] as explored by [redacted] who postulates that this increased risk can be attributed to certain genes on chromosome 21 [redacted] Their paper discusses the implications and results of the [redacted], a government funded, multi-institutional study which evaluated [redacted] individuals for signs of AD onset. Their results closely aligned with the Amyloid Cascade Hypothesis, identifying the Amyloid Precursor Protein (APP) gene [redacted] patients with AD but weakly expressed in [redacted] patients without AD. This study aligns with not only Tanzi and Bertram, but also the widely accepted explanation for the disease. Aligning with this theory is Jiang, Qian and Shengdi's paper on Oxidative stress as a cause of AD, which identified that the build-up of Amyloid-Beta and Tau proteins was tightly associated with the neural damage caused by AD, affecting signaling pathways such as Protein Kinase B and other extracellular protein Kinases (Jiang et al., 2016). Results were determined by analyzing subjects with AD (primarily the quantity of free

radicals in the form of reactive oxygen species), and observing the effects of oxidative damage to the lipids, proteins, and DNA in the central nervous system. While the paper did not identify the mechanisms by which the protein plaques affected neural pathways, it confirmed the damage vector of Amyloid plaques, allowing for investigations into future treatment methods. While the Amyloid Cascade Hypothesis has strong backing from large portions of the neuroscience community, a paper by [redacted] argued that the lack of therapies and treatments for dementia has stagnated as a result of the [redacted] Exploring the data from [redacted], the team predicated that because the Amyloid Hypothesis was so widely accepted, researchers are less willing to explore other potential avenues [redacted]. [redacted] paper proposes that exploration into other avenues should be attempted.

Research Question

Most aspects of the Amyloid Cascade Hypothesis and the impact of APP have either been exhausted as avenues of research, or are under investigation by government-backed or independently funded institutions. As a consequence, this paper is focused on identifying alternative genes which affect and/or contribute to Alzheimer's onset, as well as the manner in which these genes affect the onset of AD.

> Which genes other than APP are significantly differentially expressed in AD patients?

Hypothesis

Based on the near-universal agreement within existing scientific literature, I expect to find that if certain genes are consistently found to be significantly different between diseased and healthy sample groups, these genes will be related to the development and onset of AD.

Method

Participants

[redacted]

Data Import and Read Quality Analysis

[redacted] in the form of GZip compressed FastQSanger files, were imported [redacted] to the [redacted] bioinformatics server [redacted] Because sequencing was performed by [redacted], confirming the quality of the data is essential to the integrity of the research. To do this, each forward and reverse read file was analyzed using FastQC (Andrews, 2018), using default parameters (no contaminant list, no adapter list, no submodules, grouping of bases >50bp enabled).

Trimmomatic (Bolger et al., 2014) was used to remove low-quality reads and sequencing adaptors from each read. Trimmomatic was run using default parameters, except the initial ILLUMINACLIP step enabled, and TruSeq3 (paired-

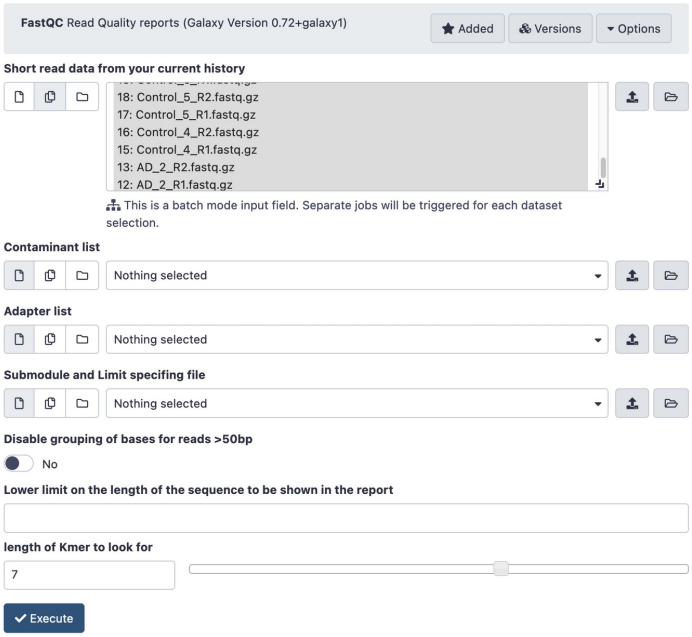

Figure 2: FastQC Setup in Galaxy

ended for MiSeq and HiSeq) adapter sequences. Low quality reads (average quality <20% across 4 reads) were removed, as well as null reads.

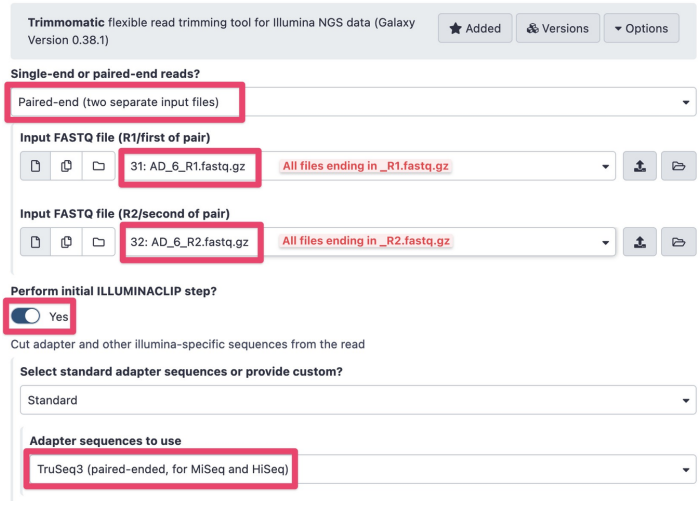

Figure 3: Changed settings for Trimmomatic.

Read Alignment

Trimmed sequencing reads from Trimmomatic were aligned to the human reference genome hg38 (Kent et al., 2002) using HISAT2 (Kim et al., 2015). The reference genome should be set to "Human Dec. 2013 (GRCh36/hg38) (hg38)". The paired library mode should be enabled, with the Trimmomatic "R1 paired" files being selected for FASTA/Q file #1, and the "R2 paired" files being selected for FASTA/Q file #2. In summary options, the alignment summary was set to print to file, and in advanced options, the GTF file with known splice sites was set to "gencode.v36.annotation.gtf.gz".

Identifying Gene Expression

The mapped gene sequences from HISAT2 were then imported into featureCounts (Liao et al., 2013), a tool which measures Gene expression in RNA-seq experiments. The reports from featureCounts provided the abundance and expression levels of each gene. All of the HISAT2 output files were selected as alignment files, and the built-in hg38 genome was used as the annotation file.

EdgeR (Robinson et al., 2009) was used to compare the differential expression of genes between the AD samples and the control samples through the featureCounts and HISAT2 outputs. By identifying the Trimmomatic inputs for each HISAT2 output, a list should be made of the featureCounts names for the control and AD sample data. In EdgeR, group 1 was labelled "AD", while group 2 was labelled "Control". Under Contrast, the contrast of interest was set to "AD-Control".

Identifying Named Genes for Simpler Processing

As EdgeR outputs the numerical IDs of genes, AnnotatemyIDs was used to convert the numerical gene IDs produced into a gene symbol format for readability and ease of use. Setting the following properties:

FileHasHeader: true

OutputColumns: SYMBOL

Allowed the software to identify the genes and label them.

Analysis of Data

EnrichR provided a set of ontologies for the genes and gene sequences of significance of p<0.05 (Chen et al., 2013). The ontologies selected derived data from WikiPathways (Martens et al., 2020) and [redacted]. From the list of differentially expressed genes, those with a p-value greater than 0.05 were discarded.

Results

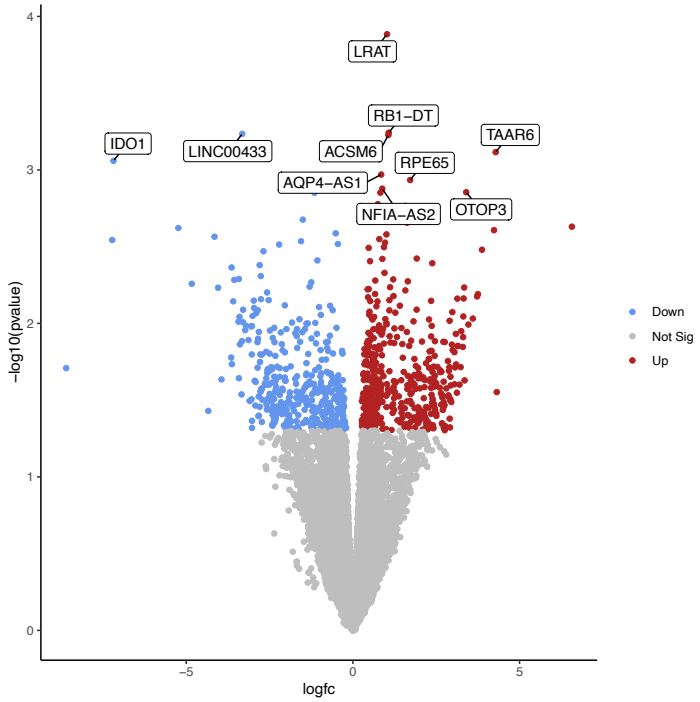

Figure 4: Volcano Plot of Differentially Expressed Genes. Top 10 most significant genes are labeled.

The list of differentially expressed genes was sorted by fold change (highest to lowest), and those with the most significant fold changes were determined to warrant further research.

The most down-regulated genes of the sample were CXCL9, IDO1, [redacted], KRT16P1, and LINC00208. The most up-regulated genes of

the sample were LINC02119, LINC00558, CT45A1, TAAR6, and [redacted]

SYMBOL	logFC	logCPM	F	PValue	FDR
CXCL9	-7.22	3.52	12.41	0.0	1
IDO1	-7.18	-0.21	15.64	0.0	1
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
KRT16P1	-4.83	-3.0	9.62	0.01	1
LINC00208	-4.34	-4.09	5.45	0.04	1
LINC02119	3.84	-4.07	15.11	0.0	1
LINC00558	4.08	-3.98	15.25	0.0	1
CT45A1	4.23	-3.95	13.79	0.0	1
TAAR6	4.28	-3.92	18.71	0.0	1
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Table 1: Summary of the most up- and down-regulated genes. The linked dataset contains the full list of significantly differentially expressed genes.

Discussion

Down-Regulated Genes of Note

CXCL9 has been identified in numerous studies (Pagoni et al., 2020; Koper et al., 2018) to be able to induce the activation of extracellular signal-regulated kinases (such as ERK1/2), resulting in the expression of inflammatory mediators in the regions of the brain affected by AD. The findings regarding its regulation and expression in the sample data appear to be in agreement with existing literature.

IDO1 is partially responsible for the regulation of the tryptophan catabolic process to kynurenine, which is known to play a role in the inflammatory response observed in AD patients (Gong et al., 2011). The heavy down-regulation of this gene may be a contributing factor of an individual's risk of Alzheimer's.

While [redacted] does not have existing literature presenting a link between it and Alzheimer's, individuals with [redacted] have been identified as being more resistant to [redacted], which causes this gene to be up-regulated [redacted]

[redacted]. Since the samples [redacted], it is highly likely [redacted]

KRT16P1 is a pseudogene, referencing positions 16000000-24000000 on chromosome 17. This region contains the gene RASD1, which is responsible for the regulation of Dexamethasone, a chemical which reduces inflammation responses in neural tissue (Wie et al., 2015). The down-regulation of this region may have resulted in inhibited production of Dexamethasone, reducing an individual's resistance to Alzheimer's, however, this conclusion would require further direct examination.

LINC00208 seems to be unrelated to Alzheimer's, with no tangential association found in existing literature. This gene is related to mean arterial pressure, myeloid white cell count, and blood pressure, none of which are directly related to AD. However, further research should be performed to ensure that there is no other connection between this gene and AD.

References

2016 Alzheimer's disease facts and figures. (2016). *Alzheimer's & Dementia*, *12*(4), 459–509. https://doi.org/10.1016/j.jalz.2016.03.001

Global regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. (2019). *The Lancet Neurology*, *18*(1), 88–106. https://doi.org/10.1016/s1474-4422(18)30403-4

Twenty Years of the Alzheimer's Disease Amyloid Hypothesis: A Genetic Perspective. (2005). *Cell*, *120*(4), 545–555. https://doi.org/10.1016/j.cell.2005.02.008

The Search for Biomarkers of Alzheimer's Disease in Down Syndrome. (2020). *American Journal on Intellectual and Developmental Disabilities*, *125*(2), 97–99. https://doi.org/10.1352/1944-7558-125.2.97

Oxidative stress: A major pathogenesis and potential therapeutic target of antioxidative agents in Parkinson's disease and Alzheimer's disease. (2016).

Progress in Neurobiology, *147*, 1–19. https://doi.org/10.1016/j.pneurobio.2016.07.005

Why has therapy development for dementia failed in the last two decades?. (2015). *Alzheimers & Dementia*, *12*(1), 60–64. https://doi.org/10.1016/j.jalz.2015.12.003

Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. (2016). *ELife*, *5*. https://doi.org/10.7554/elife.10421

The Clinical Dementia Rating (CDR). (1993). *Neurology*, *43*(11), 2412.2–2412–a. https://doi.org/10.1212/wnl.43.11.2412-a

The Galaxy platform for accessible reproducible and collaborative biomedical analyses: 2018 update. (2018). *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

FastQC: A Quality Control tool for High Throughput Sequence Data. (2018). http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimmomatic: a flexible trimmer for Illumina sequence data. (2014). *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

The Human Genome Browser at UCSC. (2002). *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

HISAT: a fast spliced aligner with low memory requirements. (2015). *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. (2013). *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. (2009). *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. (2013). *BMC Bioinformatics*, *14*(1), 128. https://doi.org/10.1186/1471-2105-14-128

WikiPathways: connecting communities. (2020). *Nucleic Acids Research*, *49*(D1), D613–D621. https://doi.org/10.1093/nar/gkaa1024

Gene Ontology: tool for the unification of biology. (2000). *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

The Gene Ontology resource: enriching a GOld mine. (2020). *Nucleic Acids Research*, *49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Causal effects of circulating cytokine concentrations on risk of Alzheimer's disease: A bidirectional two-sample Mendelian randomization study. (2020). https://doi.org/10.1101/2020.11.18.20232629

CXCL9 CXCL10, CXCL11, and their receptor (CXCR3) in neuroinflammation and neurodegeneration. (2018). *Advances in Clinical and Experimental Medicine*, *27*(6), 849–856. https://doi.org/10.17219/acem/68846

Targeting the kynurenine pathway as a potential strategy to prevent and treat Alzheimer's disease. (2011). *Medical Hypotheses*, *77*(3), 383–385. https://doi.org/10.1016/j.mehy.2011.05.022

Galectin-1 Deactivates Classically Activated Microglia and Protects from Inflammation-Induced Neurodegeneration. (2012). *Immunity*, *37*(2), 249–263. https://doi.org/10.1016/j.immuni.2012.05.023

The Roles of Rasd1 small G proteins and leptin in the activation of TRPC4 transient receptor potential channels. (2015). *Channels*, *9*(4), 186–195. https://doi.org/10.1080/19336950.2015.1058454

2016 Alzheimer's disease facts and figures. (2016). *Alzheimers & Dementia*, *12*(4), 459–509. https://doi.org/10.1016/j.jalz.2016.03.001

Boyer, P. D. (1998). Energy Life, and ATP (Nobel Lecture). *Angewandte Chemie International Edition*, *37*(17), 2296–2307. https:

//doi.org/10.1002/(sici)1521-3773(19980918)37:17<2296::aid-anie2296>3.0.co;2-w

Rasche, H. (2019). UseGalaxy.eu: Community, Training, Infrastructure, and Users. *F1000Research*. https://doi.org/10.7490/f1000research.1117097.1

Zhou, J., Xue, Z., Du, Z., Melese, T., & Boyer, P. D. (1988). Relationship of tightly bound ADP and ATP to control and catalysis by chloroplast ATP synthase. *Biochemistry*, *27*(14), 5129–5135. https://doi.org/10.1021/bi00414a027

Zinszer, K., Morrison, K., Verma, A., & Brownstein, J. S. (2017). Spatial Determinants of Ebola Virus Disease Risk for the West African Epidemic.. *PLoS Curr*, *9*.

Andrews, Simon. (2018). *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Institute for Health Metrics and Evaluation. (2021). *Global Health Data Exchange Visualisation Hub*. University of Washington. https://perma.cc/HRM3-62N9