Transcriptome analysis of zebrafish genetic models to reveal early molecular drivers of Alzheimer's disease

Nhi Hin

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics

School of Biological Sciences

The University of Adelaide

September 2020

Table of Contents

List of included publications

1. **Hin N**, Newman M, Kaslin J, Douek AM, Lumsden A, Nik SHM, et al. Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human *PSEN2*. *PLOS One*. 2020;15:e0227258. doi:10.1371/journal.pone.0227258.

2. Newman M, **Hin N**, Pederson S, Lardelli M. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish *presenilin 1* gene implies effects on energy production. *Molecular Brain*. 2019;12:43. doi:10.1186/s13041-019-0467-y.

3. **Hin N**, Newman M, Pederson SM, Lardelli MM. Iron Responsive Element (IRE)-mediated responses to iron dyshomeostasis in Alzheimer's disease. *bioRxiv*. 2020;:2020.05.01.071498. doi: 10.1101/2020.05.01.071498.

4. **Hin N**, Newman M, Pederson S, Lardelli M. Adaptive elastic-net sparse PCA for robust cross-species, cross-platform analysis of complex gene expression data in Alzheimer's disease. 2020.

Abstract

Alzheimer's disease (AD) is a complex neurodegenerative disease that still evades effective treatment. Developing treatments to slow or prevent AD will require a detailed understanding of the early stages of AD at the molecular level. Unfortunately, the pathogenesis of AD progresses silently over decades and, in the case of genetically inherited familial AD, may involve subtle changes starting during young adulthood. The difficulties associated with studying human brains at young ages have meant that transcriptome analyses of accurate animal models are essential. This is the approach taken in the work here, which uses a data-driven, bioinformatics-led approach to analyse brain transcriptomes from knock-in zebrafish models of AD developed to resemble the genetic background of human AD.

In the Introduction, the importance of transcriptome analysis in understanding AD at the molecular level is explained. Chapters 2, 3 and 4 describe the first brain transcriptome analyses of two different knock-in zebrafish mutation models ($psen1^{K97fs}$ and $psen1^{Q96\_K97del}$) modelling different aspects of human AD. In both zebrafish models, young adult brains revealed notable transcriptome changes, including changes to immune/stress responses and energy metabolism respectively. Gene network and gene set analysis approaches revealed that some of these gene expression changes were preserved in human AD datasets, suggesting the validity and utility of the approach used. The work in Chapter 4 supported an important role for iron dyshomeostasis in AD across animal models and human AD and demonstrated, for the first time, the viability of detecting iron dyshomeostasis

changes relevant to AD at the transcriptional level. This was achieved using a unique computational approach to the definition of gene sets based on predicted Iron Responsive Elements. Lastly, Chapter 5 used an advanced dimension reduction method to integrate zebrafish and mouse AD model datasets with human familial and sporadic AD. This resulted in the first preliminary comparison of human familial and sporadic AD transcriptomes as well as confirmation that the brains of a knock-in familial AD mutation-like zebrafish model (*psen1*$^{Q96\_K97del}$) and the commonly used 5XFAD mouse model show extensive differences at the molecular level.

Overall, analysis of young adult zebrafish brains revealed potential molecular mechanisms relevant to early stages of human AD that were significantly preserved in human familial and sporadic AD datasets. This work demonstrates the value to our understanding of AD of a bioinformatics-led approach involving transcriptome analysis of knock-in zebrafish models.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Nhi Hin   Date: 23 September 2020

Acknowledgements

I am particularly indebted to my supervisors **Michael Lardelli**, **Stephen Pederson**, and **David Adelson** for their generous mentorship, guidance, and support.

Michael Lardelli and **Morgan Newman**, I cannot thank you enough for welcoming me into the lab during my undergraduate studies. I am extremely blessed to have had the opportunity to build my foundation as a researcher at the Alzheimer's Disease Genetics Laboratory. **Michael**, your thoughtfulness, perceptiveness, and integrity make you a supervisor that any student would be lucky to have. Thank you for encouraging me to try new things outside of my comfort zone and for guiding me through many difficulties – I owe much of my growth both as a person and researcher over these past few years to you.

Steve, you have not only been an incredible supervisor but also an exemplar of the statistical rigor, attention to detail, and creativity that is so important in bioinformatics research. Thank you for being so patient with me. I have been very lucky to build my foundation in bioinformatics and data analysis under your guidance.

Dave, thank you for introducing me to bioinformatics, teaching me the importance of an open mind, and for having confidence in my abilities even when my imposter syndrome was at its worst. I would like to express my heartfelt appreciation for our initial discussions which shaped the trajectory of my PhD research significantly.

Chapter 1   Introduction

Transcriptome analysis enables study of Alzheimer's disease at the molecular level

1   Molecular and genetic basis of Alzheimer's disease

The cognitive decline experienced during normal aging is accompanied by structural, functional and molecular changes in the human brain [1]. These changes are exacerbated in individuals with dementia, who experience markedly accelerated cognitive decline. A complex neurodegenerative disease known as Alzheimer's disease (AD) causes the majority (50–70%) of the >50 million dementia cases worldwide [2, 3]. Aging is the most significant risk factor for AD [2], but AD brains clearly differ from those undergoing normal, healthy aging. At the structural level, AD brains exhibit atrophy, astrogliosis, neuronal loss, synaptic loss and hypometabolism [4], while at the molecular level, they often display amyloid-beta (Aβ) protein aggregates in extracellular spaces and blood vessels [5–8], and aggregates of hyperphosphorylated tau proteins within neurons [8–11]. While tau aggregates contributes to neurodegeneration in other diseases including frontotemporal dementia, argyrophilic grain disease and Huntington's disease [12–14], the accumulation of Aβ appears to be a unique molecular feature of AD. Consequently, significant efforts have been made to elucidate the "amyloid hypothesis", describing the pathological role of Aβ in AD.

The production, processing and pathological role of Aβ has been studied in detail. **Figure 1** summarises the cellular production of Aβ and several of its proposed pathological roles in AD. Aβ peptide monomers of various lengths are formed when amyloid precursor proteins (encoded by *APP*) are sequentially cleaved by β-secretase and γ-secretase enzymes. The catalytic subunit of γ-secretase is a

presenilin-1 or presenilin-2 aspartyl protease encoded by *PSEN1* or *PSEN2* (Li et al. 2000). Familial forms of AD are mainly caused by heterozygous, autosomal dominant mutations in *PSEN1*, *PSEN2* or *APP*, which seem to favour increased accumulation of the more neurotoxic Aβ-42 species relative to the more neuroprotective Aβ-40 species [15–20]. Studying Aβ kinetics suggests that soluble Aβ-42 species (including monomers, dimers and oligomers) are more neurotoxic than Aβ-40 species because Aβ-42 species readily undergo nucleation-dependent polymerisation along neuronal cell surfaces [21, 22]. The polymerisation process itself appears to permeabilise cell membranes [23] and contribute to widespread disruption of signalling pathways [24, 25]. Soluble Aβ species that re-enter neurons can also induce tau phosphorylation [26], and interfere with diverse cellular processes to further exacerbate AD progression. It is also likely that loss of the neuroprotective functions normally performed by Aβ-40 species may further worsen AD; for example, Aβ-40 is more effective at binding and transporting metal ions across membranes compared to Aβ-42 [27]. Although many mechanisms of Aβ-mediated toxicity have been described, it is evident that excess Aβ *in vivo* is neurotoxic. Extraction of Aβ oligomers from human AD brains followed by direct injection into mice hippocampi impairs long-term potentiation, synaptic function, and memory, in addition to reducing dendritic spine density [28]. Likewise, repeated injections of soluble Aβ-42 oligomers into mice hippocampi induces neuronal loss, tau hyperphosphorylation and memory deficits [29]. Injection itself is unlikely to be responsible for these neurotoxic effects, as transgenic mouse lines producing high but physiologically relevant levels of soluble Aβ-42 species also display genotype and age-dependent reductions in dendritic spine density [30].
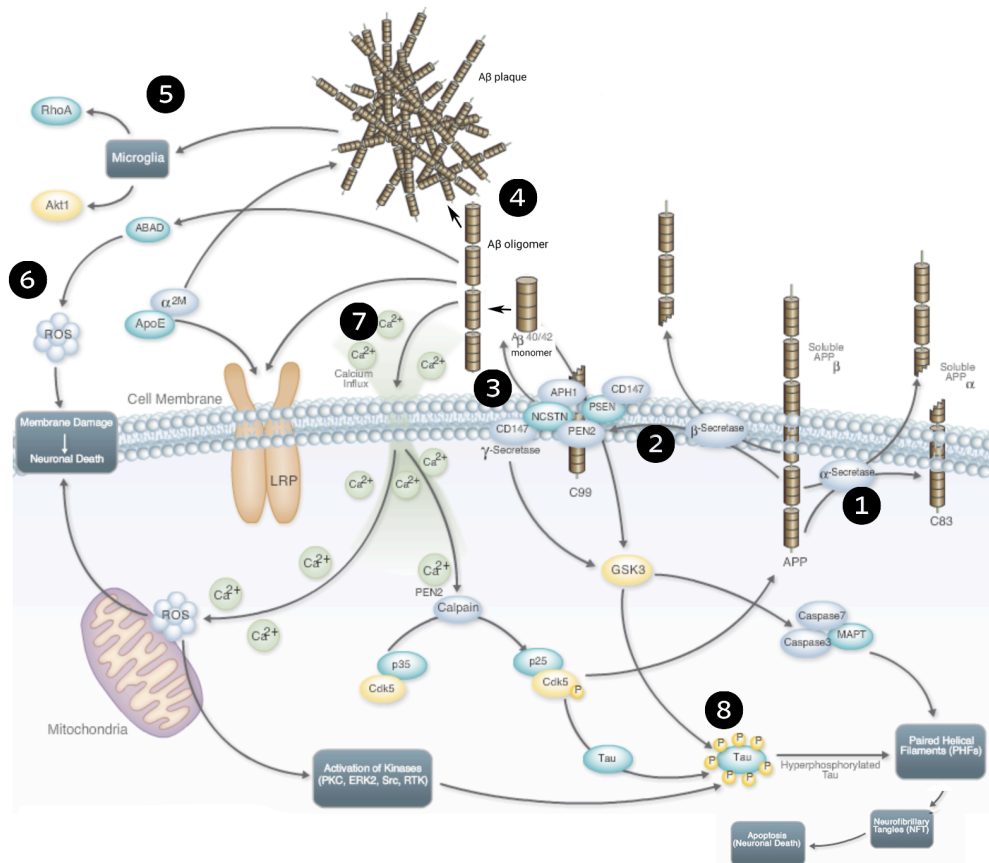
Figure 1. Cellular production of amyloid beta (Aβ) and its potential molecular mechanisms in Alzheimer's disease. Amyloid precursor protein (APP) is an integral membrane protein cleaved by two distinct pathways. (1) The non-Aβ-producing pathway involves cleavage of APP by α-secretase, producing a membrane-bound C83 fragment and soluble APPα fragment. (2) Alternatively, the pathway generating Aβ involves initial cleavage of APP by β-secretase, producing a membrane-bound C99 fragment and soluble APPβ fragment. (3) C99 is subsequently cleaved by γ-secretase to release Aβ monomers of various lengths (usually either 40 or 42 amino acids) into the extracellular space. Notably, presenilin 1 or presenilin 2 is the active protein subunit of γ-secretase and is encoded by the *PSEN1* or *PSEN2* gene. The Aβ-42 isoform is more prone to polymerisation than Aβ-40. (4) Polymerisation of Aβ forms Aβ oligomers, which aggregate to form plaques. Plaques act as reservoirs of Aβ oligomers which are neurotoxic. They can (5) activate microglia and astrocytes, resulting in neuroinflammation, (6) permeabilise plasma membranes and disrupt calcium ion homeostasis, and (7) induce generation of reactive oxygen species which damage cellular structures and impair cellular processes. A cellular cascade of events is believed to be initiated, which eventually triggers 8 phosphorylation of cytoskeletal tau proteins, formation of neurofibrillary tangles, and neuronal loss, leading to the clinical symptoms of Alzheimer's disease. Figure modified from Abcam [31].

Despite evidence supporting the neurotoxicity of aberrant Aβ accumulation and its contribution to familial AD, there is insufficient evidence to conclude that Aβ is the primary cause of all cases of AD, particularly in sporadic AD which represents over 95% of all AD cases [4]. Unlike familial AD, the etiology of sporadic AD is heterogenous and likely involves a combination of genetic and environmental factors. In contrast to familial AD genes, risk genes for sporadic AD are not necessarily related to amyloid precursor protein processing or Aβ accumulation. The most well-known genetic risk factor for sporadic AD is possession of the ε4 allele at *APOE* [32, 33]. *APOE* encodes an apolipoprotein involved in cholesterol transport in the brain. Although studies have been quick to implicate *APOE* in Aβ clearance [34–36], it is important to note that ε4 allele possession is also associated with diseases like athleroschlerosis, HIV, sleep apnea and ischemic cerebrovascular disease, where evidence supports a broader role for *APOE* in innate immunity through suppressing TNF secretion from inflammatory cells [37, 38]. Genome-wide association studies have since identified several other genes consistently associated with sporadic AD in diverse populations, including *CLU*, *CR1*, *PICALM*, *TREM2* and *CD33* [39–44]. Unfortunately, the functions of these genes are still primarily interpreted in terms of Aβ processing or clearance [45–49], although it is important to note that most do not play direct roles in Aβ processing and also participate in diverse processes like innate immunity, inflammatory processes, macroautophagy and cholesterol homeostasis in the brain [50–52]. Likewise, environmental risk factors for sporadic AD show strong association with those in varied diseases including cerebrovascular disease [53, 54], metabolic diseases like type II diabetes [55–57], and chronic pathogenic infection [58–60].

Several key lines of evidence collectively imply that aberrant Aβ accumulation may not initiate AD or contribute significantly to its pathogenesis. First, rather than being the initiator of AD, Aβ appears to have a physiological role in innate immunity. Aβ oligomerisation, which is typically viewed as a pathological event in AD, appears to be required for Aβ-mediated antimicrobial activity against gram-positive and gram-negative bacteria [61]. This response occurs *in vivo* with bacterial infection in the brain and seems to be evolutionarily conserved in organisms as diverse as nematodes and mice [62]. Association between infection and AD has been well-documented, with a greater-than-expected incidence of fungal and bacterial infections detected in post-mortem brain tissues of AD patients [58, 60]. As not all AD patients display brain infections, it is unclear whether Aβ is responding to a real or perceived pathogenic threat in AD. However, it is evident that Aβ accumulation should not be viewed as an isolated event in AD, and the causes of its accumulation should be investigated in more detail. Second, while it is evident that Aβ is correlated with AD, it may not be sufficient or necessary to account for all AD cases; most notably, a sizable number (15 – 31%) of cognitively normal individuals apparently display substantial Aβ deposits in their brains [63–65], while 16 – 25% individuals clinically diagnosed with mild to moderate AD display only low levels of soluble and insoluble Aβ [63, 66]. The existence of these individuals with Aβ pathology seemingly uncorrelated to their cognitive state raises concerns about the current use of Aβ as a diagnostic marker of AD. Lastly, no therapeutic aimed at reducing Aβ activity has improved patient outcomes to date. A long-term follow-up study with AD patients previously immunised with human Aβ-42 revealed that despite decreased mean Aβ load and plaque density in patients, AD symptoms worsened at the same

rate as the placebo group [67]. Similar results were obtained from recent Phase 3

clinical trials of two different antibodies capable of binding and reducing soluble

human Aβ, which demonstrated no differences in AD patient outcomes compared to

placebo groups [68, 69].

Collectively, these studies indicate that although Aβ is neurotoxic and tends to

accumulate in at least some portion of AD brains, it may not possess a primary role

in AD pathogenesis, underscoring the need to study other pathological mechanisms

of AD that precede both cognitive decline and Aβ accumulation.

2 Transcriptome analysis reveals the diversity of molecular pathology in Alzheimer's disease

While the amyloid hypothesis of AD is a major area of research, the evidence against Aβ as the sole cause of AD underscores the importance of investigating other potential causes or underlying factors of AD. Importantly, in humans, longitudinal studies reveal early structural and functional brain changes decades before clinical onset of familial or sporadic AD. For example, young adult (18-26 year-old) carriers of the familial AD E280A *PSEN1* mutation already display accelerated brain atrophy and increased soluble Aβ levels in blood plasma and cerebrospinal fluid, despite dementia tests indicating they are cognitively normal [70]. Similarly, young adult carriers of the ε4 allele of *APOE* (the major sporadic AD risk gene) consistently display lower glucose metabolism rates in several brain regions corresponding to those with lowered glucose metabolism in AD patients [71]. These studies are valuable as they demonstrate clear structural and functional changes in the brains of individuals long before onset of AD, presence of amyloid plaques, or any cognitive deficits.

The techniques used in these types of studies means that identifying the early cellular processes and molecular mechanisms driving these changes is beyond their scope. Instead, omics techniques that measure expression of genes (transcriptomics) or proteins (proteomics) are able to fill this need as they capture molecular changes without prior assumptions of disease etiology. This is especially

relevant for complex diseases such as AD where much of the underlying molecular pathology (aside from Aβ biology) is still not well characterised.

The transcriptome of a particular biological sample describes its global RNA transcript expression at a particular time. Two techniques are commonly used to collect transcriptome data from biological samples: cDNA microarrays, and whole-transcriptome sequencing (RNA-seq). The key differences between these methods is shown in **Figure 2**. Compared to microarrays, RNA-seq is more recent technology that allows for improved reproducibility as well as access to sequence-level information that facilitates observation of phenomena that are not readily detected by microarrays (e.g. splicing events, non-coding RNA quantification and discovery [72]). Comparing the transcriptomes of different conditions (e.g. disease state, time period, tissue type) allows changes in molecular pathways between conditions to be inferred.

Existing transcriptome analyses of brain tissue from post-mortem AD brains suggests vast dysregulation of genes with roles in diverse biological activities including membrane organisation, immune responses, inflammation, vesicle trafficking, synaptic transmission and plasticity, axon transport, neuron differentiation, protein homeostasis, cytoskeletal rearrangement, and metabolic processes (see summarised findings of relevant studies in **Table 1**). It is important to note that all gene expression studies of AD are performed on tissue samples taken from post-mortem AD brains representing very late stages of AD. During these late stages, extensive disruption to normal brain function and structure is expected. A

major limitation of this is that it is difficult to discern earlier, causative events in AD pathogenesis from later changes arising as symptoms. From present transcriptome analyses, it is evident that dysregulation of diverse processes has occurred, but their order and contribution to AD remains uncertain.
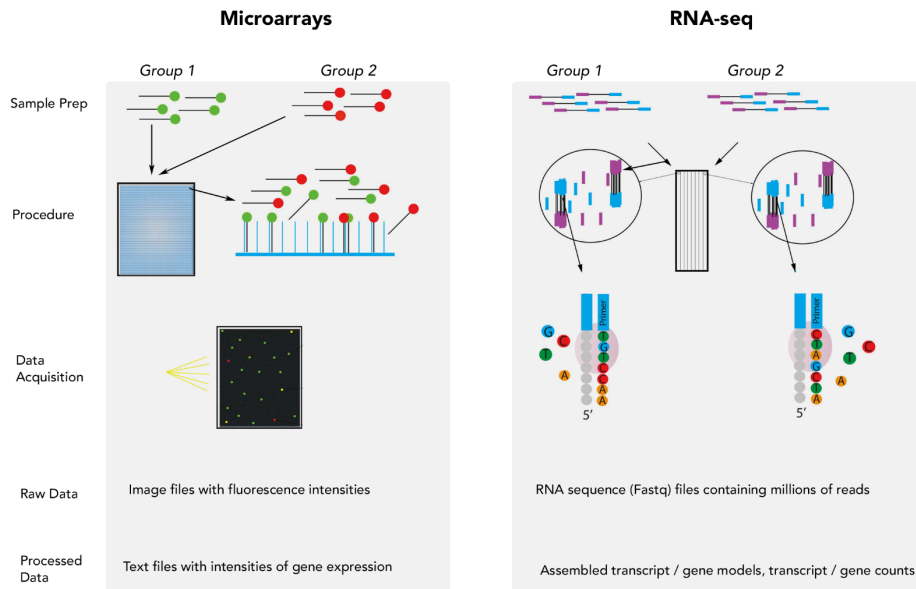
Figure 2. Comparison of Affymetrix microarray and RNA-seq technologies for transcriptomic data collection. Microarrays and RNA-seq sample preparation initially involves the extraction of RNA from a sample, followed by fragmentation and conversion to complementary DNA (cDNA) by reverse transcription. For microarrays, cDNA samples are labelled with fluorescent probes, while for RNA-seq, the RNA samples are ligated with adapters. In Affymetrix microarrays, probe sequences complimentary to target RNA sequences are placed at pre-defined locations on the microarray. For the RNA-seq system, a flow cell and sequencing-by-synthesis method is used to generate millions of sequences per sample. The raw data from a microarray workflow are CEL files containing continuous intensities corresponding to gene expression for each probe, while for RNA-seq, the raw data are the reads corresponding to RNA sequences. By mapping RNA-seq reads to a reference genome, discrete gene counts can be obtained [72].

Table 1. Selected microarray and RNA-seq studies focusing on brain transcriptome analysis of human patients with Alzheimer's disease (AD).

Technology	Brain Regions	Samples	Accession	Summarised Results
Microarray; Affymetrix Genechip	19 brain regions from each sample including prefrontal cortex, temporal lobe, amygdala, hippocampus.	125 AD patients	GEO: GSE84422 PMID: 27799057	1558 co-expressed gene network modules in AD, with enriched processes including actin cytoskeleton and axon transport [73].
Microarray; Illumina Human HT-12 V3.0 expression beadchip	Brain (also skin, muscle, blood)	>250 AD patients, >250 MCI patients, age-matched controls	GEO: GSE63060 PMID: 26343147	At least 150 gene markers of cognitive decline in MCI (Mild Cognitive Impairment) and AD implicated in aging-related processes [74].
Microarray; Affymetrix Human Genome U133 Plus 2.0 Array	hippocampus, entorhinal cortex, superior frontal cortex, post-central gyrus	55 controls, 26 AD patients	GEO: GSE48350 PMID: 23273601, 18832152	340 DE genes in AD, associated with synaptic vesicle trafficking, SNARE complex, neurotransmitter receptors, cell adhesion regulating synaptic stability, and neuromodulatory systems [75, 76].
Microarray; Affymetrix Human Exon 1.0 ST Array	temporal cortex	8 controls, 8 AD patients	GEO: GSE37264 DOI: 10.1016/j.gdata. 2014.09.002	22 genes with evidence of differential splicing in AD; one gene, FynT, was linked to an astrocytic inflammatory response [77, 78].
Microarray; Affymetrix Human Gene 1.1 ST Array	posterior cingulate	7 controls, 7 early-onset AD patients, 7 *PSEN1* familial AD patients	ArrayExpress: E-GEOD-39420 PMID: 23369545	3183-3350 DE genes in AD associated with calcium and metal ion signalling, neuroactive ligand-receptor interactions, axon guidance, and long-term potentiation [79].
Microarray; Sentrix HumanRef-8 Expression BeadChip	cortex	187 controls, 176 AD patients	GEO: GSE15222 PMID: 19361613	Dysregulation in processes including cholinergic transmission and ribosomal protein transport [80].
Microarray; Affymetrix Human Genome U133A Array	frontal cortex	8 controls, 6 AD patients	GEO: GSE12685 PMID: 19295912	1838 DE genes in AD; increased expression of genes regulating translation in the synapse [81].
Microarray; Affymetrix Human Genome U133 Plus 2.0 Array	entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate, superior frontal gyrus, primary visual cortex	74 controls, 76 AD patients	GEO: GSE5281 PMID: 17077275	At least 50-100 DE genes involved in synaptic plasticity, neuronal repair, cholinergic synaptic transmission, and phospholipase C activating pathway [82].

RNA-seq; Illumina HiSeq 2000	prefrontal cortex	6 controls, 6 late-stage AD patients	GEO: GSE48552 PMID: 24014289	32 differentially expressed micro RNAs (miRNAs), including miR-132-3p, which interacts with the influential FOXO1a transcription factor [83].
RNA-seq; Illumina Genome Analyser II	total brain, frontal lobe, temporal lobe	23 controls, 1 AD patient	SRA: SRA027308.2 PMID: 21283692	Differential splicing patterns and promoter usage of the *APOE* gene in AD compared to controls; enrichment in protein localisation, vesicle transport, phosphate metabolic processes [84].
RNA-seq; Illumina HiSeq 2000	dorsolateral prefrontal cortex	8 controls, 9 late-stage AD patients	GEO: GSE53697 PMID: 26894958	Identification of 8681 binding sites on transcripts for the Neuronal ELAV-like RNA binding proteins; enrichment of stress response genes in AD [85].
RNA-seq; Illumina HiSeq 2000	hippocampus	4 controls, 4 late-onset AD patients	GEO: GSE67333 ArrayExpress: E-GEOD-67333 PMID: 26402107	Differentially expressed genes in AD, including 143 protein-coding genes, 90 lincRNAs, 31 antisense transcripts. Enriched pathways include nerve impulse transmission, neuropeptide signalling, Aβ clearance [86].
RNA-seq; Illumina NextSeq 500	whole brain	1 control, 1 AD patient	GEO: GSE85075 PMID: 28127595	125 upregulated piRNAs responsible for downregulation of 1923 genes in AD. Enriched pathways include NMDA receptor trafficking and axonal transport [87].
RNA-seq; Illumina HiSeq 2000	whole brain	1 control (2 time points), 1 AD (2 time points)	GEO: GSE65159 PMID: 25693568	Upregulation of immune response genes, downregulation of synaptic plasticity genes [88].

3 Limitations in the previous use of animal models for studying Alzheimer's disease

Post-mortem AD brains display extensive structural and molecular pathology, so it is clear that studying young adult brains predisposed to AD would give more insight into earlier disease pathogenesis. However, it is evidently not feasible to extract brain tissue from the living brains of young adults for transcriptome analysis. Because of this, animal models have been developed to investigate AD pathogenesis at any age.

Early mouse models of AD possessed single endogenous knock-in mutations in mouse orthologs of familial AD-causing genes like *PSEN1* or *APP*. This was done in an attempt to model the genetic state of familial AD as closely as possible. Unfortunately, further research with these models was largely abandoned when it was discovered that these mouse models only recapitulated partial AD phenotypes, with mutant *APP* mice demonstrating substantial Aβ plaque formation but negligible neurodegeneration [89], and mutant *PSEN1* mice exhibiting neurodegeneration without Aβ plaques [90, 91]. Furthermore, these phenotypes were only observed in old (13-month-old) mutant mice, with young adult (6-month-old) mutant mice showing normal or only subtle phenotypes compared to age-matched, wildtype siblings. These findings were interpreted as a failure on the part of the models to fully capture human AD. At the time of these studies, transcriptome analysis was still in its infancy, so any potential changes to the brains of these models at the molecular level were not interrogated at all.

At the time, the common sentiment was that Aβ and tau aggregates comprised the characteristic molecular pathology causing AD, so the next mouse models that were developed prioritised Aβ and tau pathology over having a genetic background resembling human familial AD. To create models with extensive Aβ and tau pathology, transgenic mouse models were developed that possessed several human transgenes with multiple familial AD-linked mutations (e.g. 5XFAD [92] and 3XTG [93]). It is important to note that while these models phenotypically resemble human AD in terms of Aβ plaques and cognitive deficits being present, these characteristics are accomplished by overexpressing multiple mutant human familial AD genes [89, 94]. There have been concerns that this could induce unnatural artifacts and confound AD pathogenesis. For example, the order of appearance and degree of severity of key symptoms like neuronal loss, synaptic impairment and cognitive decline is unexpectedly variable between different models [95, 96]. These observations are consistent with results from a recent meta-analysis comparing the transcriptomes of five mouse models of AD and human AD. While human AD cases displayed similar gene expression changes that differed from cognitively-normal, age-matched controls, the transcriptomes of AD mouse models differed from both each other and human AD [97]. This lack of similarity between different mouse models as well as human AD raises concerns about whether these transgenic mouse models are able to give relevant insight into AD pathogenesis, especially when findings from these models are being applied to pharmaceutical interventions. Concerningly, there has been evidence that overexpression of human wild-type *APP* in both wildtype and transgenic mice with familial AD mutations itself induces

memory deficits at 8 months of age despite minimal Aβ-42 levels [98]. Similarly, overexpressing *APP* and *PSEN1* in at least one mouse model induces p25 generation independently of increased Aβ levels [99]. The results of the latter study are particularly concerning, given that the p25 generation detected was an artifact, while multiple studies have interpreted p25 generation in the 5xFAD mouse model as a mechanism of Aβ-mediated neurotoxicity in AD [92, 100, 101].

Collectively, these results suggest while overexpression of multiple human transgenes may result in animal phenotypes resembling AD, the underlying physiological state may not be informative for modelling human AD.

4 Recent knock-in zebrafish models of familial Alzheimer's disease

The concerns regarding transgenic models have seen an increase in popularity of knock-in mouse models in recent years [102, 103]. Aside from mouse models, zebrafish models have also been emerging in popularity as complementary model organisms of AD. They are optically transparent, undergo rapid development (reaching full maturity in 3-4 months), have comparable lifespans to mice (median lifespan of 31 months, maximum lifespan of 45 months [104], produce many offspring (50-200 embryos per week from one pair of parents), and are amenable to genetic manipulation with TALENs, CRISPR and morpholino antisense oligonucleotides [105–107]. Zebrafish share some advantages with other simpler vertebrate models (e.g. fly, nematode) that facilitate identification of highly conserved AD mechanisms [108–111], while possessing more similar brain physiology to humans [107, 112]. Approximately 70% of human genes have at least one zebrafish orthologue, so zebrafish models may be useful to independently verify genes implicated in AD from current mouse models [113]. Importantly, zebrafish express orthologues of human AD genes including *APP*, *PSEN1*, *PSEN2*, *APOE* and *SORL1* [105, 114], and their functions appear to be conserved. For example, zebrafish presenilin 1 possesses the critical aspartate residue required for cleavage of human APP to produce Aβ-42 [115]. Ultimately, applying transcriptome analysis to zebrafish models possessing endogenous familial AD mutations appears to be an unexplored and reasonable approach for investigating AD pathogenesis.

The Alzheimer's Disease Genetics Laboratory (ADGL) has developed the first zebrafish models possessing distinct endogenous familial AD mutations at the zebrafish *psen1* locus (orthologous to human *PSEN1*). Similar to humans with familial AD, these models express mutant *psen1* in a heterozygous state, resulting in physiologically relevant levels of presenilin 1 protein [116]. Like early knock-in mouse models of familial AD with endogenous familial AD mutations, these zebrafish models avoid confounding effects that might arise from overexpression of human transgenes [90, 91, 99]. Consequently, the underlying physiological state of these zebrafish models are relevant for studying human familial AD at the molecular level.

Although several zebrafish models have been developed by the ADGL, the Q96_K97del and K97fs zebrafish lines are important to introduce for the work done in the following manuscripts. In both of these lines, zebrafish are heterozygous for a different endogenous familial AD mutation at the *psen1* locus. The Q96_K97del mutation is a double-codon deletion that preserves the open-reading frame to result in full-length, mutant presenilin-1 protein, similar to mutations causing familial AD in humans. This model will be termed the fAD-mutation-like model. In contrast, the K97fs line mimics a human mutation, K115fs, found in the human *PSEN2* gene. In humans, this mutation results in increased expression of an alternative protein isoform, PS2V, that is also often seen to be increased in the brains of late onset, sporadic AD patients [117]. In previous studies by the ADGL, the PS2V isoform was shown to be evolutionarily conserved in zebrafish *psen1* rather than *psen2*, where the zebrafish PS2V-like protein still retained γ-secretase activity, Notch signalling activity, and suppression of the unfolded protein response similar to human PS2V

[117]. Consequently, rather than being a direct familial AD model, the K97fs line models the effects of forced PS2V-like expression in the brain. The location of both zebrafish *psen1* mutations on the orthologous human presenilin-1 protein are shown in **Figure 3**.
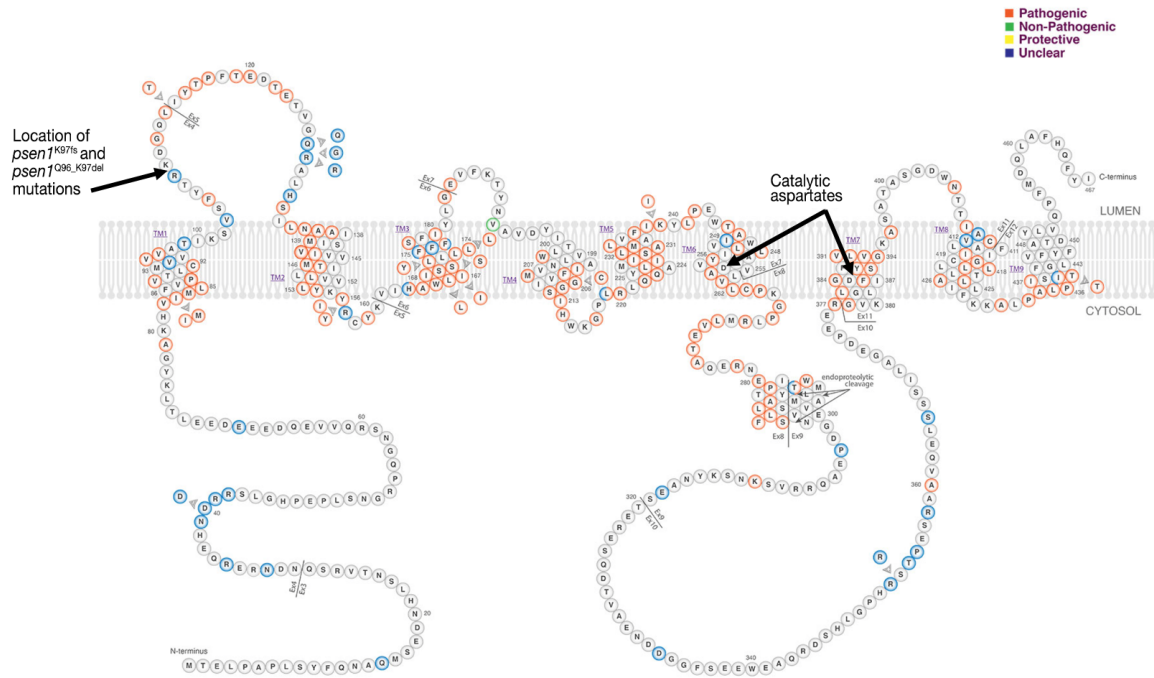
Figure 3. Human presenilin-1 protein. The location of the orthologous zebrafish *psen1* mutations Q96_K97del and K97fs in Exon 4 are indicated. Other known mutations in human presenilin-1 are also colour coded. Figure from AlzForum [118].
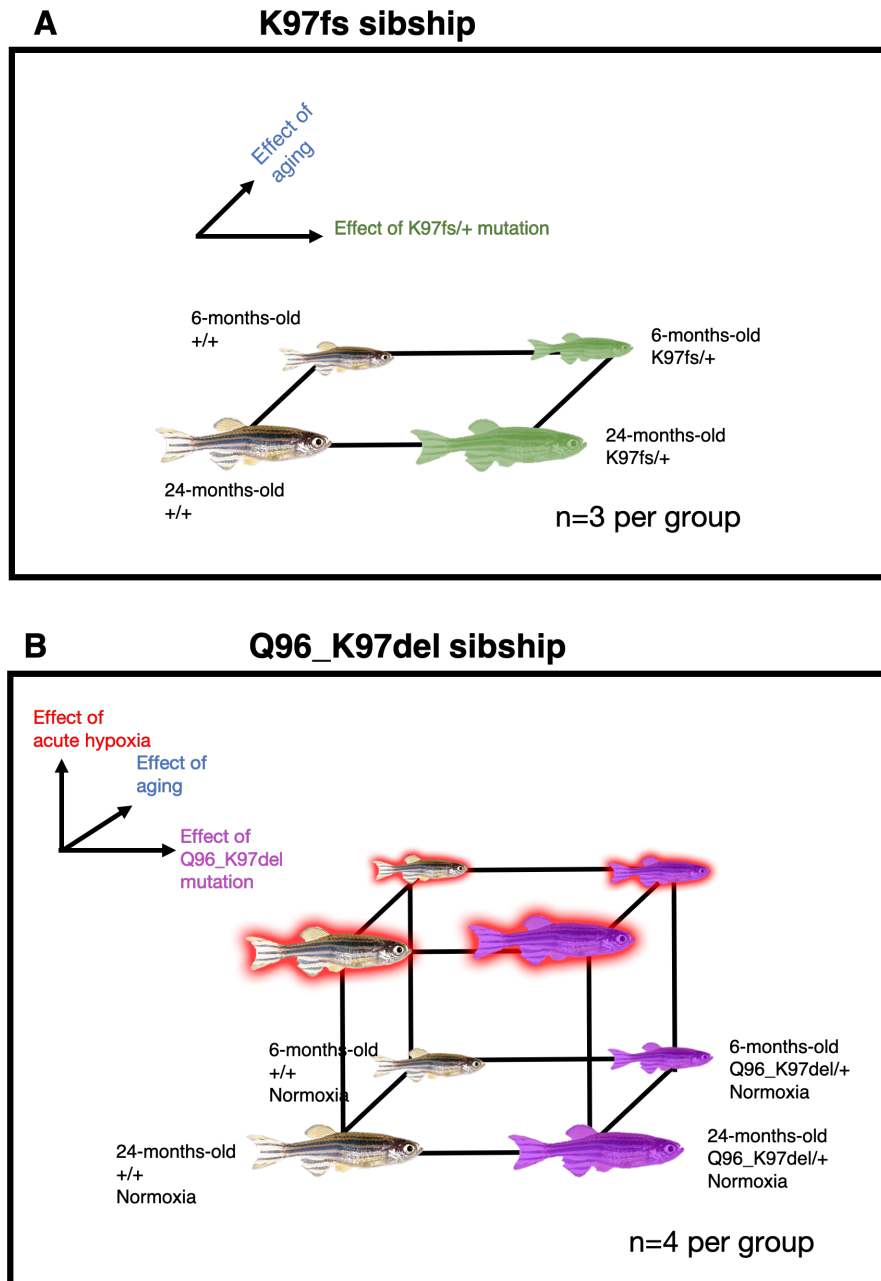
Figure 4. Summary of experimental design and statistical comparisons between groups in the (A) K97fs zebrafish sibship and (B) Q96_K97del zebrafish sibship. Each sample represents a whole zebrafish brain. Samples are derived from total RNA (K97fs sibship) or polyA enriched RNA (Q96_K97del sibship). 6-month-old zebrafish represent a young adult stage while 24-month-old zebrafish represent an aged, infertile stage. The lines connecting each biological group indicate the statistical pairwise comparisons to be performed to identify differentially expressed genes between the two groups. n is the number of biological replicates per group (n=3 for K97fs sibship and n=4 for Q96_K97del sibship).

RNA extracted from
whole zebrafish brains

Raw reads ----------> Quality control:
sequence quality,
AdapterRemoval: sequence length,
quality filtering and GC content,
removing adapters k-mer content,
duplication levels,
Trimmed reads ----------> overrepresented
sequences

HISAT2:
map reads to
reference zebrafish
salmon: genome assembly
transcript-level (GRCz10)
quantification

Aligned reads ---------->

featureCounts:
gene-level
quantification

Transcript
abundance
estimates Gene counts Quality control:
dimension reduction
(e.g. MDS) to identify
variation between
limma, voom: RNA-seq libraries and
generalised linear models any batch effects
for differential expression

Differential gene
expression analysis

Figure 5. RNA-seq data processing workflow. Raw paired-end reads from RNA-

seq libraries are processed with a range of Unix-based computational tools to

quantify gene expression in each library. At each step of the data processing

pipeline, quality control is performed to assess the impact of program parameters on

data quality. In this workflow, gene expression is quantified at both the transcript-

and gene-levels. Quantification data is typically used for differential gene expression

analysis.

5 Bioinformatics methods for transcriptome analysis

In the manuscripts later described, RNA-seq is used to capture the transcriptional states present in whole brains of knock-in mutant zebrafish models and compare them to those of healthy, wild type siblings. The comparisons of interest are summarised in the experimental design diagrams in **Figure 4**. In this section, the methods underlying RNA-seq data processing and analysis are briefly described.

A computational pipeline that can be used to process the raw RNA-seq data is shown in **Figure 5**. Raw paired-end or single-end RNA-seq reads are the initial input, which are then subject to adapter-trimming and quality-filtering using tools such as *AdapterRemoval* [119], alignment of trimmed reads to the zebrafish reference genome assembly with an alignment tool such as *HISAT2* [120] or *STAR* [121], and quantification of gene expression from the aligned reads as gene-level counts using *featureCounts* [122]. Alternatively, instead of aligning reads to the reference genome assembly, it is also possible to obtain transcript abundance estimates from trimmed reads using approaches such as *kallisto* [123] or *salmon* [124], where transcript abundance estimates can then be imported into R using tools like *tximport* [125] or *edgeR*'s "catchSalmon" or "catchKallisto" functions [126]. Importantly, the effect of each pre-processing step on the quality of the output data is assessed through a range of diagnostic plots (e.g. per-base read quality, GC content of reads, average read length, k-mer content in reads) which are implemented in *FastQC* [127] and *ngsReports* [128]. Depending on output data quality, different parameters or computational tools will be tested, and additional steps to improve data quality (e.g.

discarding reads mapping to multiple loci, filtering out PCR duplicates or other technical artifacts) may be implemented.

Gene-level counts are the primary output of the RNA-seq data pre-processing pipeline, and are typically analysed with various Bioconductor packages [129]. Bioconductor is an open-source, R-based collection of software packages focused on the statistical analysis of genomics data including microarrays, PCR, protein-arrays and RNA-seq. The aim of differential gene expression analysis is to determine genes that are differentially expressed between specific experimental conditions. Numerous Bioconductor software packages support differential expression analysis [130]. They differ in statistical approaches like normalisation methods and assumptions about the distribution of gene counts in RNA-seq libraries. In the work later described, the *limma* package is utilised to determine genes which are differentially expressed between conditions. *limma* implements weights to handle variable quality across genes and libraries, making the differential gene analysis potentially more robust to outlier samples and genes without the need for *ad hoc* data removal [131–133]. *limma* also has favourable power to detect differential expression compared to other methods, even for small sample sizes [125]. After differentially expressed genes are identified for desired comparisons, a range of analyses can be performed to gain further insight on biological changes corresponding to gene expression pattern changes. These include gene ontology over-representation analysis, promoter motif over-representation, gene set enrichment analysis, and gene co-expression network analysis are summarised in **Table 2**.

Table 2. Summary of analyses performed on a processed RNA-seq dataset.

Analysis	Input Data	Purpose	Example Methods
Differential gene expression analysis	Gene counts	Identify genes or which are differentially expressed between different conditions	*limma* [133, 134]; *edgeR* [135]
Gene ontology or KEGG over-representation analysis	List of genes differentially expressed between conditions; Access to gene ontology or KEGG terms	Identify over-represented biological activities or pathways in the differentially expressed genes	*limma*::goana and *limma*::kegga [133, 136]; *HOMER* [137]
Promoter motif over-representation analysis	List of genes differentially expressed between conditions; Access to known promoter motifs	Identify potential transcription factors regulating a set of differentially expressed genes	*HOMER* [137]; *MEME* Suite [138]
Gene set enrichment analysis (GSEA)	Gene counts	Identify overall changes in biological activities or pathways while taking into account expression of all genes	*GSEA* [139], *ROAST* [140], camera [141]
Co-expression network analysis	Gene counts	Infer relationships between genes; detect modules of co-regulated genes; infer biological functions of unknown genes	Weighted gene co-expression network analysis (*WGCNA*) [142]

6 Integration of multiple datasets

Many of the techniques described in **Table 2** are effective at providing biological insight into a single gene expression dataset. However, different techniques are required for the integrative analysis of multiple datasets simultaneously (known as data integration). There are many reasons why data integration would be useful, especially in the context of studying complex diseases such as AD. These include being able to take advantage of existing publicly available datasets from past studies; analysing different omics data types (e.g. proteomics and transcriptomics) from the same samples to give greater confidence into findings; and getting insight into whether findings from animal models also apply to human AD datasets.

Unfortunately, data integration is often not straightforward. Part of this is because the nature of high-throughput technologies like microarrays or RNA-seq depends on a complex system of reagents, hardware, and highly trained individuals to facilitate accurate measurements. Variance in these conditions (e.g. differences in RNA-seq library preparation or sequencing platform used) will cause the measurements to be affected by both technical and biological factors in ways that are not fully predictable or reproducible. This phenomenon is known as introducing batch effects into the data. Concerningly, when multiple gene expression datasets are analysed together without accounting for batch effects, the majority of differentially expressed genes tends to occur across batches rather than across biological conditions [143]. Many bioinformatics tools have been developed that use statistical approaches to account for and reduce batch effects within datasets (e.g. *RUVSeq* [144] or *surrogate variable analysis (sva)* [145]). In contrast, techniques for the integration of different

datasets must be capable of accounting for batch effects between datasets. These

techniques include network-based approaches, dimension reduction, gene set

enrichment analysis, and meta-analysis and are summarised in **Table 3**.

Table 3. Approaches for omics dataset integration.

Approach	Summary	Reduction of batch effects	Examples
Network-based approaches	Convert each dataset into graph to model interactions or correlations between genes and/or proteins. Datasets are represented as separate graphs (nodes, edges) which can then be compared using graph-based measures (e.g. degree, connectivity, subgraphs, clusters)	Filtering of genes / proteins before constructing the graph limits batch effects arising from low-expressed or low-variance genes / proteins; comparing graph-based measures (e.g. clusters) between datasets rather than individual measurements reduces impact of individual noisy genes.	Multi-omics interaction networks to infer relationships between genes and/or proteins [146]; co-expression network analysis for comparing multiple gene expression datasets [142]
Dimension reduction	Unsupervised method for summarising dataset into fewer latent variables capturing the majority of information. The small number of latent variables from multiple datasets can be simultaneously visualised and compared in an interpretable manner.	As batch effects are strong, they are typically isolated into separate latent variables away from biological sources of variation.	Non-negative matrix factorisation with latent variables representing data-driven sources of variation [147]; sparse Principal Component Analysis approaches like AES-PCA reduce the number of genes contributing to latent variables [148]
Meta-analysis	Each dataset is analysed separately, and significantly preserved changes in measurements (e.g. genes, proteins) between datasets is determined using post-hoc statistical meta-analysis methods.	Meta-analysis techniques can take direction of the change as well as dataset-specific weighting into account across multiple datasets, making the findings more robust to batch effects.	Concordance analysis for assessing whether gene expression changes are similar between datasets [149]; fixed-effects or random-effects model [150]

7 Conclusion

AD is a complex neurodegenerative disease that is still yet to respond to any pharmaceutical intervention. While current paradigms emphasise the importance of protein aggregates such as amyloid beta and tau, the molecular basis of the disease is more complicated than expected, highlighting the need for further studies to understand how the disease initially develops. While late stages of AD have been extensively characterised from studies of post-mortem AD brains, the earliest molecular changes in the brain driving incipient stages of AD are still unclear. In familial AD, these earliest changes may start during young adulthood, making it clearly unfeasible to study these early stages in humans and emphasising the necessity of accurate animal models of AD. Due to confounding artifacts associated with humanised, transgenic models of AD, knock-in models in particular appear more suitable for this purpose. Importantly, several large gaps remain in the field. For example, no transcriptome analysis has ever been performed on any knock-in model of AD, despite the earliest knock-in mouse models existing for decades. There have been no studies to assess whether the brain transcriptomes of knock-in models of AD might resemble human AD more than transgenic models. In addition, while zebrafish have several advantages compared to mice as complementary model organisms, comparisons of these AD models at the transcriptome level have not been performed. The earliest molecular events and cellular pathways in the brain driving early stages of familial AD remain unclear, along with potential differences in molecular mechanisms between different familial AD mutations. Importantly, familial and sporadic AD brains have not been subject to a single comparison at the transcriptional level, and this may be associated with the difficulties associated with

data integration. While high-throughput omics technologies and bioinformatics methods have been increasing in popularity, there remain significant gaps in their application to the context of AD. The following manuscripts aim to address these gaps through using bioinformatics-focused strategies to analyse novel zebrafish datasets from the ADGL and integrate findings with existing knowledge from publicly available mouse model and human AD datasets.

8 References

1. Fjell AM, Walhovd KB. Structural brain changes in aging: Courses, causes and cognitive consequences. Rev Neurosci. 2010;21:187–221. doi:10.1515/REVNEURO.2010.21.3.187.

2. Alzheimer's Association. 2019 Alzheimer's Disease Facts and Figures. 2019. https://www.alz.org/media/documents/alzheimers-facts-and-figures-2019-r.pdf. Accessed 7 Sep 2020.

3. Alzheimer's Disease International. World Alzheimer Report 2019. 2019. https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf. Accessed 7 Sep 2020.

4. Masters CL, Bateman R, Blennow K, Rowe CC, Sperling RA, Cummings JL. Alzheimer's disease. Nat Rev Dis Prim. 2015;1:1–18. doi:10.1038/nrdp.2015.56.

5. Portelius E, Bogdanovic N, Gustavsson MK, Volkmann I, Brinkmalm G, Zetterberg H, et al. Mass spectrometric characterization of brain amyloid beta isoform signatures in familial and sporadic Alzheimer's disease. Acta Neuropathol. 2010;120:185–93. doi:10.1007/s00401-010-0690-1.

6. Selkoe DJ, Hardy J. The amyloid hypothesis of Alzheimer's disease at 25 years. EMBO Mol Med. 2016;8:595–608.

7. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. Science (80-). 2002;297:353–6. doi:10.1126/science.1072994.

8. Arnold SE, Hyman BT, Flory J, Damasio AR, Van Hoesen GW. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with alzheimer's disease. Cereb Cortex. 1991;1:103–16. doi:10.1093/cercor/1.1.103.

9. Fitzpatrick AWP, Falcon B, He S, Murzin AG, Murshudov G, Garringer HJ, et al. Cryo-EM structures of tau filaments from Alzheimer's disease. Nature. 2017;547:185–90. doi:10.1038/nature23002.

10. Del C. Alonso A, Grundke-Iqbal I, Iqbal K. Alzheimer's disease hyperphosphorylated tau sequesters normal tau into tangles of filaments and disassembles microtubules. Nat Med. 1996;2:783–7. doi:10.1038/nm0796-783.

11. Braak H, Braak E. Staging of alzheimer's disease-related neurofibrillary changes. Neurobiol Aging. 1995;16:271–8. doi:10.1016/0197-4580(95)00021-6.

12. Frost B, Hemberg M, Lewis J, Feany MB. Tau promotes neurodegeneration through global chromatin relaxation. Nat Neurosci. 2014;17:357–66. doi:10.1038/nn.3639.

13. Ikeda C, Yokota O, Nagao S, Ishizu H, Oshima E, Hasegawa M, et al. The Relationship Between Development of Neuronal and Astrocytic Tau Pathologies in Subcortical Nuclei and Progression of Argyrophilic Grain Disease. Brain Pathol. 2016;26:488–505. doi:10.1111/bpa.12319.

14. Rodrigues FB, Byrne L, McColgan P, Robertson N, Tabrizi SJ, Leavitt BR, et al. Cerebrospinal fluid total tau concentration predicts clinical phenotype in Huntington's disease. J Neurochem. 2016;139:22–5. doi:10.1111/jnc.13719.

15. Veugelen S, Saito T, Saido TC, Chávez-Gutiérrez L, De Strooper B. Familial Alzheimer's Disease Mutations in Presenilin Generate Amyloidogenic Aβ Peptide Seeds. Neuron. 2016;90:410–6.

doi:10.1016/j.neuron.2016.03.010.

16. Nishitsuji K, Tomiyama T, Ishibashi K, Ito K, Teraoka R, Lambert MP, et al. The E693Δ mutation in amyloid precursor protein increases intracellular accumulation of amyloid β oligomers and causes endoplasmic reticulum stress-induced apoptosis in cultured cells. Am J Pathol. 2009;174:957–69. doi:10.2353/ajpath.2009.080480.

17. Walker ES, Martinez M, Brunkan AL, Goate A. Presenilin 2 familial Alzheimer's disease mutations result in partial loss of function and dramatic changes in Aβ 42/40 ratios. J Neurochem. 2005;92:294–301. doi:10.1111/j.1471-4159.2004.02858.x.

18. Nilsberth C, Westlind-Danielsson A, Eckman CB, Condron MM, Axelman K, Forsell C, et al. The "Arctic" APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. Nat Neurosci. 2001;4:887–93. doi:10.1038/nn0901-887.

19. Scheuner D, Eckman C, Jensen M, Song X, Citron M, Suzuki N, et al. Secreted amyloid β-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. Nat Med. 1996;2:864–70. doi:10.1038/nm0896-864.

20. Suzuki N, Cheung TT, Cai XD, Odaka A, Otvos L, Eckman C, et al. An increased percentage of long amyloid β protein secreted by familial amyloid β protein precursor (βAPP717) mutants. Science (80-). 1994;264:1336–40. doi:10.1126/science.8191290.

21. Jan A, Adolfsson O, Allaman I, Buccarello AL, Magistretti PJ, Pfeifer A, et al. Aβ42 neurotoxicity is mediated by ongoing nucleated polymerization process rather than by discrete Aβ42 species. J Biol Chem. 2011;286:8585–96. doi:10.1074/jbc.M110.172411.

22. Wogulis M, Wright S, Cunningham D, Chilcote T, Powell K, Rydel RE. Nucleation-dependent polymerization is an essential component of amyloid-mediated neuronal cell death. J Neurosci. 2005;25:1071–80. doi:10.1523/JNEUROSCI.2381-04.2005.

23. Reinders NR, Pao Y, Renner MC, da Silva-Matos CM, Lodder TR, Malinow R, et al. Amyloid-β effects on synapses and memory require AMPA receptor subunit GluA3. Proc Natl Acad Sci U S A. 2016;113:E6526–34. doi:10.1073/pnas.1614249113.

24. Drews A, Flint J, Shivji N, Jönsson P, Wirthensohn D, De Genst E, et al. Individual aggregates of amyloid beta induce temporary calcium influx through the cell membrane of neuronal cells. Sci Rep. 2016;6:31910. doi:10.1038/srep31910.

25. Fonseca ACRG, Moreira PI, Oliveira CR, Cardoso SM, Pinton P, Pereira CF. Amyloid-beta disrupts calcium and redox homeostasis in brain endothelial cells. Mol Neurobiol. 2015;51:610–22. doi:10.1007/s12035-014-8740-7.

26. Chabrier MA, Cheng D, Castello NA, Green KN, LaFerla FM. Synergistic effects of amyloid-beta and wild-type human tau on dendritic spine loss in a floxed double transgenic model of Alzheimer's disease. Neurobiol Dis. 2014;64:107–17. doi:10.1016/j.nbd.2014.01.007.

27. Kepp KP. Alzheimer's disease due to loss of function: A new synthesis of the available data. Prog Neurobiol. 2016;143:36–60. doi:10.1016/j.pneurobio.2016.06.004.

28. Shankar GM, Li S, Mehta TH, Garcia-Munoz A, Shepardson NE, Smith I, et al. Amyloid-beta

protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. Nat Med. 2008;14:837–42. doi:10.1038/nm1782.

29. Brouillette J, Caillierez R, Zommer N, Alves-Pires C, Benilova I, Blum D, et al. Neurotoxicity and memory deficits induced by soluble low-molecular-weight amyloid-β1-42 oligomers are revealed in vivo by using a novel animal model. J Neurosci. 2012;32:7852–61. doi:10.1523/JNEUROSCI.5901-11.2012.

30. Lanz TA, Carter DB, Merchant KM. Dendritic spine loss in the hippocampus of young PDAPP and Tg2576 mice and its prevention by the ApoE2 genotype. Neurobiol Dis. 2003;13:246–53. doi:10.1016/S0969-9961(03)00079-2.

31. Abcam. Alzheimer's disease pathway. Abcam. 2017. http://www.abcam.com/pathways/alzheimers-disease-interactive-pathway.

32. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson J V, Lince DH, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry. 2007;68:613–8. https://www.ncbi.nlm.nih.gov/pubmed/17474819.

33. Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993;43:1467–72. https://www.ncbi.nlm.nih.gov/pubmed/8350998.

34. Mattsson N, Insel PS, Palmqvist S, Stomrud E, van Westen D, Minthon L, et al. Increased amyloidogenic APP processing in APOE ε4-negative individuals with cerebral β-amyloidosis. Nat Commun. 2016;7:10918. doi:10.1038/ncomms10918.

35. Koffie RM, Hashimoto T, Tai H-C, Kay KR, Serrano-Pozo A, Joyner D, et al. Apolipoprotein E4 effects in Alzheimer's disease are mediated by synaptotoxic oligomeric amyloid-β. Brain A J Neurol. 2012;135 Pt 7:2155–68. doi:10.1093/brain/aws127.

36. Deane R, Sagare A, Hamm K, Parisi M, Lane S, Finn MB, et al. apoE isoform-specific disruption of amyloid beta peptide clearance from mouse brain. J Clin Invest. 2008;118:4002–13. doi:10.1172/JCI36663.

37. Vitek MP, Brown CM, Colton CA. APOE genotype-specific differences in the innate immune response. Neurobiol Aging. 2009;30:1350–60. doi:10.1016/j.neurobiolaging.2007.11.014.

38. Laskowitz DT, Goel S, Bennett ER, Matthew WD. Apolipoprotein E suppresses glial cell secretion of TNF alpha. J Neuroimmunol. 1997;76:70–4. https://www.ncbi.nlm.nih.gov/pubmed/9184634.

39. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009;41:1088–93. doi:10.1038/ng.440.

40. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45:1452–8. doi:10.1038/ng.2802.

41. Shen N, Chen B, Jiang Y, Feng R, Liao M, Zhang L, et al. An Updated Analysis with 85,939 Samples Confirms the Association Between CR1 rs6656401 Polymorphism and Alzheimer's Disease.

Mol Neurobiol. 2015;51:1017–23. doi:10.1007/s12035-014-8761-2.

42. Chen LH, Kao PYP, Fan YH, Ho DTY, Chan CSY, Yik PY, et al. Polymorphisms of CR1, CLU and PICALM confer susceptibility of Alzheimer's disease in a southern Chinese population. Neurobiol Aging. 2012;33:210.e1-7. doi:10.1016/j.neurobiolaging.2011.09.016.

43. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013;368:117–27. doi:10.1056/NEJMoa1211851.

44. Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 2009;41:1094–9. doi:10.1038/ng.439.

45. Biffi A, Shulman JM, Jagiella JM, Cortellini L, Ayres AM, Schwab K, et al. Genetic variation at CR1 increases risk of cerebral amyloid angiopathy. Neurology. 2012;78:334–41. doi:10.1212/WNL.0b013e3182452b40.

46. Zhu X-C, Wang H-F, Jiang T, Lu H, Tan M-S, Tan C-C, et al. Effect of CR1 genetic variants on cerebrospinal fluid and neuroimaging biomarkers in healthy, mild cognitive impairment and alzheimer's disease cohorts. Mol Neurobiol. 2017;54:551–62. doi:10.1007/s12035-015-9638-8.

47. Rogers J, Li R, Mastroeni D, Grover A, Leonard B, Ahern G, et al. Peripheral clearance of amyloid beta peptide by complement C3-dependent adherence to erythrocytes. Neurobiol Aging. 2006;27:1733–9. doi:10.1016/j.neurobiolaging.2005.09.043.

48. Thomas RS, Henson A, Gerrish A, Jones L, Williams J, Kidd EJ. Decreasing the expression of PICALM reduces endocytosis and the activity of β-secretase: implications for Alzheimer's disease. BMC Neurosci. 2016;17:50. doi:10.1186/s12868-016-0288-1.

49. Tian Y, Chang JC, Greengard P, Flajolet M. The convergence of endosomal and autophagosomal pathways: implications for APP-CTF degradation. Autophagy. 2014;10:694–6. doi:10.4161/auto.27802.

50. Golde TE, Streit WJ, Chakrabarty P. Alzheimer's disease risk alleles in TREM2 illuminate innate immunity in Alzheimer's disease. Alzheimers Res Ther. 2013;5:24. doi:10.1186/alzrt178.

51. Shapiro B, Tocci P, Haase G, Gavert N, Ben-Ze'ev A. Clusterin, a gene enriched in intestinal stem cells, is required for L1-mediated colon cancer metastasis. Oncotarget. 2015;6:34389–401. doi:10.18632/oncotarget.5360.

52. Mercer, Leib J. A Role for PICALM in Macroautophagy and Cellular Cholesterol Homeostasis. 2015. https://dukespace.lib.duke.edu/dspace/handle/10161/11317.

53. Valenti R, Pantoni L, Markus HS. Treatment of vascular risk factors in patients with a diagnosis of Alzheimer's disease: a systematic review. BMC Med. 2014;12:160. doi:10.1186/s12916-014-0160-z.

54. Arvanitakis Z, Capuano AW, Leurgans SE, Bennett DA, Schneider JA. Relation of cerebral vessel disease to Alzheimer's disease dementia and cognitive function in elderly people: a cross-sectional study. Lancet Neurol. 2016;15:934–43. doi:10.1016/S1474-4422(16)30029-1.

55. Miklossy J, McGeer PL. Common mechanisms involved in Alzheimer's disease and type 2 diabetes: a key role of chronic bacterial infection and inflammation. Aging (Albany NY). 2016;8:575–88. doi:10.18632/aging.100921.

56. Pugazhenthi S, Qin L, Reddy PH. Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease. Biochim Biophys Acta. 2017;1863:1037–45. doi:10.1016/j.bbadis.2016.04.017.

57. Huang C-C, Chung C-M, Leu H-B, Lin L-Y, Chiu C-C, Hsu C-Y, et al. Diabetes mellitus and the risk of Alzheimer's disease: a nationwide population-based study. PLoS One. 2014;9:e87095. doi:10.1371/journal.pone.0087095.

58. Bu XL, Yao XQ, Jiao SS, Zeng F, Liu YH, Xiang Y, et al. A study on the association between infectious burden and Alzheimer's disease. Eur J Neurol. 2015;22:1519–25. doi:10.1111/ene.12477.

59. Miklossy J. Emerging roles of pathogens in Alzheimer disease. Expert Rev Mol Med. 2011;13:e30. doi:10.1017/S1462399411002006.

60. Pisa D, Alonso R, Rábano A, Rodal I, Carrasco L. Different Brain Regions are Infected with Fungi in Alzheimer's Disease. Sci Rep. 2015;5:15015. doi:10.1038/srep15015.

61. Soscia SJ, Kirby JE, Washicosky KJ, Tucker SM, Ingelsson M, Hyman B, et al. The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide. PLoS One. 2010;5:e9505. doi:10.1371/journal.pone.0009505.

62. Kumar DKV, Choi SH, Washicosky KJ, Eimer WA, Tucker S, Ghofrani J, et al. Amyloid-β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease. Sci Transl Med. 2016;8:340ra72. doi:10.1126/scitranslmed.aaf1059.

63. Monsell SE, Kukull WA, Roher AE, Maarouf CL, Serrano G, Beach TG, et al. Characterizing Apolipoprotein E ε4 Carriers and Noncarriers With the Clinical Diagnosis of Mild to Moderate Alzheimer Dementia and Minimal β-Amyloid Peptide Plaques. JAMA Neurol. 2015;72:1124–31. doi:10.1001/jamaneurol.2015.1721.

64. Mormino EC, Brandel MG, Madison CM, Rabinovici GD, Marks S, Baker SL, et al. Not quite PIB-positive, not quite PIB-negative: slight PIB elevations in elderly normal control subjects are biologically relevant. Neuroimage. 2012;59:1152–60. doi:10.1016/j.neuroimage.2011.07.098.

65. Rowe CC, Ellis KA, Rimajova M, Bourgeat P, Pike KE, Jones G, et al. Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. Neurobiol Aging. 2010;31:1275–83. doi:10.1016/j.neurobiolaging.2010.04.007.

66. Serrano-Pozo A, Qian J, Monsell SE, Blacker D, Gómez-Isla T, Betensky RA, et al. Mild to moderate Alzheimer dementia with insufficient neuropathological changes. Ann Neurol. 2014;75:597–601. doi:10.1002/ana.24125.

67. Holmes C, Boche D, Wilkinson D, Yadegarfar G, Hopkins V, Bayer A, et al. Long-term effects of Abeta42 immunisation in Alzheimer's disease: follow-up of a randomised, placebo-controlled phase I trial. Lancet. 2008;372:216–23. doi:10.1016/S0140-6736(08)61075-2.

68. Salloway S, Sperling R, Fox NC, Blennow K, Klunk W, Raskind M, et al. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. N Engl J Med. 2014;370:322–33. doi:10.1056/NEJMoa1304839.

69. Doody RS, Thomas RG, Farlow M, Iwatsubo T, Vellas B, Joffe S, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N Engl J Med. 2014;370:311–21. doi:10.1056/NEJMoa1312889.

70. Reiman EM, Quiroz YT, Fleisher AS, Chen K, Velez-Pardo C, Jimenez-Del-Rio M, et al. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case-control study. Lancet Neurol. 2012;11:1048–56. doi:10.1016/S1474-4422(12)70228-4.

71. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, et al. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. Proc Natl Acad Sci U S A. 2004;101:284–9. doi:10.1073/pnas.2635903100.

72. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011;9:34. doi:10.1186/1741-7007-9-34.

73. Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Med. 2016;8:104. doi:10.1186/s13073-016-0355-3.

74. Sood S, Gallagher IJ, Lunnon K, Rullman E, Keohane A, Crossland H, et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. Genome Biol. 2015;16:185. doi:10.1186/s13059-015-0750-x.

75. Berchtold NC, Sabbagh MN, Beach TG, Kim RC, Cribbs DH, Cotman CW. Brain gene expression patterns differentiate mild cognitive impairment from normal aged and Alzheimer's disease. Neurobiol Aging. 2014;35:1961–72. doi:10.1016/j.neurobiolaging.2014.03.031.

76. Berchtold NC, Coleman PD, Cribbs DH, Rogers J, Gillen DL, Cotman CW. Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. Neurobiol Aging. 2013;34:1653–61. doi:10.1016/j.neurobiolaging.2012.11.024.

77. Lai MKP, Esiri MM, Tan MGK. Genome-wide profiling of alternative splicing in Alzheimer's disease. Genomics data. 2014;2:290–2. doi:10.1016/j.gdata.2014.09.002.

78. Lee C, Low CYB, Wong SY, Lai MKP, Tan MGK. Selective induction of alternatively spliced FynT isoform by TNF facilitates persistent inflammatory responses in astrocytes. Sci Rep. 2017;7:43651. doi:10.1038/srep43651.

79. Antonell A, Lladó A, Altirriba J, Botta-Orfila T, Balasa M, Fernández M, et al. A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. Neurobiol Aging. 2013;34:1772–8. doi:10.1016/j.neurobiolaging.2012.12.026.

80. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, et al. Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet. 2009;84:445–58. doi:10.1016/j.ajhg.2009.03.011.

81. Williams C, Mehrian Shai R, Wu Y, Hsu Y-H, Sitzer T, Spann B, et al. Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease. PLoS One. 2009;4:e4936. doi:10.1371/journal.pone.0004936.

82. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, et al. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genomics. 2007;28:311–22. doi:10.1152/physiolgenomics.00208.2006.

83. Lau P, Bossers K, Janky R, Salta E, Frigerio CS, Barbash S, et al. Alteration of the microRNA network during the progression of Alzheimer's disease. EMBO Mol Med. 2013;5:1613–34. doi:10.1002/emmm.201201974.

84. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PLoS One. 2011;6:e16266. doi:10.1371/journal.pone.0016266.

85. Scheckel C, Drapeau E, Frias MA, Park CY, Fak J, Zucker-Scharff I, et al. Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. Elife. 2016;5. doi:10.7554/eLife.10421.

86. Magistri M, Velmeshev D, Makhmutova M, Faghihi MA. Transcriptomics Profiling of Alzheimer's Disease Reveal Neurovascular Defects, Altered Amyloid-β Homeostasis, and Deregulated Expression of Long Noncoding RNAs. J Alzheimer's Dis. 2015;48:647–65. doi:10.3233/JAD-150398.

87. Roy J, Sarkar A, Parida S, Ghosh Z, Mallick B. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Mol Biosyst. 2017;13:565–76. doi:10.1039/c6mb00699j.

88. Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai L-H, et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature. 2015;518:365–9. doi:10.1038/nature14252.

89. Holcomb L, Gordon MN, McGowan E, Yu X, Benkovic S, Jantzen P, et al. Accelerated Alzheimer-type phenotype in transgenic mice carrying both mutant amyloid precursor protein and presenilin 1 transgenes. Nat Med. 1998;4:97–100. doi:10.1038/nm0198-097.

90. Chui DH, Tanahashi H, Ozawa K, Ikeda S, Checler F, Ueda O, et al. Transgenic mice with Alzheimer presenilin 1 mutations show accelerated neurodegeneration without amyloid plaque formation. Nat Med. 1999;5:560–4. doi:10.1038/8438.

91. Guo Q, Fu W, Sopher BL, Miller MW, Ware CB, Martin GM, et al. Increased vulnerability of hippocampal neurons to excitotoxic necrosis in presenilin-1 mutant knock-in mice. Nat Med. 1999;5:101–6.

92. Oakley H, Cole SL, Logan S, Maus E, Shao P, Craft J, et al. Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: potential factors in amyloid plaque formation. J Neurosci. 2006;26:10129–40. doi:10.1523/JNEUROSCI.1202-06.2006.

93. Oddo S, Caccamo A, Shepherd JD, Murphy MP, Golde TE, Kayed R, et al. Triple-transgenic model of Alzheimer's disease with plaques and tangles: intracellular Abeta and synaptic dysfunction. Neuron. 2003;39:409–21. https://www.ncbi.nlm.nih.gov/pubmed/12895417.

94. Borchelt DR, Ratovitski T, van Lare J, Lee MK, Gonzales V, Jenkins NA, et al. Accelerated amyloid deposition in the brains of transgenic mice coexpressing mutant presenilin 1 and amyloid precursor proteins. Neuron. 1997;19:939–45. doi:10.1016/S0896-6273(00)80974-5.

95. Savonenko A V, Xu GM, Price DL, Borchelt DR, Markowska AL. Normal cognitive behavior in two distinct congenic lines of transgenic mice hyperexpressing mutant APPSWE. Neurobiol Dis.

2003;12:194–211. doi:10.1016/S0969-9961(02)00012-8.

96. Kobayashi DT, Chen KS. Behavioral phenotypes of amyloid-based genetically modified mouse models of Alzheimer's disease. Genes Brain Behav. 2005;4:173–96. doi:10.1111/j.1601-183X.2005.00124.x.

97. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? Behav Brain Res. 2017;322 Pt B:311–28.

98. Simón A-M, Schiapparelli L, Salazar-Colocho P, Cuadrado-Tejedor M, Escribano L, López de Maturana R, et al. Overexpression of wild-type human APP in mice causes cognitive deficits and pathological features unrelated to Abeta levels. Neurobiol Dis. 2009;33:369–78. doi:10.1016/j.nbd.2008.11.005.

99. Saito T, Matsuba Y, Yamazaki N, Hashimoto S, Saido TC. Calpain activation in alzheimer's model mice is an artifact of APP and presenilin overexpression. J Neurosci. 2016;36:9933–6. doi:10.1523/JNEUROSCI.1907-16.2016.

100. Shukla V, Seo J, Binukumar BK, Amin ND, Reddy P, Grant P, et al. TFP5, a Peptide Inhibitor of Aberrant and Hyperactive Cdk5/p25, Attenuates Pathological Phenotypes and Restores Synaptic Function in CK-p25Tg Mice. J Alzheimer's Dis. 2017;56:335–49. doi:10.3233/JAD-160916.

101. Song W-J, Son M-Y, Lee H-W, Seo H, Kim JH, Chung S-H. Enhancement of BACE1 Activity by p25/Cdk5-Mediated Phosphorylation in Alzheimer's Disease. PLoS One. 2015;10:e0136950. doi:10.1371/journal.pone.0136950.

102. Xia D, Watanabe H, Wu B, Lee SH, Li Y, Tsvetkov E, et al. Presenilin-1 knockin mice reveal loss-of-function mechanism for familial Alzheimer's disease. Neuron. 2015;85:967–81. doi:10.1016/j.neuron.2015.02.010.

103. Saito T, Matsuba Y, Mihira N, Takano J, Nilsson P, Itohara S, et al. Single App knock-in mouse models of Alzheimer's disease. Nat Neurosci. 2014;17:661–3. doi:10.1038/nn.3697.

104. Herrera M, Jagadeeswaran P. Annual fish as a genetic model for aging. J Gerontol A Biol Sci Med Sci. 2004;59:101–7. doi:10.1093/gerona/59.2.B101.

105. Newman M, Ebrahimie E, Lardelli M. Using the zebrafish model for Alzheimer's disease research. Front Genet. 2014;5:189. doi:10.3389/fgene.2014.00189.

106. Kalueff A V, Stewart AM, Gerlai R. Zebrafish as an emerging model for studying complex brain disorders. Trends Pharmacol Sci. 2014;35:63–75. doi:10.1016/j.tips.2013.12.002.

107. Stewart AM, Braubach O, Spitsbergen J, Gerlai R, Kalueff A V. Zebrafish models for translational neuroscience research: from tank to bedside. Trends Neurosci. 2014;37:264–78. doi:10.1016/j.tins.2014.02.011.

108. Shaham S. Glial development and function in the nervous system of Caenorhabditis elegans. Cold Spring Harb Perspect Biol. 2015;7:a020578. doi:10.1101/cshperspect.a020578.

109. Struhl G, Greenwald I. Presenilin-mediated transmembrane cleavage is required for Notch signal transduction in Drosophila. Proc Natl Acad Sci. 2001;98:229–34. doi:10.1073/pnas.98.1.229.

110. Link CD. C. elegans models of age-associated neurodegenerative diseases: lessons from transgenic worm models of Alzheimer's disease. Exp Gerontol. 2006;41:1007–13.

doi:10.1016/j.exger.2006.06.059.

111. Shulman JM, Shulman LM, Weiner WJ, Feany MB. From fruit fly to bedside: translating lessons from Drosophila models of neurodegenerative disease. Curr Opin Neurol. 2003;16:443–9. doi:10.1097/01.wco.0000084220.82329.60.

112. Kalueff A V, Echevarria DJ, Stewart AM. Gaining translational momentum: more zebrafish models for neuroscience research. Prog Neuropsychopharmacol Biol Psychiatry. 2014;55:1–6. doi:10.1016/j.pnpbp.2014.01.022.

113. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503. doi:10.1038/nature12111.

114. Barthelson K, Pederson S, Newman M, Lardelli M. Brain transcriptome analysis reveals subtle effects on mitochondrial function and iron homeostasis of mutations in the SORL1 gene implicated in early onset familial Alzheimers disease. bioRxiv. 2020;:2020.07.17.207787. doi:10.1101/2020.07.17.207787.

115. Leimer U, Lun K, Romig H, Walter J, Grünberg J, Brand M, et al. Zebrafish (Danio rerio) Presenilin Promotes Aberrant Amyloid β-Peptide Production and Requires a Critical Aspartate Residue for Its Function in Amyloidogenesis†. Biochemistry. 1999;38:13602–9. doi:10.1021/bi991453n.

116. Newman M, Nornes S, Martins RN, Lardelli MT. Robust homeostasis of Presenilin1 protein levels by transcript regulation. Neurosci Lett. 2012;519:14–9. doi:10.1016/j.neulet.2012.04.064.

117. Moussavi Nik SH, Newman M, Wilson L, Ebrahimie E, Wells S, Musgrave I, et al. Alzheimer's disease-related peptide PS2V plays ancient, conserved roles in suppression of the unfolded protein response under hypoxia and stimulation of γ-secretase activity. Hum Mol Genet. 2015;24:3662–78. doi:10.1093/hmg/ddv110.

118. AlzForum. PSEN-1 Diagram with known human mutations. 2020. https://www.alzforum.org/mutations/psen-1. Accessed 8 Sep 2020.

119. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88. doi:10.1186/s13104-016-1900-2.

120. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60. doi:10.1038/nmeth.3317.

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

122. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30. doi:10.1093/bioinformatics/btt656.

123. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. doi:10.1038/nbt.3519.

124. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. doi:10.1038/nmeth.4197.

125. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4. doi:10.12688/f1000research.7563.1.

126. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. doi:10.1093/bioinformatics/btp616.

127. Andrews S. FastQC: A Quality Control tool for High Throughput Sequence Data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

128. Ward CM, To TH, Pederson SM. NgsReports: A Bioconductor package for managing FastQC reports and other NGS related log files. Bioinformatics. 2020;36:2587–8. doi:10.1093/bioinformatics/btz937.

129. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80. doi:10.1186/gb-2004-5-10-r80.

130. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA (New York). 2016;22:839–51. doi:10.1261/rna.053959.115.

131. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. doi:10.1093/nar/gkv007.

132. Law CW, Alhamdoosh M, Su S, Smyth GK, Ritchie ME. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research. 2016;5:1408. doi:10.12688/f1000research.9005.2.

133. Smyth GK. Limma: linear models for microarray data. Bioinforma Comput Biol Solut Using R Bioconductor. 2005;:397–420. doi:10.1007/0-387-29362-0_23.

134. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15.

135. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

136. Gene Ontology C, Blake JA, Dolan M, Drabkin H, Hill DP, Li N, et al. Gene Ontology annotations and resources. Nucleic Acids Res. 2013;41 Database issue:D530-5. doi:10.1093/nar/gks1050.

137. Benner C. HOMER (Hypergeometric Optimization of Motif EnRichment). http://homer.ucsd.edu/homer/.

138. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. 2009;37 SUPPL. 2:W202–8. doi:10.1093/nar/gkp335.

139. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50. doi:10.1073/pnas.0506580102.

140. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26:2176–82. doi:10.1093/bioinformatics/btq401.

141. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40:e133. doi:10.1093/nar/gks461.

142. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi:10.1186/1471-2105-9-559.

143. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9. doi:10.1038/nrg2825.

144. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902. doi:10.1038/nbt.2931.

145. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:1724–35. doi:10.1371/journal.pgen.0030161.

146. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet. 2019;10 JUN:535. doi:10.3389/fgene.2019.00535.

147. Fertig EJ, Ding J, Favorov A V., Parmigiani G, Ochs MF. CoGAPS: An R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics. 2010;26:2792–3. doi:10.1093/bioinformatics/btq503.

148. Odom G, Ban Y, Liu L, Sun X, Pico A, Zhang B, et al. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection. bioRxiv. 2019;:615435. doi:10.1101/615435.

149. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? Behav Brain Res. 2017;322 Pt B:311–28. doi:10.1016/j.bbr.2016.05.007.

150. Haidich AB. (2010). Meta-analysis in medical research. Hippokratia. 2010;14 Suppl 1:29–37. doi:10.5005/jp/books/10519.

Chapter 2

Accelerated brain aging

towards transcriptional inversion

in a zebrafish model of the K115fs

mutation of human _PSEN2_

Statement of Authorship

Title of Paper	Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human *PSEN2*
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Hin, N., Newman, M., Kaslin, J., Douek, A.M., Lumsden, A., Nik, S.H.M., Dong, Y., Zhou, X.F., Mañucat-Tan, N.B., Ludington, A. and Adelson, D.L., 2020. Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human *PSEN2*. *PLOS One*, 15(1), p.e0227258. DOI: 10.1371/journal.pone.0227258

Principal Author

Name of Principal Author (Candidate)	Nhi Hin		
Contribution to the Paper	- All bioinformatics analysis - Drafting of manuscript - Producing figures / visualisation of data - Corresponding author on manuscript - Management of deposition of transcriptome data into public database (GEO)		
Overall percentage (%)	45%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	08/09/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

    i.   the candidate's stated contribution to the publication is accurate (as detailed above);

    ii.   permission is granted for the candidate in include the publication in the thesis; and

    iii.   the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Morgan Newman		
Contribution to the Paper	- Contributed equally to this work with Nhi Hin (Overall Percentage: 45%) - Genome editing and creation of *psen1*[K97fs] zebrafish mutant line - Removal of brains from fish and purification of mRNA for RNA-seq - Producing figures / visualisation of data - Editing of manuscript drafts		
Signature		Date	10/09/2020

Name of Co-Author	Jan Kaslin		
Contribution to the Paper	- Contributed methodology and final results for examining zebrafish brains for markers of senescence - Editing of manuscript drafts		
Signature		Date	9/9/20

Name of Co-Author	Alon M. Douek		
Contribution to the Paper	- Contributed methodology and final results for examining zebrafish brains for markers of senescence - Editing of manuscript drafts		
Signature		Date	11/9/2020

Name of Co-Author	Amanda Lumsden		
Contribution to the Paper	- Contributed to methodology and results of initial promoter analysis - Editing of manuscript drafts		
Signature		Date	16/9/2020

Name of Co-Author	Seyed Hani Moussavi Nik		
Contribution to the Paper	- Contributed validation PCRs for differentially expressed genes - Editing of manuscript drafts		
Signature		Date	

Name of Co-Author	Yang Dong		
Contribution to the Paper	- Contributed validation PCRs for differentially expressed genes - Editing of manuscript drafts		
Signature		Date	11/09/2020

Name of Co-Author	Xin-Fu Zhou		
Contribution to the Paper	- Involved in initial attempts to examine the zebrafish brains for markers of senescence - Editing of manuscript drafts		
Signature		Date	11/09/2020

Name of Co-Author	Noralyn B. Mañucat-Tan		
Contribution to the Paper	- Involved in initial attempts to examine the zebrafish brains for markers of senescence - Editing of manuscript drafts		
Signature		Date	11/09/20

Name of Co-Author	Alastair Ludington
Contribution to the Paper	- Assiste ith ioin or ati s i elines an so t are use - E iting o anus ri t ra ts
Signature	Date: 11/09/2020

Name of Co-Author	David L. Adelson
Contribution to the Paper	- Project supervision - Conceptualisation of the WGCNA approach and network-based bioinformatics analysis - Editing of manuscript drafts
Signature	Date: 11/9/2020

Name of Co-Author	Stephen Pederson
Contribution to the Paper	- Supervision of bioinformatics work and statistical methodology - Editing of manuscript draft
Signature	Date: 11/09/20

Name of Co-Author	Michael Lardelli
Contribution to the Paper	- Conceptualisation - Project supervision - Editing of manuscript drafts
Signature	Date: 9/9/20

Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human *PSEN2*

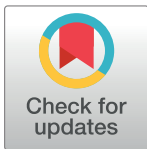Nhi Hin[1,2◉], Morgan Newman[2◉], Jan Kaslin[3], Alon M. Douek[3], Amanda Lumsden[4], Seyed Hani Moussavi Nik[1], Yang Dong[1], Xin-Fu Zhou[5], Noralyn B. Mañucat-Tan[5], Alastair Ludington[1], David L. Adelson[6], Stephen Pederson[1‡], Michael Lardelli[2‡]*

1 Bioinformatics Hub, School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia, **2** Alzheimer's Disease Genetics Laboratory, School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia, **3** Australian Regenerative Medicine Institute, Monash University, Clayton, Victoria, Australia, **4** College of Medicine and Public Health, and Centre for Neuroscience, Flinders University, Adelaide, South Australia, Australia, **5** School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia, Australia, **6** Centre for Bioinformatics and Computational Genetics, School of Bioogical Sciences, Adelaide, South Australia, Australia

◉ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* michael.lardelli@adelaide.edu.au

Abstract

Background

The molecular changes involved in Alzheimer's disease (AD) progression remain unclear since we cannot easily access antemortem human brains. Some non-mammalian vertebrates such as the zebrafish preserve AD-relevant transcript isoforms of the *PRESENILIN* genes lost from mice and rats. One example is PS2V, the alternative transcript isoform of the *PSEN2* gene. PS2V is induced by hypoxia/oxidative stress and shows increased expression in late onset, sporadic AD brains. A unique, early onset familial AD mutation of *PSEN2*, K115fs, mimics the PS2V coding sequence suggesting that forced, early expression of PS2V-like isoforms may contribute to AD pathogenesis. Here we use zebrafish to model the K115fs mutation to investigate the effects of forced PS2V-like expression on the transcriptomes of young adult and aged adult brains.

Methods

We edited the zebrafish genome to model the K115fs mutation. To explore its effects at the molecular level, we analysed the brain transcriptome and proteome of young (6-month-old) and aged (24-month-old) wild type and heterozygous mutant female sibling zebrafish. Finally, we used gene co-expression network analysis (WGCNA) to compare molecular changes in the brains of these fish to human AD.

Results

Young heterozygous mutant fish show transcriptional changes suggesting accelerated brain aging and increased glucocorticoid signalling. These early changes precede a

transcriptional 'inversion' that leads to glucocorticoid resistance and other likely pathological changes in aged heterozygous mutant fish. Notably, microglia-associated immune responses regulated by the ETS transcription factor family are altered in both our zebrafish mutant model and in human AD. The molecular changes we observe in aged heterozygous mutant fish occur without obvious histopathology and possibly in the absence of Aβ.

Conclusions

Our results suggest that forced expression of a PS2V-like isoform contributes to immune and stress responses favouring AD pathogenesis. This highlights the value of our zebrafish genetic model for exploring molecular mechanisms involved in AD pathogenesis.

Introduction

Alzheimer's disease (AD) is the leading cause of dementia, a condition characterised by the progressive decline of memory and cognition. Like other neurodegenerative diseases, AD affects diverse cellular processes in the brain, including mitochondrial function [1, 2], metal ion homeostasis [3–5], lipid metabolism [6–8], immune responses [9, 10], synaptic transmission [11], and protein folding and trafficking [12, 13]. Dysregulation of these processes eventually results in severe atrophy of several brain regions (reviewed by Braak and Braak [14] and Masters et al. [15]). Consequently, late stages of AD are likely to be much more difficult to treat than earlier stages of AD, contributing to our failure to discover ameliorative drugs [16].

The pathological processes that result in AD are likely to initiate decades before clinical symptoms arise. Decreased levels of soluble amyloid beta (Aβ) peptides in the cerebrospinal fluid is one of the earliest markers of both sporadic and familial forms of AD, preceding disease onset by 20–30 years [17, 18], while vascular changes are likely to occur even earlier [19]. Individuals possessing highly penetrant, dominant mutations in genes linked to the familial form of AD (fAD) such as *PSEN1* show structural and functional changes in their brains as early as 9 years of age, despite being cognitively normal [20, 21]. Similar findings are evident in young adults carrying the ε4 allele of *APOE*, the major risk gene for the sporadic form of AD [22]. To prevent AD, we must identify the stresses underlying these early pathological changes. However, detailed molecular analysis of the brains of asymptomatic young adult fAD mutation carriers is currently impossible.

Analysing high-throughput 'omics data (e.g. transcriptomic, proteomic) is a comprehensive and relatively unbiased approach for studying complex diseases like AD. Over the past decade, numerous post-mortem AD brains have been profiled using microarray and RNA-seq technologies, exposing an incredibly complex and interconnected network of cellular processes implicated in the disease [23, 24]. Unfortunately, analysing post-mortem AD brains does not discern which cellular processes are responsible for initiating the cascade of events leading to AD.

Animal models can assist exploration of the early molecular changes that promote AD. However, early "knock-in" mouse models that attempted to model the genetic state of human fAD showed no obvious histopathology [25–27]. Modern 'omics technologies provide molecular-level descriptions of disease states, but these technologies were not available when the early knock-in models were made. Subsequent transgenic models of AD constructed with multiple genes and/or mutations have displayed what are assumed to be AD-related histopathologies and these have also been analysed by 'omics methods. However recent analysis of brain transcriptomes from five different transgenic AD models showed

47

little concordance with human, late onset, sporadic AD brain transcriptomes. Worse still, none of the models were concordant with each other [28].

The overwhelming majority of fAD mutations are present in a heterozygous state in human patients. Despite this, there has been a lack of detailed molecular investigation of the young adult brains of any animal model closely imitating the human fAD genetic state–i.e. heterozygous for a fAD-like mutation in a single, endogenous gene. Previously, we used zebrafish to analyse the unique, frameshifting fAD mutation of human *PRESENILIN2* (*PSEN2*), K115fs, that inappropriately mimics expression of a hypoxia-induced truncated isoform of PSEN2 protein, PS2V [29–32]. Mice and rats have lost the ability to express PS2V [33] (and the fAD genes of these rodents are evolving more rapidly than in many other mammals [33]), but in zebrafish, this isoform is expressed from the animal's *psen1* gene [32]. Consequently, to model and explore early changes in the brain contributing to AD pathogenesis, we have now used gene-editing technology to introduce a K115fs-equivalent mutation into the zebrafish *psen1* gene, K97fs. In this paper, we analyse data collected from young adult (6-month-old) and aged (24-month-old) adult heterozygous mutant and wild type zebrafish brains to comprehensively assess gene and protein expression changes in the brain due to aging and this mutation. At the molecular level, we find that the young heterozygous mutant brains show elements of accelerated aging while aged heterozygous mutant brains appear to 'invert' into a distinct, and presumably pathological, state. Our results highlight the important role that non-transgenic models of fAD mutations in a heterozygous state play in elucidating mechanisms of AD pathogenesis.

Results

Gene editing in zebrafish to produce the *psen1* K97fs mutation is described in the **Materials and Methods** and in **Fig A** in S1 File. To confirm that the K97fs mutation of *psen1* forces measurable expression of a PS2V-like transcript under normoxic conditions we performed digital quantitative PCR (dqPCR) specifically detecting either heterozygous mutant or wild type transcript sequences in cDNA synthesised from the brains of female 6-month-old (young) and 24-month-old (aged) *psen1*K97fs/+ (heterozygous mutant) and *psen1*+/+ (wild type) zebrafish (Fig 1). We only included female fish to reduce variability between samples and minimise confounding by potential gender-specific gene expression patterns, given that females are more vulnerable to AD and that gender-specific changes have been documented in AD [34, 35]. K97fs transcripts constitute approximately 30% of the *psen1* transcripts detected in young brains and over 70% of the detected transcripts in aged brains. Despite these different biases in heterozygous mutant and wild type transcript expression, the total levels of *psen1* transcript appeared similar between heterozygous mutant and wild type fish at either age. This supports that the K97fs mutant transcript (like PS2V transcripts in humans) is not completely degraded by nonsense mediated decay despite possession of a premature termination codon [29]. PCR tests on cDNA from heterozygous mutant brains did not detect aberrant splicing of the *psen1* gene due to the K97fs mutation. We currently have no explanation for the observed bias, or its age-dependent change, between the expression of the heterozygous mutant versus wild type *psen1* transcripts. The extent of the decrease in the wild type *psen1* transcript in the aged heterozygous mutant brains means that this may contribute to any molecular phenotype caused by heterozygosity for the K97fs mutation in addition to the effects of the PS2V-like transcripts.

To determine whether the K97fs mutation in the zebrafish *psen1* gene induces changes in the expression of other genes and proteins, we removed entire brains of heterozygous mutant and wild type adult zebrafish for total RNA sequencing (RNA-seq) and label-free tandem mass
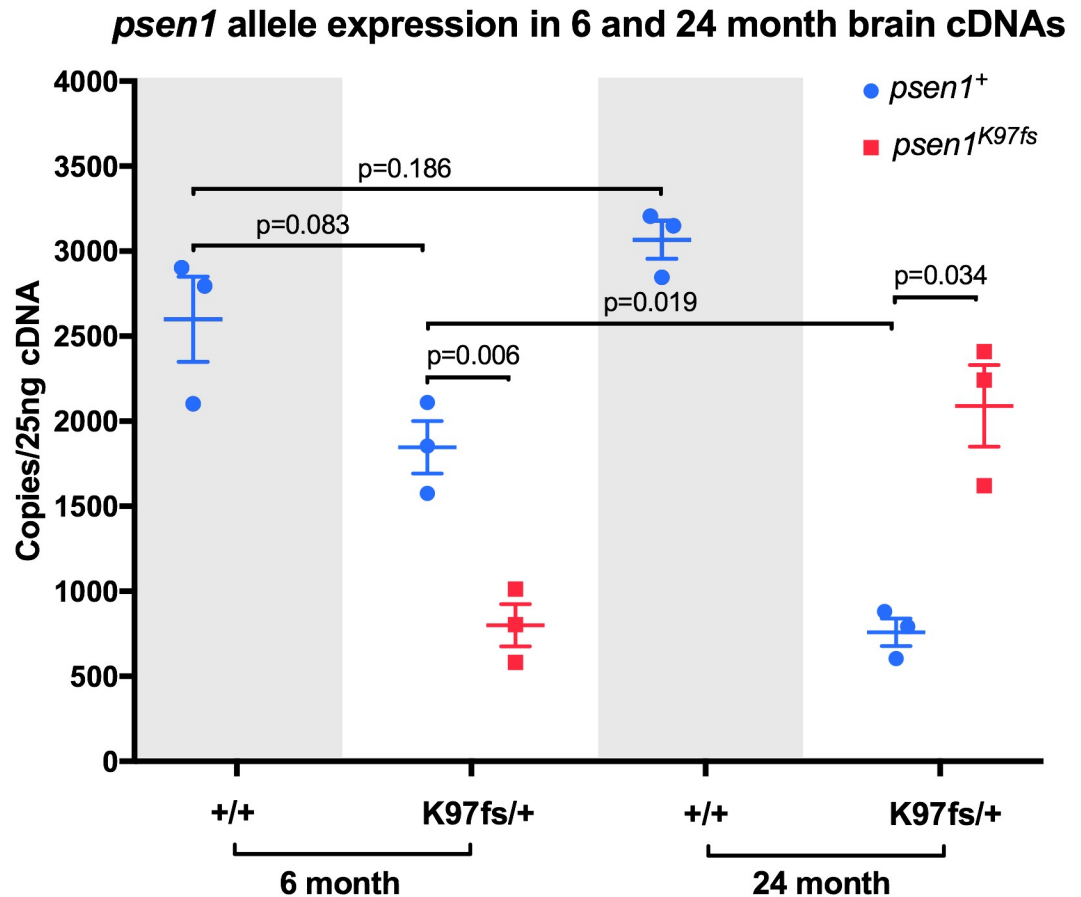
48

psen1 allele expression in 6 and 24 month brain cDNAs

Fig 1. Quantification of heterozygous mutant and wild type allele relative transcript expression. Digital quantitative PCRs specifically detecting transcripts from the heterozygous mutant (K97fs) or wild type (+) alleles of *psen1* were performed using cDNA synthesised from total brain mRNA from fish at 6 and 24 months of age. Means and standard error of the means are indicated, and *p*-values are from two-sample *t*-tests assuming unequal variances.

https://doi.org/10.1371/journal.pone.0227258.g001

spectroscopy (LC-MS/MS) when zebrafish were 6 months (young adult) and 24 months (aged adult) old. We used three biological replicates to represent each of the four experimental conditions (young wild type, young heterozygous mutant, aged wild type, aged heterozygous mutant), and performed pairwise comparisons between experimental conditions to determine differentially expressed (DE) genes and differentially abundant (DA) proteins (Fig 2). Full lists of DE genes and DA proteins are provided in S1 and S2 Tables.

Gene expression changes in the heterozygous mutant zebrafish reveal accelerated brain aging followed by inversion into a presumably pathological state

The brains of children or young adults carrying fAD mutations display morphological and functional differences compared to age-matched individuals without these mutations [20, 21]. Consequently, we hypothesised that gene expression in the brains of young adult (6-month-old) zebrafish carrying this K115fs-like mutation would also be altered when compared to wild type zebrafish siblings. Overall, we find supporting evidence for 105 genes that are differentially expressed in young heterozygous mutant brains relative to wild type brains (65 up-regulated, 40 down-regulated; FDR-adjusted *p*-value < 0.05) (**Fig B in** S1 File).
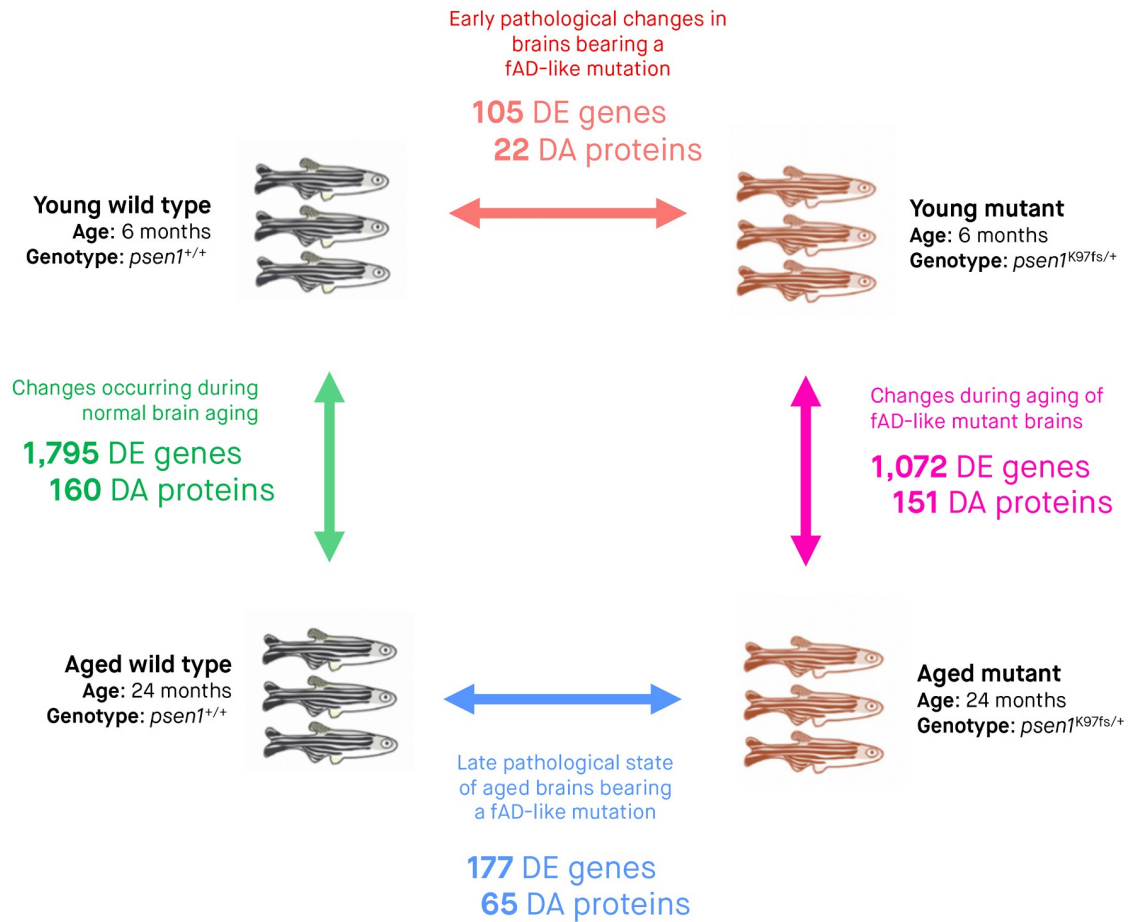
49

Fig 2. Summary of experimental groups, differentially expressed (DE) genes and differentially abundant (DA) proteins. Three biological replicates (entire zebrafish brains) were subjected to RNA-seq and LC-MS/MS for each of the four experimental conditions. Arrows indicate pairwise comparisons (to identify DE genes and DA proteins) between experimental conditions. The numbers of DE genes and DA proteins determined from RNA-seq and LC-MS/MS analyses are indicated underneath the arrow for each comparison. We considered genes to be DE and proteins to be DA if the False Discovery Rate [FDR]-adjusted *p*-value of their moderated *t*-test (*limma*) was below 0.05. All zebrafish of the same age are siblings raised in the same tank.

https://doi.org/10.1371/journal.pone.0227258.g002

Of these 105 genes, 65 have an estimated $\log_2$ fold change greater than 0.5 (or less than -0.5) in the 'young heterozygous mutant vs. young wild type' comparison (Fig 3A). By examining the expression of these genes in the other three comparisons described in Fig 2, we observe two important phenomena:

1. **Accelerated aging genes are associated with increased immune response**: 62% (65/105) of the genes that are DE in 6-month-old heterozygous mutant brains ('young heterozygous mutant vs. young wild type') show the same direction of expression change during normal aging ('aged wild type vs. young wild type'). However, far more genes are DE during normal aging (1,795 compared to 105). This suggests that the 6-month-old heterozygous mutant brains may demonstrate accelerated aging for a subset of cellular functions. As an initial step to explore these altered cellular functions, we applied functional enrichment analysis on these 65 genes and discovered significant enrichment in an MSigDB gene set relating to immune response genes that are up-regulated following lipopolysaccharide treatment "GSE9988 LPS VS VEHICLE TREATED MONOCYTE UP" (Bonferroni adjusted *p*-value 0.000948) (S3 Table).
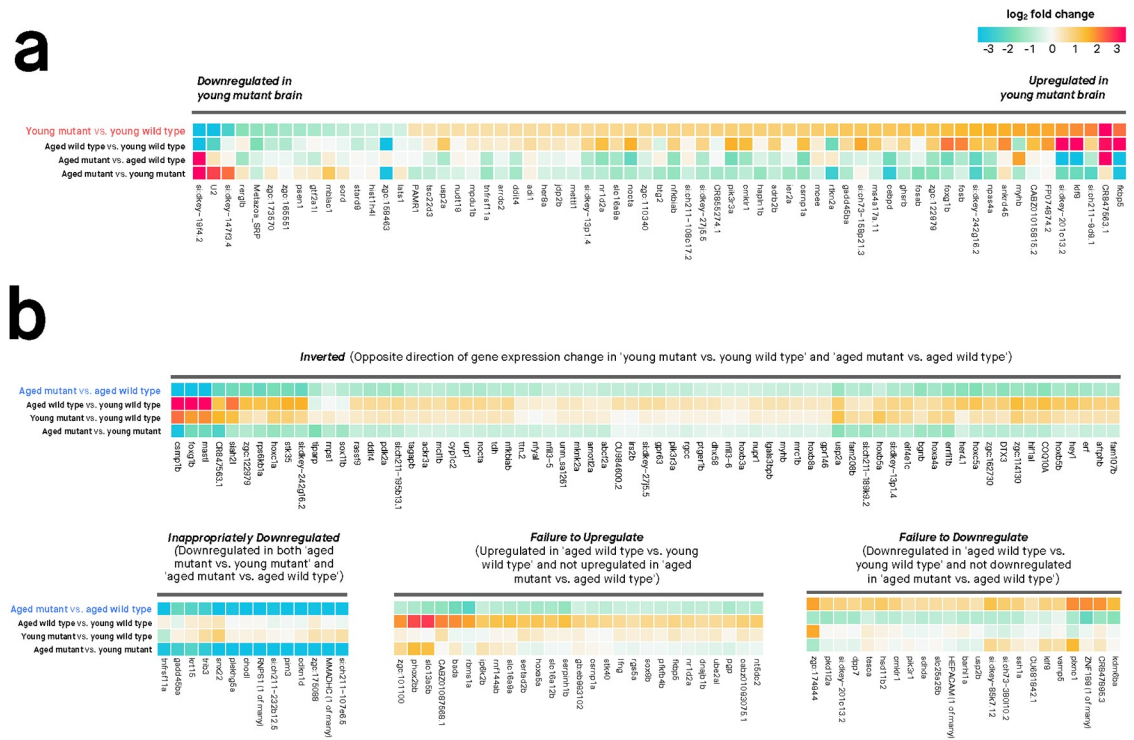
50

Fig 3. Differential gene expression between heterozygous mutant (*psen1*^K97fs/+^) and wild type (*psen1*^+/+^) zebrafish brains at 6 months (young) and 24 months (aged). Only genes with absolute log$_2$ fold change > 0.5 are shown. Genes were considered differentially expressed if their moderated *t*-test FDR-adjusted *p*-value was below 0.05. **(A) Differentially expressed genes at 6 months. (B) Differentially expressed genes at 24 months**. The differentially expressed genes are grouped into clusters based on gene expression changes across the four comparisons. Overall, note the similar expression changes in 'young heterozygous mutant vs. young wild type' and 'aged wild-type vs. young wild-type' and the contrast of these to comparisons involving aged heterozygous mutants. This illustrates the accelerated brain aging in young heterozygous mutant brains and the "inverted" gene expression pattern of aged heterozygous mutant brains.

https://doi.org/10.1371/journal.pone.0227258.g003

2. **Age-dependent 'inversion' pattern**: A subset of 63 genes with increased expression in 6-month-old heterozygous mutant brains ('young heterozygous mutant vs. young wild type') show decreased expression in 24-month-old heterozygous mutant brains ('aged heterozygous mutant vs. aged wild type'). We call this expression pattern an age-dependent 'inversion' between heterozygous mutant and wild type brains, and explore the biological relevance of the genes involved in this inversion pattern later.

By comparing gene expression in 24-month-old heterozygous mutant and wild type zebrafish brains, we can gain insight into a putatively pathological transcriptomic state present in the brains of aged zebrafish carrying this mutation. We find supporting evidence for 177 genes that are differentially expressed in heterozygous mutant brains relative to wild type brains (139 down-regulated, 38 up-regulated; FDR-adjusted *p*-value < 0.05) (Fig 3B; **Fig B in** S1 File). Note that not all of these genes are shown in Fig 3B, which only includes genes with log$_2$ fold change values greater than 0.5 or less than -0.5. To allow for easier interpretation of these 177 genes, we used hierarchical clustering to separate them into groups with distinct expression patterns based on all four brain-types:

• **Inverted (63 genes)**: Defined as genes showing opposite fold-changes in young heterozygous mutant brains ('young heterozygous mutant vs. young wild type') compared to aged

51

heterozygous mutant brains ('aged heterozygous mutant vs. aged wild type'). To be included in this group, genes were required to have an FDR-adjusted *p*-value < 0.05 in either the 'young heterozygous mutant vs. young wild type' or 'aged heterozygous mutant vs. aged wild type' comparison and an unadjusted *p*-value < 0.05 in the other comparison.

- **Inappropriately down-regulated (57 genes)**: Defined as genes that are down-regulated in the 'aged heterozygous mutant vs. young heterozygous mutant' and 'aged heterozygous mutant vs. aged wild type' comparisons (FDR-adjusted *p*-value < 0.05 in both).

- **Failure to up-regulate (94 genes)**: Defined as genes that are up-regulated during normal aging (FDR-adjusted *p*-value < 0.05 in the 'aged wild type vs. young wild type' comparison) but not up-regulated in the 'aged heterozygous mutant vs. aged wild type' comparison.

- **Failure to down-regulate (26 genes)**: Defined as genes that are down-regulated during normal aging (FDR-adjusted *p*-value < 0.05 in the 'aged wild type vs. young wild type' comparison) but not down-regulated in the 'aged heterozygous mutant vs. aged wild type' comparison.

To determine whether these different component groups of the gene expression patterns are biologically relevant, we assessed each group's functional enrichment using Gene Ontology terms, MSigDB gene sets, and Reactome and Interpro pathways (summarised in S3 Table; full results in S4 Table). Overall, we find statistically significant enrichment (Bonferroni adjusted *p*-value < 0.05) for all groups except for the 'failure to down-regulate' group. The 'inverted' group is significantly enriched in several gene sets related to stress and immune response; the 'inappropriately down-regulated' group is significantly enriched in developmental transcription factors including homeobox genes; the 'failure to up-regulate' group is significantly enriched in immune responses.

Gene expression changes during aging of the heterozygous mutant zebrafish brains partially overlap with normal brain aging

Gene expression changes associated with aging in the wild type and heterozygous mutant zebrafish brains only partially overlap. When comparing 24-month-old and 6-month-old wild type zebrafish brains, 1,795 genes show differential expression. However, when comparing 24-month-old and 6-month-old heterozygous mutant zebrafish brains, 1,072 genes show altered expression (FDR-adjusted *p*-value < 0.05). When comparing these two sets of genes, only 525 genes show fold-changes in the same direction during wild type and heterozygous mutant aging. These genes can be considered an 'aging signature' and are functionally enriched in gene ontology terms relating to immune function (S3 Table). This suggests that the heterozygous mutant fish still preserve some immune-related gene expression changes that occur during normal aging.

Gene expression changes are likely not due to changes in proportions of brain cell types

It is possible that changes in the proportions of different cell types in the brain could result in genes being falsely interpreted as differentially expressed. As a preliminary test to see whether our observations of differential gene expression were artefacts of change in the proportions of major brain cell types (e.g. astrocytes, microglia, neurons, oligodendrocytes), we checked for noticeable changes in the average expression for sets of marker genes characteristic of each of the major brain cell types across the samples in each experimental condition (young wild type, young heterozygous mutant, aged wild type, aged heterozygous mutant). Representative

52

marker genes for microglia were obtained from Oosterhof et al. [36] while gene markers for astrocytes, neurons, and oligodendrocytes were obtained from Lein et al. [37] The number of genes used to calculate the average gene expression (in logCPM) was 41 (astrocyte), 99 (microglia), 77 (neuron) and 78 (oligodendrocyte). Although this method is limited in that it does not account for the significant diversity within these broader cell types nor regional brain differences, this level of analysis suggests that broadly, the average expression of gene markers for the major neural cell types does not appear to change much across experimental conditions. In addition, no obvious outlier samples were evident (**Fig C in** S1 File).

Regulation of gene expression in the heterozygous mutant zebrafish brains differs from normal brain aging

A transcription factor can regulate gene expression by binding to a specific DNA motif in the promoter region of a gene. We hypothesised that changes in gene expression during normal aging or differences in gene expression between heterozygous mutant and wild type brains could be driven by differences in transcription factor activity. To test this, we examined gene promoter regions for enriched motifs corresponding to known transcription factor binding sites (summarised in S5 Table; full results in S6 Table). Overall, we find:

1. **Numerous known transcription factors likely drive the gene expression changes that occur during normal zebrafish brain aging**. As wild type brains age, the genes which are differentially expressed are significantly enriched in many known motifs. These motifs correspond to binding sites for interferon regulatory factors (e.g. IRF1, IRF2, IRF8); a binding site for the PU.1-IRF8 complex; an interferon-stimulated response element (ISRE); and binding sites for various transcription factors important for essential cellular processes like proliferation, differentiation, and apoptosis (Atf3, Fra2, Ets-distal, AP-1, Fra1, JunB, BATF, and ZNF264).

2. **Altered glucocorticoid signalling in heterozygous mutant zebrafish brains is likely to contribute to a pathological state**. Promoters of genes that are differentially expressed in the 'aged heterozygous mutant vs. aged wild type' comparison are significantly enriched in the glucocorticoid receptor element motif (GRE) (Bonferroni p-value = 0.0057). Interestingly, the subset of genes showing inappropriate downregulation (down-regulated in the 'aged heterozygous mutant vs. young heterozygous mutant' and 'aged heterozygous mutant vs. aged wild type' comparisons) is more strongly enriched again in the GRE motif (Bonferroni p-value = 0.0001), suggesting that genes that are normally activated by glucocorticoid signalling during aging may not be activated in aged heterozygous mutant brains. This altered glucocorticoid signalling appears to be present even in young zebrafish brains, as genes showing inverted behaviour (opposite direction of differential expression in 'young heterozygous mutant vs. young wild type' and 'aged heterozygous mutant vs. aged wild type' comparisons) are also enriched in the GRE motif (Bonferroni p-value = 0.0047). Because these inverted genes tend to show high expression in young heterozygous mutant brains (i.e. up-regulated in the 'young heterozygous mutant vs. young wild type' comparison) and low expression in aged heterozygous mutant brains (i.e. down-regulated in the 'aged heterozygous mutant vs. aged wild type' comparison), this suggests that young heterozygous mutant zebrafish brains may initially exhibit abnormally increased glucocorticoid signalling, while aged heterozygous mutant brains later exhibit abnormally decreased glucocorticoid signalling. Notably, the inverted genes containing a GRE motif in their promoters include *COQ10A* (encodes Coenzyme Q10, a key component of the electron transport chain and free-radical scavenging antioxidant); *pik3r3a* (encodes regulatory subunit

53

gamma of phosphoinositide 3-kinase, an enzyme that interacts with insulin growth factor 1 receptor among other proteins); *mmadhc* (encodes a protein involved in an early and essential step of vitamin B12 metabolism), *plk3* (polo-like kinase 3, involved in stress response and double-stranded DNA repair), and *fkbp5* (encodes FK506 binding protein, involved in regulating immune and stress responses, protein trafficking and folding, and glucocorticoid receptor regulation). A list of zebrafish genes containing the GRE promoter motif is provided in S7 Table.

Gene expression changes in the heterozygous mutant zebrafish indicate vast changes to cellular processes and pathways

A gene set is a group of genes that contribute to a known biological function, pathway, or state. A gene set test is an analysis used to evaluate whether a particular gene set is differentially expressed for a particular comparison. We used the FRY method to test whether 'Hallmark' gene sets from the Molecular Signatures Database, MSigDB [38] were associated with differential expression in each of the four comparisons (Fig 4 and S8 Table). Using an FDR-adjusted *p*-value < 0.05 to define a gene set as differentially expressed, we find:

1. **50 gene sets are differentially expressed during normal brain aging ('aged wild type vs. young wild type')** (middle row of heatmap, Fig 4A). This supports that many biological functions and pathways are altered during normal aging. For some gene sets, the proportion of genes that are up-regulated and down-regulated is similar (e.g. interferon alpha response, E2F targets, early estrogen response). However, other gene sets contain a predominance of up-regulated genes (e.g. epithelial mesenchymal transition, TNFA signalling via NFKB) or down-regulated genes (e.g. coagulation, reactive oxygen species pathway).

2. **22 gene sets are differentially expressed in young heterozygous mutant brains ('young heterozygous mutant vs. young wild type')** (top row of heatmap, Fig 4A). These 22 gene sets may represent earlier functional changes in the brain that occur due to this mutation. The gene sets implicate diverse processes including Wnt/β-catenin signalling, early estrogen response, DNA repair, hedgehog signalling and fatty acid metabolism. Similar to the pattern of accelerated aging observed in Fig 3, we also observe that most of the gene sets up-regulated in young heterozygous mutant brains are regulated in the same direction during normal aging. This is consistent with the idea that the biological changes in young heterozygous mutant brains may partially recapitulate those that occur during normal brain aging.

3. **44 gene sets are differentially expressed between aged heterozygous mutant and aged wild type brains** (bottom row, Fig 4A). These differentially expressed gene sets may represent the pathological state of aged zebrafish brains bearing this mutation. Importantly, 21 of the 22 gene sets that were differentially expressed in young heterozygous mutant brains ('young heterozygous mutant vs. young wild type') remain altered also when these are aged ('aged heterozygous mutant vs. aged wild type'). However, the proportions of up- and down-regulated genes tend to differ; notably, several gene sets containing a predominance of up-regulated genes in the young heterozygous mutant brains contain a predominance of down-regulated genes in the old heterozygous mutant brains. These 'inverted' gene sets include biological functions and pathways as diverse as Wnt/β-catenin signalling, early estrogen response, hedgehog signalling, androgen response, epithelial mesenchymal transition, DNA repair, apical surface, and TGF-β signalling.

Fig 4. Differential gene set expression in heterozygous mutant (*psen1*^{K97fs/+}) zebrafish brains compared to wild type siblings. Values in each cell are the estimated proportions of up- and down-regulated genes for each gene set, for any particular pairwise comparison shown to the left of the cells. A missing cell indicates that the particular gene set is not differentially expressed for that particular pairwise comparison. Colours of cells are proportional to the difference between the proportion of up- and down-regulated genes in a gene set. Differentially expressed gene sets have Mixed FDR below 0.05, indicating genes within the gene set show statistically significantly altered (up and/or down) expression for a particular comparison. The genes in each gene set are defined using the "Hallmark" gene set collection at the Molecular Signatures Database (MSigDB). **(A) Gene sets showing differential expression between heterozygous mutant (*psen1*^{K97fs/+}) and wild type (*psen1*^{+/+}) zebrafish brains at 6 months (young) and 24 months (aged)**. The comparison representing normal aging (aged wild type vs. young wild type) is also shown to highlight the 'accelerated aging' phenomenon in the young heterozygous mutants. **(B) Gene sets showing differential expression during normal aging**. The aged K97fs/+ vs. young K97fs/+) comparison is also shown to highlight the phenomenon of aberrant aging in the heterozygous mutants.

https://doi.org/10.1371/journal.pone.0227258.g004

4. **Aging in the heterozygous mutant brains is similar but distinct from aging in wild type brains**. The 50 gene sets differentially expressed during normal brain aging are also differentially expressed during heterozygous mutant brain aging ('aged heterozygous mutant vs. young heterozygous mutant') (Fig 4B). However, proportions of up- and down-regulated genes differ from those in normal brain aging. This suggests that zebrafish brains bearing this mutation may not properly regulate certain gene sets during aging (e.g. cholesterol homeostasis, adipogenesis, DNA repair, hypoxia, Wnt/β-catenin signalling).

Altered protein abundance in the heterozygous mutant zebrafish brains

Despite its high sensitivity, estimating gene expression does not capture regulatory processes or post-transcriptional modifications that might affect the amount of active protein. Correlation between gene expression and protein abundance in samples from multicellular organisms has been notoriously low [39], but analysing proteomics data alongside gene expression data has been shown to be an effective complementary approach [40]. Because of this, we decided to use LC-MS/MS to compare protein abundance in heterozygous mutant zebrafish brains relative to wild type siblings. Overall, 323 proteins were reliably quantified across all samples aged 6 or 24 months. Testing for differential protein abundance was done analogously to
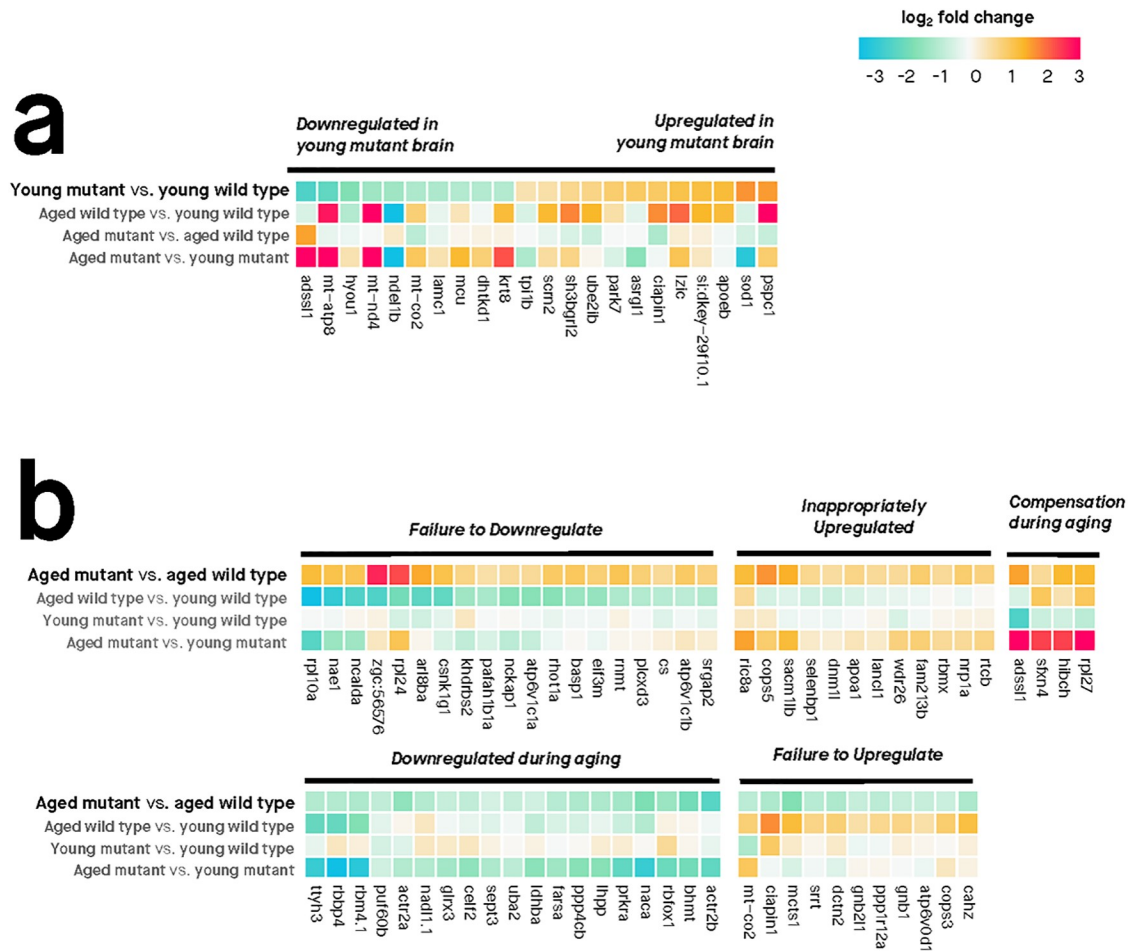
55

Fig 5. Protein abundance changes in the brains of heterozygous mutant (*psen1*[K97fs/+]) zebrafish compared to wild type (*psen1*[+/+]) siblings at 6 months (young) and 24 months (aged). Protein abundance was quantified at the peptide-level with LC-MS/MS (liquid chromatography tandem mass spectrometry) and differential abundance was assessed using moderated *t*-tests (*limma*). Differentially abundant proteins are defined as those with FDR-adjusted *p*-value < 0.05. Protein names were used to retrieve equivalent gene symbols for display purposes on these heatmaps. **(A) Differentially abundant proteins between young heterozygous mutant and wild type zebrafish brains. (B) Differentially abundant proteins between aged heterozygous mutant and wild type zebrafish brains**. The proteins have been clustered according to their abundance changes across the four comparisons.

https://doi.org/10.1371/journal.pone.0227258.g005

testing for differential gene expression, with differences at FDR-adjusted *p*-value < 0.05 considered statistically significant. 22 proteins were differentially abundant between 6-month-old heterozygous mutant and wild type brains, while 65 proteins were differentially abundant between 24-month-old heterozygous mutant and wild type brains (Fig 5; **Fig D in** S1 File). Unexpectedly, three proteins found to be differentially abundant between 6-month-old heterozygous mutant and wild type zebrafish have causative roles in human neurodegenerative diseases: apolipoprotein Eb (encoded by the zebrafish *apoeb* gene, orthologous to the major human genetic risk factor for sporadic AD, *APOE*), superoxide dismutase (encoded by the zebrafish *sod1* gene, orthologous to the human *SOD1* gene mutated in familial amyotrophic lateral sclerosis), and protein DJ-1 (encoded by the zebrafish *park7* gene, orthologous to the human *PARK7* gene mutated in familial Parkinson's disease). Overall, correlation between gene expression and protein abundance was low with $r_s = 0.4$ at 6 months of age and $r_s = 0.28$ at

56

24 months of age (**Figs E** and **F** in S1 File). However, this is overall consistent with previously reported correlation coefficients in multicellular organisms that range from 0.09 to 0.68 [39]).

Gene expression changes in the heterozygous mutant zebrafish brains can be compared to those in human AD

Our results indicate that gene expression changes involving diverse cellular processes occur in aged heterozygous mutant zebrafish brains. The K115fs mutation is a human fAD mutation, but the majority of human AD cases are sporadic, arise from diverse environmental and genetic risk factors, and can involve heterogenous pathological changes in the brain. Nevertheless, it may be informative to explore the extent to which the changes in aged heterozygous mutant zebrafish can model those in human AD.

To assess the similarity of these changes to human brains with AD, we compared gene expression patterns in our zebrafish RNA-seq dataset and an independent human RNA-seq dataset from the Mayo RNA-seq study. The Mayo RNA-seq dataset includes not only patients with AD (defined as having dementia symptoms, Braak neurofibrillary tangle stage IV or greater, and presence of amyloid pathology) and similarly aged controls, but also patients with other brain afflictions that recapitulate aspects of AD ("pathological aging" patients possessing amyloid pathology without dementia symptoms, and progressive supranuclear palsy patients possessing neurofibrillary tangle pathology but no amyloid pathology) [41].

We constructed separate gene co-expression networks from the zebrafish and human RNA-seq datasets. Each network only included genes that were orthologs in humans and zebrafish. Whilst there are many methods for constructing a co-expression network of gene expression [42], we used the weighted gene co-expression network analysis (WGCNA) method [43], which has previously been used to group genes expressed in the brain into "modules" associated with biological functions or activities [44–48]. The zebrafish brain co-expression network is shown in **Fig 6**, and the human brain co-expression network is provided in **Fig G in** S1 File.

We identified 30 modules (i.e. groupings of genes) in the zebrafish brain co-expression network containing between 54 and 1221 genes each and 27 modules in the human brain co-expression network containing between 62 and 921 genes each. We used two methods to confirm that most modules represented functional relationships between genes: enrichment analysis (for identifying enriched biological functions and enriched promoter motifs), and correlating modules with particular traits of interest (age and/or *psen1* genotype). By correlating modules with particular zebrafish traits (age and *psen1* genotype), we identified 13 (out of 30) modules showing evidence of altered expression patterns in heterozygous mutant zebrafish brains (Fig 6B). Using enrichment analysis, we identified the biological relevance of each module in the zebrafish co-expression network (see Table 1 for a summary, and full enrichment analysis results are shown in S9–S11 Tables). Overall, the majority of modules in the zebrafish and human networks show significant enrichment in known functional annotations (e.g. Gene ontology terms, MSigDB gene sets, KEGG pathways, with Bonferroni-adjusted *p*-value < 0.05), supporting the idea that these modules are likely to represent biologically relevant groupings of genes. Some of the biological functions represented by different modules in the zebrafish brain include: G-protein coupled receptor signalling pathway (represented by module **13**), TGF-β and Wnt/β-catenin signaling (represented by module **15**), PI3K/AKT activation (represented by module **18**), immune response (represented by module **20**), regulation of MAPK cascade (represented by module **26**), and oxidative phosphorylation (represented by module **27**). The genes in several modules in the zebrafish network were also significantly enriched in promoter motifs including the glucocorticoid
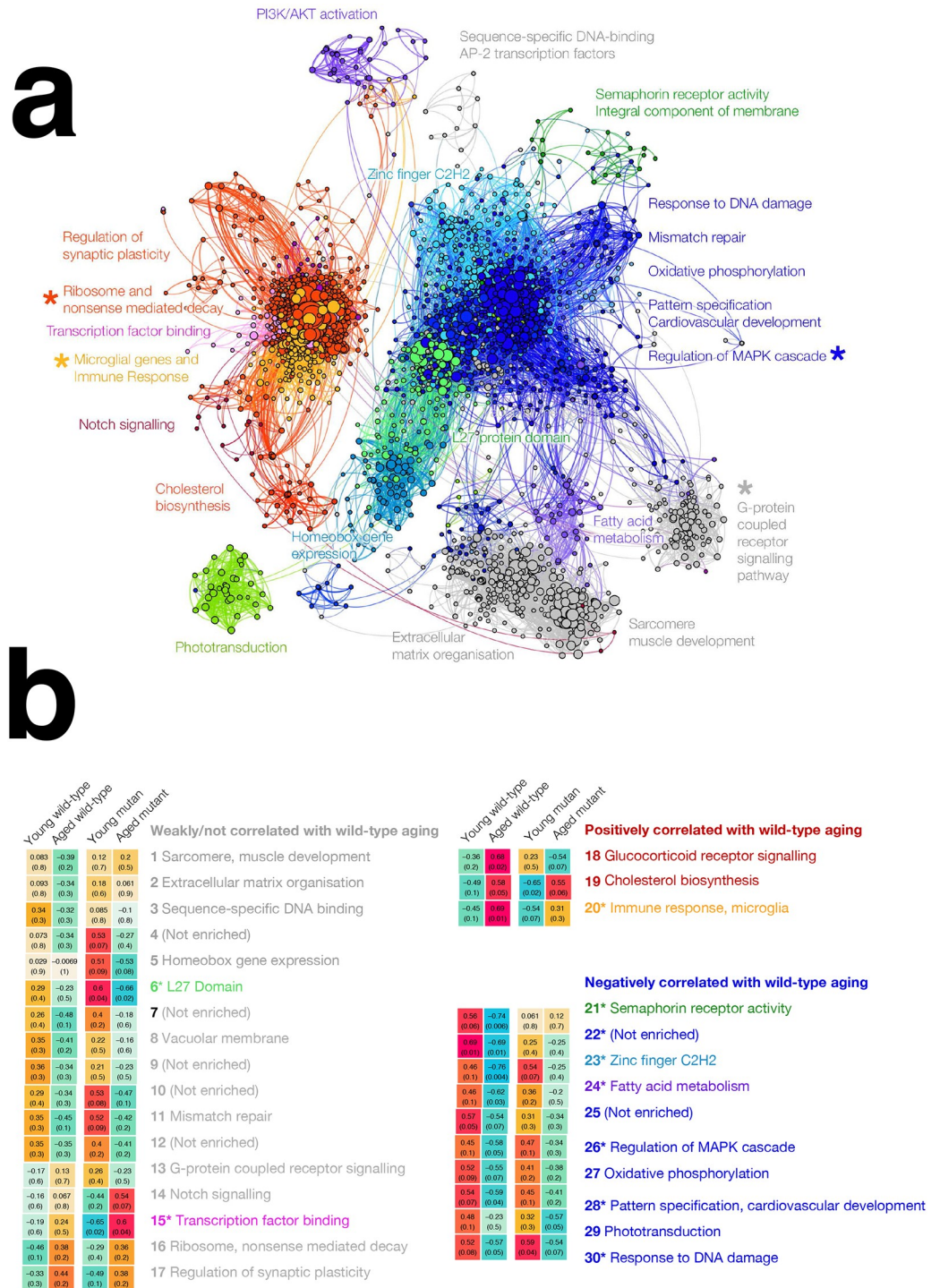
57

Fig 6. Zebrafish brain gene co-expression network. (A) Gene co-expression network visualisation. Each node represents one gene, with node size proportional to the number of connected nodes (co-expressed genes). Edges represent co-expression between two genes, with edge weight proportional to the strength of co-expression. The co-expression network is a signed adjacency matrix constructed from RNA-seq data from wild type and heterozygous mutant zebrafish brains at 6 and 24 months of age. Only nodes with at least four connections are shown. Gene "modules" are groups of genes with similar expression patterns across heterozygous mutant and wild type zebrafish brains. In this network, 30 gene modules were identified using a hierarchical clustering and branch cutting method. Modules showing no significant changes in expression are coloured grey, modules showing significantly increased expression during wild-type brain aging are coloured red, while modules showing significantly decreased expression during wild-type brain aging are

58

coloured blue. Modules with other colours also show signficantly altered expression during heterozygous mutant brain aging. See **B** for details. Asterisks indicate zebrafish brain gene modules which are significantly preserved in a co-expression network constructed from an independent human brain dataset. **(B) Gene expression patterns of modules in the gene co-expression network across heterozygous mutant and wild type zebrafish brains at 6 months and 24 months of age**. Values shown in cells are hybrid Pearson-robust correlations between the overall gene expression in a module (summarised using the first principal component) and experimental condition encoded as a binary variable (6-month-old heterozygous mutant, 24-month-old heterozygous mutant, 6-month-old wild type, 24-month-old wild type). Values in parentheses are unadjusted Student correlation p-values. Modules showing potentially altered expression patterns during heterozygous mutant aging compared to wild-type aging are labelled with coloured text, with colours corresponding to module colours in **(A)**. Asterisks indicate zebrafish brain gene modules which are significantly preserved in a co-expression network constructed from an independent human brain dataset.

receptor element (GRE) motif (for genes in module **18**), GATA3 motif (for genes in module **23**), and numerous ETS transcription factor motifs (for genes in module **20**) (Bonferroni-adjusted *p*-values < 0.05) (S12 Table).

Several pathological changes in the heterozygous mutant zebrafish brains are similar to those in human AD brains

There are several methods for assessing whether modules are preserved across two independent gene co-expression networks constructed using the same genes [49]. The most easily interpretable method is to compare directly the assignment of equivalent genes to modules identified in each network. The resulting overlap in gene co-expression patterns across the two networks can be visualised using a Sankey diagram (Fig 7). Overall, the gene co-expression patterns in the zebrafish brain appear to be broadly similar to the gene co-expression patterns in the human brain, despite differences in RNA-seq platform and brain regions used, which would be expected to make the networks less comparable. A more sophisticated method of assessing module preservation involves using permutation-based *Z*-statistics to test whether certain properties of modules (e.g. density, connectivity) defined in one co-expression network are preserved in another network [49]. *Z*-statistics for each module property can be summarised into a *Z*-summary score, with *Z*-summary scores less than 2 indicating no module preservation, scores between 2 and 10 indicating weak to moderate module preservation, and scores above 10 indicating strong preservation [49]. When comparing zebrafish and human brain co-expression networks, four of the 30 zebrafish modules (**16**, **20**, **13**, **26**) have *Z*-summary scores between 2 and 10, indicating weak to moderate preservation in the human co-expression network (Table 1, S9 Table). While modules **16** (enriched in functions relating to ribosome and nonsense mediated decay) and **13** (enriched in G-protein coupled receptor activity) do not show significant differences between heterozygous mutant and wild-type brains as they age, modules **20** and **26** display distinct coordinated changes in expression during aging of wild-type brains. Module **20** genes are enriched in immune response functional terms and tend to be up-regulated with aging (correlation *p*-value 0.01), while module **26** genes are enriched in terms relating to regulation of the MAPK cascade and tend to be down-regulated with aging (correlation *p*-value 0.05, Fig 6). Importantly, these coordinated gene expression changes appear to be lost in aged heterozygous mutant brains (correlation *p*-values of 0.3 and 0.3 respectively), suggesting the K97fs mutation in *psen1* may contribute to alterations in at least these biological functions. Notably, module **26** which is enriched in immune response functions also displays significant enrichment in ETS and IRF promoter motifs (see Table 1, all FDR-adjusted *p*-values < 0.05). The equivalent module in the human co-expression network also displays enrichment in these particular motifs, suggesting that the regulation of immune and microglial gene expression responses is likely well conserved between aged zebrafish and human brains.

59

Table 1. Summary of modules in a co-expression gene expression network constructed from zebrafish RNA-seq data and their preservation in an independent human brain microarray data set.

Module ID	Number of Genes	Z-Summary Score	Top Functional Enrichment Terms (FDR *p*-value < 0.05)	Promoter Motif Enrichment (FDR *p*-value < 0.05)	Cell Type Marker Enrichment (FDR *p*-value < 0.05)
Random	1000	-0.450372394	-	-	-
1	262	0.951510976	Sarcomere, muscle structure development	-	-
2	209	1.967152607	Extracellular matrix organisation	EKLF(Zf)	-
3	137	0.864232356	Sequence specific DNA binding, AP2 transcription factors	-	-
4	54	0.097738923	-	-	-
5	133	0.342079228	Homeobox genes	-	-
6	212	-0.201216548	L27 protein domain	-	-
7	98	1.004032939	-	-	-
8	67	0.62219677	Vacuolar membrane	-	-
9	1493	3.032384176	-	-	-
10	57	-0.720174553	-	-	-
11	319	-0.828726101	Mismatch repair	-	-
12	163	0.101290609	-	-	-
13	668	3.95602259	G-protein coupled receptor signalling pathway	-	Neuron
14	81	0.842726312	Gland morphogenesis, Notch	Hoxb4(Homeobox)	-
15	82	1.741071825	Transcription factor binding	-	-
16	62	5.804071733	Ribosome, nonsense mediated decay	GFX, ZBTB33(Zf), ERG(ETS)	-
17	59	1.92925347	Synapse part, regulation of synaptic plasticity	-	-
18	189	-0.849139273	PI3K/AKT activation	GRE(NR)	-
19	1221	0.552830789	Cholesterol biosynthesis	-	-
20	381	5.071793968	Immune response	SpiB(ETS), ELF3(ETS), PU.1(ETS), IRF1(IRF), IRF8(IRF), PU.1-IRF (ETS:IRF), EWS:ERG-fusion(ETS), ELF5(ETS), IRF3(IRF), IRF2(IRF), ISRE(IRF), EHF(ETS), PU.1:IRF8(ETS:IRF), EBF(EBF), SPDEF(ETS)	Microglia
21	97	-0.689019253	Semaphorin receptor activity, integral component of membrane	-	-
22	127	0.53275753	-	-	-
23	309	1.081006115	Zinc Finger C2H2	GATA3(Zf)	-
24	103	0.322883371	Oxidative phosphorylation, fatty acid metabolism	-	-
25	55	-0.250719822	-	-	-
26	90	2.432423753	Regulation of MAPK cascade	-	-
27	69	-0.222869961	Oxidative phosphorylation	-	-
28	59	0.023104171	Pattern specification process, cardiovascular system development	-	-
29	171	0.553682784	Phototransduction	-	-
30	549	-0.624680799	Response to DNA damage	ETS1	-

The *Z*-Summary preservation score is a statistic that aggregates various *Z*-statistics obtained from permutation tests of the coexpression network to test whether network properties such as density and connectivity in the zebrafish co-expression network are preserved in an independent co-expression network constructed from human brain gene expression data. In this analysis, 200 permutations were used. *Z*-summary scores less than 2 indicate no preservation, while scores between 2 and 10 indicate weak-to-moderate evidence of preservation. The top functional enrichment and cell type marker enrichment terms are used to give insight into possible biological functions represented within each module. Cell type marker enrichment gene sets are from MSigDB, while functional enrichment terms are from Gene Ontology and MSigDB gene sets. The "Random" module is a random sample of 1,000 genes in the zebrafish co-expression network expected to show non-significant preservation (*Z*-summary < 2) in the human co-expression network. Shaded rows indicate zebrafish gene modules identified as showing significant preservation in the human network.
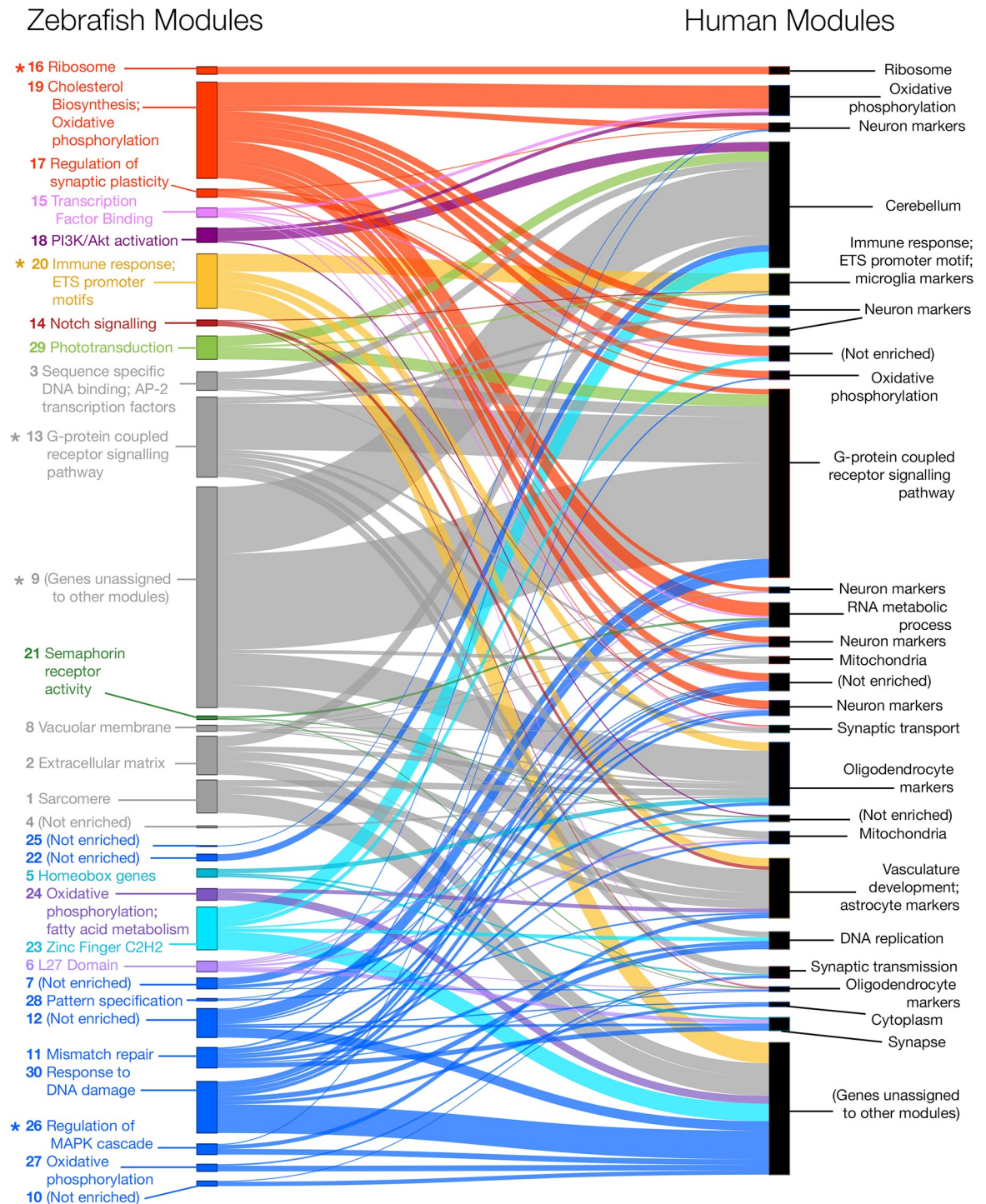
https://doi.org/10.1371/journal.pone.0227258.t001

60

Fig 7. Module overlap between co-expression networks constructed using zebrafish and human brain gene expression data.
Zebrafish and human co-expression networks were constructed using 7,118 genes that were orthologs in zebrafish and humans and expressed in brain gene expression data. Modules of co-expressed genes were separately identified for both the zebrafish and human co-expression networks, resulting in 30 modules in the zebrafish network (left) and 27 modules in the human network (right). Several zebrafish modules (indicated with asterisks) were found to have Z-summary preservation score > 2, indicating statistically significant weak-to-moderate preservation of these modules (i.e. genes in these modules still tend to be co-expressed) in the human brain co-expression network. Four out of five of these modules also showed statistically significant functional enrichment. See Table 1 for more details on the Z-summary preservation scores and functional enrichment for each module in the zebrafish co-expression network.

https://doi.org/10.1371/journal.pone.0227258.g007

61

Aged heterozygous mutant brains possess increased abundance of microglia

The changes we observed in immune-microglia gene co-expression in the aged heterozygous mutant brains prompted us to ask whether differences might be observable in microglial form or even abundance. We used immunostaining for the pan-leukocyte marker L-plastin to detect microglia on sections of fixed brain material from 24-month-old wild type and heterozygous mutant zebrafish (Fig 8). An increased abundance of cells expressing L-plastin was evident in *psen1*^K97fs/+^ heterozygotes in both ventricular (Fig 8C.i and 8E.i) and parenchymal (Fig 8C.ii and 8F.i) regions compared to wild type brains (Fig 8D.i and 8D.ii). We observed significant differences in mean fluorescent intensity (MFI) of the image in the L-plastin channel indicating increased abundance of cells expressing L-plastin in the forebrain, midbrain and hindbrain regions of heterozygous mutant and wild type fish (Fig 8G, $p = 0.0048$, $p = 0.0005$, $p<0.0001$ respectively; two-way ANOVA with Sidak's multiple comparisons test). This immunostaining was also capable of distinguishing between distinct morphologies of microglia in the ventricular (amoeboid "activated" morphology) and parenchymal (ramified morphology) regions in the zebrafish brain (**Fig H in** S1 File) although there was no obvious variation in morphology observed between heterozygous mutant and wild type brains.

Molecular changes in the aged heterozygous mutant zebrafish brains occur without obvious histopathology

Teleosts (bony fish) such as the zebrafish show impressive regenerative ability following tissue damage that includes repair of nervous tissue. Previous attempts to model neurodegenerative
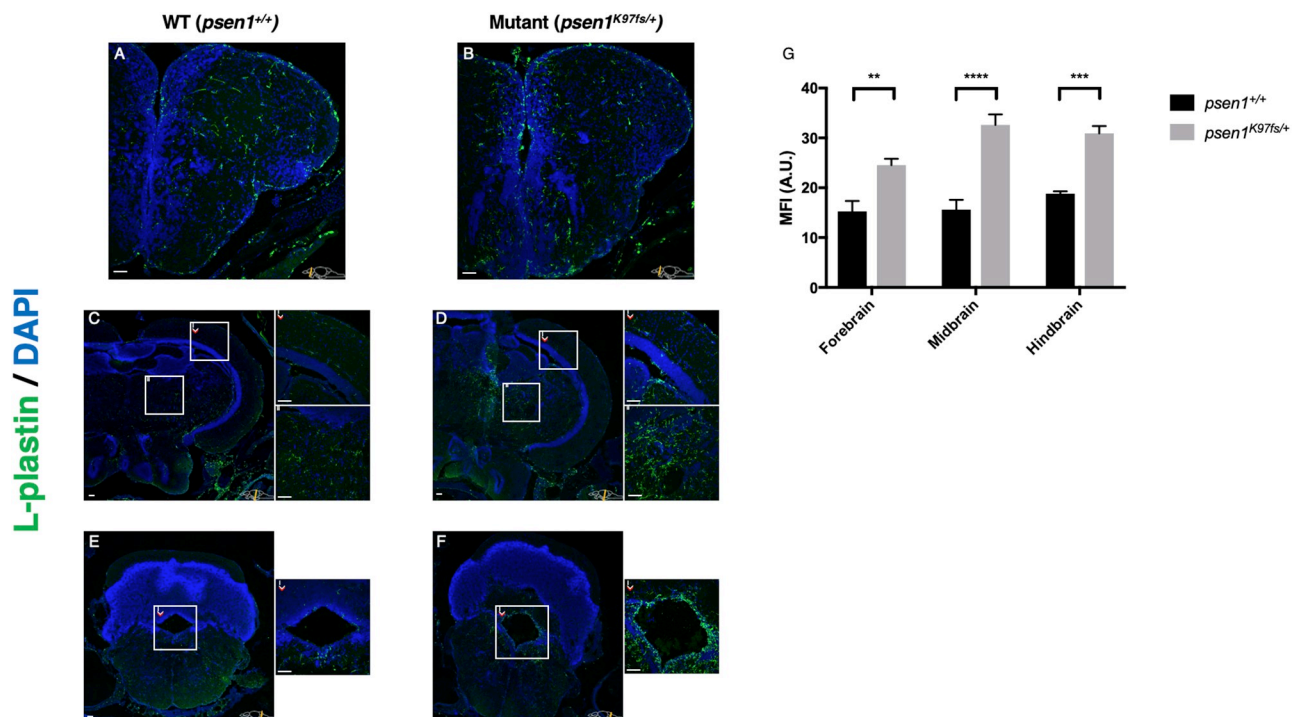

Fig 8. Cells expressing L-plastin are more abundant across the heterozygous mutant (*psen1*^K97fs/+^) zebrafish brain than in wild type siblings at 24 months. Immunostaining for the pan-leukocyte marker L-plastin supports increased numbers of microglia in the forebrain (**A-B**), midbrain (**C-D**) and hindbrain (**E-F**). Increased microglial abundance is evident in *psen1*^K97fs/+^ heterozygotes in both ventricular (**D.i, F.i**) and parenchymal (**D.ii**) regions compared to wild types (**C, E**). (**G**) Significant differences in MFI were observed between the forebrain, midbrain and hindbrain of psen1K97fs/+ and psen1+/+ fish; **p = 0.0048, ***p = 0.0005, ****p < 0.0001; two-way ANOVA with Sidak's multiple comparisons test. Data presented as means with SEM. Scale bar 50 μm in all images.

https://doi.org/10.1371/journal.pone.0227258.g008

62

diseases in adult zebrafish have failed to show cellular phenotypes [50]. Also, zebrafish are thought unlikely to produce the Aβ peptide [51] that many regard as central to AD pathological mechanisms [52]. The analyses described in this paper support that a fAD mutation mimicking PS2V formation may accelerate aspects of brain aging and promote a shift in aged heterozygous mutant brains towards an altered, pathological state of gene and protein expression. We therefore made histopathological comparisons of aged (24 months) wild type and heterozygous mutant brains equivalent to those used in our 'omics analyses. Analysis of various brain regions using markers of aging, senescence and amyloid accumulation (lipofuscin, senescence-associated β-galactosidase, and Congo Red staining respectively) revealed no discernible differences (see Materials and methods and **Figs I-K in** S1 File). This is consistent with the lack of neurodegenerative histopathology observed in a heterozygous knock-in model of a *PSEN1* fAD mutation in mice [25].

Discussion

Fig 9 summarises the main molecular changes that occur with aging and heterozygous mutation.

Evidence of increased stress long preceding AD

We identified a subset of 'inverted' genes that are up-regulated in young heterozygous mutant brains, but down-regulated in aged heterozygous mutant brains. Although this pattern might be overlooked, similar patterns have been observed in human cases. Patients with Mild Cognitive Impairment, pre-clinical AD, or Down Syndrome (who often develop AD in adulthood) initially display increased expression of particular genes, which show decreased expression when AD symptoms become more severe [23, 53–55]. Collectively, results from these studies and our heterozygous mutant zebrafish suggest that early increases in brain activity likely precede AD symptoms in both *PSEN1*-mutation carriers and more general cases of AD. Evidently, to find strategies for preventing AD progression while patients are still asymptomatic, it is important to understand the causes of this increased gene activity in the brain.

Our results suggest that stress responses likely contribute to early increases in brain activity for fAD mutation carriers. In heterozygous mutant zebrafish, the inverted gene expression pattern seems to arise from altered glucocorticoid signalling. In humans, chronically increased glucocorticoid signalling in the brain can lead to glucocorticoid resistance, whereby the brain is unable to increase glucocorticoid signalling even during stressful conditions [56, 57]. We did not confirm whether glucocorticoid signalling and cortisol levels were altered in zebrafish brains *in vivo*. However, many of the inverted genes possess glucocorticoid receptor elements in their promoters, with one particular inverted gene (*fkbp5*) encoding a protein which is known to bind directly to the glucocorticoid receptor to negatively regulate its activity. Previous studies in humans demonstrate that *fkbp5* levels are highly responsive to chronic stress and stress-related diseases (e.g. bipolar disorder; depression in AD [58]), implying that *fkbp5* expression is a sensitive marker of glucocorticoid signalling. Our analysis supports this idea, with *fkbp5* mRNAs showing a significant difference in expression between heterozygous mutant and wild type brains (logFC = 2.1, FDRp = 1.77e-06 in young heterozygous mutant vs wild type; logFC = -3.9, FDRp = 3.16e-08 in aged heterozygous mutant vs wild type). Aside from altered glucocorticoid signalling, we also found altered gene expression patterns associated with diverse biological changes in heterozygous mutant zebrafish brains. If we assume that these heterozygous mutant zebrafish model some aspects of human AD, then these alterations may offer insight into early changes in the brains of human fAD-mutation carriers and, potentially, other individuals predisposed to AD. The brains of young heterozygous mutant

63

Fig 9. Summary of the molecular changes in the brains of zebrafish due to aging and/or the K115fs-like mutation (*psen1*^{K97fs/+}). For each of the four pairwise comparisons shown, the summarised molecular changes (↑ = overall increased, ↓ = overall decreased, • = significant alterations but not in an overall direction) were inferred from a combination of the following analyses: functional enrichment analysis of differentially expressed genes and proteins, promoter motif enrichment analysis of differentially expressed genes, gene set enrichment analysis of differentially expressed genes, and weighted co-expression network analysis of the gene expression data.

https://doi.org/10.1371/journal.pone.0227258.g009

zebrafish exhibit changes relating to developmental signalling pathways (Wnt/β-catenin signalling, hedgehog signalling, TGF-β signalling), stress and immune responses (DNA repair, IL2-STAT5 signalling, complement system, IFN-γ response, inflammatory response), hormonal changes (early and late estrogen responses, androgen response), and energy metabolism (glycolysis, oxidative phosphorylation). Appropriate regulation of these biological processes is critical for brain function, so it is unsurprising that disruption of these processes in the brain

has been linked previously to various pathological states, including early stages of neurodegeneration [59–62].

Quantifying protein abundance in young heterozygous mutant zebrafish brains revealed additional sources of early-life stress. In young heterozygous mutant zebrafish brains, proteins associated with oxidative stress responses and energy metabolism in mitochondria already demonstrated altered abundance. Overall, stress responses were increased, consistent with the RNA-seq data, and decreased abundance of metabolic and antioxidant proteins imply mitochondrial function was likely already impaired. Both increased oxidative stress and altered energy metabolism are known to be early events in AD [2, 63–69], consistent with the idea that these events may contribute to early stress responses in the brain.

Involvement of microglia-mediated immune responses in AD

Our analysis identified two modules with altered gene co-expression patterns in both the aged heterozygous mutant zebrafish brains and post-mortem human AD brains. These modules demonstrated significant functional enrichment in immune and microglial responses (module **20** in the zebrafish network) and regulation of the MAPK cascade (module **26** in the zebrafish network), consistent with their well-established dysfunction in human AD [9, 36, 70, 71]. Gene co-expression changes associated with the immune-microglia responses and the MAPK cascade were evident in aged but not young heterozygous mutant brains, suggesting that these changes are likely to occur in later stages of AD pathogenesis. Our results are consistent with two independent studies involving co-expression analysis of AD brains by Miller et al. [47] and Zhang et al. [48] which also identified a prominent immune-microglia module demonstrating similar changes in gene co-expression in AD patients, despite differences in patient cohorts used, brain regions and tissue types sampled, RNA-seq or microarray platforms, and methodology used to construct the gene co-expression networks. Collectively, the results from these studies and our analysis support the involvement of microglia-mediated immune responses in late stages of AD pathogenesis.

Our analysis reveals additional insights that help explain the involvement of the immune-microglia module in AD. Promoter enrichment analysis of genes in the immune-microglia module indicates statistically significant enrichment in several known motifs. Interestingly, all of these motifs are binding sites for transcription factors from either the ETS (SpiB, ELF3, ELF5, PU.1, EHF) or IRF (IRF3, IRF8, IRF1) families. This finding is important, because 1) ETS and IRF transcription factor motifs are also enriched in the promoters of genes that are up-regulated with brain aging in wild type zebrafish, but not in genes that are up-regulated with brain aging in heterozygous mutant zebrafish. This suggests that the genes they regulate are important during normal brain aging and that their dysregulation may contribute to pathology. 2) ETS and IRF transcription factors are known to mediate critical biological functions, with ETS factors regulating cellular differentiation, proliferation, cell-cycle control, apoptosis, migration and mesenchymal-epithelial interactions [72, 73], and IRF factors mediating immune and other stress responses. Our results are consistent with those in a previous study by Gjoneska et al. [74] that analysed RNA-seq and ChIP-seq (chromatin immunoprecipitation sequencing) data from mouse and human brain tissues, which found that immune response genes were up-regulated in both the CK-p25 mouse model and in human sporadic AD, that these genes were enriched in ChIP-seq peaks corresponding to ETS and IRF transcription factor motifs, and that microglia-specific activation was likely responsible for these gene expression changes.

Immunohistochemistry on sections from aged brains to identify L-plastin-expressing cells, (thought to represent microglia), revealed an increased abundance of these cells in

65

heterozygous mutant fish compared to wild type fish but no obvious genotype-dependent differences in cell morphologies. The concentration of these cells in ventricle-proximal regions suggests an involvement with neural cell proliferation [75] which might occur in the regenerative zebrafish brain if the rate of cell turnover was increased due to pathological processes and this deserves future investigation. The increased abundance of L-plastin-expressing cells was not reflected in a noticeable increase in the mean expression of multiple microglial marker genes from the RNA-seq data. However, the RNA-seq data was derived from entire zebrafish brains and this may have obscured region-specific differences in microglial abundance, morphology, and activation.

Heterozygous mutant zebrafish in our study overall appear to recapitulate partially certain transcriptional and molecular changes that occur in more general cases of sporadic AD. Although revealing valuable insights, our comparison of the gene co-expression patterns in the zebrafish and human datasets is limited by inherent differences in species-level gene expression, differences in the brain regions and tissues sampled in each dataset, and difference in the RNA-seq platforms used to collect data, which has been previously shown to affect network properties including connectivity and density of modules [42]. In addition, the heterogeneity of sporadic AD would likely result in variation in gene expression patterns which may also confound our ability to identify reliably gene modules showing similar expression patterns across all samples. All of these differences would likely have contributed to decreasing our ability to detect preservation of modules between the zebrafish and human co-expression networks.

AD-like gene expression changes can occur without amyloid pathology typically associated with AD

Somewhat surprisingly, the gene and protein expression changes observed in our aged heterozygous mutant zebrafish were not reflected in an obvious histopathology. However, this is consistent with an attempt to model neuronal ceroid lipofuscinosis in adult zebrafish [50] and with observations from heterozygous fAD mutation knock-in models in mice [25–27] (although, in general, mouse single heterozygous mutation brain histology phenotypes have not been reported). It is important to realise that differences in scale between the mass of a human brain and the brains of mice and zebrafish, (~1,000-fold and ~200,000 fold respectively) mean that any metabolic or other stresses in the small brains of the genetic models are likely exacerbated in the huge human brain [76]. Human brains also lack the regenerative ability of zebrafish, while mice and zebrafish both show sequence divergences in the Aβ regions of their APP orthologous genes greater than seen in most mammals [33, 77, 78]. Nevertheless, the heterozygous fAD-like mutation models of mice and (with this paper) zebrafish are probably the closest one can come to modelling AD in these organisms without subjectively imposing an opinion of what AD is by addition of further mutations or transgenes.

It is important to remember that the pathological role in AD of Aβ, neuritic plaques, and neurofibrillary tangles is still debated and that around one quarter of people clinically diagnosed with AD lack typical amyloid pathology upon post-mortem examination [79]. By the current definition, these people do not have AD [80] although this restrictive definition has been questioned [81, 82]. Many people also have brains containing high levels of Aβ [83] or Braak stage III to VI neurodegeneration [79] without obvious dementia. Thus the connection between amyloid pathology, histopathological neurodegeneration and Alzheimer's disease dementia is unclear. Our data indicate that the AD cellular pathologies may occur subsequent to cryptic but dramatic changes in the brain's molecular state (gene and protein expression) that are the underlying drivers of AD.

Finally, it is also important to acknowledge that the specific fAD mutation modelled in this study (K115fs of *PSEN2*) is an uncommon fAD mutation which produces novel alternative transcripts and splice isoforms, and it is unclear how pathogenic effects of this mutation might compare to other more common fAD mutations [84]. When beginning this research, we initially hypothesised that the alternative protein product PS2V (exon 6 deletion) produced from mutant K115fs *PSEN2* played a pathogenic role in AD. PS2V has previously been detected in sAD brains [29], and our previous research suggested dominant-negative effects of the functionally similar PS2V-like protein product produced from zebrafish *psen1* [32, 85]. Recent research suggests that some aberrant transcripts derived from the human K115fs mutant allele may, in fact, follow the "fAD mutation reading frame preservation rule" that is obeyed by all other fAD mutations in the *PSEN* genes [84]. If this is true, then our zebrafish model of K115fs is best regarded as illuminating the contribution that PS2V-mimicry by the K115fs mutation can make to its overall fAD phenotype. Nevertheless, our results indicate that the contribution made by such PS2V-mimicry is likely to be very significant. Our laboratory has been developing additional heterozygous mutant zebrafish modelling other forms of fAD mutation [86], and future analysis incorporating these zebrafish to produce a consensus co-expression network should help to identify and refine a "signature" of the transcriptome and proteome changes that cause fAD.

Materials and methods

Zebrafish husbandry and animal ethics

This study was approved under permits S-2014-108 and S-2017-073 issued by the Animal Ethics Committee of the University of Adelaide. Tübingen strain zebrafish were maintained in a recirculated water system.

Generation of TALEN coding sequences and single stranded oligonucleotide

TALEN coding sequences were designed by, and purchased from, Zgenebio (Taipai City, Taiwan). The DNA binding sites for the TALEN pair targeting *psen1* were (5' to 3'): left site, CAAATCTGTCAGCTTCT and right site, CCTCACAGCTGCTGTC (**Fig A in** S1 File). The coding sequences of the TALENs were provided in the pZGB2 vector for mRNA *in-vitro* synthesis. The single stranded oligonucleotide (ssoligo) sequence was designed such that the dinucleotide 'GA' deletion was in the centre of the sequence with 26 and 27 nucleotides of homology on either side of this site (**Fig A in** S1 File). The ssoligo was synthesized by Sigma-Aldrich (St. Louis, Missouri, USA) and HPLC purified. The oligo sequence was (5' to 3'): CCATCAAA TCTGTCAGCTTCTACACACAAGGACGGACAGCAGCTGTGAGGAGC (**Fig A in** S1 File).

In-vitro mRNA synthesis

Each TALEN plasmid was linearized with *Not* I. Purified linearized DNA was used as a template for *in-vitro* mRNA synthesis using the mMESSAGE mMACHINE SP6 transcription kit (Thermo Fisher, Waltham, USA) as per the manufacturer's instructions as previously described [85].

Microinjection of zebrafish embryos

Embryos were collected from natural mating and, at the 1-cell stage, were microinjected with a ~3nl mixture of 250ng/μl of left and right TALEN mRNA and 200ng/μl of the ssoligo.

Genomic DNA extraction of zebrafish tissue

Embryos.   A selection of 10–20 embryos were collected at 24 hpf and placed in 150μl of a 50mM NaOH 1xTE solution and then incubated at 95˚C until noticeably dissolved (10-20mins). The lysis solution was cooled to 4˚C and 50μl of Tris solution (pH 8) was added. The mixture was then centrifuged at maximum speed for 2 mins to pellet cellular debris. The supernatant was transferred into a fresh microfuge tube ready for subsequent PCR.

Adult fin clips.   For fin clips, adult fish were first anesthetised in a 0.16 mg/mL tricaine solution and a small section of the caudal fin was removed with a sharp blade. Fin clips were placed in 50μl of a 1.7 μg/ml Proteinase K 1xTE solution and then incubated at 55˚C until noticeably dissolved (2-3hours). The lysis solution was then placed at 95˚C for 5mins to inactivate the Proteinase K.

Genomic DNA PCR and sequencing for mutation detection

To genotype by PCR amplification, 5 μl of the genomic DNA was used with the following primer pairs as relevant. Primers to detect wild type (WT) sequence at the mutation site: primer psen1WTF: (5ʹ TCTGTCAGCTTCTACACACA*GA*AGG3ʹ) (GA nucleotides in italics) with primer psen1WTR: (5ʹ AGTAGGAGCAGTTTAGGGATGG3ʹ). Primers to detect the presence of the GA dinucleotide deletion: primer psen1GAdelF: (5ʹ AATCTGTCAGCTTCTACAC ACAAGG3ʹ) with primer psen1WTR. To confirm the presence of the GA dinucleotide deletion mutation by sequencing of extracted genomic DNA, PCR primers were designed to amplify a 488 bp region around the GA mutation site: primer psen1GAsiteF: (5ʹ GGCACA CAAGCAGCACCG3ʹ) with primer psen1GAsiteR: (5ʹ TCCTTTCCTGTCATTCAGACCTG CGA3ʹ). This amplified fragment was purified and sequenced using the primer psen1seqF: (5ʹ AGCCGTAATGAGGTGGAGC 3ʹ). All primers were synthesized by Sigma-Aldrich. PCRs were performed using GoTaq polymerase (Promega, Madison, USA) for 30 cycles with an annealing temperature of 65˚C (for the mutation-detecting PCR) or 61˚C (for the WT sequence-detecting PCR) for 30 s, an extension temperature of 72˚C for 30 s and a denaturation temperature of 95˚C for 30 s. PCR products were assessed on 1% TAE agarose gels run at 90V for 30 mins and subsequently visualized under UV light.

Whole brain removal from adult zebrafish

Adult fish were euthanized by sudden immersion in an ice water slurry for at least ~30 seconds before decapitation and removal of the entire brain for immediate RNA or protein extraction. All fish brains were removed during late morning/noon to minimise any influence of circadian rhythms.

RNA extraction from whole brain

Total RNA was isolated from heterozygous mutant and WT siblings using the *mir*Vana miRNA isolation kit (Thermo Fisher). RNA isolation was performed according to the manufacturer's protocol. First a brain was lysed in a denaturing lysis solution. The lysate was then extracted once with acid-phenol:chloroform leaving a semi-pure RNA sample. The sample was then purified further over a glass-fiber filter to yield total RNA. This procedure was formulated specifically for miRNA retention to avoid the loss of small RNAs. Total RNA was then sent to the ACRF Cancer Genomics Facility (Adelaide, Australia) to assess RNA quality and for subsequent RNA sequencing on the Illumina NextSeq platform as paired-end 100bp reads.

RNA extraction and cDNA synthesis from entire brains for digital PCR.   Total RNA was extracted using the QIAGEN RNeasy Mini Kit according to the manufacturer's protocol. The

68

RNA was DNase-treated using RQ1 DNase (Promega) according to the manufacturer's protocol prior to cDNA synthesis. Equal concentrations of total RNA from each brain were used to synthesise first-strand cDNA by reverse transcription with random priming (Superscript III kit; Invitrogen). cDNA was RNaseH treated before use in 3D Quant Studio Digital PCR.

Allele-specific digital quantitative PCR. Digital PCR was performed on a QuantStudio™ 3D Digital PCR System (Life Technologies, Carlsbad, California, USA). 20μL reaction mixes were prepared containing 9 μL 1X QuantStudio™3D digital PCR Master Mix (Life Technologies), 2 μL of 20X Sybr$^{®}$ dye in TE buffer, 25ng cDNA per total reaction (determined from the RNA concentration under the assumption that single strand cDNA synthesis from total RNA was complete), 200nM of specific primers and 6.3 μL of nuclease-free water (Qiagen). 14.5μL of the reaction mixture was loaded onto a QuantStudio™3D digital PCR 20 K chip (Life Technologies) using an automatic chip loader (Life Technologies) according to manufacturer's instructions. Loaded chips underwent thermo-cycling on the Gene Amp 9700 PCR system under the following conditions: 96˚C for 10 min; 45 cycles of 60˚C for 2 min and 98˚C for 30 sec; followed by a final extension step at 60˚C for 2 min. After thermo-cycling, the chips were imaged on a QuantStudio™ 3D instrument [87, 88]. Primers used for *psen1* allele detection were: wild-type allele forward 5' CTACACACAGAAGGACGGACAGC 3', K97fs allele forward 5' TCTGTCAGCTTCTACACACAAGGA 3' and both were paired with a common reverse primer 5' GCCAGGCTTGAATCACCTTGTA 3'.

PCR test for aberrant splicing in the region of the K97fs mutation in zebrafish psen1. Total RNA was extracted from each 24-month-old zebrafish brain using the QIAGEN RNeasy mini Kit (QIAGEN, Hilden, Germany). 250ng of total RNA from each brain was then used to synthesise 20μL of first-strand cDNA by reverse transcription (SuperScript III kit, Invitrogen, Camarillo, California, USA). 10ng of each cDNA preparation (a quantity calculated from the RNA concentration on the assumption that reverse transcription of RNA into cDNA was complete) was used to perform PCR using Phusion high-fidelity DNA polymerase (New England Biolabs, Ipswich, Massachusetts). Each 25μL PCR reaction contained 0.2mM of deoxyribonucleotide triphosphates (dNTPs), 0.4μM of each PCR primer, 1 unit of Phusion polymerase and 10ng of zebrafish brain cDNA template. PCR cycling was performed with 35 cycles of a denaturation temperature of 95˚C for 30s, then an annealing temperature of 60˚C for 30s and then an extension temperature of 72˚C for 2 minutes. PCR products were electrophoresed through a 1% agarose gel in 1×TAE buffer for separation and identification.

Protein extraction and proteomic analysis of adult brain

Sample preparation. Freshly removed entire adult zebrafish brains were lysed under denaturing conditions in 7 M urea (Merck, Darmstadt, Germany) plus complete protease inhibitors (Roche) using a Bioruptor (Diagenode, Seraing, Belgium) in ice cold water. Samples were quantified using the EZQ protein assay (Life Technologies) and the extracts were trypsin-digested using the FASP method [89]. Protein samples were then sent to the Adelaide Proteomics Centre (Adelaide, Australia) for quantification and data acquisition.

Data acquisition. Nano-LC-ESI-MS/MS was performed using an Ultimate 3000 RSLC system (Thermo Fisher Scientific) coupled to an Impact HD™ QTOF mass spectrometer (Bruker Daltonics, Bremen, Germany) via an Advance Captive Spray source (Bruker Daltonics). Peptide samples were pre-concentrated onto a C18 trapping column (THC164535, Thermo Fisher) at a flow rate of 5 μL/min in 2% (v/v) ACN 0.1% (v/v) FA for 10 minutes. Peptide separation was performed using a 75μm ID 50 cm C18 column (THC164540, Thermo Fisher) at a flow rate of 0.2 μL/minute using a linear gradient from 5 to 45% B (A: 5% (v/v) ACN 0.1% (v/v) FA, B: 80% (v/v) ACN 0.1% (v/v) FA) over 180 minutes. MS scans were

69

acquired in the mass range of 300 to 2,200 m/z in a data-dependent fashion using Bruker's Shotgun Instant Expertise™ method (singly charged precursor ions excluded from acquisition, CID from 23% to 65% as determined by the m/z of the precursor ion).

Data analysis. The acquired peptide spectra were identified and quantified using the mass spectrometry software MaxQuant with the Andromeda search engine against all entries in the non-redundant UniProt database (protein and peptide false discovery rate set to 1%). The MaxQuant software allows for the accurate and robust proteomewide quantification of label-free mass spectrometry data [90].

RNA-seq analysis

Data processing. We used *FastQC* [91] to evaluate the quality of the raw paired-end reads. Using *AdapterRemoval* [92], we trimmed reads and removed adapter sequences. From the *FastQC* reports, some over-represented sequences in the raw and trimmed reads corresponded to ribosomal RNA, possibly from insufficient depletion during library preparation. We removed ribosomal RNA sequences *in silico* by aligning all trimmed reads to known zebrafish ribosomal RNA sequences and discarding all reads that aligned. Next, we used *HISAT2* [93] to align reads to the Ensembl zebrafish genome assembly (GRCz10). Using *Picard* [94] and the MarkDuplicates function, we removed optical and PCR duplicates from the aligned reads. To quantify gene expression, we used *FeatureCounts* [95], resulting in a matrix of gene expression counts for 32,266 genes for the 12 RNA-seq libraries.

Differential gene expression analysis. Differential gene analysis was performed in R [96] using the packages *edgeR* [97] and *limma* [98–100]. We retained 18,296 genes with >1.5 cpm in at least 6 of the 12 RNA-seq libraries. We then calculated TMM-normalisation factors to account for differences in library sizes and applied the RUVs method from the *RUVseq* package [101] to account for a batch effect with one factor of unwanted variation ($k = 1$). Differential gene expression analysis was performed using *limma*. We considered genes differentially expressed if the FDR-adjusted *p*-value associated with their moderated *t*-test was below 0.05. We used the *pheatmap* R package [102] to produce all heatmaps.

Gene set testing. We downloaded the Hallmark gene set collection from MSigDB v6.1 [38]. Using *biomaRt* [103, 104], we converted human Entrezgene identifiers to zebrafish Entrezgene identifiers. To perform gene set testing, we applied the fast rotation gene set testing (FRY) method [105] for each comparison. We considered all gene sets with non-directional (Mixed) FDR < 0.05 as differentially expressed. To obtain estimates of the proportions of up-regulated and down-regulated genes for each significant gene set, we used the ROAST [106] method with 9,999 rotations with the 'set.statistic' option set to 'mean', to maintain consistency with the results obtained from FRY.

Promoter motif analysis. We performed promoter motif enrichment analysis using *HOMER* [107, 108] and downloaded a set of 364 zebrafish promoter motifs from published ChIP-seq experiments, as collated by *HOMER* authors, using the command 'configureHomer.pl -install zebrafish-p'. We retained default parameters with the findMotifs.pl program with the following modifications: the 18,296 Ensembl genes considered as expressed in the differential gene expression analysis were specified as the background; and promoter regions were defined as 1500 bp upstream and 200 bp downstream of the transcription start site. We defined motifs as being significantly enriched in a set of genes if the Bonferroni-adjusted *p*-value was less than 0.05.

LC-MS/MS analysis

Data processing. Raw MS/MS spectra were analysed using *MaxQuant* (V. 1.5.3.17). A False Discovery Rate (FDR) of 0.01 for peptides and a minimum peptide length of 7 amino

70

acids was specified. MS/MS spectra were searched against the zebrafish UniProt database. *MaxQuant* output files for the 6-month-old and 24-month-old samples were processed in separate batches with the *MSStats* R package [109] due to an unresolvable batch effect. Briefly, peptide intensities were $\log_2$-transformed and quantile normalised, followed by using an accelerated failure time model to impute censored peptides. Peptide-level intensities were summarised to protein-level intensities using Tukey's median polish method. This resulted in 2,814 peptides (summarised to 534 proteins) for the 6-month-old data and 3,378 peptides (summarised to 582 proteins) for the 24-month-old data. After summarisation, both sets of protein intensities were combined, quantile normalised and filtered to retain the 323 proteins that were detected across all samples.

Differential protein analysis.   Differential protein abundance analysis was performed using *limma* [110] using moderated *t*-tests. Proteins were identified as being differentially abundant if FDR-adjusted p-values were below 0.05. Over-representation analysis using the 'goana' and 'kegga' functions from *limma* were used to test for enriched gene ontology terms and KEGG pathways respectively.

Gene co-expression network analysis

Network construction.   We used the *WGCNA* R package to construct co-expression networks for our zebrafish RNA-seq data and a processed human RNA-seq dataset from the Mayo RNAseq study [41]. The human RNA-seq data consists of 101 bp paired-end reads sequenced with the Illumina HiSeq 2000 platform and derived from cerebellum and temporal cortex samples from North American Caucasian subjects with either AD (n = 86), progressive supranuclear palsy (PSP, n = 84), pathological aging (PA, n = 28) or controls lacking neurodegeneration (n = 80). The Mayo RNAseq study authors performed read alignment and counting using the *SNAPR* software with the GRCh38 reference human genome and Ensembl v77 gene models, and provided TMM-normalised gene counts as output by the *edgeR* package [97, 111]. We matched zebrafish genes to human homologous genes via orthologous Ensembl gene identifiers and retained genes that were expressed in both the human and zebrafish datasets, leaving 8,396 genes for network construction. To reduce noise during network construction, we calculated connectivities for each gene in each dataset and retained 7,576 genes with connectivities above the 10th percentile of all connectivities. To construct approximately scale-free weighted networks, the Pearson correlation was calculated between each pair of genes, and the resulting correlation matrix was raised to the soft-thresholding power of 14 to produce a signed adjacency matrix for each dataset [43]. Next, we applied a transformation to obtain a measure of topological overlap for each pair of genes. Lastly, we hierarchically clustered genes in each dataset based on the measure 1—Topological Overlap. To identify modules of co-expressed genes, we used the Hybrid Tree Cut method from the *dynamicTreeCut* package [112] with default parameters except for the following modifications: minimum module size set at 40 genes, 0.90 as the maximum distance to assign previously unassigned genes to modules during PAM (Partioning Around Medoids) stage, and the deepSplit parameter to 1 for both the human and zebrafish datasets.

Network analysis.   We assessed functional enrichment of each module using default settings in the *anRichment* R package. We assessed promoter motif enrichment using HOMER as described earlier. To calculate the correlation between modules and phenotypic traits, we calculated the hybrid-robust correlation between the first principal component of each module and four binary variables defining the experimental conditions [113]. We evaluated the preservation of zebrafish modules in the human network and vice versa using the modulePreservation function from *WGCNA*, which uses a permutation-based approach to determine whether

71

module properties (e.g. density, connectivity) are preserved in another network [49]. We also used the Sankey diagram functionality in the *networkD3* package to visualise overlap between zebrafish and human modules [114].

Network visualisation. To visualise networks, we imported edges and nodes into *Gephi* and applied the OpenOrd algorithm with default settings [115]. We coloured the nodes (genes) based on their assigned modules from *WGCNA*.

psen1[K97fs/+] vs. wild type L-plastin immunostain

Anti-L-plastin immunostain. Frozen cryosections of adult *psen1*[K97fs/+] (n = 3 fish) or *psen1*[+/+] (n = 4 fish) brains were dried for >1hr at room temperature, then rehydrated in 1x PBS for >30 min. Sections were then washed twice with 0.3% Triton X-100 in 1x PBS (PBS-Tx 0.3%) for 15 min at room temperature (RT). Sections were subsequently incubated with polyclonal rabbit anti-L-plastin primary antibody [116, 117] (a kind gift from Prof. Dr. Michael Brand, Centre for Regenerative Therapies, Technische Universität Dresden) at 1:2500 concentration in PBS-Tx 0.3%, overnight at 4˚C in a humid chamber. Sections were then washed 3 x 20 min in PBS-Tx 0.3% at RT, followed by 1 h incubation at RT with goat anti-rabbit Alexa 488 secondary antibody (Thermo Fisher, 1:750 in PBS-Tx 0.3%) alongside DAPI at 1:5000 concentration. Sections were subsequently washed once for 10 min with PBS-Tx 0.3%, then twice for 20 min with 1x PBS at RT, then mounted with 50% glycerol in 1 x PBS.

Imaging. Z-stacks of brain sections were acquired on a Leica TCS SP8 confocal microscope equipped with a HyD detector, and using the Leica LASX software suite. Stacks were captured at 1024 x 1024 resolution at scanning speed of 600, with bidirectional X scanning active. No averaging or accumulation was applied to stack acquisitions, and laser power and gain were kept constant throughout image acquisition. Overview stacks were captured with either a 10x dry objective (midbrain, hindbrain overviews) or a 20x oil-immersion objective (forebrain overviews), while higher magnification stacks were acquired with a 40x water-immersion objective. ~3 stacks were captured per fish (one of each forebrain, midbrain and hindbrain at approximately equal levels).

Image processing and statistical analysis. Stacks were opened in Fiji 2 (https://imagej. net/Fiji/Downloads), and split into individual channels. The green channel (corresponding to L-plastin in all stacks) was max-projected and the MFI of the entire image was recorded using the Measure function. Statistical analysis was conducted in Prism 7 (GraphPad); MFI between brain regions in *psen1*[K97fs/+] and *psen1*[+/+] fish was compared via two-way ANOVA with Sidak's multiple comparisons test. Statistical significance was defined as $p<0.05$, with all data presented as means with SEMs.

Histological analysis

Tissue preparation. Two-year-old adult zebrafish heterozygous for the psen1K97fs mutation and their wild type siblings were sacrificed by immersion in ice-water, then tails were nicked to exsanguinate the fish and prevent blood clotting on neural tissue. The dorsal neurocranium was subsequently resected to expose the brain. Fish were then decapitated and heads were incubated in a decalcification solution (100 ml 0.5M EDTA, 22 g sucrose, 11 ml 10x phosphate buffered saline solution, PBS) for four hours on a slow shaker at room temperature. Decalcified heads were then fixed overnight in 4% paraformaldehyde in phosphate buffer (PFA in PB), at 4˚C on a slow shaker. Heads were then embedded in a sucrose-gelatin medium (20% sucrose, 8% cold-water fish gelatin in 1x PBS), frozen on dry ice and cryosectioned at 16 µm thickness on a Leica CM3050-S cryostat. Serial sections were

72

mounted on SuperFrost Plus microscope slides (Menzel-Gläser). Sections were subsequently dried at room temperature for four hours, and then stored at -20˚C until staining. Prior to all stains, sections were retrieved from -20˚C and brought to RT, then rehydrated in 1x PBS.

Senescence-associated β-galactosidase (saβgal) staining. Sections were prefixed with 4% PFA in PB for one hour in a humid chamber at room temperature, then washed twice for 15 minutes with 0.3% Triton X-100 in 1x PBS (PBS-Tx (0.3%)). The pH of sections was then equilibrated with two 30 minute washes with 1x PBS at pH 5.5 (all washes were performed in a humid chamber). Sections were then stained for 16 hours at 37˚C in a humid chamber in staining solution (2 mM $MgCl_2$, 5 mM $K_3Fe(CN)_6$, 5 mM $K_3Fe(CN)_6 \cdot 3H_2O$, 1 mg/ml X-gal, with the remaining volume made up of 1x PBS at pH 5.5). Following incubation, staining was arrested by a 20-minute wash in 4% PFA in PB at room temperature. Sections were then washed well in PBS at pH 5.5 and mounted in 50% glycerol. Sections were then imaged on an Olympus Provis AX70 widefield microscope with an Olympus DP70 camera, with images acquired at 4080x3072 resolution.

Congo Red staining for amyloid. Following rehydration in 1x PBS, sections were stained for 20 minutes in Congo Red staining solution (0.5% Congo Red in 50% ethanol). Sections were then rinsed in distilled water, and quickly differentiated by dipping five times in an alkaline alcohol solution (1% NaOH in 50% ethanol). Sections were rinsed for 1 minute in distilled water, then mounted in 50% glycerol. Sections were imaged in both brightfield and birefringence on a Leica Abrio polarising microscope at 1024x1024 resolution.

Autofluorescent detection of lipofuscin. Following rehydration in 1x PBS, sections were mounted in 50% glycerol and imaged confocally on a Leica TCS SP8 invert confocal laser scanning microscope with a Leica HyD hybrid detector using 20x and 63x oil-immersion objectives. Lipofuscin has an emission maximum at 590 nm when excited at 488 nm; samples were thus excited with a 488nm laser at power 5.00, and the detector was gated for emission wavelengths between 560–700 nm. Images were acquired at 1024x1024 resolution with gain at 137.8%.

Supporting information

S1 File. Supporting figures and text.
(PDF)

S1 Table. Differential gene expression results for each comparison.
(XLSX)

S2 Table. Differential protein abundance results for each comparison.
(XLSX)

S3 Table. Summary of functional enrichment results for groups of genes identified in differential gene expression analysis.
(PDF)

S4 Table. Full gene ontology analysis results for groups of genes identified in differential gene expression analysis.
(XLSX)

S5 Table. Summary of promoter motif enrichment analysis in groups of genes identified from differential gene expression analysis.
(PDF)

73

S6 Table. Full enrichment results of known zebrafish promoter motifs in groups of genes identified from differential gene expression analysis.
(XLSX)

S7 Table. List of zebrafish genes containing the GRE (glucocorticoid receptor element) motif. These genes have a 90% match for the GRE motif in the region 1500bp upstream to 200bp downstream of the transcription start site.
(XLSX)

S8 Table. Differentially expressed Hallmark MSigDB gene sets for each comparison identified using the FRY (fast rotation) method. Proportions of up- and downregulated genes in each set were estimated using ROAST (which FRY approximates). Differentially expressed gene sets were defined as having "Mixed FDR" < 0.05.
(XLSX)

S9 Table. Zebrafish module functional enrichment results using anRichment R package.
(XLSX)

S10 Table. Zebrafish module promoter motif enrichment results using HOMER software.
(XLSX)

S11 Table. Human module functional enrichment results using anRichment R package.
(XLSX)

S12 Table. Promoter motif enrichment analysis of the immune-microglia enriched module in the zebrafish network.
(XLSX)

S13 Table. *Z*-score calculation results for preservation of zebrafish co-expression network properties in human co-expression network.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: David L. Adelson, Stephen Pederson, Michael Lardelli.

Data curation: Morgan Newman, Seyed Hani Moussavi Nik, Yang Dong, Michael Lardelli.

Formal analysis: Nhi Hin, Morgan Newman, Jan Kaslin, Alon M. Douek, Amanda Lumsden, Xin-Fu Zhou, Stephen Pederson.

Funding acquisition: Morgan Newman, Michael Lardelli.

Investigation: Nhi Hin, Morgan Newman, Jan Kaslin, Yang Dong, Xin-Fu Zhou, Stephen Pederson, Michael Lardelli.

Methodology: Nhi Hin, Morgan Newman, Jan Kaslin, Alon M. Douek, Amanda Lumsden, Seyed Hani Moussavi Nik, Xin-Fu Zhou, Alastair Ludington, Michael Lardelli.

Project administration: Michael Lardelli.

74

Resources: David L. Adelson, Michael Lardelli.

Software: Alastair Ludington, David L. Adelson, Stephen Pederson.

Supervision: David L. Adelson, Stephen Pederson, Michael Lardelli.

Validation: Jan Kaslin, Stephen Pederson.

Visualization: Nhi Hin, Morgan Newman, Jan Kaslin, Michael Lardelli.

Writing – original draft: Nhi Hin, Morgan Newman, Jan Kaslin, Stephen Pederson, Michael Lardelli.

Writing – review & editing: Nhi Hin, Morgan Newman, Jan Kaslin, Alon M. Douek, Amanda Lumsden, Seyed Hani Moussavi Nik, Yang Dong, Xin-Fu Zhou, Noralyn B. Mañucat-Tan, Alastair Ludington, David L. Adelson, Stephen Pederson, Michael Lardelli.

References

1. Cadonic C, Sabbir MG, Albensi BC. Mechanisms of Mitochondrial Dysfunction in Alzheimer's Disease. Mol Neurobiol. 2015. https://doi.org/10.1007/s12035-015-9515-5 PMID: 26537901.

2. Castellani R, Hirai K, Aliev G, Drew KL, Nunomura A, Takeda A, et al. Role of mitochondrial dysfunction in Alzheimer's disease. J Neurosci Res. 2002; 70(3):357–60. Epub 2002/10/23. https://doi.org/10.1002/jnr.10389 PMID: 12391597.

3. Fedrizzi L, Carafoli E. Ca2+ dysfunction in neurodegenerative disorders: Alzheimer's disease. Biofactors. 2011; 37(3):189–96. Epub 2011/06/24. https://doi.org/10.1002/biof.157 PMID: 21698698.

4. Mills E, Dong XP, Wang F, Xu H. Mechanisms of brain iron transport: insight into neurodegeneration and CNS disorders. Future Med Chem. 2010; 2(1):51–64. https://doi.org/10.4155/fmc.09.140 PMID: 20161623.

5. Moir RD, Atwood CS, Huang X, Tanzi RE, Bush AI. Mounting evidence for the involvement of zinc and copper in Alzheimer's disease. Eur J Clin Invest. 1999; 29(7):569–70. Epub 1999/07/20. https://doi.org/10.1046/j.1365-2362.1999.00472.x PMID: 10411660.

6. Arimon M, Takeda S, Post KL, Svirsky S, Hyman BT, Berezovska O. Oxidative stress and lipid peroxidation are upstream of amyloid pathology. Neurobiol Dis. 2015; 84:109–19. https://doi.org/10.1016/j.nbd.2015.06.013 PMID: 26102023.

7. Oresic M, Hyotylainen T, Herukka SK, Sysi-Aho M, Mattila I, Seppanan-Laakso T, et al. Metabolome in progression to Alzheimer's disease. Transl Psychiatry. 2011; 1:e57. Epub 2011/01/01. https://doi.org/10.1038/tp.2011.55 PMID: 22832349.

8. Poirier J, Miron J, Picard C, Gormley P, Theroux L, Breitner J, et al. Apolipoprotein E and lipid homeostasis in the etiology and treatment of sporadic Alzheimer's disease. Neurobiol Aging. 2014; 35 Suppl 2:S3–10. Epub 2014/06/29. https://doi.org/10.1016/j.neurobiolaging.2014.03.037 PMID: 24973118.

9. Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015; 14(4):388–405. Epub 2015/03/21. https://doi.org/10.1016/S1474-4422(15)70016-5 PMID: 25792098.

10. Zotova E, Nicoll JA, Kalaria R, Holmes C, Boche D. Inflammation in Alzheimer's disease: relevance to pathogenesis and therapy. Alzheimers Res Ther. 2010; 2(1):1. Epub 2010/02/04. https://doi.org/10.1186/alzrt24 PMID: 20122289.

11. Abramsson A, Kettunen P, Banote RK, Lott E, Li M, Arner A, et al. The zebrafish amyloid precursor protein-b is required for motor neuron guidance and synapse formation. Dev Biol. 2013; 381(2):377–88. https://doi.org/10.1016/j.ydbio.2013.06.026 PMID: 23850871.

12. De Strooper B. Proteases and proteolysis in Alzheimer disease: a multifactorial view on the disease process. Physiol Rev. 2010; 90(2):465–94. https://doi.org/10.1152/physrev.00023.2009 PMID: 20393191.

13. Hipp MS, Park SH, Hartl FU. Proteostasis impairment in protein-misfolding and -aggregation diseases. Trends Cell Biol. 2014. Epub 2014/06/21. https://doi.org/10.1016/j.tcb.2014.05.003 PMID: 24946960.

14. Braak H, Braak E. Diagnostic criteria for neuropathologic assessment of Alzheimer's disease. Neurobiol Aging. 1997; 18(4 Suppl):S85–8. Epub 1997/07/01. https://doi.org/10.1016/s0197-4580(97)00062-6 PMID: 9330992.

75

15. Masters CL, Bateman RJ, Blennow K, Rowe CC, Sperling RA, Cummings JL. Alzheimer's disease. Nature reviews Disease Primers. 2015; 1:1–8. https://doi.org/10.1038/nrdp.2015.56 PMID: 27188934

16. Mehta D, Jackson R, Paul G, Shi J, Sabbagh M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010–2015. Expert Opin Investig Drugs. 2017; 26(6):735–9. https://doi.org/10.1080/13543784.2017.1323868 PMID: 28460541.

17. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, et al. Amyloid beta deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol. 2013; 12(4):357–67. Epub 2013/03/13. https://doi.org/10.1016/S1474-4422(13)70044-9 PMID: 23477989.

18. Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med. 2012; 367(9):795–804. https://doi.org/10.1056/NEJMoa1202753 PMID: 22784036.

19. Iturria-Medina Y, Sotero RC, Toussaint PJ, Mateos-Perez JM, Evans AC, Alzheimer's Disease Neuroimaging I. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. Nat Commun. 2016; 7:11934. https://doi.org/10.1038/ncomms11934 PMID: 27327500.

20. Quiroz YT, Schultz AP, Chen K, Protas HD, Brickhouse M, Fleisher AS, et al. Brain Imaging and Blood Biomarker Abnormalities in Children With Autosomal Dominant Alzheimer Disease: A Cross-Sectional Study. JAMA Neurol. 2015; 72(8):912–9. https://doi.org/10.1001/jamaneurol.2015.1099 PMID: 26121081.

21. Reiman EM, Quiroz YT, Fleisher AS, Chen K, Velez-Pardo C, Jimenez-Del-Rio M, et al. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case-control study. Lancet Neurol. 2012; 11(12):1048–56. https://doi.org/10.1016/S1474-4422(12)70228-4 PMID: 23137948.

22. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, et al. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. Proc Natl Acad Sci U S A. 2004; 101(1):284–9. Epub 2003/12/23. https://doi.org/10.1073/pnas.2635903100 PMID: 14688411

23. Berchtold NC, Sabbagh MN, Beach TG, Kim RC, Cribbs DH, Cotman CW. Brain gene expression patterns differentiate mild cognitive impairment from normal aged and Alzheimer's disease. Neurobiol Aging. 2014; 35(9):1961–72. https://doi.org/10.1016/j.neurobiolaging.2014.03.031 PMID: 24786631.

24. Antonell A, Llado A, Altirriba J, Botta-Orfila T, Balasa M, Fernandez M, et al. A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. Neurobiol Aging. 2013; 34(7):1772–8. Epub 2013/02/02. https://doi.org/10.1016/j.neurobiolaging.2012.12.026 PMID: 23369545.

25. Guo Q, Fu W, Sopher BL, Miller MW, Ware CB, Martin GM, et al. Increased vulnerability of hippocampal neurons to excitotoxic necrosis in presenilin-1 mutant knock-in mice. Nat Med. 1999; 5(1):101–6. https://doi.org/10.1038/4789 PMID: 9883847.

26. Kawasumi M, Chiba T, Yamada M, Miyamae-Kaneko M, Matsuoka M, Nakahara J, et al. Targeted introduction of V642I mutation in amyloid precursor protein gene causes functional abnormality resembling early stage of Alzheimer's disease in aged mice. Eur J Neurosci. 2004; 19(10):2826–38. https://doi.org/10.1111/j.0953-816X.2004.03397.x PMID: 15147316.

27. Siman R, Reaume AG, Savage MJ, Trusko S, Lin YG, Scott RW, et al. Presenilin-1 P264L knock-in mutation: differential effects on abeta production, amyloid deposition, and neuronal vulnerability. J Neurosci. 2000; 20(23):8717–26. https://doi.org/10.1523/JNEUROSCI.20-23-08717.2000 PMID: 11102478.

28. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? Behav Brain Res. 2017; 322(Pt B):311–28. Epub 2016/05/04. https://doi.org/10.1016/j.bbr.2016.05.007 PMID: 27155503.

29. Sato N, Hori O, Yamaguchi A, Lambert JC, Chartier-Harlin MC, Robinson PA, et al. A novel presenilin-2 splice variant in human Alzheimer's disease brain tissue. Journal of Neurochemistry. 1999; 72 (6):2498–505. https://doi.org/10.1046/j.1471-4159.1999.0722498.x PMID: 10349860

30. Jayadev S, Leverenz JB, Steinbart E, Stahl J, Klunk W, Yu CE, et al. Alzheimer's disease phenotypes and genotypes associated with mutations in presenilin 2. Brain. 2010; 133:1143–54. https://doi.org/10.1093/brain/awq033 PMID: 20375137

31. Newman M, Wilson L, Verdile G, Lim A, Khan I, Moussavi Nik SH, et al. Differential, dominant activation and inhibition of Notch signalling and APP cleavage by truncations of PSEN1 in human disease. Hum Mol Genet. 2014; 23(3):602–17. Epub 2013/10/09. https://doi.org/10.1093/hmg/ddt448 PMID: 24101600.

76

32. Moussavi Nik SH, Newman M, Wilson L, Ebrahimie E, Wells S, Musgrave I, et al. Alzheimer's disease-related peptide PS2V plays ancient, conserved roles in suppression of the unfolded protein response under hypoxia and stimulation of gamma-secretase activity. Hum Mol Genet. 2015. Epub 2015/03/31. https://doi.org/10.1093/hmg/ddv110 PMID: 25814654.

33. Sharman MJ, Moussavi Nik SH, Chen MM, Ong D, Wijaya L, Laws SM, et al. The Guinea Pig as a Model for Sporadic Alzheimer's Disease (AD): The Impact of Cholesterol Intake on Expression of AD-Related Genes. PLoS One. 2013; 8(6):e66235. https://doi.org/10.1371/journal.pone.0066235 PMID: 23805206.

34. Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, et al. Lifetime risk of dementia and Alzheimer's disease. The impact of mortality on risk estimates in the Framingham Study. Neurology. 1997; 49(6):1498–504. https://doi.org/10.1212/wnl.49.6.1498 PMID: 9409336.

35. Gamberger D, Ženko B, Mitelpunkt A, Shachar N, Lavrač N. Clusters of male and female Alzheimer's disease patients in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Brain Inform. 2016; 3(3):169–79. Epub 2016/03/30. https://doi.org/10.1007/s40708-016-0035-5 PMID: 27525218.

36. Oosterhof N, Holtman IR, Kuil LE, van der Linde HC, Boddeke EW, Eggen BJ, et al. Identification of a conserved and acute neurodegeneration-specific microglial transcriptome in the zebrafish. Glia. 2017; 65(1):138–49. Epub 2016/10/19. https://doi.org/10.1002/glia.23083 PMID: 27757989.

37. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007; 445(7124):168–76. Epub 2006/12/06. https://doi.org/10.1038/nature05453 PMID: 17151600.

38. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021.

39. de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009; 5(12):1512–26. Epub 2009/10/01. https://doi.org/10.1039/b908315d PMID: 20023718.

40. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D. Integrating transcriptome and proteome profiling: Strategies and applications. Proteomics. 2016; 16(19):2533–44. Epub 2016/08/25. https://doi.org/10.1002/pmic.201600140 PMID: 27343053.

41. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016; 3:160089. Epub 2016/10/11. https://doi.org/10.1038/sdata.2016.89 PMID: 27727239.

42. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015; 31(13):2123–30. https://doi.org/10.1093/bioinformatics/btv118 PMID: 25717192.

43. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology2005.

44. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012; 489(7416):391–9. https://doi.org/10.1038/nature11405 PMID: 22996553.

45. Winden KD, Oldham MC, Mirnics K, Ebert PJ, Swan CH, Levitt P, et al. The organization of the transcriptional network in specific neuronal classes. Mol Syst Biol. 2009; 5:291. https://doi.org/10.1038/msb.2009.46 PMID: 19638972.

46. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences. 2006; 103 (47):17973–8. https://doi.org/10.1073/pnas.0605938103 PMID: 17101986

47. Miller JA, Horvath S, Geschwind DH. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proceedings of the National Academy of Sciences. 2010; 107 (28):12698–703.

48. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2013; 153 (3):707–20. https://doi.org/10.1016/j.cell.2013.03.030 PMID: 23622250.

49. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011; 7(1):e1001057. https://doi.org/10.1371/journal.pcbi.1001057 PMID: 21283776.

50. Solchenberger B, Russell C, Kremmer E, Haass C, Schmid B. Granulin knock out zebrafish lack frontotemporal lobar degeneration and neuronal ceroid lipofuscinosis pathology. PLoS One. 2015; 10(3): e0118956. https://doi.org/10.1371/journal.pone.0118956 PMID: 25785851.

77

51. Moore DB, Gillentine MA, Botezatu NM, Wilson KA, Benson AE, Langeland JA. Asynchronous evolutionary origins of Abeta and BACE1. Mol Biol Evol. 2014; 31(3):696–702. https://doi.org/10.1093/molbev/mst262 PMID: 24361992.

52. Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. Science. 1992; 256 (5054):184–5. https://doi.org/10.1126/science.1566067 PMID: 1566067.

53. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A. 2004; 101(7):2173–8. https://doi.org/10.1073/pnas.0308512100 PMID: 14769913.

54. Sperling RA, Dickerson BC, Pihlajamaki M, Vannini P, LaViolette PS, Vitolo OV, et al. Functional alterations in memory networks in early Alzheimer's disease. Neuromolecular Med. 2010; 12(1):27–43. https://doi.org/10.1007/s12017-009-8109-7 PMID: 20069392.

55. Head E, Lott IT, Patterson D, Doran E, Haier RJ. Possible compensatory events in adult Down syndrome brain prior to the development of Alzheimer disease neuropathology: targets for nonpharmacological intervention. J Alzheimers Dis. 2007; 11(1):61–76. https://doi.org/10.3233/jad-2007-11110 PMID: 17361036.

56. Du X, Pang TY. Is Dysregulation of the HPA-Axis a Core Pathophysiology Mediating Co-Morbid Depression in Neurodegenerative Diseases? Front Psychiatry. 2015; 6:32. https://doi.org/10.3389/fpsyt.2015.00032 PMID: 25806005.

57. Silverman MN, Sternberg EM. Glucocorticoid regulation of inflammation and its functional correlates: from HPA axis to glucocorticoid receptor dysfunction. Ann N Y Acad Sci. 2012; 1261:55–63. https://doi.org/10.1111/j.1749-6632.2012.06633.x PMID: 22823394.

58. Arlt S, Demiralay C, Tharun B, Geisel O, Storm N, Eichenlaub M, et al. Genetic risk factors for depression in Alzheimer's disease patients. Curr Alzheimer Res. 2013; 10(1):72–81. PMID: 23157339.

59. Lin JX, Leonard WJ. The role of Stat5a and Stat5b in signaling by IL-2 family cytokines. Oncogene. 2000; 19(21):2566–76. https://doi.org/10.1038/sj.onc.1203523 PMID: 10851055.

60. Moon RT, Bowerman B, Boutros M, Perrimon N. The promise and perils of Wnt signaling through beta-catenin. Science. 2002; 296(5573):1644–6. https://doi.org/10.1126/science.1071549 PMID: 12040179

61. Rapoport SI, Hatanpaa K, Brady DR, Chandrasekaran K. Brain energy metabolism, cognitive function and down-regulated oxidative phosphorylation in Alzheimer disease. Neurodegeneration. 1996; 5 (4):473–6. https://doi.org/10.1006/neur.1996.0065 PMID: 9117565.

62. Barron AM, Fuller SJ, Verdile G, Martins RN. Reproductive hormones modulate oxidative stress in Alzheimer's disease. Antioxid Redox Signal. 2006; 8(11–12):2047–59. https://doi.org/10.1089/ars.2006.8.2047 PMID: 17034349.

63. Schapira AH. Oxidative stress and mitochondrial dysfunction in neurodegeneration. Curr Opin Neurol. 1996; 9(4):260–4. https://doi.org/10.1097/00019052-199608000-00003 PMID: 8858182.

64. Perry G, Nunomura A, Hirai K, Takeda A, Aliev G, Smith MA. Oxidative damage in Alzheimer's disease: the metabolic dimension. Int J Dev Neurosci. 2000; 18(4–5):417–21. Epub 2000/05/19. https://doi.org/10.1016/s0736-5748(00)00006-x PMID: 10817925

65. Nunomura A, Perry G, Aliev G, Hirai K, Takeda A, Balraj EK, et al. Oxidative damage is the earliest event in Alzheimer disease. J Neuropathol Exp Neurol. 2001; 60(8):759–67. Epub 2001/08/07. https://doi.org/10.1093/jnen/60.8.759 PMID: 11487050.

66. Perry G, Nunomura A, Hirai K, Zhu X, Perez M, Avila J, et al. Is oxidative damage the fundamental pathogenic mechanism of Alzheimer's and other neurodegenerative diseases? Free Radic Biol Med. 2002; 33(11):1475–9. Epub 2002/11/26. https://doi.org/10.1016/s0891-5849(02)01113-9 PMID: 12446204

67. Scherz-Shouval R, Elazar Z. ROS, mitochondria and the regulation of autophagy. Trends Cell Biol. 2007; 17(9):422–7. Epub 2007/09/07. https://doi.org/10.1016/j.tcb.2007.07.009 PMID: 17804237.

68. Daulatzai MA. Death by a thousand cuts in Alzheimer's disease: hypoxia-the prodrome. Neurotox Res. 2013; 24(2):216–43. https://doi.org/10.1007/s12640-013-9379-2 PMID: 23400634.

69. Daulatzai MA. Cerebral hypoperfusion and glucose hypometabolism: Key pathophysiological modulators promote neurodegeneration, cognitive impairment, and Alzheimer's disease. J Neurosci Res. 2017; 95(4):943–72. https://doi.org/10.1002/jnr.23777 PMID: 27350397.

70. Zhao WQ, Ravindranath L, Mohamed AS, Zohar O, Chen GH, Lyketsos CG, et al. MAP kinase signaling cascade dysfunction specific to Alzheimer's disease in fibroblasts. Neurobiol Dis. 2002; 11(1):166–83. https://doi.org/10.1006/nbdi.2002.0520 PMID: 12460556.

78

71. Drewes G, Lichtenberg-Kraag B, Döring F, Mandelkow EM, Biernat J, Goris J, et al. Mitogen activated protein (MAP) kinase transforms tau protein into an Alzheimer-like state. EMBO J. 1992; 11(6):2131–8. PMID: 1376245.

72. Maroulakou IG, Bowe DB. Expression and function of Ets transcription factors in mammalian development: a regulatory network. Oncogene. 2000; 19(55):6432–42. https://doi.org/10.1038/sj.onc.1204039 PMID: 11175359.

73. Kar A, Gutierrez-Hartmann A. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. Crit Rev Biochem Mol Biol. 2013; 48(6):522–43. https://doi.org/10.3109/10409238.2013.838202 PMID: 24066765.

74. Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai LH, et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature. 2015; 518(7539):365–9. https://doi.org/10.1038/nature14252 PMID: 25693568.

75. Morgan SC, Taylor DL, Pocock JM. Microglia release activators of neuronal proliferation mediated by activation of mitogen-activated protein kinase, phosphatidylinositol-3-kinase/Akt and delta-Notch signalling cascades. J Neurochem. 2004; 90(1):89–101. https://doi.org/10.1111/j.1471-4159.2004.02461.x PMID: 15198670.

76. Herculano-Houzel S. Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution. PLoS One. 2011; 6(3):e17514. https://doi.org/10.1371/journal.pone.0017514 PMID: 21390261.

77. Yamada T, Sasaki H, Furuya H, Miyata T, Goto I, Sakaki Y. Complementary DNA for the mouse homolog of the human amyloid beta protein precursor. Biochem Biophys Res Commun. 1987; 149(2):665–71. https://doi.org/10.1016/0006-291x(87)90419-0 PMID: 3322280.

78. Musa A, Lehrach H, Russo VA. Distinct expression patterns of two zebrafish homologues of the human APP gene during embryonic development. Dev Genes Evol. 2001; 211(11):563–7. Epub 2002/02/28. https://doi.org/10.1007/s00427-001-0189-9 PMID: 11862463.

79. Monsell SE, Kukull WA, Roher AE, Maarouf CL, Serrano G, Beach TG, et al. Characterizing Apolipoprotein E epsilon4 Carriers and Noncarriers With the Clinical Diagnosis of Mild to Moderate Alzheimer Dementia and Minimal beta-Amyloid Peptide Plaques. JAMA Neurol. 2015; 72(10):1124–31. https://doi.org/10.1001/jamaneurol.2015.1721 PMID: 26302353.

80. Jack CR Jr., Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 2011; 7(3):257–62. PMID: 21514247.

81. Whitehouse PJ, George DR. A tale of two reports: what recent publications from the Alzheimer's Association and Institute of Medicine say about the state of the field. J Alzheimers Dis. 2016; 49(1):21–5. https://doi.org/10.3233/JAD-150663 PMID: 26444797.

82. Morris GP, Clark IA, Vissel B. Questions concerning the role of amyloid-β in the definition, aetiology and diagnosis of Alzheimer's disease. Acta Neuropathol. 2018; 136(5):663–89. Epub 2018/10/22. https://doi.org/10.1007/s00401-018-1918-8 PMID: 30349969.

83. Jansen WJ, Ossenkoppele R, Knol DL, Tijms BM, Scheltens P, Verhey FR, et al. Prevalence of cerebral amyloid pathology in persons without dementia: a meta-analysis. JAMA. 2015; 313(19):1924–38. https://doi.org/10.1001/jama.2015.4668 PMID: 25988462.

84. Braggin JE, Bucks SA, Course MM, Smith CL, Sopher B, Osnis L, et al. Alternative splicing in a presenilin 2 variant associated with Alzheimer disease. Ann Clin Transl Neurol. 2019; 6(4):762–77. Epub 2019/03/10. https://doi.org/10.1002/acn3.755 PMID: 31020001.

85. Nornes S, Newman M, Verdile G, Wells S, Stoick-Cooper CL, Tucker B, et al. Interference with splicing of Presenilin transcripts has potent dominant negative effects on Presenilin activity. Hum Mol Genet. 2008; 17(3):402–12. Epub 2007/11/06. https://doi.org/10.1093/hmg/ddm317 PMID: 17981814.

86. Newman M, Hin N, Pederson S, Lardelli M. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production. Mol Brain. 2019; 12(1):43. Epub 2019/05/03. https://doi.org/10.1186/s13041-019-0467-y PMID: 31053140.

87. Huggett JF, Whale A. Digital PCR as a novel technology and its potential implications for molecular diagnostics. Clin Chem. 2013; 59(12):1691–3. Epub 2013/10/07. https://doi.org/10.1373/clinchem.2013.214742 PMID: 24100808.

88. de St Groth Fazekas. The evaluation of limiting dilution assays. J Immunol Methods. 1982; 49(2):R11–23. https://doi.org/10.1016/0022-1759(82)90269-1 PMID: 7040548.

89. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009; 6(5):359–62. https://doi.org/10.1038/nmeth.1322 PMID: 19377485.

79

90. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014; 13(9):2513–26. https://doi.org/10.1074/mcp.M113.031591 PMID: 24942700.

91. Andrews S. FastQC. 0.11.5 ed2010.

92. Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 2012; 5:337. https://doi.org/10.1186/1756-0500-5-337 PMID: 22748135.

93. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12(4):357–60. https://doi.org/10.1038/nmeth.3317 PMID: 25751142.

94. Broad Institute. Picard. 2.14.0 ed2017. p. A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

95. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677.

96. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2017.

97. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308.

98. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res. 2015; 43(15):e97. https://doi.org/10.1093/nar/gkv412 PMID: 25925576.

99. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. Ann Appl Stat. 2016; 10(2):946–63. https://doi.org/10.1214/16-AOAS920 PMID: 28367255.

100. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7):e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792.

101. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; 32(9):896–902. https://doi.org/10.1038/nbt.2931 PMID: 25150836.

102. Kolde R. pheatmap: Pretty Heatmaps. 1.0.8 ed2015.

103. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005; 21 (16):3439–40. https://doi.org/10.1093/bioinformatics/bti525 PMID: 16082012.

104. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009; 4(8):1184–91. https://doi.org/10.1038/nprot.2009.97 PMID: 19617889.

105. Giner G, Smyth GK. FRY: a fast approximation to ROAST gene set test with mean aggregated set statistics [version 1; not peer reviewed]. F1000Research. 2016; 5(2605). https://doi.org/10.7490/f1000research.1113351.1

106. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010; 26(17):2176–82. https://doi.org/10.1093/bioinformatics/btq401 PMID: 20610611.

107. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38(4):576–89. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432.

108. Benner C. HOMER (Hypergeometric Optimization of Motif EnRichment). v4.9 ed2017. p. Software for motif discovery and next generation sequencing analysis.

109. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014; 30(17):2524–6. https://doi.org/10.1093/bioinformatics/btu305 PMID: 24794931.

110. Kammers K, Cole RN, Tiengwe C, Ruczinski I. Detecting Significant Changes in Protein Abundance. EuPA Open Proteom. 2015; 7:11–9. https://doi.org/10.1016/j.euprot.2015.02.002 PMID: 25821719.

111. Magis AT, Funk CC, Price ND. SNAPR: a bioinformatics pipeline for efficient and accurate RNA-seq alignment and analysis. IEEE Life Sci Lett. 2015; 1(2):22–5. Epub 2015/08/28. https://doi.org/10.1109/LLS.2015.2465870 PMID: 29270443.

80

112. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008; 24(5):719–20. https://doi.org/10.1093/bioinformatics/btm563 PMID: 18024473.

113. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Softw. 2012; 46(11). PMID: 23050260.

114. Allaire JJ, Gandrud C, Russell K, Yetman CJ. networkD3: D3 JavaScript Network Graphs from R. 2017.

115. Martin S, Brown WM, Klavans R, Boyack KW, editors. OpenOrd: an open-source toolbox for large graph layout. IS&T/SPIE Electronic Imaging; 2011: SPIE.

116. Kroehne V, Freudenreich D, Hans S, Kaslin J, Brand M. Regeneration of the adult zebrafish brain from neurogenic radial glia-type progenitors. Development. 2011:dev.072587. https://doi.org/10.1242/dev.072587 PMID: 22007133

117. Kaslin J, Kroehne V, Ganz J, Hans S, Brand M. Distinct roles of neuroepithelial-like and radial glia-like progenitor cells in cerebellar regeneration. Development. 2017; 144(8):1462–71. Epub 2017/03/16. https://doi.org/10.1242/dev.144907 PMID: 28289134.

81

Chapter 3

Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish *presenilin 1* gene implies effects on energy production

Statement of Authorship

Title of Paper	Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Newman, M., Hin, N., Pederson, S. and Lardelli, M., 2019. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production. *Molecular Brain*, 12(1), p.43.

Principal Author

Name of Principal Author (Candidate)	Nhi Hin		
Contribution to the Paper	- All bioinformatics analysis - Editing of manuscript drafts		
Overall percentage (%)	45%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	08/09/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

    i.   the candidate's stated contribution to the publication is accurate (as detailed above);

    ii.   permission is granted for the candidate in include the publication in the thesis; and

    iii.   the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Morgan Newman		
Contribution to the Paper	- Contributed equally to this work with Nhi Hin (Overall Percentage: 45%) - Genome editing and creation of *psen1*^{Q96K97} zebrafish mutant line - Removal of brains from fish and purification of mRNA for RNA-seq - Editing of manuscript drafts		
Signature		Date	10/11/2020

Name of Co-Author	Stephen Pederson		
Contribution to the Paper	- Supervision of bioinformatics work and statistical methodology - Editing of manuscript drafts		
Signature		Date	

Name of Co-Author	Michael Lardelli		
Contribution to the Paper	- Conceptualisation - Project supervision - Drafting of manuscript - Corresponding author		
Signature		Date	9/9/20

MICRO REPORT

Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish *presenilin 1* gene implies effects on energy production

Morgan Newman[†], Nhi Hin[†], Stephen Pederson and Michael Lardelli[*]

Abstract

To prevent or ameliorate Alzheimer's disease (AD) we must understand its molecular basis. AD develops over decades but detailed molecular analysis of AD brains is limited to postmortem tissue where the stresses initiating the disease may be obscured by compensatory responses and neurodegenerative processes. Rare, dominant mutations in a small number of genes, but particularly the gene PRESENILIN 1 (*PSEN1*), drive early onset of familial AD (EOfAD). Numerous transgenic models of AD have been constructed in mouse and other organisms, but transcriptomic analysis of these models has raised serious doubts regarding their representation of the disease state. Since we lack clarity regarding the molecular mechanism(s) underlying AD, we posit that the most valid approach is to model the human EOfAD genetic state as closely as possible. Therefore, we sought to analyse brains from zebrafish heterozygous for a single, EOfAD-like mutation in their *PSEN1*-orthologous gene, *psen1*. We previously introduced an EOfAD-like mutation (Q96_K97del) into the endogenous *psen1* gene of zebrafish. Here, we analysed transcriptomes of young adult (6-month-old) entire brains from a family of heterozygous mutant and wild type sibling fish. Gene ontology (GO) analysis implies effects on mitochondria, particularly ATP synthesis, and on ATP-dependent processes including vacuolar acidification.

Keywords: Alzheimer's disease, Presenilin 1, Mutation, Transcriptome, Brain, ATP synthesis, Mitochondria, Vacuolar acidification, Zebrafish, Genome editing

Background

AD is the most common form of dementia with severe personal, social, and economic impacts. Rare, familial forms of AD exist caused by autosomal dominant mutations in single genes (reviewed by [1]). The majority of these mutations occur in the gene PRESENILIN 1 (*PSEN1*) that encodes a multipass integral membrane protein involved in intra-membrane cleavage of numerous proteins [1].

A wide variety of transgenic models of AD have been created and studied. These are aimed at reproducing histopathologies posited to be central to the disease process, i.e. amyloid plaques and neurofibrillary tangles

of the protein MAPT [2]. However, analysis of the effects on the brain transcriptome of the transgenes driving a number of these mouse models showed little concordance with transcriptomic differences between human AD brains and age-matched controls [3] (although a recent study asserts that this lack of concordance for the popular "5XFAD" transgenic mouse model is due to previous failure to analyse the effects of its transgenes in a variety of genetic backgrounds [4]). We posit that, in the absence of an understanding of the molecular mechanism(s) underlying AD, the most objective approach to modeling this disease (or, at least, modeling its genetic form, EOfAD) is to create a genetic state as similar as possible to the EOfAD state in humans. Mouse "knock-in" models of EOfAD mutations were created over a decade ago and showed subtle phenotypic effects but not the desired histopathologies (e.g. [5, 6]).

* Correspondence: Michael.lardelli@adelaide.edu.au
[†]Morgan Newman and Nhi Hin are equal first authors.
Department of Molecular and Biomedical Science, University of Adelaide, School of Biological Sciences, North Terrace, Adelaide, SA 5005, Australia

Table 1 GOs enriched for genes differentially expressed between heterozygous mutant and wild type sibling fish brains

Gene Ontology Term	Ontology	Total Genes	DE Genes	p-value	FDR *p*-value
ATP biosynthetic process	BP	29	7	3.48987E-08	0.00041
ribonucleoside triphosphate biosynthetic process	BP	49	8	9.41317E-08	0.00045
nucleoside triphosphate biosynthetic process	BP	54	8	2.06555E-07	0.00060
purine nucleoside triphosphate biosynthetic process	BP	41	7	4.46237E-07	0.00060
purine ribonucleoside triphosphate biosynthetic process	BP	41	7	4.46237E-07	0.00060
hydrogen transport	BP	60	8	4.783E-07	0.00060
proton transport	BP	60	8	4.783E-07	0.00060
energy coupled proton transport, down electrochemical gradient	BP	27	6	5.89038E-07	0.00060
ATP synthesis coupled proton transport	BP	27	6	5.89038E-07	0.00060
transport	BP	2072	48	2.11748E-06	0.00165
purine nucleoside monophosphate biosynthetic process	BP	54	7	3.09019E-06	0.00172
purine ribonucleoside monophosphate biosynthetic process	BP	54	7	3.09019E-06	0.00172
hydrogen ion transmembrane transport	BP	54	7	3.09019E-06	0.00172
ribonucleoside triphosphate metabolic process	BP	133	10	3.8448E-06	0.00178
establishment of localization	BP	2123	48	4.20295E-06	0.00182
ATP metabolic process	BP	109	9	5.50772E-06	0.00230
nucleoside triphosphate metabolic process	BP	140	10	6.08925E-06	0.00245
cation transport	BP	452	18	6.61154E-06	0.00258
monovalent inorganic cation transport	BP	219	12	1.10729E-05	0.00392
ribonucleoside monophosphate biosynthetic process	BP	65	7	1.08944E-05	0.00392
nucleoside monophosphate biosynthetic process	BP	68	7	1.47269E-05	0.00492
purine ribonucleoside triphosphate metabolic process	BP	125	9	1.68142E-05	0.00546
purine nucleoside triphosphate metabolic process	BP	126	9	1.79263E-05	0.00552
transmembrane transport	BP	654	21	2.93288E-05	0.00797
purine nucleoside monophosphate metabolic process	BP	136	9	3.2951E-05	0.00837
purine ribonucleoside monophosphate metabolic process	BP	136	9	3.2951E-05	0.00837
energy coupled proton transmembrane transport, against electrochemical gradient	BP	35	5	5.20342E-05	0.01106
ATP hydrolysis coupled proton transport	BP	35	5	5.20342E-05	0.01106
ATP hydrolysis coupled transmembrane transport	BP	35	5	5.20342E-05	0.01106
ATP hydrolysis coupled ion transmembrane transport	BP	35	5	5.20342E-05	0.01106
ATP hydrolysis coupled cation transmembrane transport	BP	35	5	5.20342E-05	0.01106
ion transport	BP	737	22	5.61478E-05	0.01152
localization	BP	2621	52	6.0913E-05	0.01207
ribonucleoside monophosphate metabolic process	BP	147	9	6.06496E-05	0.01207
nucleoside monophosphate metabolic process	BP	150	9	7.09445E-05	0.01360
single-organism localization	BP	819	23	9.51294E-05	0.01738
single-organism transport	BP	776	22	0.000119082	0.02109
ribonucleotide biosynthetic process	BP	129	8	0.000143028	0.02423
ribose phosphate biosynthetic process	BP	129	8	0.000143028	0.02423
vacuolar acidification	BP	11	3	0.000246582	0.04101
ribonucleotide metabolic process	BP	220	10	0.000281352	0.04506
proton-transporting two-sector ATPase complex, proton-transporting domain	CC	25	6	3.59375E-07	0.00060
proton-transporting two-sector ATPase complex	CC	45	7	8.65692E-07	0.00078
mitochondrial membrane	CC	285	15	1.42199E-06	0.00119

Table 1 GOs enriched for genes differentially expressed between heterozygous mutant and wild type sibling fish brains *(Continued)*

Gene Ontology Term	Ontology	Total Genes	DE Genes	p-value	FDR p-value
mitochondrial envelope	CC	303	15	3.0322E-06	0.00172
membrane part	CC	4868	85	1.1722E-05	0.00403
organelle membrane	CC	789	24	1.84982E-05	0.00555
mitochondrial inner membrane	CC	195	11	1.97958E-05	0.00579
integral component of membrane	CC	4419	78	2.52479E-05	0.00720
intrinsic component of membrane	CC	4453	78	3.37749E-05	0.00840
organelle envelope	CC	420	16	3.76291E-05	0.00917
envelope	CC	422	16	3.98337E-05	0.00950
organelle inner membrane	CC	215	11	4.86028E-05	0.01106
Cul2-RING ubiquitin ligase complex	CC	7	3	5.4156E-05	0.01131
proton-transporting ATP synthase complex	CC	19	4	6.25883E-05	0.01220
mitochondrial membrane part	CC	117	8	7.21148E-05	0.01360
mitochondrial part	CC	404	15	8.83156E-05	0.01639
membrane	CC	5379	88	0.000106964	0.01924
vacuolar proton-transporting V-type ATPase, V0 domain	CC	9	3	0.000127733	0.02229
mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	CC	12	3	0.000325933	0.04885
proton-transporting V-type ATPase, V0 domain	CC	12	3	0.000325933	0.04885
ATPase activity, coupled to transmembrane movement of ions, rotational mechanism	MF	34	7	1.1446E-07	0.00045
hydrogen ion transmembrane transporter activity	MF	84	9	6.11883E-07	0.00060
ATPase activity, coupled to transmembrane movement of substances	MF	98	9	2.27123E-06	0.00166
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	MF	101	9	2.92425E-06	0.00172
primary active transmembrane transporter activity	MF	104	9	3.73269E-06	0.00178
P-P-bond-hydrolysis-driven transmembrane transporter activity	MF	104	9	3.73269E-06	0.00178
cation-transporting ATPase activity	MF	56	7	3.96731E-06	0.00178
ATPase coupled ion transmembrane transporter activity	MF	56	7	3.96731E-06	0.00178
ATPase activity, coupled to movement of substances	MF	112	9	6.88692E-06	0.00260
active ion transmembrane transporter activity	MF	96	8	1.72916E-05	0.00546
active transmembrane transporter activity	MF	281	13	2.87859E-05	0.00797
proton-transporting ATP synthase activity, rotational mechanism	MF	16	4	3.02121E-05	0.00803
transporter activity	MF	991	25	0.000249051	0.04101
substrate-specific transmembrane transporter activity	MF	709	20	0.000263528	0.04279
ion transmembrane transporter activity	MF	660	19	0.000293184	0.04572
substrate-specific transporter activity	MF	828	22	0.000297009	0.04572
monovalent inorganic cation transmembrane transporter activity	MF	264	11	0.000297217	0.04572

GOs are grouped by ontology (BP, CC or MF) and ranked by FDR-corrected *p*-value

However, at that time, researchers did not have access to RNA-Seq technology. To the best of our knowledge, transcriptome analysis of the EOfAD mutation knock-in mouse models was never performed.

In humans, AD is thought to develop over decades and the median survival to onset age for EOfAD mutations in human *PSEN1* considered collectively is 45 years [7]. Functional MRI of human children carrying EOfAD mutations in *PSEN1* has revealed differences in brain activity compared to non-carriers in individuals as young as 9 years of age [8]. Presumably therefore, heterozygosity for EOfAD mutations in *PSEN1* causes early molecular changes/stresses that eventually lead to AD.

Transcriptome analysis is currently the most detailed molecular phenotypic analysis possible on cells or tissues. Here we present an initial analysis of the transcriptomic differences caused in young adult (6-month-old) zebrafish brains by the presence of an EOfAD-like mutation in the

gene *psen1* that is orthologous to the human *PSEN1* gene. GO analysis supports very significant effects on mitochondrial function, especially synthesis of ATP, and on ATP-dependent functions such as the acidification of lysosomes that are critical for autophagy.

Materials and methods

The mutant allele, Q96_K97del, of *psen1* was a byproduct identified during our introduction of the K97fs mutation into *psen1* (that models the K115fs mutation of human *PSEN2* – see [9] for an explanation).

Q96_K97del is a deletion of 6 nucleotides from the coding sequence of the *psen1* gene. This is predicted to distort the first lumenal loop of the Psen1 protein. In this sense, it is similar to a number of EOfAD mutations of human *PSEN1* [10]. Also, in common with all the widely distributed EOfAD mutations in *PSEN1*, (and consistent with the PRESENILIN EOfAD mutation "reading frame preservation rule" [1]), the Q96_K97del allele is predicted to encode a transcript that includes the C-terminal sequences of the wild type protein. Therefore, as a model of an EOfAD mutation, it is superior to the K97fs mutation in *psen1* [9].

To generate a family of heterozygous Q96_K97del allele (i.e. *psen1$^{Q96\_K97del}$*/+) and wild type (+/+) sibling fish, we mated a *psen1$^{Q96\_K97del}$*/+ individual with a +/+ individual and raised the progeny from a single spawning event together in one tank. Zebrafish can live for up to 5 years but, in our laboratory, typically show greatly reduced fertility after 18 months. The fish become fertile after around 3 months of age, so we regard 6-month-old fish as equivalent to young adult humans. Therefore we analysed the transcriptomes of entire young adult, 6-month-old fish brains using poly-A enriched RNA-seq technology, and estimated gene expression from the resulting single-end 75 bp reads using the reference GRCz11 zebrafish assembly transcriptome [11, 12]. Each zebrafish brain has a mass of approximately 7 mg. Since AD is more prevalent in human females than males, and to further reduce gene expression "noise" in our analyses, we obtained brain transcriptome data from four female wild type fish and four female heterozygous mutant fish. This data has been made publicly available at the Gene Expression Omnibus (GEO, see under *Availability of data and materials* below).

Results

Differentially expressed genes (DE genes)

Genes differentially expressed between wild type and heterozygous mutant sibling fish were identified using moderated *t*-tests and a false discovery rate (FDR)-adjusted *p*-value cutoff of 0.05 as previously described [9, 13, 14]. In total, 251 genes were identified as differentially expressed (see Additional file 1). Of these, 105 genes showed increased expression in heterozygous mutant brains relative to wild type sibling brains while 146 genes showed decreased expression.

GO analysis

To understand the significance for brain cellular function of the differential gene expression identified in young adult heterozygous mutant brains we used the *goana* function [15] of the *limma* package of Bioconductor software [14] to identify GOs in which the DE genes were enriched at an FDR-corrected *p*-value of less than 0.05. Seventy-eight GOs were identified (Table 1) of which 20 addressed cellular components (CC). Remarkably, most of these CCs concerned the mitochondrion, membranes, or ATPases. Seventeen GOs addressed molecular functions (MF) and largely involved membrane transporter activity, particularly ion transport and ATPase activity coupled to such transport. Forty-one GOs addressed biological processes (BP) and involved ATP metabolism, ribonucleoside metabolism, and transmembrane transport processes including vacuolar acidification (that has previously been identified as affected by EOfAD mutations in *PSEN1* [16]). Overall, our GO analysis indicates that this EOfAD-like mutation of zebrafish *psen1* has very significant impacts on cellular energy metabolism and transmembrane transport processes.

Additional file

Additional file 1: Genes differentially expressed between heterozygous mutant and wild type brains at 6 months. Lists the genes identified as differentially expressed between the brains of heterozygous *psen1$^{Q96\_K97del}$* mutant fish and the brains of their wild type siblings at an age of 6 months. Genes are ranked according to FDR-corrected *p*-value. Only genes with a FDR-corrected p-value less than 0.05 are shown. "FC" denotes fold change. "DE" denotes differential expression. For DE_Direction, "1" denotes increased expression in the mutant and "-1" denotes decreased expression in the mutant. (XLSX 39 kb)

Abbreviations

AD: Alzheimer's disease; ATP: Adenosine triphosphate; BP: Biological process (GO term); CC: Cellular component (GO term); DE genes: Differentially expressed genes; EOfAD: Early onset familial Alzheimer's disease; FDR: False discovery rate; GEO: Gene Expression Omnibus; GO: Gene ontology; MAPT: MICROTUBULE-ASSOCIATED PROTEIN TAU (human protein); MF: Molecular function (GO term); mg: Milligrams; MRI: Magnetic resonance imaging; *PSEN1*: PRESENILIN 1 (human gene); PSEN1: PRESENILIN 1 (human protein); *psen1*: presenilin 1 (zebrafish gene); Psen1: Presenilin 1 (zebrafish protein)

Newman *et al. Molecular Brain*   (2019) 12:43

Page 5 of 5

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE126096.
$psen1^{Q96\_K97del}$ mutant zebrafish are available upon request. However, due to Australia's strict quarantine and export regulations, export of fish involves considerable effort and expense and these costs must be borne by the party requesting the fish.

Authors' contributions
MN conceived the project, sought funding, generated the $psen1^{Q96\_K97del}$ mutant zebrafish, identified the genotype of individuals, and isolated mRNA from zebrafish brains. NH processed the RNA-seq data and performed bioinformatics analysis to identify DE genes and GOs. SP supervised the work of NH and performed data quality checks. ML conceived the project, sought funding, coordinated the project, and drafted this research report. All authors contributed to interpretation of data and to reviewing and editing drafts of the submitted manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
This study was conducted under the auspices of the Animal Ethics Committee of the University of Adelaide, under permits S-2014-108 and S-2017-073.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Jayne T, Newman M, Verdile G, Sutherland G, Munch G, Musgrave I, et al. Evidence for and against a pathogenic role of reduced gamma-Secretase activity in familial Alzheimer's disease. J Alzheimers Dis. 2016;52(3):781–99.
2. Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018;14(4):535–62.
3. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: what are our rodent models telling us? Behav Brain Res. 2017;322(Pt B):311–28.
4. Neuner SM, Heuer SE, Huentelman MJ, O'Connell KMS, Kaczorowski CC. Harnessing genetic complexity to enhance translatability of Alzheimer's disease mouse models: a path toward precision medicine. Neuron. 2019; 101(3):399–411 e5.
5. Guo Q, Fu W, Sopher BL, Miller MW, Ware CB, Martin GM, et al. Increased vulnerability of hippocampal neurons to excitotoxic necrosis in presenilin-1 mutant knock-in mice. Nat Med. 1999;5(1):101–6.
6. Siman R, Reaume AG, Savage MJ, Trusko S, Lin YG, Scott RW, et al. Presenilin-1 P264L knock-in mutation: differential effects on abeta production, amyloid deposition, and neuronal vulnerability. J Neurosci. 2000; 20(23):8717–26.
7. Ryman DC, Acosta-Baena N, Aisen PS, Bird T, Danek A, Fox NC, et al. Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. Neurology. 2014;83(3):253–60.
8. Quiroz YT, Schultz AP, Chen K, Protas HD, Brickhouse M, Fleisher AS, et al. Brain imaging and blood biomarker abnormalities in children with autosomal dominant Alzheimer disease: a cross-sectional study. JAMA Neurol. 2015;72(8):912–9.
9. Hin N, Newman M, Kaslin J, Douek AM, Lumsden A, Xin-Fu Z, et al. Accelerated brain aging towards transcriptional inversion in a zebrafish model of familial Alzheimer's disease. bioRxivorg. 2018.
10. Cruts M, Theuns J, Van Broeckhoven C. Locus-specific mutation databases for neurodegenerative brain diseases. Hum Mutat. 2012;33(9):1340–4.
11. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
12. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754–D61.
13. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res. 2016;5:1408.
14. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
15. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.
16. Lee JH, Yu WH, Kumar A, Lee S, Mohan PS, Peterhoff CM, et al. Lysosomal proteolysis and autophagy require presenilin 1 and are disrupted by Alzheimer-related PS1 mutations. Cell. 2010;141(7):1146–58.

Chapter 4

Iron Responsive Element (IRE)-mediated

responses to iron dyshomeostasis

in Alzheimer's disease

Statement of Authorship

Title of Paper	Iron Responsive Element (IRE)-mediated responses to iron dyshomeostasis in Alzheimer's disease
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Hin, N., Newman, M., Pederson, S. and Lardelli, M., 2020. Iron Responsive Element (IRE)-mediated responses to iron dyshomeostasis in Alzheimer's disease. *bioRxiv*. DOI: 10.1101/2020.05.01.071498

Principal Author

Name of Principal Author (Candidate)	Nhi Hin
Contribution to the Paper	- All bioinformatics analysis - Drafting of manuscript - Producing figures / data visualisation - Corresponding author on manuscript - Management of deposition of transcriptome data into public database (GEO)
Overall percentage (%)	90%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 08/09/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Morgan Newman
Contribution to the Paper	- Genome editing and creation of *psen1*[Q96K97del] zebrafish mutant line. - Removal of brains from fish and purification of mRNA for RNA-seq - Editing of manuscript drafts
Signature	Date 10/11/2020

Name of Co-Author	Stephen Pederson
Contribution to the Paper	- Supervision of bioinformatics work and statistical methodology - Editing of manuscript drafts
Signature	Date

Name of Co-Author	Michael Lardelli
Contribution to the Paper	- Project supervision - Conceptualisation - Editing of manuscript drafts
Signature	Date 9/9/20

Iron Responsive Element (IRE)-mediated responses to iron dyshomeostasis in Alzheimer's disease

Author Details:

Nhi Hin*[1,2], Morgan Newman[2], Stephen Pederson[1], Michael Lardelli*[2]

1. Bioinformatics Hub, School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia
2. Alzheimer's Disease Genetics Laboratory, School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia

* corresponding authors

Author Full Names and Email Addresses:

Nhi Hin: nhi.hin@adelaide.edu.au

Morgan Newman: morgan.newman@adelaide.edu.au

Stephen Pederson: stephen.pederson@adelaide.edu.au

Michael Lardelli: michael.lardelli@adelaide.edu.au

Abstract

Background: Iron trafficking and accumulation has been associated with Alzheimer's disease (AD) pathogenesis. However, the role of iron dyshomeostasis in early disease stages is uncertain. Currently, gene expression changes indicative of iron dyshomeostasis are not well characterised, making it difficult to explore these in existing datasets.

Methods: We identified sets of genes predicted to contain Iron Responsive Elements (IREs) and used these to explore iron dyshomeostasis responses in gene expression datasets.

Results: IRE gene sets were sufficiently sensitive to distinguish not only between iron overload and deficiency in cultured cells, but also between AD and other pathological brain conditions in different brain regions. Notably, we see changes in IRE transcript abundance as amongst the earliest observable changes in zebrafish familial AD (fAD)-like brains, preceding other AD-typical pathologies such as inflammatory changes. Unexpectedly, while some IREs in the 3' untranslated regions of transcripts show significantly increased stability under iron deficiency in line with current assumptions, many such transcripts instead show decreased stability, indicating that this is not a generalizable paradigm.

Conclusions: Our results reveal iron dyshomeostasis as a likely early driver of fAD and as able to distinguish AD from other brain pathologies. Our work demonstrates how differences in the stability of IRE-containing transcripts can be used to explore

and compare iron dyshomeostasis responses in different species, tissues, and conditions.

Key words

Iron responsive element, iron homeostasis, Alzheimer's disease, gene expression, RNA-seq, gene set enrichment analysis, familial Alzheimer's disease

Background

The pathological processes underlying Alzheimer's Disease (AD) commence decades before symptoms become evident [1,2]. Since the early days of AD research, iron trafficking and accumulation has been observed to be altered in AD [3–7]. However, it is still unclear whether iron dyshomeostasis represents a late symptom or an early pathological driver of AD. Iron homeostasis is closely linked to many critical biological processes including cellular respiration, hypoxia, and immune responses, all of which are disrupted in AD. Consistent with this, discoveries over the past decade have shown that disruptions to iron homeostasis can drive feedback loops that worsen AD pathology [8,9]. However, evidence for iron dyshomeostasis as an early driver of the disease is only just emerging. Recently, amyloidogenic processing of the Amyloid Precursor Protein (APP) was shown to impair the export of iron from cells, potentially disrupting iron homoeostasis to drive excessive iron accumulation during aging with pathological consequences[10]. While iron accumulation associated with amyloid plaque formation has been observed in

various transgenic mouse models of AD [11–14], the age at which iron

dyshomeostasis first occurs is uncertain [15].

Currently, gene expression patterns representing responses to iron dyshomeostasis

are not well-characterised. Cellular responses to iron dyshomeostasis are complex

and involve several systems and layers of regulation. The stability of the transcription

factor HIF1α, (a component of HIF1, a master regulator of responses to hypoxia) is

regulated by an iron-dependent mechanism so that transcriptional responses to iron

deficiency can resemble hypoxia responses [16]. However, cellular iron homeostasis

is also regulated at the post-transcriptional level by the IRP/IRE system [17–19]. In

this system, altered levels of available ferrous iron (Fe^{2+}) cause Iron Responsive

Proteins (IRP1 or IRP2) to change conformation or stability respectively [20,21]. This

alters their ability to bind *cis*-regulatory Iron Responsive Element (IRE) stem-loop

motifs in the 3' or 5' untranslated regions (UTRs) of genes encoding products related

to iron metabolism. Only a few IRE-containing genes have been characterised in

detail, including TfR1 (transferrin receptor 1; 3' UTR IRE), DMT1 (divalent metal

transporter 1; 3' UTR IRE), H- and L-ferritin (both 5' UTR IRE), and ferroportin (5'

UTR IRE) [19]. In general, these well-characterised IRE-containing genes suggest

that IRPs binding to 3' IREs tend to stabilise transcripts to increase protein

translation, while binding to 5' IREs suppresses translation [18]. Currently however,

global gene expression changes mediated by the IRP/IRE system have not been

well-defined, and the overall expression patterns of IRE-containing genes have not

been explored in the context of AD. In addition, it is unclear how expression of these

genes might differ between AD and other neurodegenerative diseases, or how AD risk factors such as aging and hypoxia might contribute.

In this study, we utilised the SIREs (Searching for Iron Responsive Elements) tool [22] to predict and identify sets of IRE-containing genes in human, mouse, and zebrafish. We then applied these gene sets to explore overall IRP/IRE-mediated iron dyshomeostasis responses in datasets involving: (1) a cultured cell line subjected to iron overload and deficiency treatments, (2) a cohort of AD patients, healthy controls, and two other pathological conditions affecting the brain, (3) 5XFAD mice used to model the amyloid and tau pathology seen in AD, and (4) a zebrafish knock-in model possessing a familial AD (fAD)-like mutation.

Our IRE gene sets displayed significant enrichment in AD, the 5XFAD mouse model, and an fAD-like zebrafish model, demonstrating for the first time the early and extensive involvement of IRP/IRE-mediated iron dyshomeostasis responses in the context of AD. IRE gene sets were sufficiently sensitive to distinguish not only between iron overload and deficiency in a cultured cell line dataset, but also between AD and other pathological conditions affecting the brain (pathological aging and progressive supranuclear palsy), implying that the dysregulation of IRE-containing genes and iron dyshomeostasis responses in AD could differ from other conditions. Iron dyshomeostasis was already evident in young adult brains (3 month old FXFAD mice, 6 month old fAD-like zebrafish). Overall, our observations do not support the current assumption that IRP binding to 3'IREs generally stabilizes transcripts as we

observed both increases and decreases in abundance of transcripts with either 3' or 5' IREs under most conditions.

Results

Defining sets of genes containing Iron Responsive Elements (IREs)

We utilised the SIREs (Searching for Iron Responsive Elements) tool, to define species-specific IRE gene sets by searching for IRE and IRE-like motifs in the 3' and 5' untranslated regions (UTRs) of the reference transcriptomes of human (hg38), mouse (mm10), and zebrafish (z11). *SIREs* assigns a quality-score to all detected IREs, with "high-quality" scores corresponding to canonical IREs and "medium-quality" or "low-quality" scores reflecting deviations from the canonical IRE (alternative nucleotide composition in the apical loop, one bulge at the 3' or one mismatch in the upper stem) that would still produce an IRE-like motif with the potential to be functional [22–26]. **Figure 1** summarises the IRP/IRE interaction effects on transcripts and gives examples of canonical and non-canonical IREs.

We defined four gene sets for each species as follows: ***HQ 3' IREs*** (high-quality predicted 3' IRE genes), ***HQ 5' IREs*** (high-quality predicted 5' IRE genes)*, **all 3' IREs*** (including all low, medium, and high-quality predicted 3' IRE genes), and ***all 5' IREs*** (including all low, medium, and high-quality predicted 5' IRE genes). The size of these gene sets for each species and the overlap present is shown in **Figure 2** and the gene sets are provided in **Supplementary Table 1**. Overall, searching

through human UTRs uncovered the largest number of predicted IRE genes, followed by mouse, and then zebrafish. Overlap between homologous genes from the IRE gene sets of different species was generally poor. The largest sets of genes shared between species were consistently found between human and mouse, which is reflective of their closer evolutionary divergence. While not many high-quality IRE genes were identified across all three species, the few we identified are consistent with known and well-characterised IREs in the literature [27,28]. For example, the single shared **_HQ 3' IRE_** gene between the three species was *TFRC* (transferrin receptor), while the shared **_HQ 5' IRE_** genes between the three species included *FTH1* (ferritin heavy chain 1) and *ALAS2* (5'-aminolevulinate synthase 2).

IRE gene sets are over-represented within up-regulated AD genes, but overall not well-represented in existing gene sets

We explored the biological relevance of the predicted IRE gene sets described above by testing whether genes within them were over-represented in existing gene sets from the Molecular Signatures Database (MSigDB; available at https://www.gsea-msigdb.org/gsea/msigdb). MSigDB represents one of the largest publically-available collection of gene sets collated from existing studies and other biological databases [29]. We limited our analysis to gene sets from the following collections: Hallmark (non-redundant sets of ~200 genes each representing various biological activities), C2 (gene sets from databases including KEGG and Reactome and published studies), C3 (gene sets containing genes with motif elements), and C5 (gene sets based on gene ontology terms) (see **Methods**). We performed over-representation analysis for the predicted IRE gene sets (**_all 3' IREs_** and **_all 5' IREs_**)

separately for each species (human, mouse, zebrafish), and used Wilkinson's meta-analytic approach to determine gene sets that were significantly over-represented across the three species' IRE gene sets. Our results indicated that 1,148 of 10,427 tested gene sets displayed over-representation of IRE gene sets across all three species (Bonferroni-adjusted Wilkinson's p-value < 0.05) (**Supplementary Table 2**). Remarkably, the "Blalock Alzheimer's Disease Up" gene set from the C2 collection was significantly over-represented in both the sets of ***all 3' IREs*** (Bonferroni-adjusted Wilkinson's p-value = 3.9e-14) and ***all 5' IREs*** (Bonferroni-adjusted Wilkinson's p-value = 2.1e-26) in the meta-analysis, and was also the gene set with the most significant over-representation of the human ***all 3' IREs*** set (Bonferroni-adjusted Fisher's exact test p-value = 7.8e-58) (**Supplementary Table 2**). This supports that disturbance of iron homeostasis particularly distinguishes Alzheimer's disease from other disease conditions and pathways represented within the C2 collection. In addition, the top 15 MSigDB gene sets showing the most significant over-representation generally differed between species. However, in all cases, a large proportion of IRE genes from the predicted IRE gene sets were not contained within any of these top-ranked MSigDB gene sets or within the Heme Metabolism geneset belonging to the Hallmark collection (**Figure 3**). This demonstrates that the predicted IRE genes we defined are not fully captured by existing gene sets and so may be uniquely useful for investigating gene expression changes during the IRE-IRP response to iron dyshomeostasis.

Involvement of transcription factor regulation within IRE gene sets

To be confident that IRE gene sets accurately capture information about the IRP/IRE-mediated responses, we needed to investigate possible co-ordinate regulation by other factors. As a starting point, we examined whether known binding motifs for transcription factors were significantly over-represented in the promoter regions of genes within each IRE gene set (*all 3' IREs*, *all 5' IREs*, *HQ 3' IREs*, and *HQ 5' IREs*) for each species. We detected significant over-representation of several transcription factors including the Klf14 motif in the zebrafish *all 3' IREs* set (FDR-adjusted p-value = 0.049), the E2F motif in the human *all 5' IREs* set (FDR-adjusted p-value = 0.049) and the Bach1 (FDR-adjusted p-value = 0.012), Nrf2 (FDR-adjusted p-value = 0.013), and NF-E2 (FDR-adjusted p-value = 0.019) motifs in the zebrafish *HQ 5' IREs* set (**Supplementary Table 3**). This suggests that the expression of subsets of genes in some of the IRE gene sets may be influenced by other factors.

Gene set enrichment testing approach

Our predicted IRE gene sets can be used in any standard gene set enrichment testing approach to detect potential changes in iron homeostasis between conditions. Our workflow, which we later successfully apply on human, mouse, and zebrafish datasets, is shown in **Figure 4**. Due to variability in the results produced by different gene set enrichment testing methods, we were inspired by the EGSEA framework [30] to combine the results from different methods. Based on an initial analysis using EGSEA, we chose *fry/mroast* [31,32], *camera* [33], and *fgsea* [34,35] as the representative methods to use. (See **Supplementary Figure 1** for a principal component analysis of results from different gene set enrichment analysis

approaches.) A summary of the different characteristics of the three methods is shown in **Table 1**. The raw enrichment *p*-values from these approaches can be combined to obtain an overall enrichment *p*-value for each gene set. In accordance with EGSEA default parameters, we used Wilkinson's method to combine raw *p*-values, and then applied adjustment for multiple testing on all combined *p*-values. Along with performing this gene set enrichment testing on our IRE gene sets, we also recommend using the same method to test for enrichment for the MSigDB Hallmark gene sets as the diverse biological activities they represent help to provide context for interpreting the IRE enrichment results. To explore further the results of IRE gene set enrichment analysis, we use UpSet plots to display the overlap between sets of "leading-edge" genes in the ***all 3' IREs*** and ***all 5' IREs*** gene sets. The leading-edge genes can be interpreted as the core (often biologically important) genes of a gene set that account for the significant enrichment as calculated by GSEA [35].

Differences in IRE gene set enrichment during iron deficiency and iron overload in a cultured cell line

We first tested how our enrichment approach illuminated the effects of iron overload and deficiency in a cultured cell line microarray dataset from Caco-2 cells (GEO accession: GSE3573). As only one cell type contributed to this dataset, interpretation of the IRE enrichment results is simplified by not having to consider the differing iron requirements of different cell types. This allowed us to focus on whether the iron dyshomeostasis treatments (iron overload and iron deficiency) could be detected and distinguished in terms of their IRP/IRE system-driven transcript abundance

response. Because the dataset was from a microarray experiment, we performed only the differential gene expression and gene set enrichment testing portions of our workflow. The results of Principal Component Analysis and differential gene expression analysis are provided in **Supplementary Figure 2**.

In general, we found iron deficiency and iron overload treatments resulted in different gene expression responses. In terms of differential gene expression, iron deficiency was associated with 96 differentially expressed genes of which 10 possess predicted IREs while iron overload was associated with 212 differentially expressed genes (FDR-adjusted p-value from *limma* < 0.05) of which 33 possess predicted IREs (**Supplementary Figure 2**). There were 17 differentially expressed genes in common between iron deficiency and iron overload treatments, and all moved in opposite directions according to the treatment (i.e. increased abundance under iron deficiency and decreased abundance under iron overload, or vice versa). These differences between iron deficiency and iron overload were reflected in gene set enrichment analyses using the MSigDB Hallmark gene sets, where the gene sets involved and the proportions of gene transcripts with increased or decreased abundance differed (**Figure 5A**). As expected, IRE gene sets also showed significant enrichment under iron deficiency and overload conditions (**Figure 5B**) (Bonferroni adjusted p-value < 0.05).

We expected that iron deficiency would result in increased expression of genes with 3' IREs under the IRP/IRE paradigm. However, both the 3' and 5' IRE gene sets displayed mixed patterns of increased and decreased expression under both the iron

deficiency and iron overload treatments (**Figure 5B**). This indicates it would be difficult to distinguish between these conditions based purely on overall increases or decreases in the expression of IRE gene sets. Despite this, we see that the iron deficiency and iron overload treatments can be distinguished by their "leading-edge" genes (those genes contributing most to the enrichment signal for the predicted IRE gene sets (*all 3' IREs* and *all 5' IREs*) (**Figure 5C**). This supports that gene set enrichment using our predicted IRE gene sets is sufficiently sensitive to detect whether iron dyshomeostasis is present and to distinguish between different IRE-mediated gene expression responses in iron deficiency and iron overload treatments.

A distinct iron homeostasis response in human AD patients compared to other neuropathologies

Given that IRE gene sets could distinguish between iron overload and deficiency in a cultured cell line, we next tested our gene sets on a more complex data set including cerebellum and temporal cortex tissue samples from post-mortem human brains. The brains originated from either healthy controls or patients with one of three conditions: AD; pathological aging (PA), a condition involving amyloid pathology but no significant dementia symptoms; or progressive supranuclear palsy (PSP), a tauopathy without amyloid pathology [36]. An important characteristic of the dataset is that both cerebellum and temporal cortex tissue samples were available from each patient. A summary of the 236 patients whose data we included in this analysis is shown in **Table 2** and the results of differential gene expression analysis and IRE gene set enrichment analysis are shown in **Table 3** and **Figure 6A**. In our analyses,

we focus mainly on comparing conditions within each tissue rather than between

tissues. This is because we found significant differences in the AD vs. control

comparison in the temporal cortex compared to the cerebellum (see **Supplementary**

Figure 3).

Overall, our IRE enrichment analyses indicate significant enrichment of IRE gene

sets in all pathological conditions (AD, PA or PSP) compared to healthy controls

within the cerebellum and temporal cortex (**Figure 6A**). In all pathological conditions,

3' and 5' IRE gene transcripts show overall mixed patterns of abundance (e.g.

increased and decreased). Further examination of the leading-edge genes from

these IRE gene sets gives more insight regarding potential differences and

similarities between AD, PA, and PSP (**Figure 6B**). Overall, AD, PSP, and PA

appear to involve distinct yet partially overlapping IRE gene expression responses in

a tissue-specific manner. Within the temporal cortex, there are 435 3' IRE leading-

edge genes and 178 5' IRE leading-edge genes exclusively present in the "AD vs.

control" comparison. These are greater numbers than for the other two conditions.

Interestingly, AD and PA share only relatively small numbers of leading-edge genes

despite the fact that many regard PA as a prodrome of AD [37]. These observations

suggest that iron dyshomeostasis may be an essential component of AD cognitive

change. Interestingly, in the cerebellum, while AD and PSP share many 3' and 5'

IRE leading-edge genes, PA is associated with a large number of unique leading-

edge 3' and 5' IRE genes, further emphasising its difference from AD and also PSP.

In general, our IRE gene sets appear sufficiently sensitive to discern and identify

potentially interesting biological differences between these different pathological conditions affecting the brain.

Age-dependent disruption of IRE-driven iron homeostasis in the 5XFAD mouse model

Alzheimer's disease is thought to develop over decades [38–40]. However, detailed molecular studies of AD brains can only be based on post-mortem tissues. To reveal the early molecular changes that initiate the progression to AD we must study the brains of animal disease models. Given that the IRE gene sets appear to work well in human datasets, we then tested our mouse IRE gene sets on an RNA-seq dataset derived from brain cortex tissue of the 5XFAD transgenic mouse AD model (GEO: GSE140286). The 5XFAD mouse is one of the most common systems used to model the amyloid beta and tau histopathologies of AD brains. It possesses two transgenes that include a total of five different mutations, each of which individually causes fAD in humans. In this dataset, the mice analysed were either 3, 6 or 12 months of age.

Using gene set enrichment testing methods as before, we observed significant enrichment of the *all 3' IREs* and *all 5' IREs* gene sets in several comparisons. These included 5XFAD vs. wild type comparisons and wild type aging comparisons (Bonferroni-adjusted enrichment p-value < 0.05) (**Figure 7A**). Notably, even the youngest age group of 5XFAD mutant mice (3 months) displayed significant enrichment of genes containing 3' or 5' IREs compared to age-matched wild types. This is consistent with an enrichment of immune-related Hallmark gene sets that we

observed in this age group (see **Figure 7B**) and with a previous transcriptome analysis suggesting immune activation in 5XFAD mice as early as 2-4 months [41]. UpSet plots of overlapping leading-edge genes suggest that the IRE responses due to aging and due to the 5XFAD transgenes may involve partially overlapping gene expression changes (**Figure 7C**). In addition, the UpSet plots reveal subsets containing large numbers of IRE-containing genes which uniquely contribute to enrichment of IRE gene sets in only one comparison (e.g. 256 3' IRE genes only contribute to enrichment in the "*5xFAD vs WT at 6 months*" comparison). Notably, there are 126 shared 3' IRE genes contributing to enrichment in the "*5xFAD vs WT at 6 months*" and "*5xFAD vs WT at 3 months*" comparisons, but no genes shared between these comparisons and the "*5xFAD vs WT at 12 months*" comparison. These 126 genes were found to show over-representation of Gene Ontology terms including "Hsp70 protein binding" and "nuclear import signal receptor" (FDR-adj. *p*-value < 0.05) (**Supplementary Table 4**), suggesting that subsets of leading-edge genes may have biological relevance, and that these may give clues into the biologically distinct age-dependent iron dyshomeostasis responses caused by the 5XFAD transgenes. Although beyond the scope of our current analysis, further investigation into these subsets of genes may further help to define differences in iron dyshomeostasis responses under different biological conditions of interest.

Similarities in IRE-driven iron homeostasis responses during hypoxia and a familial AD-like mutation in a zebrafish model

Concerns have been raised over the relevance of transgenic mouse models in modelling the early stages of AD (reviewed by [42]). In contrast, knock-in models of

familial AD (fAD) only exhibit subtle pathological changes with little to no visible amyloid or tau pathology present [43]. However, because they more closely mimic the genetic state of fAD, knock-in models may reveal the early molecular changes that drive later fAD pathology. We had access to whole-brain RNA-seq data from a knock-in zebrafish model of fAD possessing a single fAD-like mutation in its endogenous *psen1* gene (*psen1*$^{Q96\_K97del}$/+). Previous analysis of a subset of this dataset involving young adults (6-month-old brains) revealed gene expression changes related to altered energy metabolism [44]. (In contrast, the 3-month-old young adult 5XFAD mouse brain dataset is dominated by immune/inflammation responses, **Figure 7B**). Considering the critical role of iron homeostasis in energy metabolism, we decided to revisit this zebrafish dataset. We performed IRE gene set enrichment on the entire dataset in order to include exploration of the effects of aging and acute hypoxia (two important risk factors for sporadic late onset AD) and to analyse how these effects interact with the fAD-like *psen1*$^{Q96\_K97del}$/+ mutant genotype. The experimental design and the results of the differential gene expression analyses and gene set enrichment tests are summarised in **Figure 8**.

We first turned our attention to the comparison between young adult (6-month-old) *psen1*$^{Q96\_K97del}$/+ mutant zebrafish and their wild-type siblings. At this age, gene expression changes in the mutant fish likely represent early stresses driving the development of fAD in humans. Gene set enrichment tests in this comparison identify alteration of energy metabolism-related Hallmark gene sets (e.g. OXIDATIVE PHOSPHORYLATION, GLYCOLYSIS) (**Figure 8B**) which is consistent with a previous analysis of gene ontology terms with this dataset [44]. In addition, we see

enrichment of other gene sets including FATTY ACID METABOLISM, PI3K AKT MTOR SIGNALLING, MTORC1 SIGNALLING, and HEME METABOLISM. All of these gene sets which show enrichment in 6-month-old $psen1^{Q96\_K97del}$/+ mutants relative to wild-type siblings are also enriched in 24-month-old $psen1^{Q96\_K97del}$/+ mutants relative to their wild-type siblings (**Figure 8B**). This supports that the biological activities represented in these gene sets are amongst the earliest altered in this particular fAD mutation model.

In Lumsden et al. [15] we predicted that fAD mutations in the major locus *PSEN1* would cause an early cellular deficiency of ferrous iron due to the observation of insufficient acidification of the endolysosomal pathway in in-vitro *PSEN1* mutation studies [45,46]. In Newman et al. [44] we saw that GO analysis of 6 month old $psen1^{Q96\_K97del}$/+ zebrafish brain supported that lysosomal acidification was affected. Therefore, we decided to apply our IRE enrichment analysis to test for evidence of iron dyshomeostasis in these fish. The enrichment of the ***all 3' IREs*** set in 6 month old $psen1^{Q96\_K97del}$/+ zebrafish brains supports that iron dyshomeostasis is an important stress in the early stages of fAD (**Figure 8C**).

While almost all pairwise comparisons in the zebrafish dataset show significant enrichment of at least one IRE gene set (**Figure 8C**), the expression of IRE genes appears to differ in terms of the proportions of IRE-containing transcripts which show increased versus decreased abundance (**Figure 8D**). In addition, the Principal Component Analysis plot of expression of the ***all 3' IREs*** set over all samples shown in **Figure 8E** suggests that different conditions appear to have distinct expression

patterns of these genes. Across the first principal component, different age groups

(6- and 24-month-old brains) differ in their expression of predicted 3' IRE genes,

while the second principal component appears to separate *psen1*[Q96_K97del]/+ mutants

from their wild-type siblings. This separation between *psen1*[Q96_K97del]/+ mutants and

wild type siblings is even more pronounced when both are exposed to hypoxia.

To gain more insight into the similarities and differences between the IRE responses

in the *psen1*[Q96_K97del]/+ vs. wild type comparison, we plotted UpSet plots of

overlapping leading-edge genes (**Figure 8F**). These plots suggest that the IRE

responses during hypoxia, aging, and due to this fAD-like mutation are mostly

distinct from each other with unique leading-edge genes. However, similarities in the

IRE response between different conditions are suggested by the existence of some

shared leading-edge genes. For example, for the set of **all 3' IREs**, the "*6-month-old*

psen1[Q96_K97del]/+ *vs. wild type*" comparison shares 19 leading-edge genes with the

"*24-month-old psen1*[Q96_K97del]/+ *vs. wild type*" comparison. As an initial exploration of

the biological relevance of these genes, we submitted them to the STRINGR tool.

This indicated that the proteins products of these genes were significantly associated

with each other (e.g. in terms of text-mining from Pubmed articles, experimentally

determined interactions, and co-expression). These proteins were significantly over-

represented in the sets "MAPK pathway", "AP-1 transcription factor", "Jun-like

transcription factor", and "signaling by TGF-beta family members" (FDR-adjusted

over-representation p-value < 0.05; **Supplementary Figure 4**). The AP-1 and MAPK

pathways have previously been shown to be stimulated by iron depletion [47,48].

Therefore, mechanistically, it is possible that the IRE-IRP response to iron

dyshomeostasis in the fAD-like zebrafish mutant might involve these pathways, although this requires confirmation *in vivo*.

The fact that aging, hypoxia, and a fAD-like mutation all cause changes in the abundance of IRE-containing transcripts in zebrafish raised concerns regarding the specificity of such changes. Therefore, as a negative control, we examined changes in IRE transcript abundance in a brain transcriptome dataset derived from zebrafish heterozygous for a *psen1* mutation now thought not to be fAD-like, *psen1*^{K97fs} [49]. In this dataset, *psen1*^{K97fs}/+ zebrafish are compared to their wild type siblings at 6 and 24 months. We tested this dataset for enrichment using our zebrafish 3' and 5' IRE gene sets but found no significant enrichment of any of our predicted IRE gene sets in *psen1*^{K97fs}/+ vs. wild type comparisons at any age (**Table 4**; **Supplementary Figure 5**). Reassuringly, we still observed significant enrichment of both 3' and 5' IRE gene sets during wild type aging (*24-month-old wild types vs. 6-month-old* wild types), consistent with the equivalent comparison in the *psen1*$^{Q96\_K97del}$/+ dataset. These results support that IRE-containing transcript abundance changes are sufficiently sensitive to reflect differences in iron homeostasis between different mutation models.

Simultaneous stabilisation of some 3' IRE transcripts and destabilisation of others

In the cultured cell line dataset analysed above, we noticed that even a straightforward iron deficiency treatment resulted in the simultaneous increase and decrease in expression of 3' IRE-containing genes. These findings are difficult to

reconcile with the current paradigm that stabilisation of 3' IRE-containing genes occurs under iron deficiency and suggest that the current model of the IRP/IRE system may be incomplete or insufficient for describing the regulation of non-canonical IREs. Given that many predicted 3' IRE genes with non-canonical IREs (e.g. in the **all 3' IREs** set) displayed enrichment and different gene expression patterns in the fAD-like zebrafish dataset, we decided to explore further the stability changes of these genes by comparing the expression of spliced and unspliced transcripts for each gene (**Supplementary Text 1**). We found that transcripts of some predicted 3' IRE genes were significantly increased in stability while others were significantly decreased in stability (**Supplementary Figure 6**; **Supplementary Table 5**).

Discussion

In this study we successfully identified sets of genes predicted to contain IREs in human, mouse, and zebrafish. We found that IRE genes were generally not well-represented in the existing gene sets in MSigDB. Importantly, these genes were most significantly over-represented in an existing gene set previously shown to be up-regulated in AD brains (the "Blalock Alzheimer's Disease Up" gene set from MSigDB). This supports the importance of iron dyshomeostasis in sporadic late onset Alzheimer's disease. Furthermore, our IRE gene sets displayed significant enrichment in postmortem brains from a human AD cohort, from the 5XFAD mouse model, and from a zebrafish model of a fAD-like mutation. This demonstrates for the first time the involvement of a coordinated IRE-containing gene expression response

to iron dyshomeostasis in the context of AD.

Overall, IRE-containing gene sets were sensitive enough to capture several interesting phenomena. The relatively tightly controlled conditions in the mouse and zebrafish model datasets revealed a strong age-dependent effect on the transcript abundances of these predicted IRE-containing genes. Notably, 3' IRE gene expression changes were amongst the earliest changes observable in the zebrafish fAD-like mutation model alongside changes in energy metabolism. These 3' IRE gene expression changes preceded other signals of pathological change in the transcriptome (such as altered expression of inflammatory response pathways) commonly associated with AD. In addition, IRE gene sets were sufficiently sensitive to distinguish not only between iron overload and deficiency in a cultured cell line dataset, but also between AD and other pathological conditions affecting the brain (i.e. pathological aging and progressive supranuclear palsy). This suggests that the dysregulation of IRE-containing genes and iron homeostasis in AD may differ from other conditions.

Whether iron deficiency or iron overload was present in the AD brain tissue samples (taken from either temporal cortex or cerebellum) was unclear. Most previous work has assumed that accumulation of iron in the brain with age (a phenomenon observed broadly across animal phyla [50,51]) is indicative of cellular iron overload driving oxidative stress[52]. Indeed, iron accumulation and oxidative stress apparently contribute to the insolubility of amyloid plaques and neurofibrillary tangles in AD brains [53]. However, a recent publication by [54] showed that disturbed

lysosomal function leading to increased lysosomal pH causes a deficiency of functional ferrous iron (Fe^{2+}) while non-functional ferric iron (Fe^{3+}) accumulated in lysosomes. The deleterious effects of this on the brains of mice (defective mitochondrial biogenesis and function and stimulation of inflammatory responses) could be alleviated by increasing the levels of iron in their diet. The observations of Nixon and colleagues that acidification of the endolysomal pathway is affected both by fAD mutations in *PSEN1* [45] and excessive dosage of the *APP* gene [55], together with our observations from our fAD-like *psen1*$^{Q96\_K97del}$/+ mutant zebrafish, support the possibility that fAD brains may suffer a ferrous iron deficiency in a background of ferric iron overload. Intriguingly, the greatest genetic risk factor for late onset AD, the ε4 allele of the gene *APOE*, appears to increase lysosomal pH [56] but ε4's increased risk of AD is alleviated in individuals who possess the *HFE* 282Y allele that predisposes to the iron overload disease hemochromatosis [57]. Given that many other risk loci for sporadic late onset AD also affect endolysosomal pathway function (reviewed in [58]) it is reasonable to suggest that disturbed of iron homeostasis may afflict brains with this disease.

Outside our interest in AD, our analyses also revealed the surprising finding that 3' IRE-containing genes could be both upregulated and downregulated by iron deficiency (at least in a cultured cell line). We also observed simultaneous stabilisation of some 3'IRE transcripts and destabilisation of others in our analysis of zebrafish brains. This challenges the simplistic paradigm of stabilisation of 3' IRE-containing transcripts under ferrous iron deficiency. Previous research holds that under iron deficiency, 3' IRE-containing genes such as *DMT1* (*divalent metal*

transporter 1) and *TFRC* (*transferrin receptor protein*) are stabilised and hence increased in expression [21] in order to increase ferrous iron availability. However, our findings reveal that this same principle may not apply to other, less-characterised 3' IRE-containing genes, including IREs deviating from the canonical IRE sequence. Many of these IRE-like sequences are likely to have some functionality [24–26], and their expression and stability changes in our analyses indicate that they are likely to be important in iron dyshomeostasis responses in AD. We emphasise the need to characterise further the stability of these predicted 3' IRE-containing transcripts under conditions of altered iron availability to better understand iron dyshomeostasis responses at the gene expression level.

The limitations of our study, the conservation of IRE gene sets between species, and the role of iron dyshomeostasis in AD are discussed further in **Supplementary Text 2**.

Conclusions

Our results demonstrate for the first time the involvement of a coordinated IRE-containing gene expression response to iron dyshomeostasis in the context of AD. By searching entire human, mouse, and zebrafish transcriptomes for all genes containing potential IREs, we formed comprehensive gene sets effective for exploring iron homeostasis in any human, mouse, or zebrafish gene expression dataset. While, for the first time, our work suggests the existence of iron dyshomeostasis in the young, pre-pathology brains of PSEN1 fAD mutation carriers, it also reinforces the critical importance of iron dyshomeostasis in the development

of the late onset, sporadic form of AD. More broadly, our approach highlights how

changes in the stability and abundance of IRE-containing transcripts can be used to

give insight into iron dyshomeostasis responses in different species, tissues, and

conditions.

Methods

Defining IRE gene sets for human, mouse, and zebrafish

We extracted all 3' and 5' UTR sequences from the human, mouse and zebrafish genome assemblies using the Bioconductor packages BSgenome.Hsapiens.UCSC.hg38, BSgenome.Mmusculus.UCSC.mm10 and BSgenome.Drerio.UCSC.danRer11, and gene definitions from the Ensembl 94 release. Each set of UTR sequences was then submitted input into the SIREs web server (v.2.0 [22]). The SIREs algorithm assigns quality scores to predicted IREs taking into account whether the sequence is canonical, whether it contains any mismatches and bulges, and the free energy of the secondary structure. Canonical sequences are tagged by SIREs as being high-quality, while IRE-like sequences are tagged as low or medium quality. Given that High-quality IRE predictions miss the majority of true IREs, this enables a more comprehensive sampling of IRE motifs. For human, mouse, and zebrafish, we separately defined the following four gene sets: **HQ 3' IREs** and **HQ 5' IREs** (representing genes with high-quality predicted IREs), along with **all 3' IREs** and **all 5' IREs**, which included genes containing any predicted IRE in the respective UTR. Comparisons between gene sets were performed using the UpSetR package (v.1.4.0 [59]) with mappings between species obtained by BioMart [60].

Over-representation of IRE gene sets in existing MSigDB gene sets

We downloaded the following gene set collections from MSigDB (v.6.0 [29]): Hallmark, C2 (gene sets from online pathway databases and biomedical literature,

including KEGG and the REACTOME databases), C3 (motif gene sets based on regulatory targets), and C5 (gene sets defined by Gene Ontology terms). We excluded the following collections from analysis: C1 (specific to human chromosomes and cytogenetic bands, while our analysis involves different species), C4 (computationally-defined cancer-focused gene sets), C6 (oncogenic signatures) and C7 (immunologic signatures). C4, C6, and C7 were not included as the level of detail in the gene sets in these specific collections is more domain-specific rather than broad-level. We used Fisher's exact test to determine whether any IRE geneset was significantly over-represented in each MSigDB gene set. Gene sets were defined as having significant enrichment for IRE gene sets if the FDR-adjusted p-value from Fisher's exact test was below 0.05. UpSet plots were produced using the UpSetR package (v.1.4.0 [59]) while network representations were produced in Gephi (v.0.9.3 [61]). To produce network visualisations, we exported node and edge tables from R. The nodes table contained the following gene sets: top 15 gene sets (ranked by Fisher's exact test p-value), Hallmark Heme Metabolism gene set, *all 3' IREs*, *all 5' IREs*, and all genes contained within these gene sets. The edges table contained gene – gene set edges which indicated the gene set(s) that genes belonged to. To create the network plots, we used "Force Atlas 2" as the initial layout algorithm, followed by the "Yifan Hu" [62] layout algorithm to improve the separation between groups of genes.

Over-representation analysis of transcription factor motifs in IRE gene promoters

Defining promoter regions as being 1500 bp upstream and 200 bp downstream of the transcription start site for each gene, we used the findMotifs.pl script from HOMER (v.4.11) [63,64] to search for known transcription factor binding site (TFBS) motifs in the promoters of each IRE gene set. The HOMER Motif database contains 363 vertebrate transcription factor binding motifs based on analysis of high-quality public ChIP-seq datasets (http://homer.ucsd.edu/homer/motif/HomerMotifDB/homerResults.html). We considered TFBS motifs as being significantly enriched in a gene set if the FDR-adjusted p-value was less than 0.05.

Gene set enrichment testing

We performed all gene set enrichment tests in R v3.6.1 [65] using *fry* [31,32], *camera* [33], and *fgsea* [34,35]. For *fry*, and *camera*, we used model fits obtained using *limma* [66,67], whilst for *fgsea*, a ranked list was obtained using *moderated t*-statistics taken from *limma*. All genes were used in gene set enrichment tests (i.e. not just DE genes). We combined the raw p-values from *fry, camera*, and *fgsea* using Wilkinson's method [68] with default parameters, followed by FDR-adjustment. When performing gene set enrichment testing on the MSigDB Hallmark gene sets, we applied FDR-adjustment to combined p-values and defined significant enrichment as gene sets having an adjusted p-value < 0.05. When performing gene set enrichment on the four IRE gene sets (***all 3' IREs***, ***all 5' IREs***, ***HQ 3' IREs***, ***HQ 5' IREs***), we applied Bonferroni-adjustment to combined p-values to further protect

against Type I errors and defined significant enrichment as gene sets having an adjusted p-value < 0.05. Depending on the species in the dataset being analysed, we used the respective IRE gene sets defined for human, mouse, or zebrafish.

Analysis of the Caco-2 cultured cell line dataset

We downloaded processed microarray data from the GEO dataset GSE3573. This study investigated gene expression responses to iron treatments, including iron deficiency (cells treated with ferric ammonium citrate), and iron overload (cells grown in DMEM-FBS medium with hemin) [69]. We performed differential gene expression analysis using the "lmFit" and eBayes" functions in *limma* [66]. Genes were defined as differentially expressed when their FDR-adjusted p-value < 0.05.

Analysis of the Mayo Clinic RNA-seq dataset

We downloaded processed CPM count data from Synapse (https://www.synapse.org/#!Synapse:syn5550404). We matched cerebellum and temporal cortex samples by their patient ID, and only retained genes which were present across all samples and patients for which there were both cerebellum and temporal cortex samples (n=236 patients with measurements for cerebellum and temporal cortex, 472 samples in total). We performed analysis using *limma* [66,67] and determined differentially expressed genes between conditions. In addition, we used the "duplicateCorrelation" function in *limma*, setting the "block" parameter to the patient ID. Genes were considered differentially expressed if their FDR-adjusted p-value < 0.05.

fAD-like *psen1*^{Q96_K97del/+} zebrafish

The isolation of the *psen1*^{Q96_K97del} mutation has previously been described [44].
Mutations were only analysed in the heterozygous state in this study.

Zebrafish husbandry and animal ethics

This study was approved under permits S-2014-108 and S-2017-073 issued by the
Animal Ethics Committee of the University of Adelaide. Tübingen strain zebrafish
were maintained in a recirculated water system.

Hypoxia treatment of female adult zebrafish

psen1^{Q96_K97del/+} mutants and their wild-type siblings were treated in low oxygen
levels by placing zebrafish in oxygen-depleted water for 3 hours (oxygen
concentration of 6.6 ± 0.2 mg/L in normoxia and 0.6 ± 0.2 mg/L in hypoxia.

Whole brain removal from adult zebrafish

After normoxia or hypoxia treatment adult fish were euthanized by sudden immersion
in an ice water slurry for at least ~30 seconds before decapitation and removal of the
entire brain for immediate RNA or protein extraction. All fish brains were removed
during late morning/noon to minimise any influence of circadian rhythms.

RNA extraction from whole brain

Total RNA was isolated from heterozygous mutant and WT siblings using the
*mir*Vana miRNA isolation kit (Thermo Fisher). RNA isolation was performed
according to the manufacturer's protocol. First a brain was lysed in a denaturing lysis

solution. The lysate was then extracted once with acid-phenol:chloroform leaving a semi-pure RNA sample. The sample was then purified further over a glass-fiber filter to yield total RNA. Total RNA was DNase treated using the DNA-free™ Kit from Ambion, Life Technologies according to the manufacturer's instructions. Total RNA was then sent to the Genomics Facility at the South Australian Health and Medical Research Institute (Adelaide, Australia) to assess RNA quality and for subsequent RNA sequencing *(using poly-A enriched RNA-seq technology, and estimated gene expression from the resulting single-end 75 bp reads using the reference GRCz11 zebrafish assembly transcriptome)*

Pre-processing of the fAD-like *psen1*^{Q96_K97del/+} zebrafish dataset:

RNA-seq libraries contained single-end 75bp Illumina NextSeq reads. We performed quality trimming with *AdapterRemoval* using default parameters, followed by quality assessment with FastQC and ngsReports. Trimmed reads were pseudo-aligned to the reference zebrafish transcriptome using *Kallisto* (v.0.45) [70] and transcript descriptions from Ensembl release 94. The "catchKallisto" function from *edgeR* [71] was used to import and summarise counts from transcript-level to gene-level, with all subsequent analyses performed at the gene-level.

Differential gene expression analysis for fAD-like zebrafish dataset

For differential gene expression analysis, we retained all genes with expression of at least 1 cpm in 4 or more samples, and used voomWithQualityWeights to downweight lower quality samples [67]. Contrasts were defined to include all relevant pairwise comparisons between conditions, and genes were considered as differentially expressed using an FDR-adjusted *p*-value < 0.05.

Estimation of spliced and unspliced gene expression in fAD-like zebrafish dataset

For spliced transcripts in Ensembl release 94, we additionally defined unspliced genes including intronic regions. Unspliced transcripts were appended to the end of the reference transcriptome and used to build a new Kallisto [70] index. Estimated counts for spliced transcripts and unspliced genes were imported into R using the "catchKallisto" function from *edgeR* [71].

Gene set enrichment tests for non-fAD-like (K97fs/+) zebrafish dataset

Please refer to [44] for details on RNA-seq data processing and analysis of the non-fAD-like dataset. In the current work, we used the gene expression counts matrix with *limma*. The voom, design, and contrasts objects produced as part of the *limma* analysis were used for gene set enrichment analysis with the zebrafish IRE gene sets we defined as well as the MSigDB Hallmark gene sets. Significantly enriched Hallmark gene sets had FDR-adjusted p-value < 0.05 while IRE gene sets were considered significantly enriched if the Bonferroni-adjusted p-value < 0.05.

Gene set enrichment tests for 7-day-old Q96_K97del/+ dataset

Please refer to Dong et al. (GEO accession: GSE148631, manuscript in preparation) for details on RNA-seq data processing and analysis. In the current work, we used the gene expression counts matrix with *limma* to perform gene set enrichment analysis. IRE gene sets we defined as well as the MSigDB Hallmark gene sets. Significantly enriched Hallmark gene sets had FDR-adjusted p-value < 0.05 while

IRE gene sets were considered significantly enriched if the Bonferroni-adjusted *p*-value < 0.05.

Differential transcript stability analysis

The estimated spliced and unspliced transcript count estimates from *kallisto* [70] were imported into R using the catchKallisto function from *edgeR* [71]. We used *limma* [66] to determine the logFC of spliced transcripts and unspliced transcripts for each comparison. To test for whether there was a significant difference in the logFC of the spliced and unspliced transcripts, we used Welch's *t*-test with the *s2.prior* values from *limma* as the variances of the spliced and unspliced transcripts. We defined the null (no stabilisation of transcript) and alternate (stabilisation of transcript) hypotheses for each gene as follows, where s and u refer to the spliced and unspliced versions of a particular gene:

$$H_0: \ \mathrm{logFC_s} = \mathrm{logFC_u}$$

$$H_a: \ \mathrm{logFC_s} \neq \mathrm{logFC_u}$$

We defined genes with FDR-adjusted *p*-values < 0.05 as having differential stability.

Checks that observed gene expression differences are not artifacts of differences in proportions of cell types

The Mayo Clinic RNA-seq study, 5XFAD mice, and fAD-like zebrafish datasets are bulk RNA-seq datasets. To confirm that any gene expression changes were likely due to altered transcriptional programs rather than changes in cell type proportions, we compared expression of marker genes for four common neural cell types

(astrocytes, neurons, oligodendrocytes, microglia) in conditions within each dataset. The marker genes for astrocytes, neurons, and oligodendrocytes were obtained from MSigDB gene sets from [72] which were based on studies in mice. The marker genes for microglia were derived from [73] which was based on studies in human and mouse. All gene IDs were converted to human, mouse, or zebrafish Ensembl IDs using BioMart [60] for each dataset. Please refer to **Supplementary Text 3** and **Supplementary Figures 7-9** for details.

Abbreviations

AD: Alzheimer's disease

fAD: familial Alzheimer's disease

sAD: sporadic Alzheimer's disease

FDR: False discovery rate

GSEA: Gene set enrichment analysis

IRE: Iron responsive element

IRP: Iron responsive protein

MSigDB: Molecular Signatures Database

PA: Pathological aging

PSP: Progressive supranuclear palsy

UTR: Untranslated region

Declarations

edited the manuscript. ML conceived the project, supervised all aspects of it, and edited the manuscript. All authors read and approved the manuscript.

References

1. Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's Dement. Elsevier Inc.; 2018;14:535–62.

2. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. Elsevier Inc.; 2011;7:280–92.

3. Bartzokis G, Sultzer D, Mintz J, Holt LE, Marx P, Kelly Phelan C, et al. In vivo evaluation of brain iron in Alzheimer's disease and normal subjects using MRI. Biol Psychiatry. 1994;35:480–7.

4. Bartzokis G, Tishler TA. MRI evaluation of basal ganglia ferritin iron and neurotoxicity in Alzheimer's and Huntingon's disease. Cell Mol Biol (Noisy-le-grand). 2000;46:821–33.

5. Ding B, Chen KM, Ling HW, Sun F, Li X, Wan T, et al. Correlation of iron in the hippocampus with MMSE in patients with alzheimer's disease. J Magn Reson Imaging. 2009;29:793–8.

6. Langkammer C, Ropele S, Pirpamer L, Fazekas F, Schmidt R. MRI for iron mapping in Alzheimer's disease. Neurodegener Dis. 2014;13:189–91.

7. Van Duijn S, Bulk M, Van Duinen SG, Nabuurs RJA, Van Buchem MA, Van Der Weerd L, et al. Cortical Iron Reflects Severity of Alzheimer's Disease. J Alzheimer's Dis. IOS Press; 2017;60:1533–45.

8. Mills E, Dong XP, Wang F, Xu H. Mechanisms of brain iron transport: Insight into neurodegeneration and CNS disorders. Future Med Chem. NIH Public Access; 2010;2:51–64.

9. Yamamoto A, Shin R-W, Hasegawa K, Naiki H, Sato H, Yoshimasu F, et al. Iron (III) induces aggregation of hyperphosphorylated τ and its reduction to iron (II) reverses the aggregation: implications in the formation of neurofibrillary tangles of Alzheimer's disease. J Neurochem. Wiley; 2004;82:1137–47.

10. Tsatsanis A, Wong BX, Gunn AP, Ayton S, Bush AI, Devos D, et al. Amyloidogenic processing of Alzheimer's disease β-amyloid precursor protein induces cellular iron retention. Mol Psychiatry [Internet]. Springer Nature; 2020 [cited 2020 Jun 23];1–9. Available from: https://www.nature.com/articles/s41380-020-0762-0

11. Telling ND, Everett J, Collingwood JF, Dobson J, van der Laan G, Gallagher JJ, et al. Iron Biochemistry is Correlated with Amyloid Plaque Morphology in an Established Mouse Model of Alzheimer's Disease. Cell Chem Biol. Elsevier Ltd; 2017;24:1205-1215.e3.

12. Leskovjan AC, Kretlow A, Lanzirotti A, Barrea R, Vogt S, Miller LM. Increased brain iron coincides with early plaque formation in a mouse model of Alzheimer's disease. Neuroimage. Academic Press; 2011;55:32–8.

13. Gong NJ, Dibb R, Bulk M, van der Weerd L, Liu C. Imaging beta amyloid aggregation and iron accumulation in Alzheimer's disease using quantitative susceptibility mapping MRI. Neuroimage [Internet]. Academic Press Inc.; 2019 [cited 2020 Jun 24];191:176–85. Available from: https://pubmed.ncbi.nlm.nih.gov/30739060/

14. El Tannir El Tayara N, Delatour B, Le Cudennec C, Guégan M, Volk A, Dhenain M. Age-related evolution of amyloid burden, iron load, and MR relaxation times in a transgenic mouse model of Alzheimer's disease. Neurobiol Dis. Academic Press; 2006;22:199–208.

15. Lumsden AL, Rogers JT, Majd S, Newman M, Sutherland GT, Verdile G, et al. Dysregulation of neuronal iron homeostasis as an alternative unifying effect of mutations causing familial alzheimer's disease. Front Neurosci. Frontiers Media S.A.; 2018;12:533.

16. Lane DJR, Merlot AM, Huang MLH, Bae DH, Jansson PJ, Sahni S, et al. Cellular iron uptake, trafficking and metabolism: Key molecules and mechanisms and their roles in disease. Biochim Biophys Acta - Mol Cell Res. Elsevier; 2015;1853:1130–44.

17. Gray NK, Hentze MW. Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs. EMBO J. Wiley; 1994;13:3882–91.

18. Hentze MW, Muckenthaler MU, Andrews NC. Balancing acts: Molecular control of mammalian iron metabolism. Cell. Cell Press; 2004;117:285–97.

19. Muckenthaler MU, Galy B, Hentze MW. Systemic Iron Homeostasis and the Iron-Responsive Element/Iron-Regulatory Protein (IRE/IRP) Regulatory Network. Annu Rev Nutr [Internet]. Annual Reviews; 2008 [cited 2020 Apr 30];28:197–213. Available from: http://www.annualreviews.org/doi/10.1146/annurev.nutr.28.061807.155521

20. Basilion JP, Rouault TA, Massinople CM, Klausner RD, Burgesst WH. The iron-responsive element-binding protein: Localization of the RNA-binding site to the aconitase active-site cleft (UV cross-linking/base hydrolysis). Biochemistry. 1994;91:574–8.

21. Pantopoulos K. Iron metabolism and the IRE/IRP regulatory system: An update. Ann. N. Y. Acad. Sci. New York Academy of Sciences; 2004. p. 1–13.

22. Campillos M, Cases I, Hentze MW, Sanchez M. SIREs: searching for iron-responsive elements. Nucleic Acids Res [Internet]. 2010 [cited 2020 Apr 30];38:W360-367. Available from: http://ccbg.imppc.org/sires/index.html

23. Goforth JB, Anderson SA, Nizzi CP, Eisenstein RS. Multiple determinants within iron-responsive elements dictate iron regulatory protein binding and regulatory hierarchy. RNA. 2010;16:154–69.

24. Henderson BR, Menotti E, Kühn LC. Iron regulatory proteins 1 and 2 bind distinct sets of RNA target sequences. J Biol Chem. 1996;271:4900–8.

25. Henderson BR, Menotti E, Bonnard C, Kuhn LC. Optimal sequence and structure of iron-responsive elements. Selection of RNA stem-loops with high affinity for iron regulatory factor. J Biol Chem. 1994;269:17481–9.

26. Butt J, Kim HY, Basilion JP, Cohen S, Iwai K, Philpott CC, et al. Differences in the RNA binding sites of iron regulatory proteins and potential target diversity. Proc Natl Acad Sci U S A. National Academy of Sciences; 1996;93:4345–9.

27. Surinya KH, Cox TC, May BK. Identification and characterization of a conserved erythroid-specific enhancer located in intron 8 of the human 5-aminolevulinate synthase 2 gene. J Biol Chem. American Society for Biochemistry and Molecular Biology Inc.; 1998;273:16798–809.

28. Theil EC. Regulation of ferritin and transferrin receptor mRNAs. J Biol Chem. 1990;265:4771–4.

29. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP, et al. Databases and ontologies Molecular signatures database (MSigDB) 3.0.

Bioinformatics [Internet]. 2011 [cited 2020 Apr 30];27:1739–40. Available from:

http://www.broadinstitute.org/msigdb.

30. Alhamdoosh M, Law CW, Tian L, Sheridan JM, Ng M, Ritchie ME. Easy and

efficient ensemble gene set testing with EGSEA. F1000Research. Faculty of 1000

Ltd; 2017;6:2010.

31. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. ROAST:

rotation gene set tests for complex microarray experiments. Bioinformatics [Internet].

2010 [cited 2020 Apr 30];26:2176–82. Available from: www.bioconductor.org

32. Giner G, Smyth GK, Giner G, Smyth GK. <p>FRY: a fast approximation to

ROAST gene set test with mean aggregated set statistics</p>. F1000Research.

2016;5.

33. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene

correlation. Nucleic Acids Res [Internet]. 2012 [cited 2020 Apr 30];40:e133. Available

from: https://academic.oup.com/nar/article-abstract/40/17/e133/2411151

34. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis

using cumulative statistic calculation. bioRxiv [Internet]. Cold Spring Harbor

Laboratory; 2016 [cited 2020 Apr 30];060012. Available from:

http://dx.doi.org/10.1101/060012%0Ahttps://www.biorxiv.org/content/10.1101/06001

2v1

35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et

al. Gene set enrichment analysis: A knowledge-based approach for interpreting

genome-wide expression profiles. Proc Natl Acad Sci U S A. National Academy of

Sciences; 2005;102:15545–50.

36. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al.

Human whole genome genotype and transcriptome data for Alzheimer's and other

neurodegenerative diseases. Sci Data. Nature Publishing Groups; 2016;3:1–10.

37. Ossenkoppele R, Van Berckel BN, Prins ND. Amyloid imaging in prodromal

Alzheimer's disease. Alzheimer's Res Ther. BioMed Central; 2011;3:26.

38. Protas HD, Chen K, Langbaum JBS, Fleisher AS, Alexander GE, Lee W, et al.

Posterior cingulate glucose metabolism, hippocampal glucose metabolism, and

hippocampal volume in cognitively normal, late-middle-aged persons at 3 levels of

genetic risk for alzheimer disease. JAMA Neurol. American Medical Association;

2013;70:320–5.

39. Caldwell CC, Yao J, Brinton RD. Targeting the Prodromal Stage of Alzheimer's

Disease: Bioenergetic and Mitochondrial Opportunities. Neurotherapeutics [Internet].

Springer New York LLC; 2015 [cited 2020 Apr 30];12:66–80. Available from:

http://link.springer.com/10.1007/s13311-014-0324-8

40. Vermunt L, Sikkes SAM, van den Hout A, Handels R, Bos I, van der Flier WM, et

al. Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in

relation to age, sex, and APOE genotype. Alzheimer's Dement. Elsevier Inc.;

2019;15:888–98.

41. Bundy JL, Vied C, Badger C, Nowakowski RS. Sex-biased hippocampal

pathology in the 5XFAD mouse model of Alzheimer's disease: A multi-omic analysis.

J Comp Neurol. Wiley-Liss Inc.; 2019;527:462–75.

42. Drummond E, Wisniewski T. Alzheimer's disease: experimental models and

reality. Acta Neuropathol. Springer Verlag; 2017. p. 155–75.

43. Xia D, Watanabe H, Wu B, Lee SH, Li Y, Tsvetkov E, et al. Presenilin-1 knockin

mice reveal loss-of-function mechanism for familial alzheimer's disease. Neuron. Cell

Press; 2015;85:967–81.

44. Newman M, Hin N, Pederson S, Lardelli M. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production. Mol Brain [Internet]. BioMed Central Ltd.; 2019 [cited 2020 Apr 30];12:43. Available from:

https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-019-0467-y

45. Lee JH, Yu WH, Kumar A, Lee S, Mohan PS, Peterhoff CM, et al. Lysosomal proteolysis and autophagy require presenilin 1 and are disrupted by Alzheimer-related PS1 mutations. Cell. Elsevier; 2010;141:1146–58.

46. Lee JH, McBrayer MK, Wolfe DM, Haslett LJ, Kumar A, Sato Y, et al. Presenilin 1 Maintains Lysosomal Ca2+ Homeostasis via TRPML1 by Regulating vATPase-Mediated Lysosome Acidification. Cell Rep. Elsevier B.V.; 2015;12:1430–44.

47. Yu Y, Richardson DR. Cellular iron depletion stimulates the JNK and p38 MAPK signaling transduction pathways, dissociation of ASK1-thioredoxin, and activation of ASK1. J Biol Chem. 2011;286:15413–27.

48. Riera H, Afonso V, Collin P, Lomri A. A Central Role for JNK/AP-1 Pathway in the Pro-Oxidant Effect of Pyrrolidine Dithiocarbamate through Superoxide Dismutase 1 Gene Repression and Reactive Oxygen Species Generation in Hematopoietic Human Cancer Cell Line U937. Papa S, editor. PLoS One [Internet]. Public Library of Science; 2015 [cited 2020 Apr 30];10:e0127571. Available from:

https://dx.plos.org/10.1371/journal.pone.0127571

49. Hin N, Newman M, Kaslin J, Douek AM, Lumsden A, Nik SHM, et al. Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human PSEN2. PLoS One. Public Library of Science;

2020;15:e0227258.

50. Zecca L, Youdim MBH, Riederer P, Connor JR, Crichton RR, Stankowski JN, et al. Iron-export ferroxidase activity of β-amyloid precursor protein is inhibited by zinc in Alzheimer's disease. Cell [Internet]. BioMed Central Ltd; 2011 [cited 2020 Apr 30];142:359. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17987043

51. Jacomin A-C, Geraki K, Brooks J, Tjendana-Tjhin V, Collingwood JF, Nezis IP. Impact of Autophagy and Aging on Iron Load and Ferritin in Drosophila Brain. Front Cell Dev Biol [Internet]. Frontiers Media S.A.; 2019 [cited 2020 Apr 30];7:142. Available from: https://www.frontiersin.org/article/10.3389/fcell.2019.00142/full

52. Ayton S, Lei P, Bush AI. Metallostasis in Alzheimer's disease. Free Radic Biol Med. Elsevier Inc.; 2013;62:76–89.

53. Smith MA, Harris PLR, Sayre LM, Perry G. Iron accumulation in Alzheimer disease is a source of redox-generated free radicals. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 1997 [cited 2020 Jun 24];94:9866–8. Available from: https://www.pnas.org/content/94/18/9866

54. Yambire KF, Rostosky C, Watanabe T, Pacheu-Grau D, Torres-Odio S, Sanchez-Guerrero A, et al. Impaired lysosomal acidification triggers iron deficiency and inflammation in vivo. Elife. eLife Sciences Publications Ltd; 2019;8.

55. Jiang Y, Sato Y, Im E, Berg M, Bordi M, Darji S, et al. Lysosomal dysfunction in down syndrome is app-dependent and mediated by APP-βCTF (c99). J Neurosci. Society for Neuroscience; 2019;39:5255–68.

56. Prasad H, Rao R. Amyloid clearance defect in ApoE4 astrocytes is reversed by epigenetic correction of endosomal pH. Proc Natl Acad Sci U S A. National Academy of Sciences; 2018;115:E6640–9.

57. Tisato V, Zuliani G, Vigliano M, Longo G, Franchini E, Secchiero P, et al. Gene-gene interactions among coding genes of iron-homeostasis proteins and APOE-alleles in cognitive impairment diseases. Ginsberg SD, editor. PLoS One [Internet]. Public Library of Science; 2018 [cited 2020 Apr 30];13:e0193867. Available from: https://dx.plos.org/10.1371/journal.pone.0193867

58. Van Acker ZP, Bretou M, Annaert W. Endo-lysosomal dysregulations and late-onset Alzheimer's disease: Impact of genetic risk factors. Mol Neurodegener. BioMed Central Ltd.; 2019;14:1–20.

59. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics [Internet]. [cited 2020 Apr 30]; Available from: https://gehlenborglab.shinyapps.io/

60. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart - Biological queries made easy. BMC Genomics [Internet]. BioMed Central; 2009 [cited 2020 Apr 30];10:22. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-22

61. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs. Third Int AAAI Conf Weblogs Soc Media [Internet]. 2009 [cited 2020 Apr 30]; Available from: www.aaai.org

62. Hu Y. Algorithms for Visualizing Large Networks. 2011.

63. Benner C. HOMER (Hypergeometric Optimization of Motif EnRichment) [Internet]. Available from: http://homer.ucsd.edu/homer/

64. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory

Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010;38:576–89.

65. Team RC. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.r-project.org/

66. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. Oxford University Press; 2015;43:e47.

67. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15.

68. Wilkinson B. A statistical consideration in psychological research. Psychol Bull. 1951;48:156–8.

69. Chicault C, Toutain B, Monnier A, Aubry M, Fergelot P, Le Treut A, et al. Iron-related transcriptomic variations in CaCo-2 cells, an in vitro model of intestinal absorptive cells. Physiol Genomics. 2006;26:55–67.

70. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. Nature Publishing Group; 2016;34:525–7.

71. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Oxford University Press; 2009;26:139–40.

72. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.

73. Bonham LW, Sirkis DW, Yokoyama JS. The transcriptional landscape of microglial genes in aging and neurodegenerative disease. Front Immunol. Frontiers Media S.A.; 2019;10:1170.

Tables

Table 1. Summary of gene set testing approaches used in our analyses.

Method	Hypothesis	Focused / battery approach	Accounts for inter-gene correlation	R implementation
Fry / Roast	Self-contained null hypothesis (genes within a set have no association with experimental condition)	Focused (each gene set is tested on its own terms, and multiple testing adjustment is applied afterwards)	Yes	fry() and mroast() functions in the *limma* Bioconductor package (Wu et al. 2010)
Camera	Competitive null hypothesis (genes within a set do not have a stronger association with experimental condition compared to a random gene set)	Focused (each gene set is tested on its own terms, and multiple testing adjustment is applied afterwards)	Yes	camera() function in the *limma* Bioconductor package (Wu & Smyth 2012)
fgsea (fast implementation of GSEA in R)	Self-contained null hypothesis (genes within a set have no association with experimental condition)	Battery (gene sets are pitted against each other to determine which ones are more significantly enriched)	No	fgsea() function in the *fgsea* Bioconductor package (Subramanian et al. 2005)

Table 2. Samples analysed from Mayo Clinic RNA-seq study.

Group	Total number	% Female	Mean age ± s.d (years)	Diagnosis
Alzheimer's disease (AD)	134	56.7	82.6 ± 7.3	Braak Stage ≥ 4, diagnosis according to NINCDS-ADRDA criteria
Control	130	46.2	82.9 ± 8.3	Braak Stage ≤ 3, No or sparse CERAD neuritic and cortical plaque density, no diagnoses for any neurodegenerative disease
Pathological Aging (PA)	44	54.5	84.5 ± 4.3	Braak Stage ≤ 3, CERAD neuritic and cortical plaque density of 2 or more, no diagnoses for any neurodegenerative disease or mild cognitive impairment
Progressive supranuclear palsy (PSP)	164	40.2	73.8 ± 6.5	Braak Stage ≤ 3, no or sparse CERAD neuritic and cortical plaque density

Table 3. Differential gene expression and IRE gene set enrichment results from Mayo Clinic RNA-seq study. Asterisks (*) indicate significant enrichment (Bonferroni-adjusted Wilkinson's *p*-value from *fry*, *camera*, and *fgsea* < 0.05).

	Comparison	No. of DE genes (FDR p-value < 0.05 and abs(logFC) > 0.5)	All 3' IREs enrichment *p*-value	All 5' IREs enrichment *p*-value	HQ 3' IREs enrichment *p*-value	HQ 5' IREs enrichment *p*-value
Effect of Alzheimer's disease	AD vs. control (cerebellum)	Down: 1,422 Up: 1,308	0.0109*	0.128	0.770	0.770
	AD vs. control (temporal cortex)	Down: 1,650, Up: 1,929	9.03E-15*	2.44E-14*	4.50E-15*	8.95E-08*
Effect of pathological aging (amyloid pathology, no dementia symptoms)	PA vs. control (cerebellum)	Down: 254, Up: 463	0.000116*	0.00139*	0.0201*	0.770
	PA vs. control (temporal cortex)	Down: 466, Up: 512	0.000521*	0.0461*	0.000065*	0.0000682*
Effect of progressive supranuclear palsy (PSP)	PSP vs. control (cerebellum)	Down: 2,550, Up: 1,080	1.81E-22*	3.45E-22*	1.43E-12*	0.0399*
	PSP vs. control (temporal cortex)	Down: 1,271, Up: 622	3.58E-11*	5.17E-13*	1.39E-06*	0.0399*
Differences between Alzheimer's disease and pathological aging	AD vs. PA (cerebellum)	Down: 1,728, Up: 1,006	0.000116*	0.00139*	0.0201*	0.770
	AD vs. PA (temporal cortex)	Down: 2,633, Up: 2,174	0.000521*	0.0461*	0.000065*	0.0000682*
Differences between Alzheimer's disease and progressive supranuclear palsy	AD vs. PSP (cerebellum)	Down: 108, Up: 659	0.0019*	0.00184*	0.158	0.158
	AD vs. PSP (temporal cortex)	Down: 2,052, Up: 2,354	9.12E-22*	2.10E-22*	5.35E-16*	0.040*
Differences between cerebellum and temporal cortex tissue	Cerebellum vs. temporal cortex in controls	Down: 10,925, Up: 9,302	0.00E+00*	0.00E+00*	2.82E-269*	1.60E-81*
	Cerebellum vs. temporal cortex in AD	Down: 11,966, Up: 9,950	0.00E+00*	0.00E+00*	7.71E-302*	9.35E-92*
	Cerebellum vs. temporal cortex in PA	Down: 10,200, Up: 8,737	1.87E-153*	5.88E-160*	1.22E-89*	1.31E-88*
	Cerebellum vs. temporal cortex in PSP	Down: 11,552, Up: 9,576	0.00E+00*	0.00E+00*	0.00E+00*	4.74E-104*

Table 4. Enrichment of Iron Responsive Element (IRE) gene sets in fAD-like zebrafish dataset. Raw *p*-values from *fry*, *camera* and *fgsea* were combined with Wilkinson's method, with combined *p*-values then Bonferroni-adjusted for multiple testing. The same process was repeated for the K97fs/+ dataset, which involves an independent family of fish (†).

	Comparison	*All 3' IREs* enrichment *p*-value	*All 5' IREs* enrichment *p*-value	*HQ 3' IREs* enrichment *p*-value	*HQ 5' IREs* enrichment *p*-value
Effect of *psen1* mutation	6-month-old Q96_K97del/+ vs. 6-month-old +/+ all under normoxia	0.0109	0.128	0.770	0.770
	6-month-old Q96_K97del/+ vs. 6-month-old +/+ all under hypoxia	9.03E-15	2.44E-14	4.50E-15	8.95E-08
	24-month-old Q96_K97del/+ vs. 24-month-old +/+ all under normoxia	0.000116	0.00139	0.0201	0.770
	24-month-old Q96_K97del/+ vs. 24-month-old +/+ all under hypoxia	0.000521	0.0461	0.000065	0.0000682
	6-month-old K97fs/+ vs. 6-month-old +/+ all under normoxia †	0.199	0.289	0.0197	0.289
	24-month-old K97fs/+ vs. 24-month-old +/+ all under normoxia †	0.116	0.115	0.00000118	0.979
Effect of hypoxia	6-month-old +/+ under hypoxia vs. 6-month-old +/+ under normoxia	1.10E-18	1.10E-18	8.19E-13	1.27E-08
	24-month-old +/+ under hypoxia vs. 24-month-old +/+ under normoxia	3.67E-15	7.98E-11	2.94E-07	3.31E-05
	6-month-old Q96_K97del/+ under hypoxia vs. 6-month-old Q96_K97del/+ under normoxia	1.13E-24	9.88E-22	2.28E-13	2.86e- 3
	24-month-old Q96_K97del/+ under hypoxia vs. 24-month-old Q96_K97del/+ under normoxia	0.000521	0.015	0.128	0.770
Effect of aging	24-month-old +/+ vs. 6-month-old +/+ all under normoxia	6.80E-35	4.82E-27	3.41E-18	1.26E-21
	24-month-old +/+ vs. 6-month-old +/+ all under hypoxia	1.55E-35	6.84E-31	2.36E-20	7.40E-23
	24-month-old Q96_K97del/+ vs. 6-month-old Q96_K97del/+ all under normoxia	1.85E-27	1.03E-21	8.09E-17	1.07E-17
	24-month-old Q96_K97del/+ vs. 6-month-old Q96_K97del/+ all under hypoxia	3.77E-30	1.59E-23	2.11E-19	2.46E-16
	24-month-old +/+ vs. 6-month-old +/+ all under normoxia (K97fs family) †	1.69E-10	3.96E-08	5.95E-10	4.59E-06
	24-month-old K97fs/+ vs. 6-month-old K97fs/+ all under normoxia †	0.199	0.289	0.0197	0.289
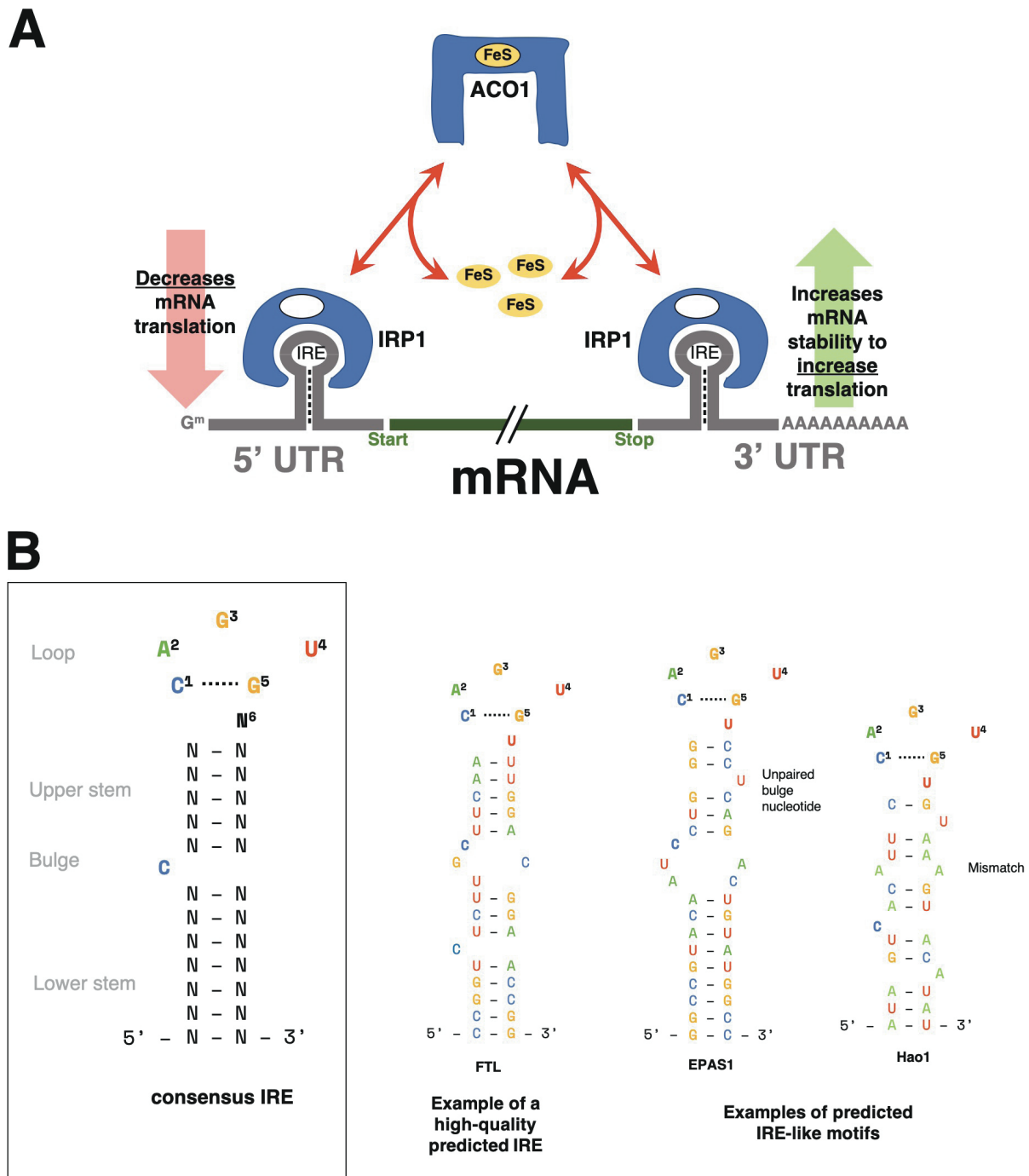
Figure 1. IRE background. A. Model of altered transcript stability due to the IRP/IRE system under iron dyshomeostasis. When cellular iron levels (through FeS) are low, ACO1 will undergo a conformational transformation and act as an IRP to bind IREs in the 5' or 3' untranslated region (UTR) of genes involved in iron homeostasis. In general, genes with IREs in their 3' UTR will be stabilised and increased in expression while genes with IREs in their 5' UTR will have their translation inhibited by the IRP. **B. Consensus IRE secondary structure and examples of high-quality and IRE-like motifs predicted by SIREs.** IRE-like motifs with non-canonical structure are able to be detected by SIREs if they have up to one mismatch pair in the upper stem (e.g. Hao1) or 1 unpaired bulge nucleotide on the 3' strand of the upper stem (e.g. EPAS1). For more details on the prediction of non-canonical IRE motifs, please refer to Figure 1 of Campillos et al. [24].
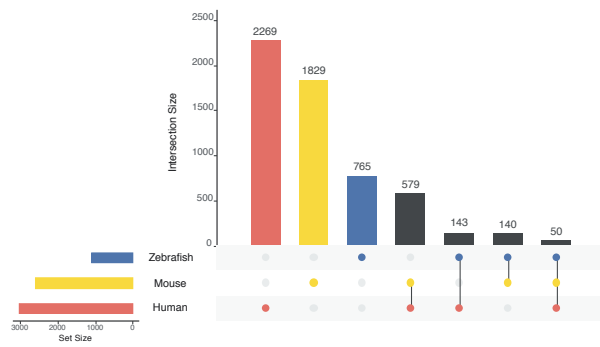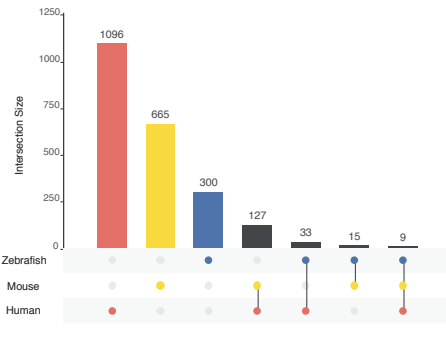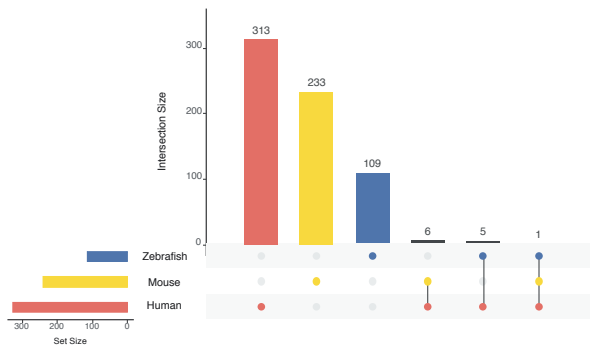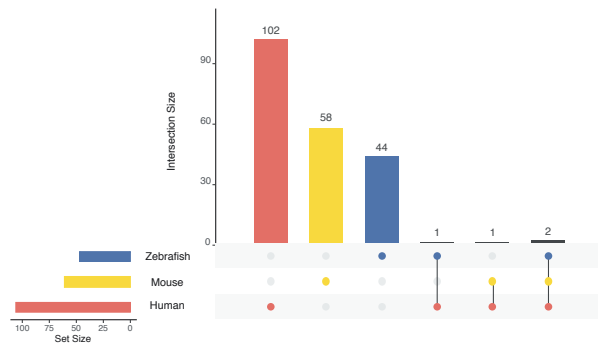
Figure 2. Overlap between predicted IRE gene sets for human, mouse, and zebrafish. The number of genes in the gene set for each species is shown at the bottom-left bars of each UpSet plot, while genes with shared homologs across species are indicated in the main plot region. IRE genes in mouse and zebrafish gene sets were excluded from this plot if they did not have a human homolog.
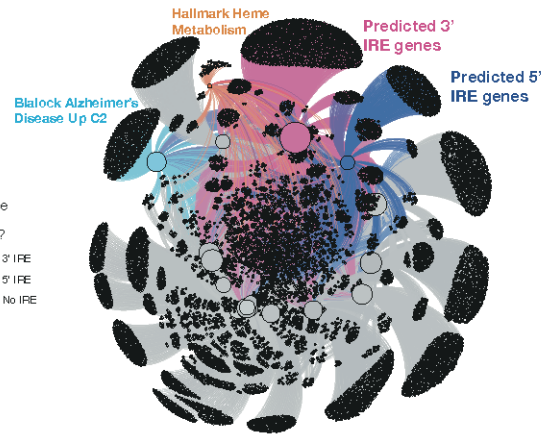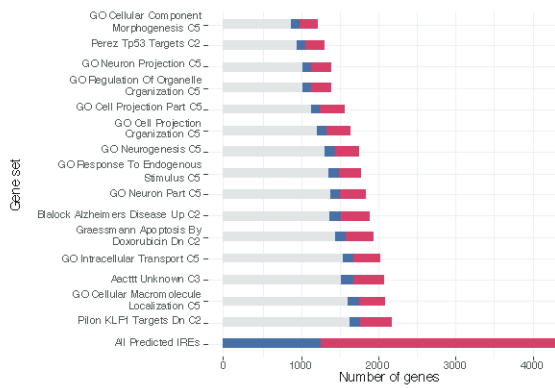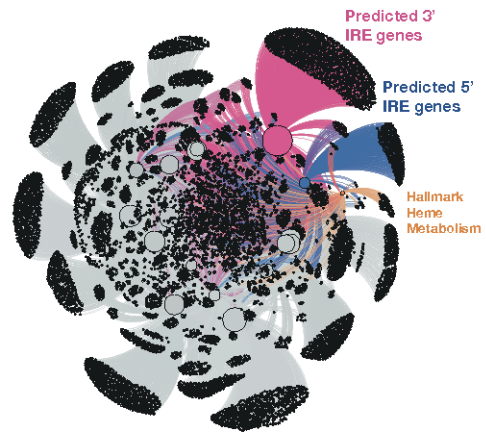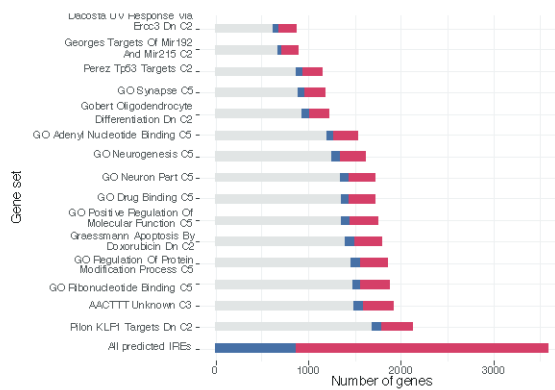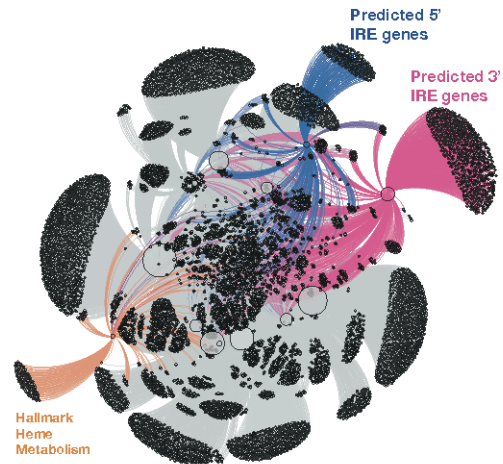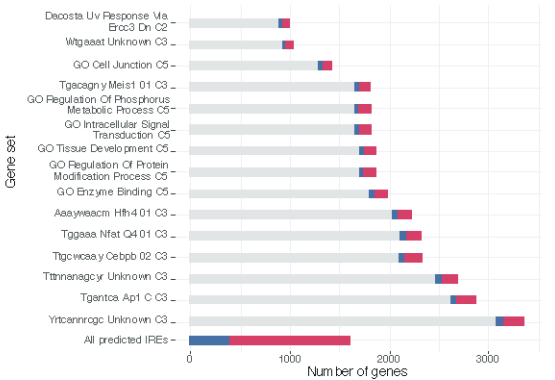
Figure 3. MSigDB gene sets showing over-representation of predicted 3' and 5' IRE genes in human, mouse, and zebrafish. The top 15 MSigDB gene sets ranked by Fisher's exact test *p*-value (testing for over-representation of the ***all 3' IREs*** and/or ***all 5' IREs*** sets) are shown for each species. In the network plots, the top 15 MSigDB gene sets are shown as large nodes, with genes represented as small nodes. Edges connecting genes to gene sets indicate the gene set(s) that a gene belongs to. Overall, the ***all 3' IREs*** and ***all 5' IREs*** gene sets have a large proportion of genes which are not included in any of the top ranked MSigDB gene sets for each species.

Inputs

Gene counts matrix

- Predicted IRE gene sets
- MSigDB Hallmark gene sets

Gene counts matrix including both spliced and unspliced genes

Differential gene expression analysis with *limma*

Gene set enrichment tests with *mroast*/*fry*, *camera*, and *fgsea*

Combine raw *p*-values with Wilkinson's method + adjustment for multiple tests

Identification of candidate 3' IRE genes stabilised under iron deficiency

Welch's *t*-test on logFC of spliced and unspliced genes to determine if changes in stability between experimental conditions are significant

Results

- Differentially expressed genes
- *limma* objects (design, contrasts, and either fit and expression matrix, or voom-adjusted counts)

Significantly enriched gene sets
- Leading edge genes contributing to *fgsea* enrichment signal for IRE gene sets
- Proportions of genes in each gene set showing increased or decreased expression

3' IRE genes showing significant increase in expression and stability

Figure 4. IRE-containing gene expression analysis workflow. The section including identification of candidate 3' IRE genes stabilised under iron deficiency was only applied to the fAD-like zebrafish dataset due to unavailable raw RNA-seq reads for the other datasets needed to identify expression of unspliced genes.
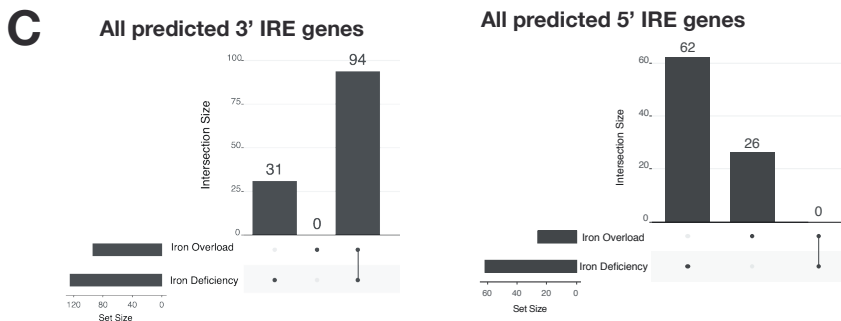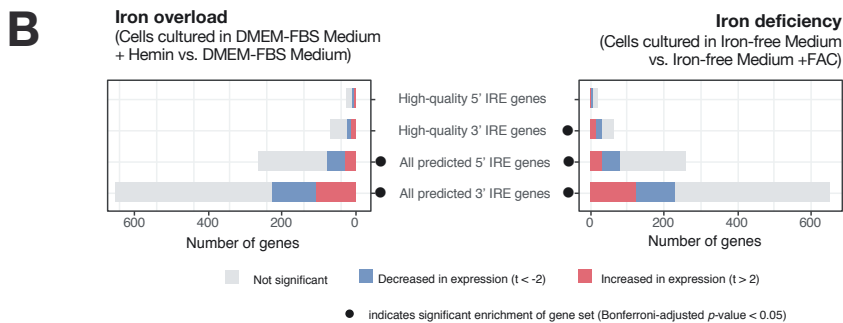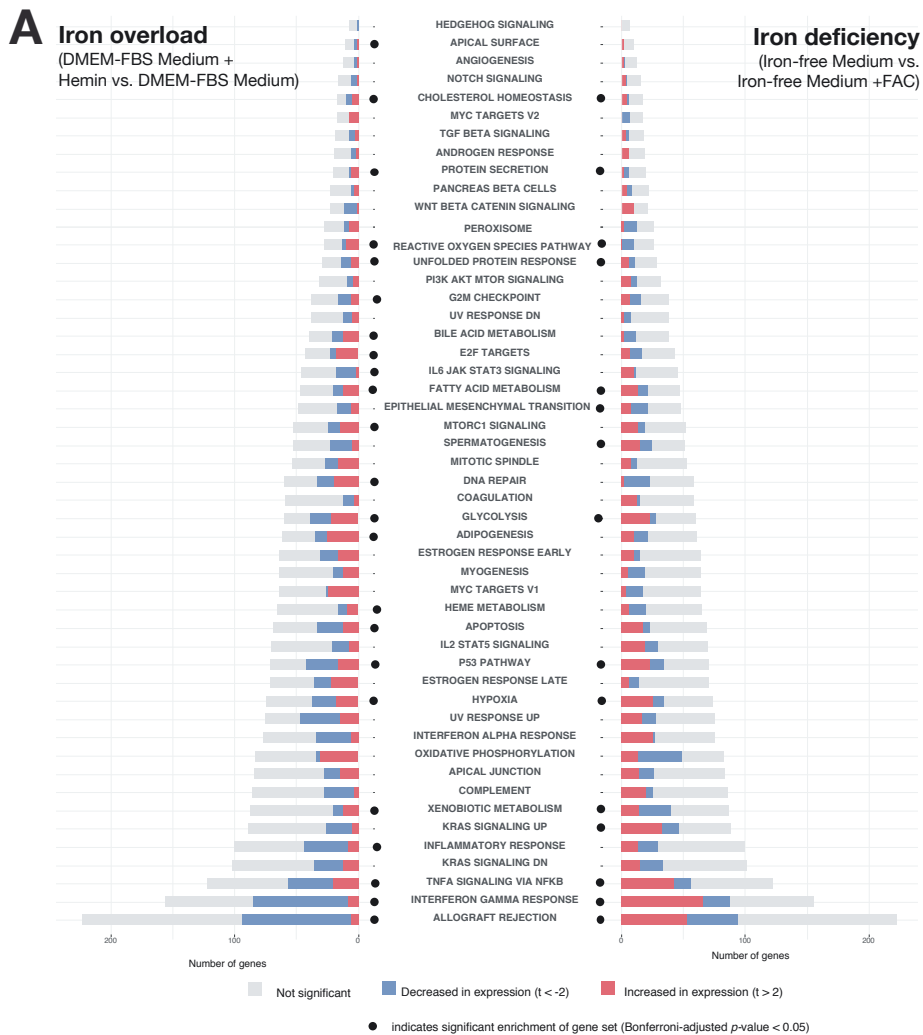
Figure 5. Analysis of Caco-2 cultured cell line microarray dataset. A. Gene set enrichment testing results for human predicted IRE gene sets for the iron overload and iron deficiency treatments. Dots indicate if a gene set was considered significantly enriched (Bonferroni-adjusted *p*-value < 0.05) in the iron overload (on left) or iron deficiency (on right) treatment. **B. Gene set enrichment testing results for MSigDB Hallmark gene sets in the iron overload and iron deficiency treatments.** Dots indicate if a gene set was considered significantly enriched (FDR-adjusted p-value < 0.05) in the iron overload (on left) or iron deficiency (on right) treatments. C. **UpSet plots showing overlap between iron overload and iron deficiency treatments in GSEA leading-edge genes for the "All predicted 3' IRE genes" and "All predicted 5' IRE genes" gene sets.** The bars to the lower-left indicate the number of the leading-edge genes for iron overload and iron deficiency treatments, while the bars in the main plot region indicate the number of leading-edge genes which are unique or shared between the treatments. 144
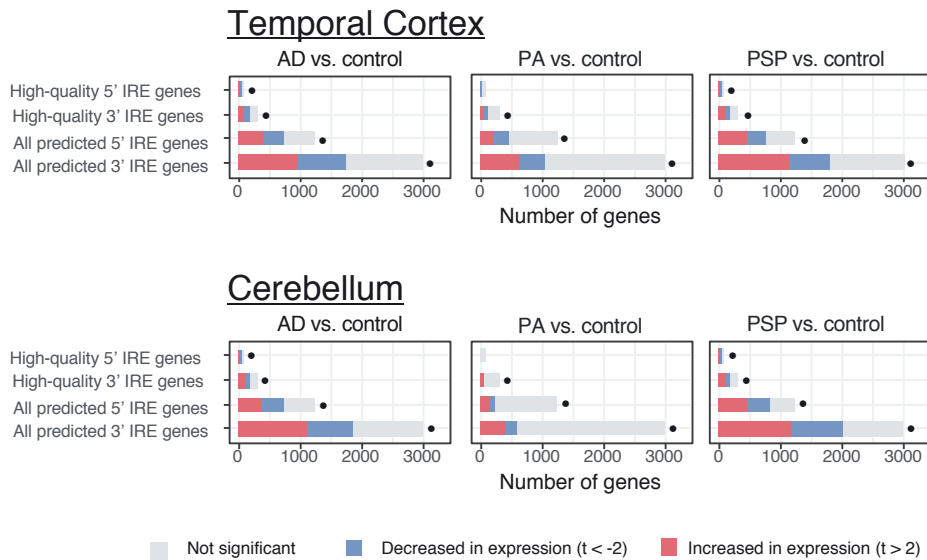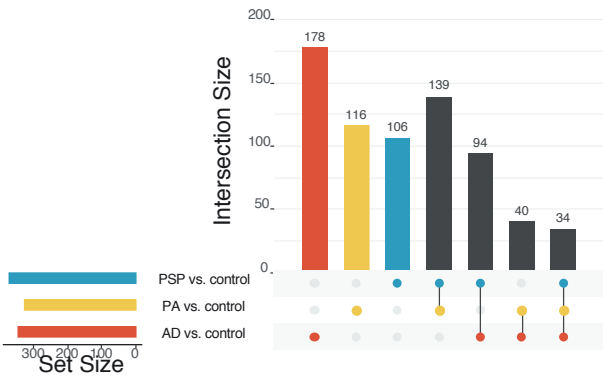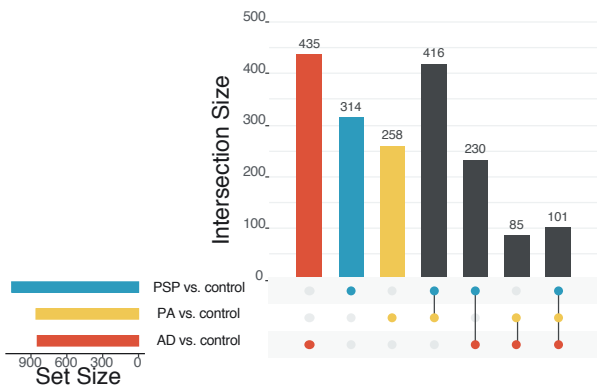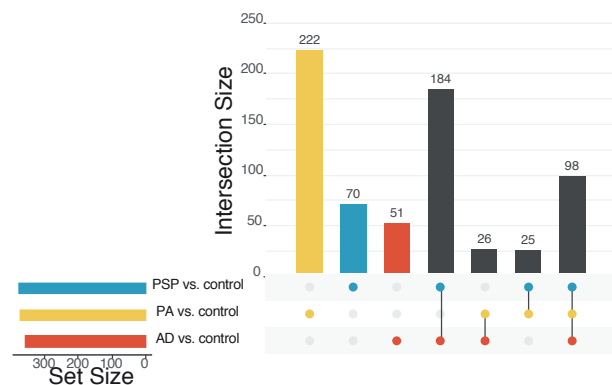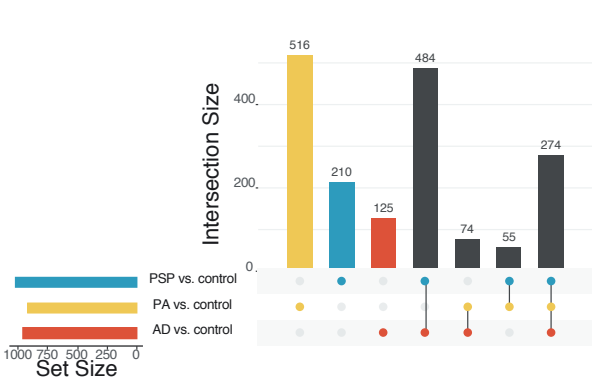
Figure 6. Analysis of Mayo Clinic RNA-seq dataset with human IRE gene sets. A. Enrichment analysis results of IRE gene sets in comparisons of AD (Alzheimer's Disease), PA (Pathological Aging), and PSP (Progressive Supranuclear Palsy) vs. control. The proportions of IRE genes within each IRE gene set with increased, decreased, or unchanged expression are indicated as coloured bars. The dots indicate that the IRE gene set was significantly enriched (Bonferroni-adjusted enrichment p-value < 0.05). **B. UpSet plots showing overlap between leading-edge IRE genes in comparisons of AD (Alzheimer's Disease), PA (Pathological Aging), and PSP (Progressive Supranuclear Palsy) vs. control.** The bars to the lower-left region of each UpSet plot indicate the number of leading-edge IRE genes for each comparison, while bars in the main plot region indicate the number of leading-edge IRE genes which are unique or shared between comparisons.

145
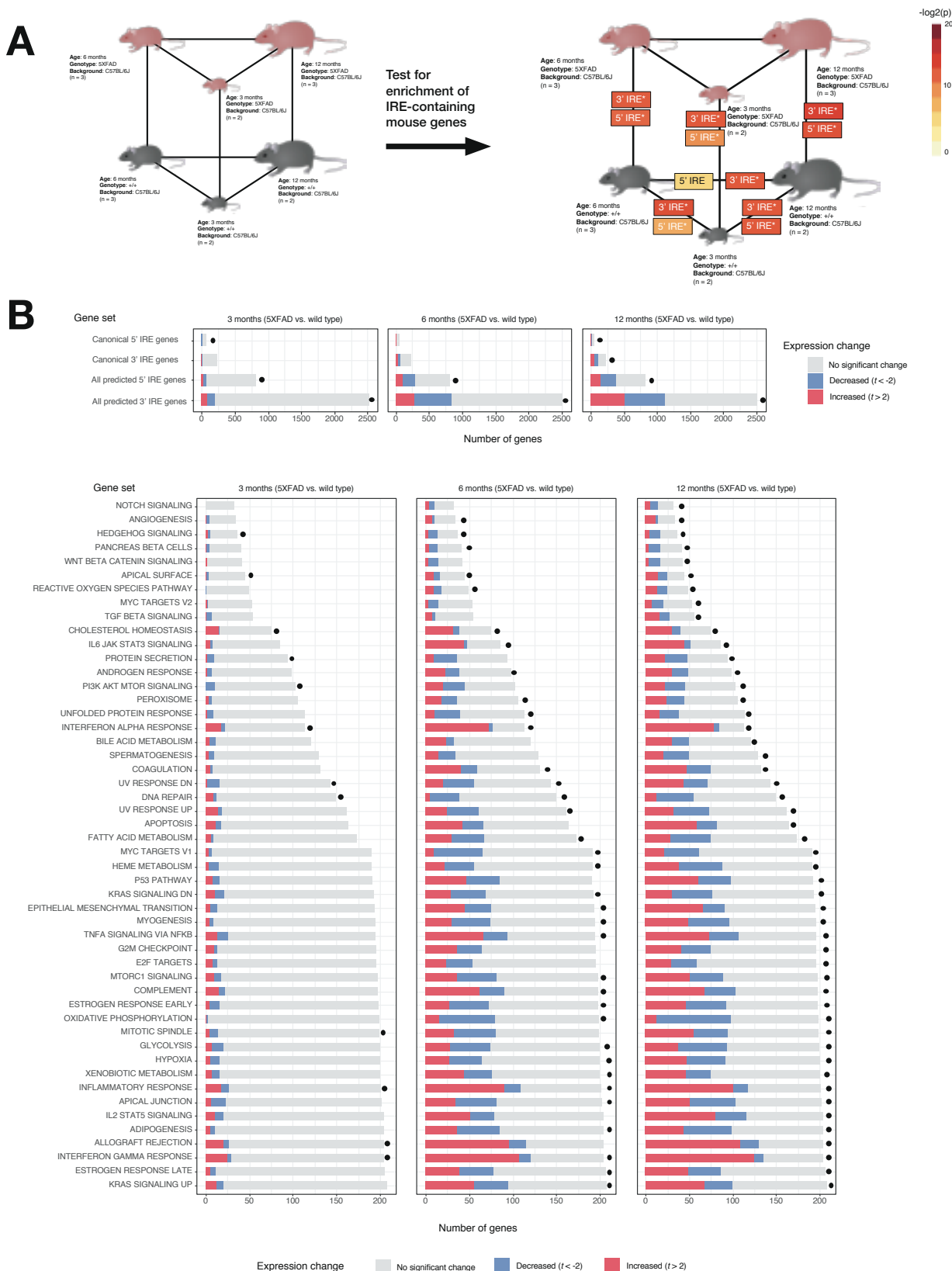
Figure 7. Analysis of 5XFAD mouse dataset. A. Experimental design and results of IRE gene set enrichment analysis. The gene sets **all 3' IREs** and all **5' IREs** derived from searching for IRE sequences in the UTRs of genes in the reference mouse genome mm10 are represented here as "3' IRE" and "5' IRE" respectively. Asterisks (*) indicate that the gene set was significantly enriched in a particular comparison (Bonferroni-adjusted p-value < 0.05). **B. Proportions of genes in IRE and MSigDB Hallmark gene sets which are increased (t > 2) or decreased (t < -2) in expression in all "5XFAD vs. wild type" comparisons.** A dot next to a bar indicates that the gene set was significantly enriched (FDR-adjusted p-value < 0.05).
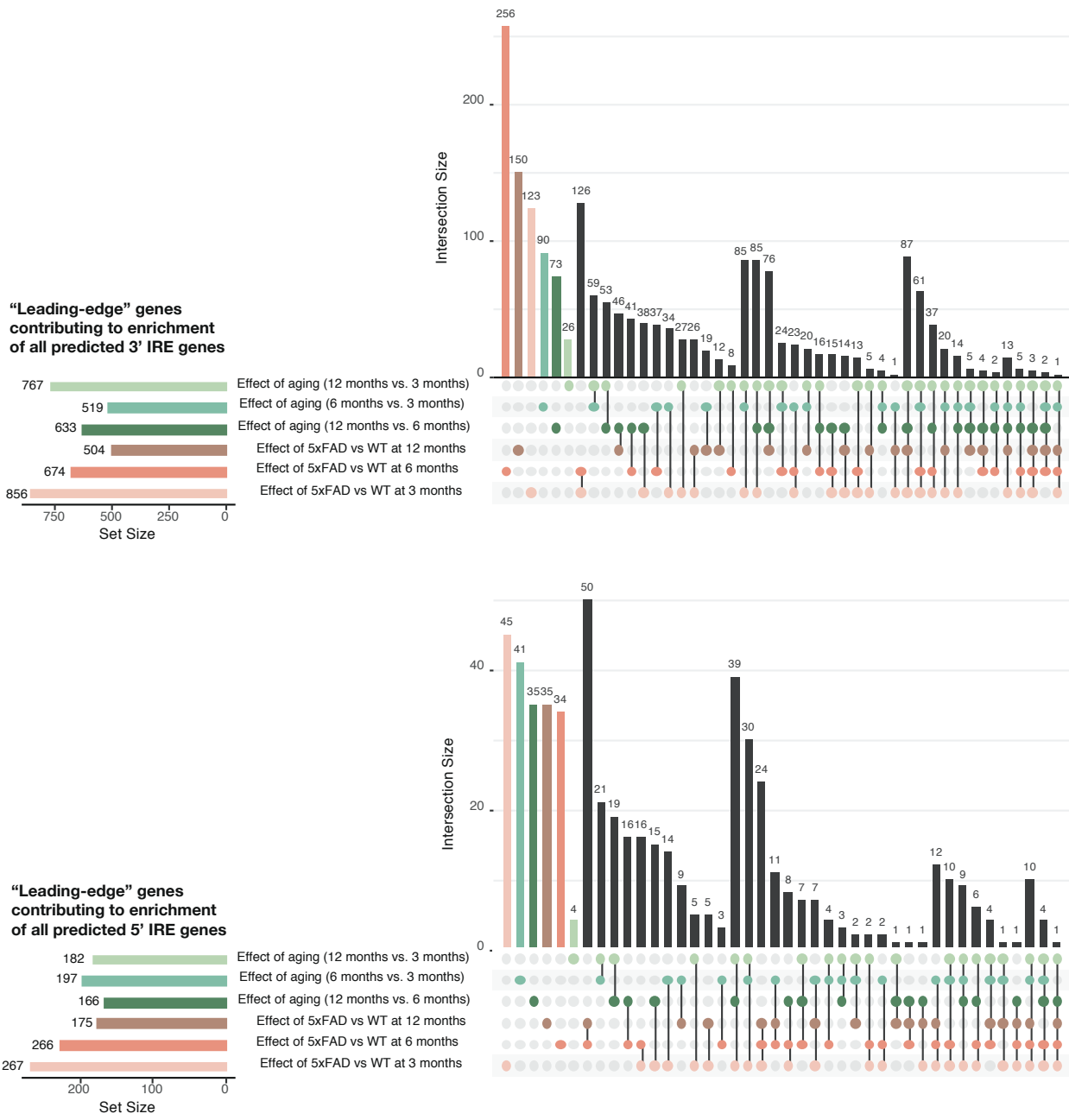
Figure 8. UpSet plots showing the overlap in GSEA leading-edge genes between all comparisons in the 5XFAD mouse datasets for the gene sets *all 3' IREs* and *all 5' IREs*. Numbers of genes for each intersection are shown above intersection bars.
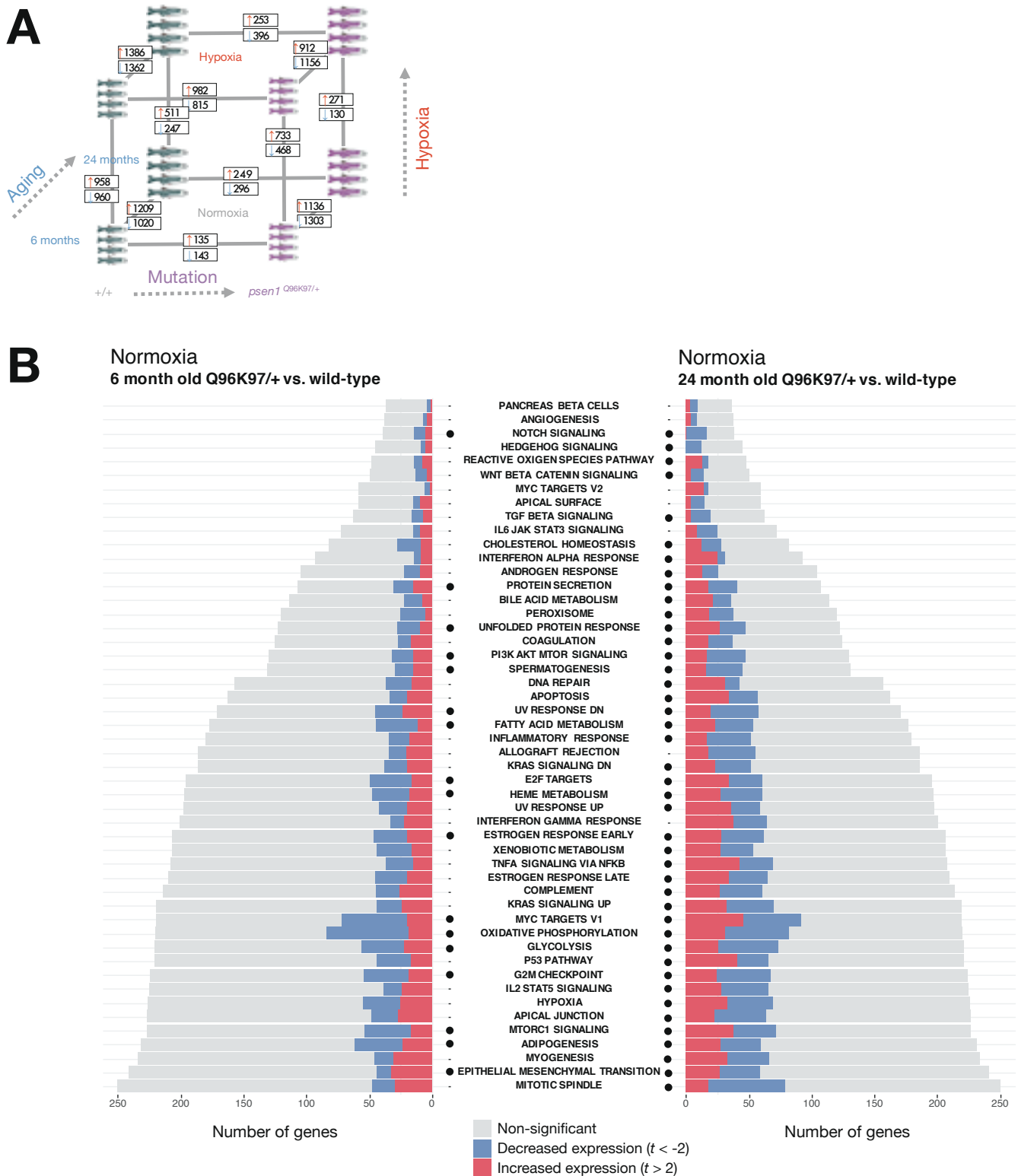
Figure 9. Differential gene expression and gene set enrichment analysis in the fAD-like zebrafish dataset. A. Results of differential gene expression analysis. Genes which were significantly increased or decreased in expression are indicated in boxes. These differentially expressed genes have FDR-adjusted p-value < 0.05. **B. Gene set enrichment with MSigDB Hallmark gene sets. The comparisons between Q96_K97del/+ fAD-like mutants and their wild-type siblings are shown for the 6-month-old (young adult) and 24-month-old (infertile adult) age groups.** Dots indicate gene sets which are significantly enriched (FDR-adjusted p-value < 0.05).
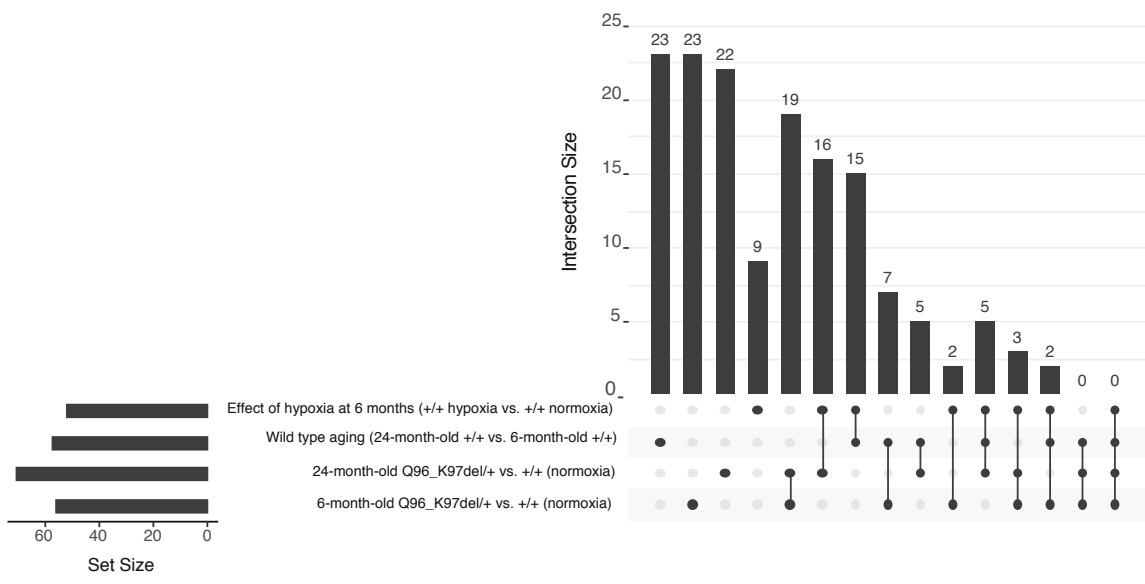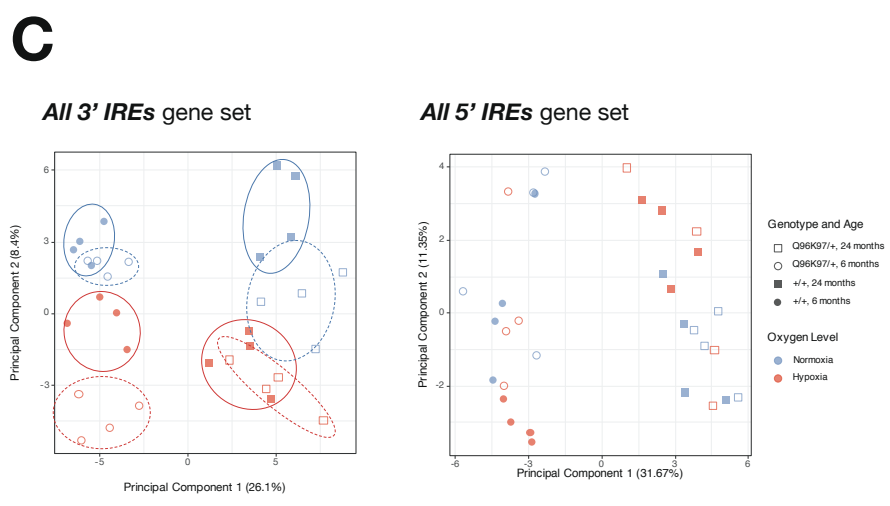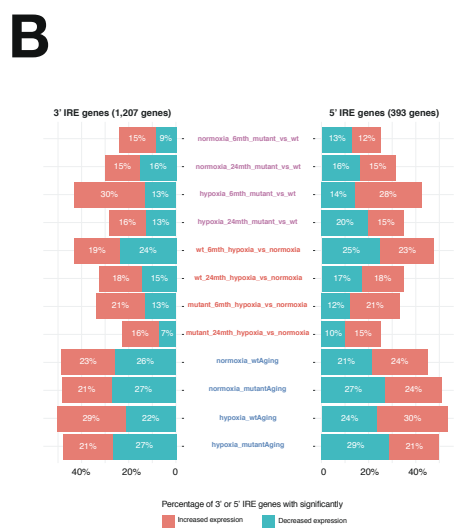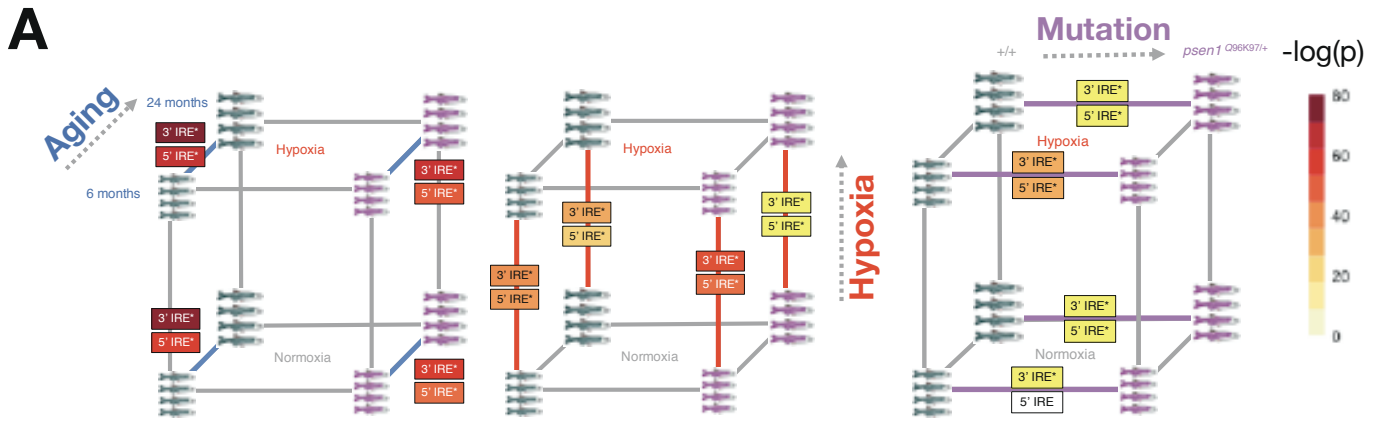
Figure 10. Iron Responsive Element (IRE)-containing gene expression in the fAD-like zebrafish dataset. A. Results of gene set enrichment testing using predicted IRE gene sets. We represent the gene sets *all 3' IREs* and *all 5' IREs* derived from searching for IRE and IRE-like sequences from z11 reference zebrafish gene UTRs as "3' IRE" and "5' IRE" in the panel. **B. Proportions of predicted IRE genes which are increased (t > 2) or decreased (t < -2) in expression for each pairwise comparison in the dataset. C. Principal component analysis of all genes in the sets *all 3' IREs* and *all 5' IREs* for all samples.** Circles on the 3' IRE plot show that different conditions generally have distinct expression of genes in the *all 3' IREs* set but not in the *all 5' IREs* set. **D. UpSet plot showing overlap in leading-edge genes for the "all predicted 3' IRE genes" for select comparisons.**

149

Supplementary Information Legends

Note: Supplementary Tables and Text files are hosted on Figshare

(https://figshare.com/s/104097febc1b8970fc5d).

Supplementary Table 1. Iron Responsive Element (IRE) gene sets defined from human, mouse, and zebrafish transcriptomes. The untranslated regions (UTR) of all known genes were searched for IRE and IRE-like motifs. Four gene sets are included for each species: ***all 3' IREs*** (all genes with predicted IRE in their 3' UTR), ***all 5' IREs*** (all genes with predicted IRE in their 5' UTR), ***HQ 3' IREs*** (genes with high-quality predicted IRE in 3' UTR), and ***HQ 5' IREs*** (genes with high-quality predicted IRE in 5' UTR).

Supplementary Table 2. MSigDB gene sets with significant over-representation of IRE gene sets in human, mouse, and zebrafish. Gene sets were defined to have significant over-representation of IRE gene sets if the FDR-adjusted Fisher's exact test p-value was < 0.05.

Supplementary Table 3. Promoter motif over-representation analysis results for all IRE gene sets in human, mouse, and zebrafish. Promoter regions were defined as being 1500 bp upstream and 500 bp downstream of genes. See methods for details. Significant over-representation of a promoter occurred when the FDR-adjusted enrichment p-value < 0.05.

Supplementary Text 1. Identifying predicted 3' IRE genes with altered stability.

Supplementary Text 2. Additional discussion.

Supplementary Text 3. Proportions of neural cell type markers (in astrocytes, neurons, oligodendrocytes, and microglia) in datasets analysed.

Supplementary Figure 1. Principal Component Analysis of results from different gene set testing methods.
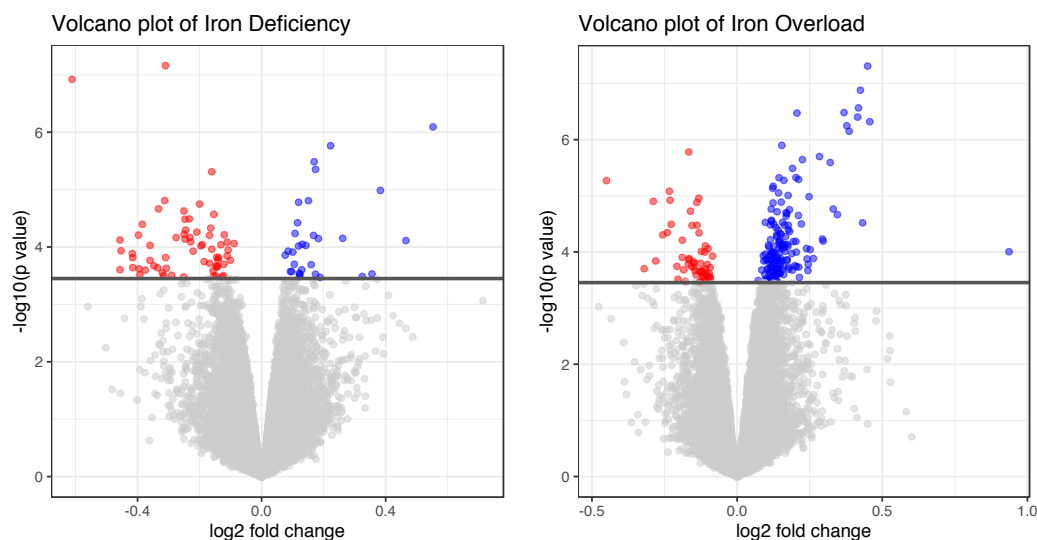

Supplementary Figure 2. Analysis of cultured Caco-2 cell line dataset. A. Principal Component Analysis plot of gene expression in the cultured Caco-2 cell line dataset. **B.** Volcano plots indicating differential gene expression due to iron overload (DMEM-FBS Medium + Hemin vs. DMEM-FBS Medium) and iron deficiency (Iron-free medium vs. Iron-free medium + FAC) treatments in the Caco-2 cell line dataset.

Supplementary Figure 3. Gene set enrichment testing results for the AD vs. control comparison in cerebellum and temporal cortex.

Supplementary Figure 4. STRINGR protein-protein interaction network plot between the 19 shared leading-edge genes between the psen1Q96_K97del/+ and psen1+/+ comparisons at 6 and 24-months-old.

Supplementary Figure 5. Principal Component Analysis plots demonstrating the minimal association between IRE gene expression and the *psen1*[K97fs/+] mutant genotype.

Supplementary Figure 6. Predicted 3' IRE containing transcripts with significant differences in stability between conditions in the fAD-like zebrafish dataset.

Supplementary Figure 7. Age-dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in human Mayo Clinic RNA-seq dataset analysed.

Supplementary Figure 8. Age-dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in 5XFAD mouse cortex RNA-seq dataset.

Supplementary Figure 9. Age- and hypoxia- dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in fAD-like zebrafish whole-brain RNA-seq dataset analysed.
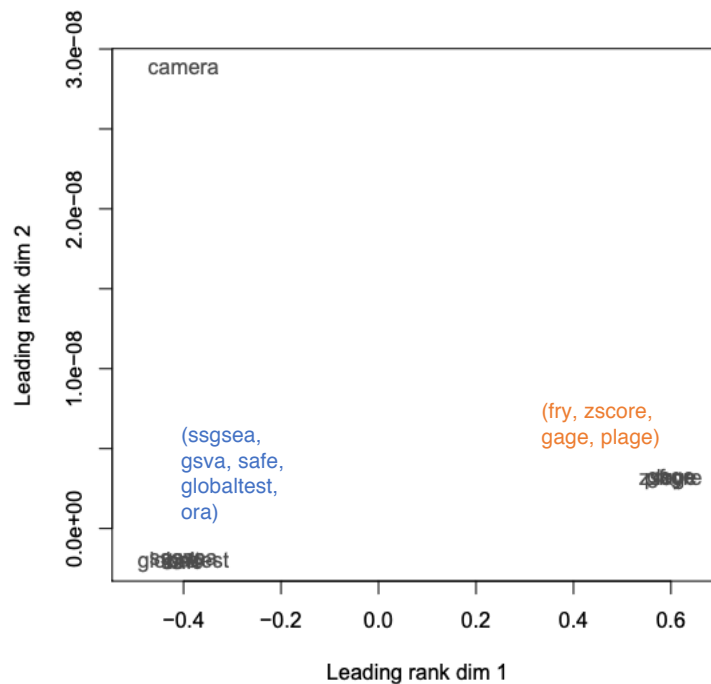
Principal Component Analysis of results from different gene set testing methods.

We used EGSEA (v.1.10.1) with default settings with R objects (design, contrasts, voom) from *limma* analysis of the fAD-like zebrafish dataset (n=32). EGSEA was run with the following methods: camera (limma v.3.38.3), safe (safe v.3.22.0), gage (gage v.2.32.1), plage (GSVA v.1.30.0), zscore (GSVA v.1.30.0), gsva (GSVA:1.30.0), ssgsea (GSVA v.1.30.0), globaltest (globaltest v.5.36.0), ora (stats v.3.5.2), fry (limma v.3.38.3). The Principal Component Analysis plot shows the relative similarity of the results obtained from running different methods, overall revealing three main groupings. To minimise the chance of *p*-value inflation from combining multiple methods that give essentially the same results, we decided to use one representative method from each group in our final analyses: camera, fry, and fgsea. Although fgsea is not a method included in EGSEA, we chose to include it as it implements the classic GSEA algorithm and also includes information about leading-edge genes which we make use of in our analyses.

A

B

A. **Principal Component Analysis plot of gene expression in the cultured Caco-2 cell line dataset.** The plot uses pre-processed and normalised expression values for 22,153 probesets across all samples.

B. **Volcano plots indicating differential gene expression due to iron overload (DMEM-FBS Medium + Hemin vs. DMEM-FBS Medium) and iron deficiency (Iron-free medium vs. Iron-free medium + FAC) treatments in the Caco-2 cell line dataset.** Horizontal lines indicate an FDR-adjusted *p*-value of 0.05, with all genes with FDR-adjusted *p*-value < 0.05 considered significantly differentially expressed (DE). In the iron deficiency treatment, this resulted in 96 significantly DE genes (65 down, 31 up), while in the iron overload treatment there were 212 significantly DE genes (55 down, 157 up).

Supplementary Figure 3

A
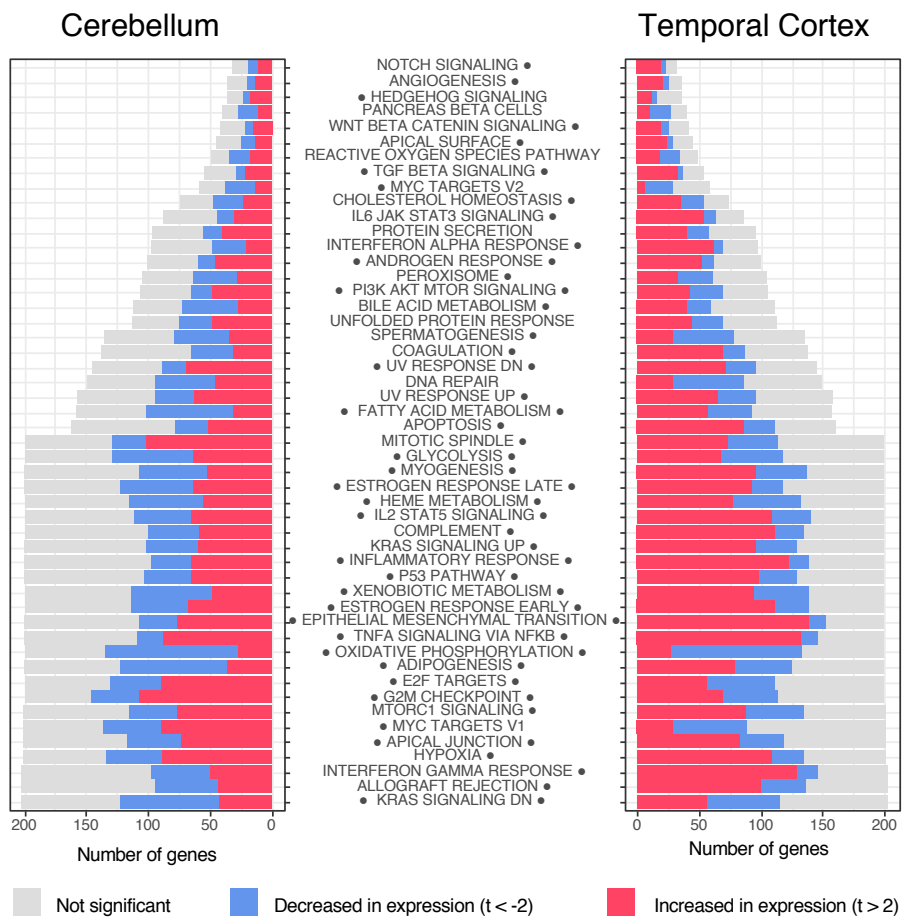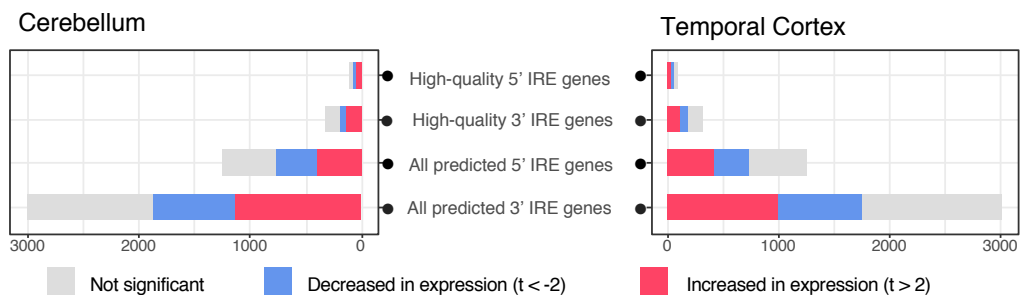
Gene set enrichment testing results for the AD vs. control comparisons in cerebellum and temporal cortex.

A. Enrichment results for MSigDB Hallmark gene sets
B. Enrichment results for human IRE gene sets
Dots to either the left or right side of the gene set name indicate that the gene set is significantly enriched in either the "AD vs. control with cerebellum tissue" or "AD vs. control with temporal cortex tissue" comparisons respectively (Bonferroni-adjusted *p*-value < 0.05).
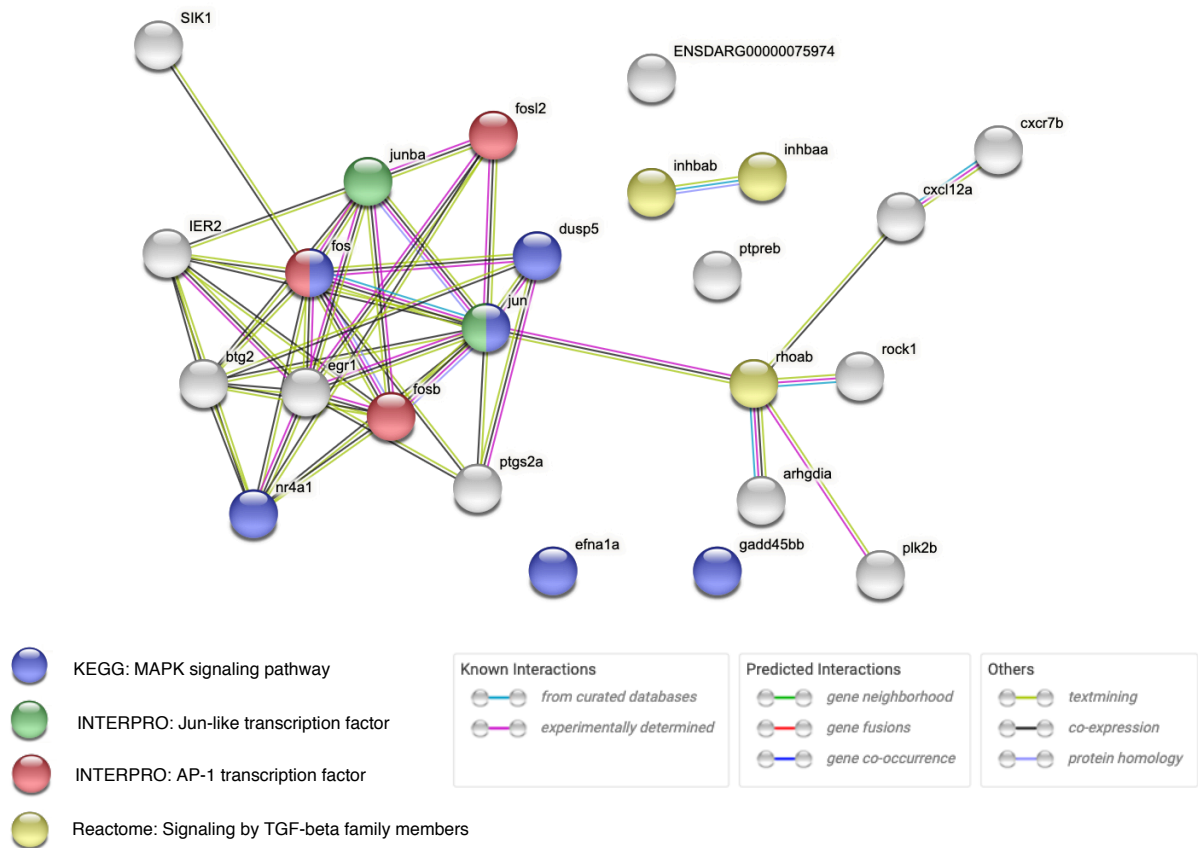
STRINGR protein-protein interaction network plot between the 19 shared leading-edge genes between the *psen1*^Q96_K97del/+ and *psen1*^+/+ comparisons at 6 and 24-months-old. Nodes are named with gene symbols and edges indicate a known or predicted protein-protein interaction. Coloured nodes indicate genes contributing to significant over-representation of particular gene sets (FDR-adjusted *p*-value < 0.05).
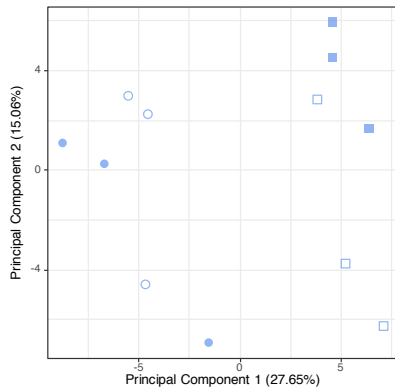
A 3' IRE genes

5' IRE genes
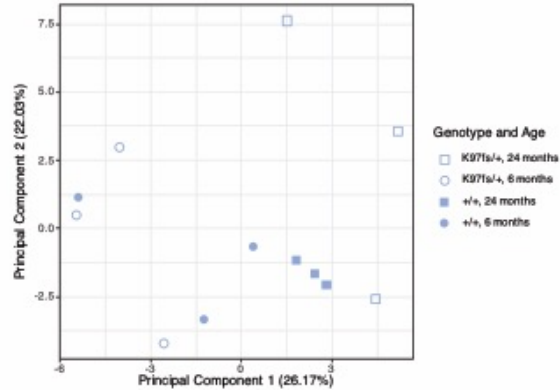
B All genes (18,296 genes)
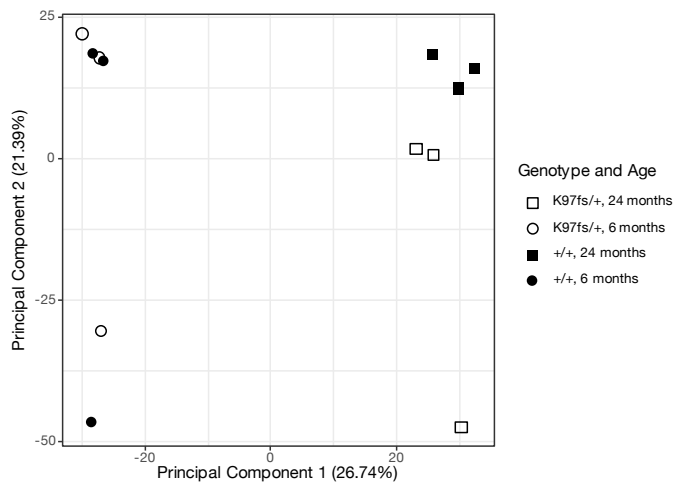
Principal Component Analysis plots demonstrating the minimal association between IRE gene expression and the *psen1*$^{K97fs/+}$ mutant genotype. A. Principal Component Analysis plots using expression of the zebrafish "all predicted 3' IRE genes" and "all predicted 5' IRE genes" gene sets. B. Principal Component Analysis plot of all genes detected in the dataset. Unlike for the fAD-like *psen1*$^{Q96\_K97del/+}$ mutant, the non fAD-like *psen1*$^{K97fs/+}$ mutant does not show significant changes in gene expression of 3' and 5' IRE genes.

6-month-old Q96_K97del/+ vs. +/+ (under normoxia)
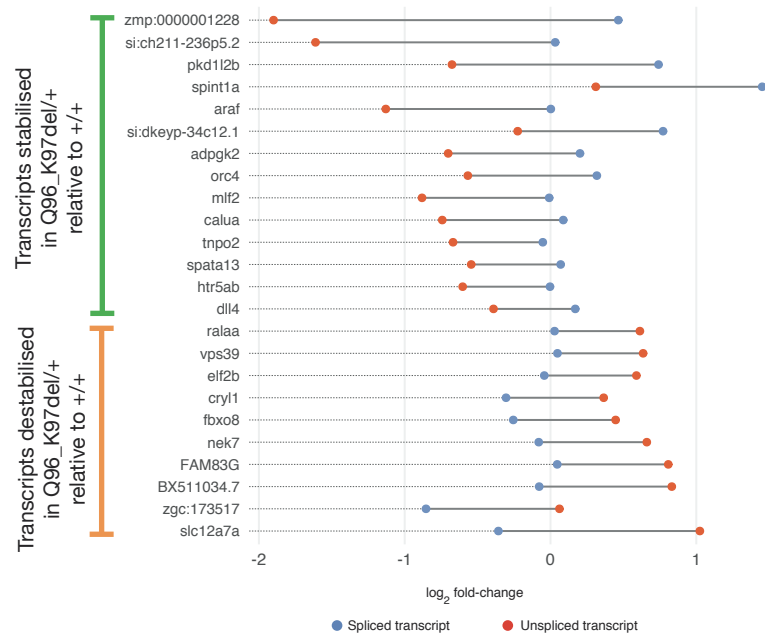
24-month-old Q96_K97del/+ vs. +/+ (under normoxia)
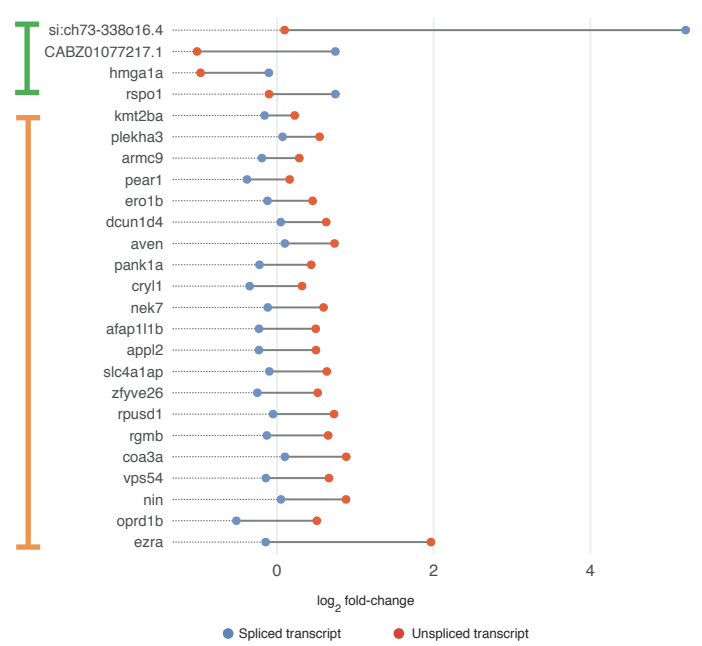
Response to hypoxia at 6-months (6-month-old +/+ under hypoxia vs. 6-month-old +/+ under normoxia)
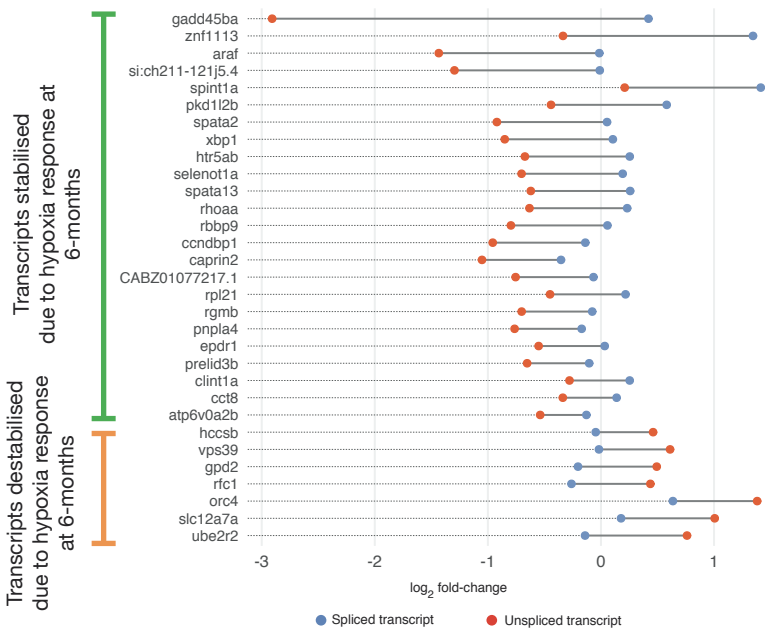
**Wild type aging (24-month-old +/+ vs. 6-month-old +/+)**
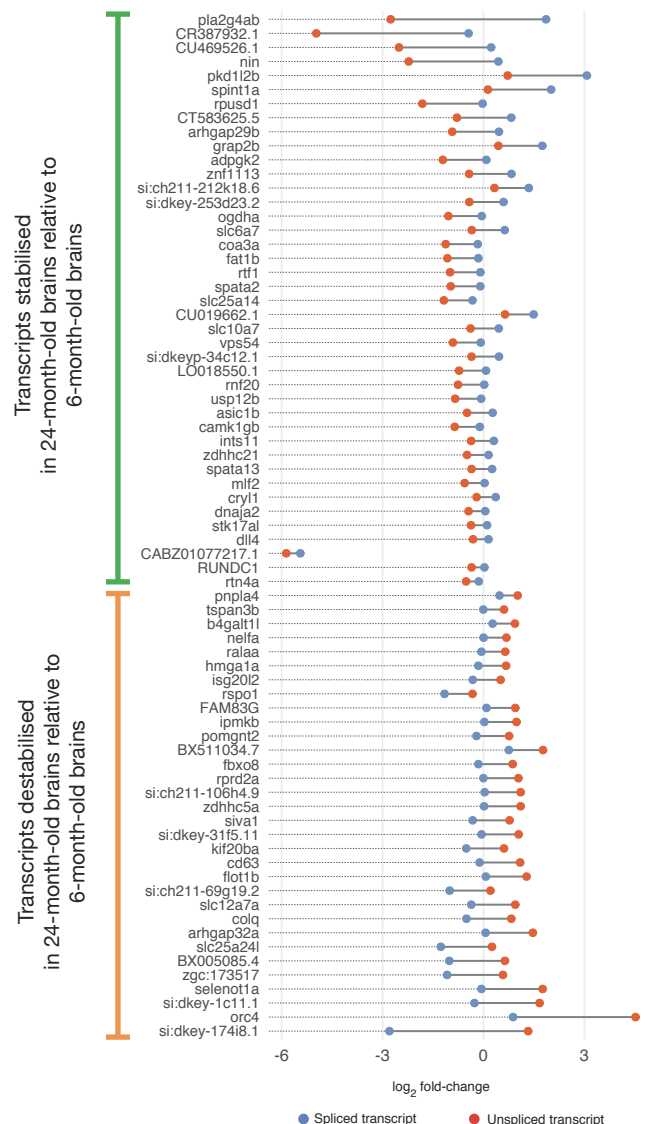
Supplementary Figure 6. Predicted 3' IRE-containing transcripts with significant differences in stability between conditions in the fAD-like zebrafish dataset. Transcripts are defined to be stabilised if the log2 fold-change of the spliced transcript is significantly greater than the log2 fold-change of the unspliced transcript. Likewise, transcripts are defined to be destabilised if the log2 fold-change of the spliced transcript is significantly less than the log2 fold-change of the unspliced transcript. All tests were done using Welch's t-test with FDR-adjusted p-value < 0.05.
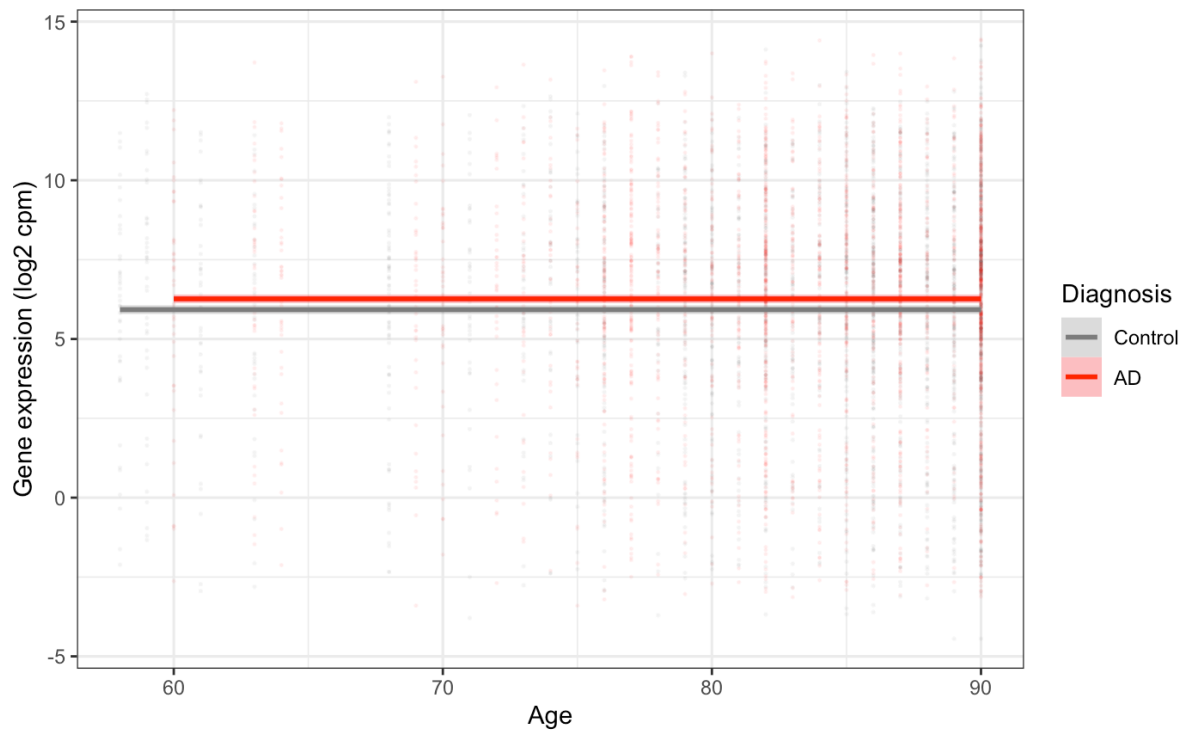
Supplementary Figure 7.

Age-dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in human Mayo Clinic RNA-seq dataset analysed.
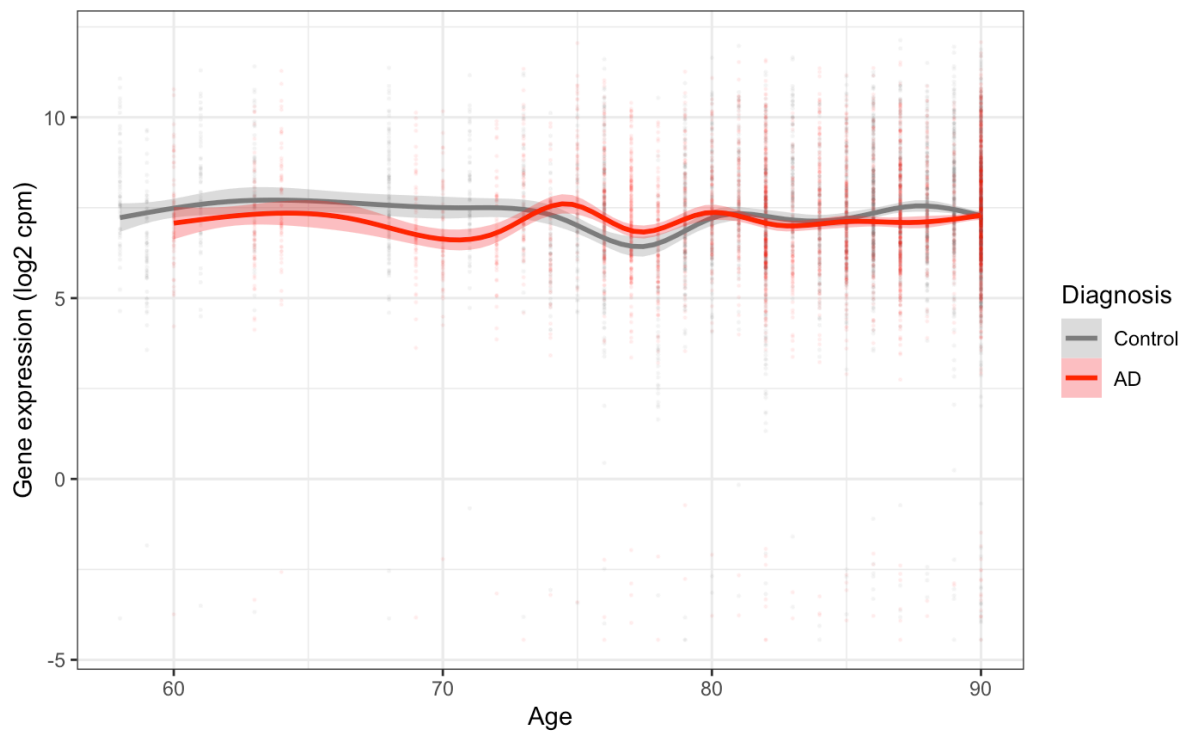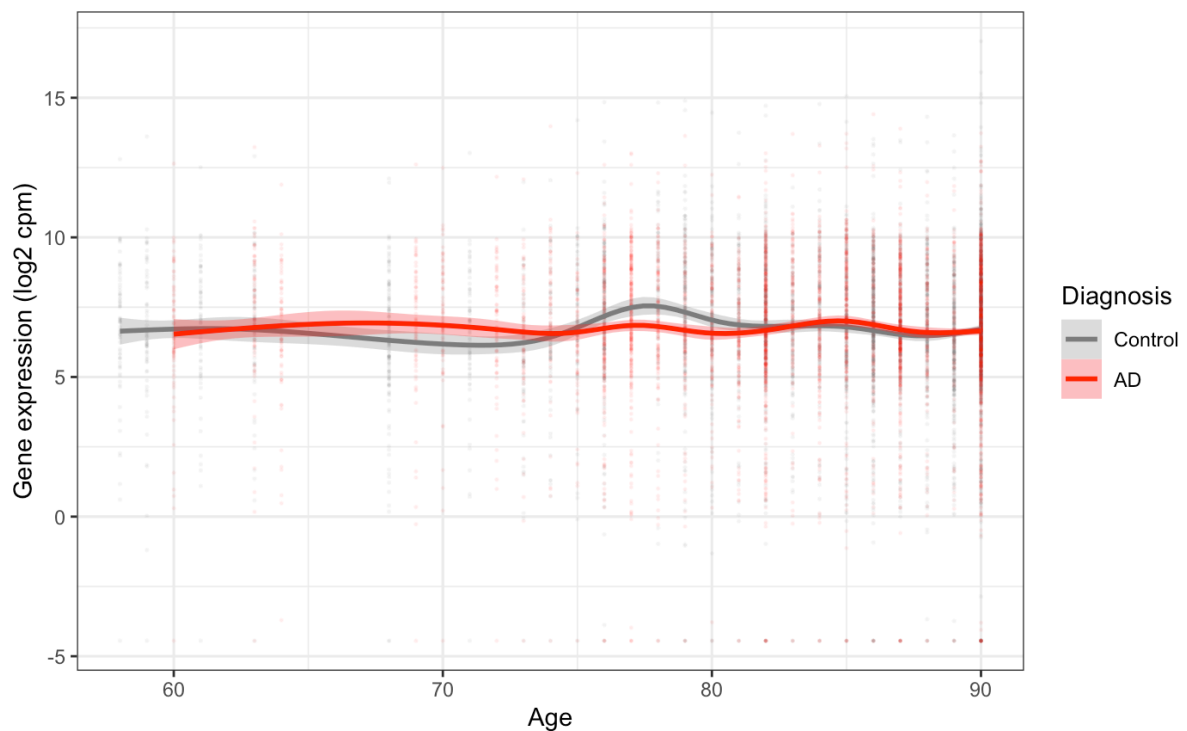
Microglia marker gene expression (20 genes)
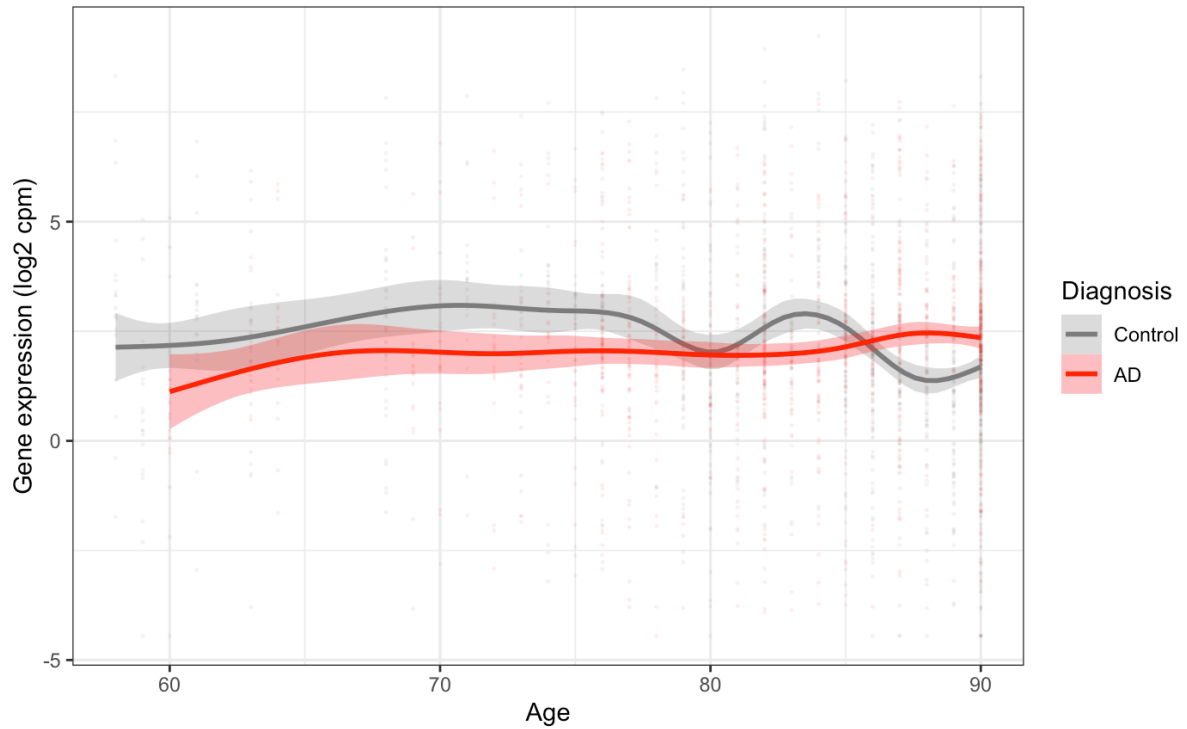
Astrocyte marker gene expression (44 genes)

Neuron marker gene expression (74 genes)

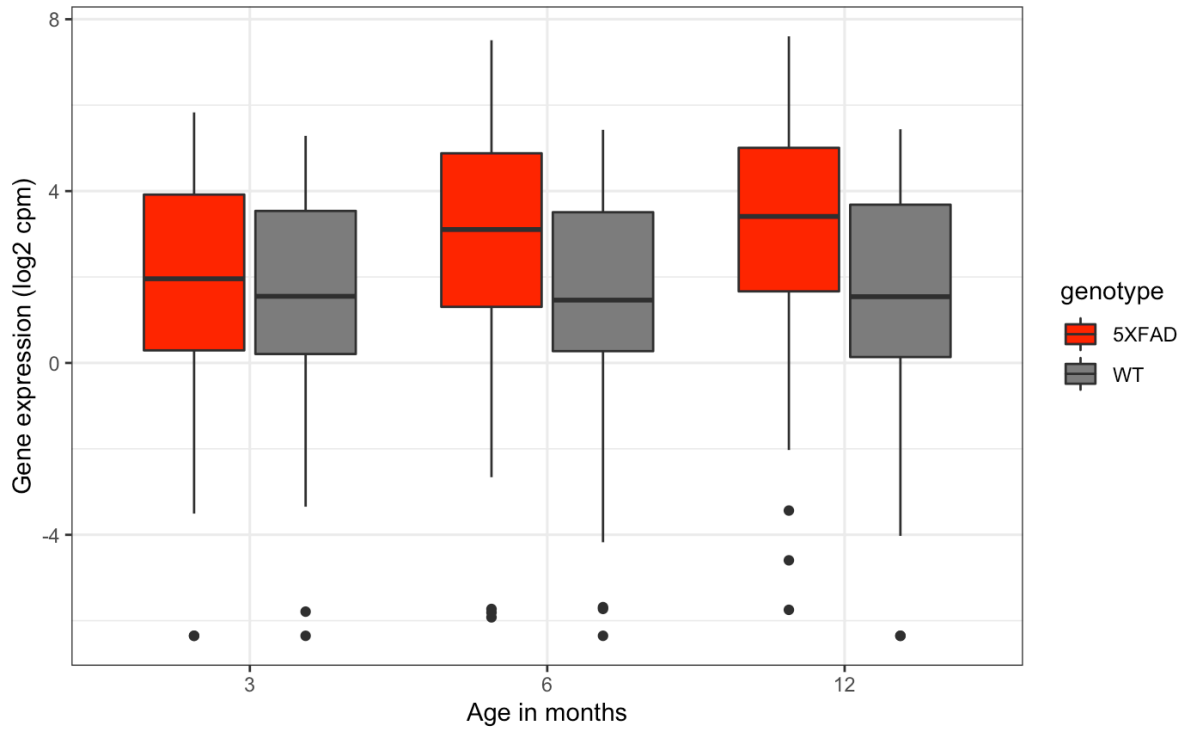## Oligodendrocyte marker gene expression (83 genes)

Error bars represent the 95% confidence interval. Overall, while we see differences in expression, there does not appear to be an overall systematic difference between AD vs. control for oligodendrocyte, neuron and microglial marker gene expression. For the astrocyte markers, AD brains appear to systemically have higher expression of these genes.
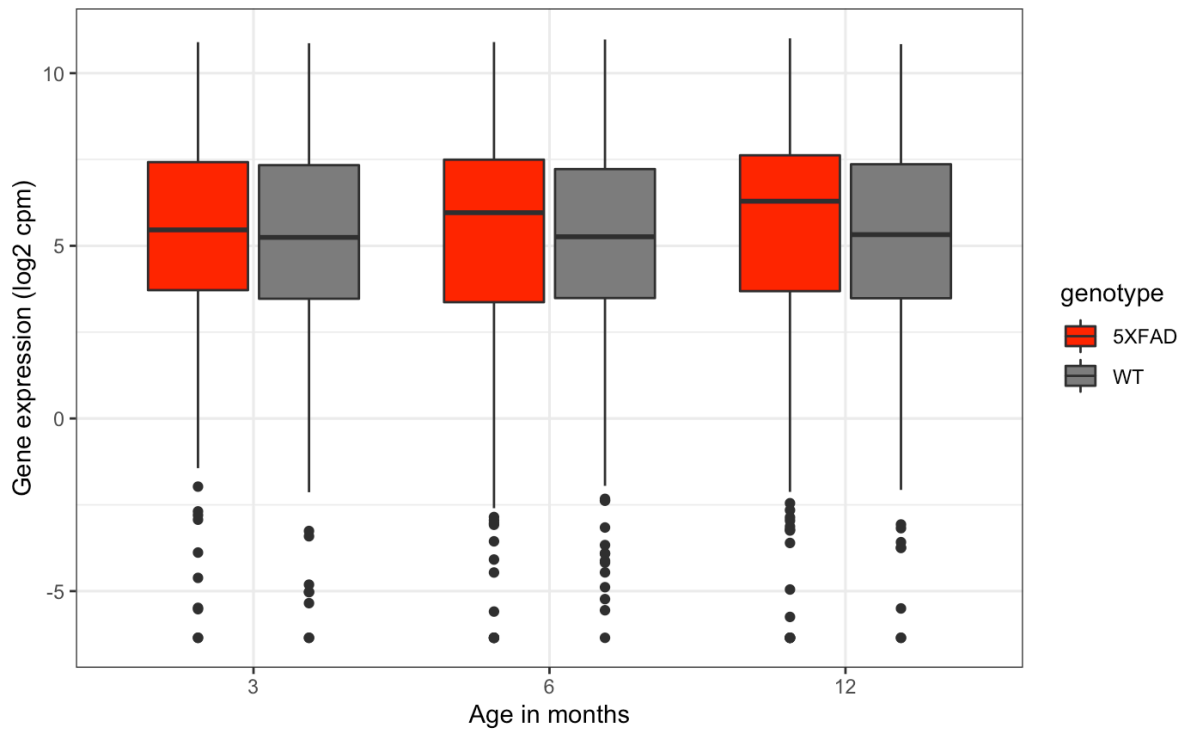
Age-dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in 5XFAD mouse cortex RNA-seq dataset.

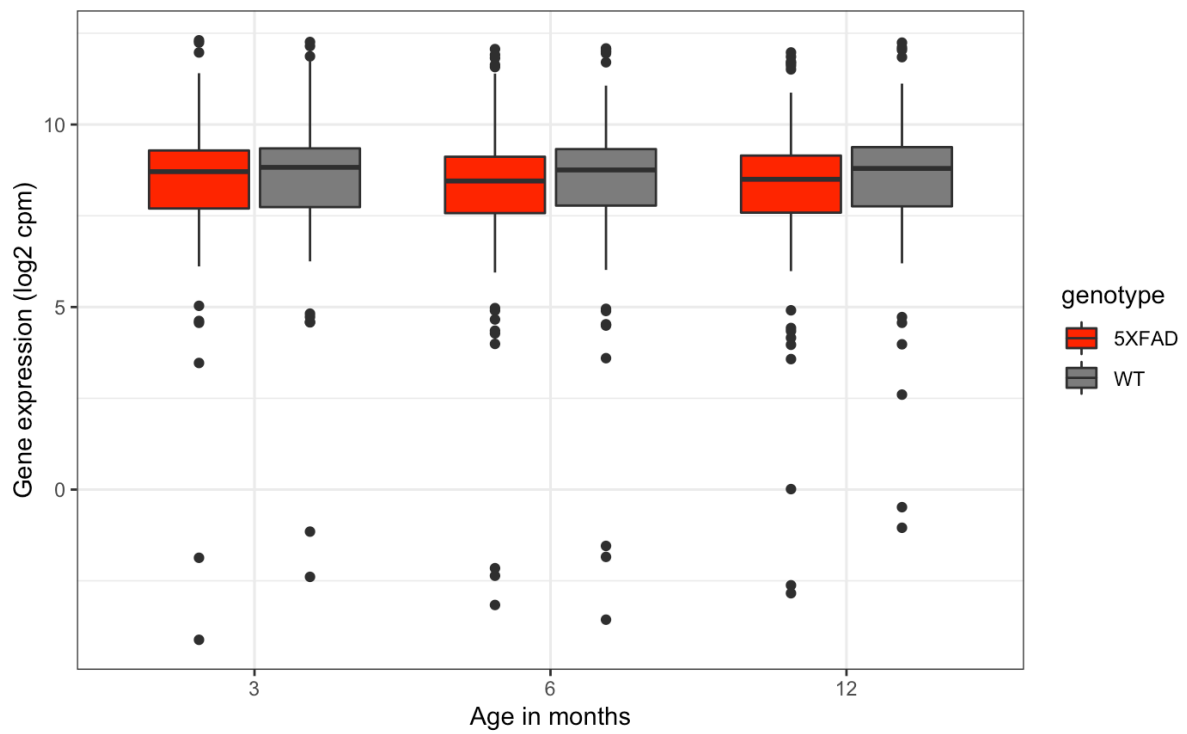Microglial marker gene expression (22 genes)

Astrocyte marker gene expression (51 genes)

Neuron marker gene expression (67 genes)

Oligodendrocyte marker gene expression (77 genes)

Supplementary Figure 9.

Age- and hypoxia- dependent expression of neural marker genes (microglia, astrocyte, neuron, oligodendrocyte) in fAD-like zebrafish whole-brain RNA-seq dataset analysed.
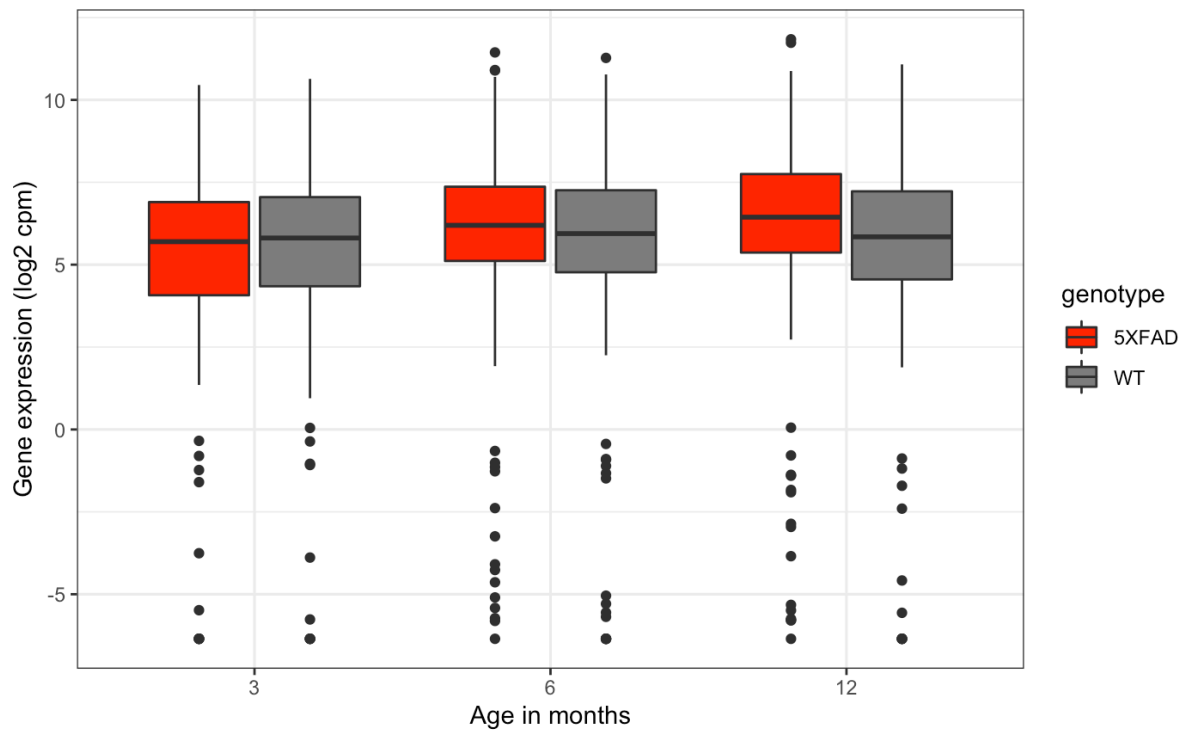
Microglial marker gene expression (16 genes)

Astrocyte marker gene expression (61 genes)

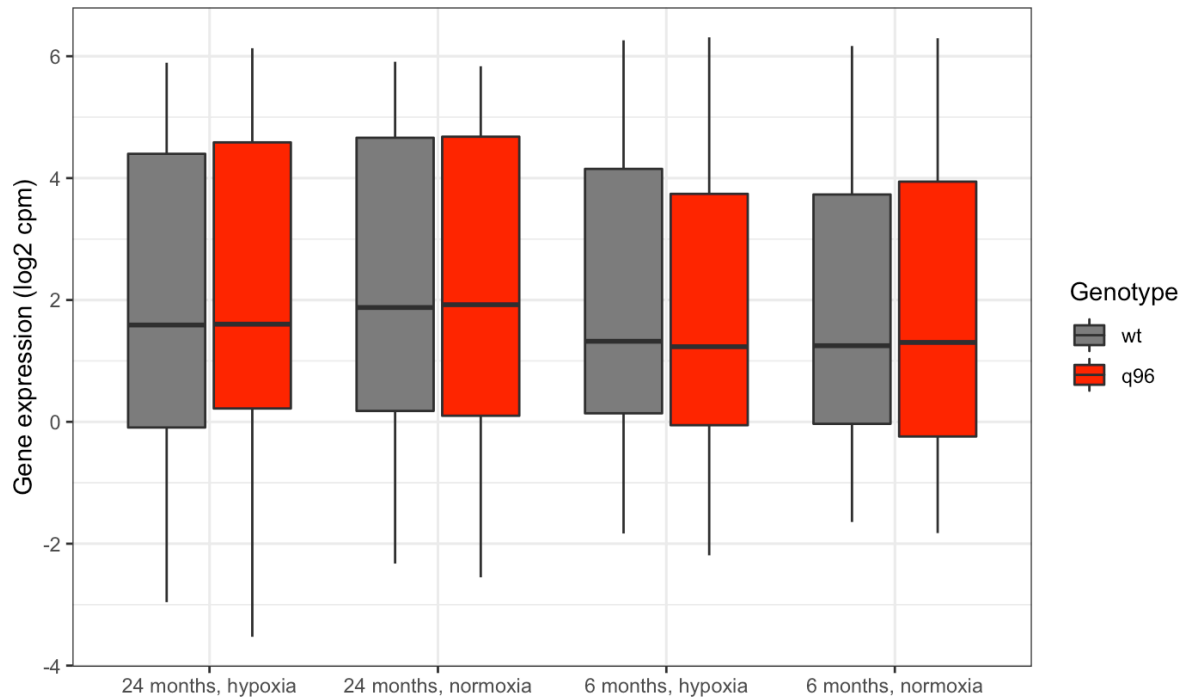## Neuron marker gene expression (87 genes)
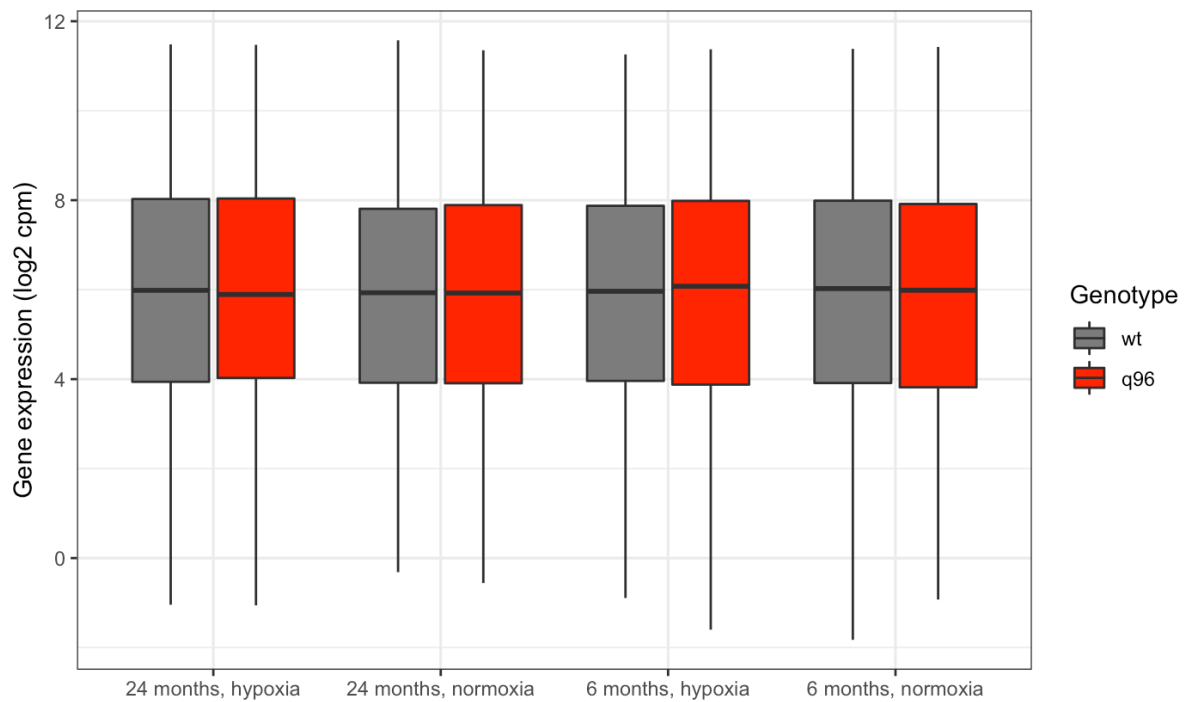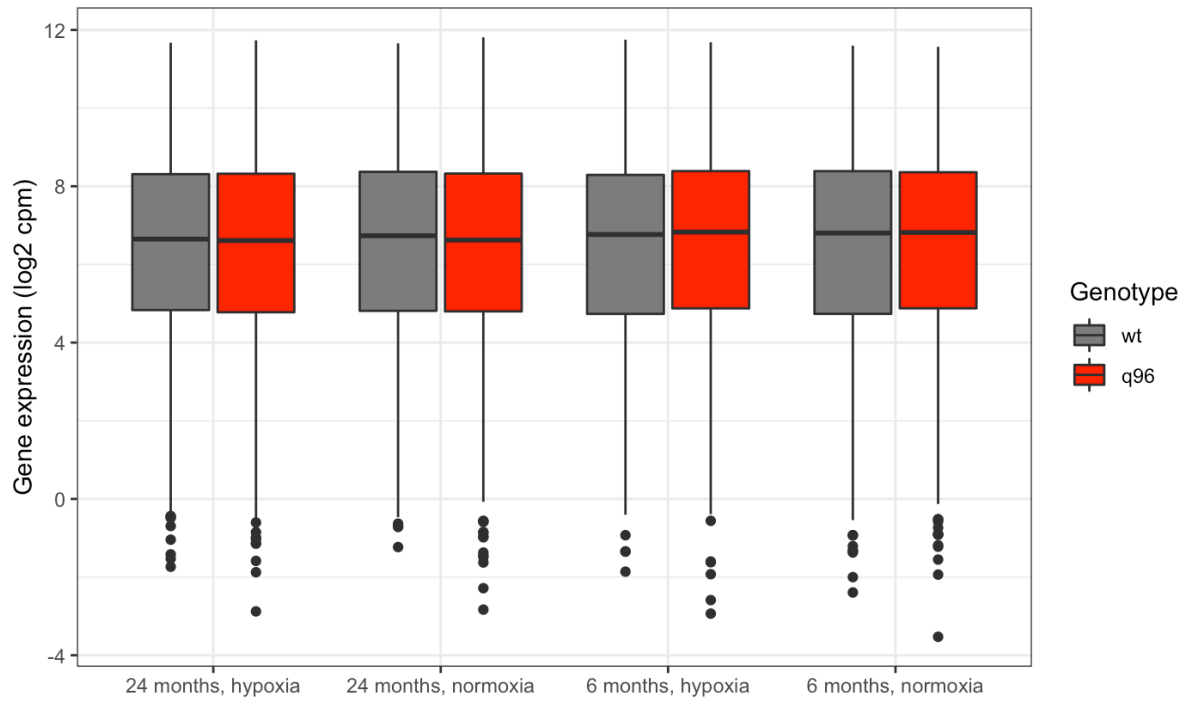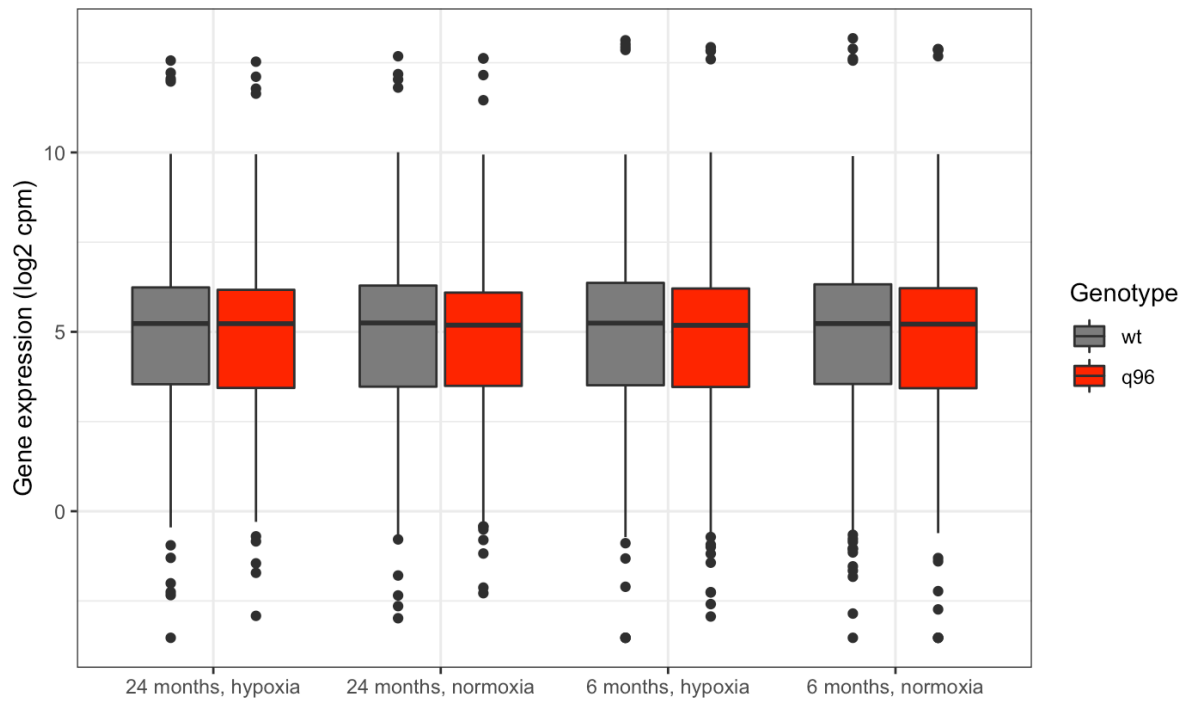

Oligodendrocyte marker gene expression (107 genes)

Chapter 5

Adaptive elastic-net sparse PCA for robust cross-species, cross-platform analysis of complex gene expression data in Alzheimer's disease

Statement of Authorship

Title of Paper	Adaptive elastic-net sparse PCA for robust cross-species, cross-platform analysis of complex gene expression data in Alzheimer's disease.
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Hin, N., Newman, M., Pederson, S. and Lardelli, M., 2020. Adaptive elastic-net sparse PCA for robust cross-species, cross-platform analysis of complex gene expression data in Alzheimer's disease.

Principal Author

Name of Principal Author (Candidate)	Nhi Hin
Contribution to the Paper	- All bioinformatics analysis - Drafting of manuscript - Corresponding author on manuscript
Overall percentage (%)	95%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 08/09/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

  i.   the candidate's stated contribution to the publication is accurate (as detailed above);

  ii. permission is granted for the candidate in include the publication in the thesis; and

  iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Morgan Newman
Contribution to the Paper	- Genome editing and creation of $psen1^{Q96K97del}$ zebrafish mutant line - Removal of brains from fish and purification of mRNA for RNA-seq - Editing of manuscript drafts
Signature	Date 10/11/2020

Name of Co-Author	Stephen Pederson
Contribution to the Paper	- Supervision of bioinformatics work and statistical methodology - Editing of manuscript drafts
Signature	Date

Name of Co-Author	Michael Lardelli
Contribution to the Paper	- Project supervision - Editing of manuscript drafts
Signature	Date 9/9/20

Adaptive elastic-net sparse PCA for robust cross-species, cross-platform analysis of complex gene expression data in Alzheimer's disease

Author Details:

Nhi Hin*[1,2], Morgan Newman[2], Stephen Pederson[1], Michael Lardelli*[2]

1. Bioinformatics Hub, School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia
2. Alzheimer's Disease Genetics Laboratory, School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia

* corresponding authors

Author Full Names and Email Addresses:

Nhi Hin: nhi.hin@adelaide.edu.au

Morgan Newman: morgan.newman@adelaide.edu.au

Stephen Pederson: stephen.pederson@adelaide.edu.au

Michael Lardelli: michael.lardelli@adelaide.edu.au

Abstract

Characterisation of the global transcriptional states underlying Alzheimer's disease (AD) is an ongoing effort, challenged by the complexity and heterogeneity of the disease as well as access to samples. Many publicly available gene expression datasets representing different animal models of AD and human cohorts are not immediately applicable to general research due to technical and biological batch effects. Here, we describe the application of a advanced dimensionality reduction technique, Adaptive, Elastic-Net Principal Component Analysis (AES-PCA) on the integrative analysis of several animal model and human familial and sporadic AD datasets. Notably, we find that latent variables captured by AES-PCA and the Hallmark gene set collection effectively reduce noise within datasets and batch effects between datasets, thus enabling (a) interpretable visualisation of key differences underlying the early pathology of two different animal models of AD (*psen1*$^{Q96\_K97del/+}$ fAD mutation-like zebrafish and 5XFAD mice), (b) cross-species comparison of alterations in these animal models to a human familial AD dataset, and (c) integrative analysis of human familial and sporadic AD in a biologically interpretable manner despite major platform-specific differences.

Introduction

Conventional approaches to gene expression analysis have focused on identifying marker (differentially expressed) genes to gain insight into the molecular biology and gene expression changes underlying a particular biological condition (reviewed in [1, 2]). This has posed difficult in the context of Alzheimer's disease (AD), where typically, human brain tissue available for analysis is derived from post-mortem brains displaying significant pathology and alteration of diverse biological activities [3–6]. Fortunately, utilising animal models of AD allows for the extraction of brain tissue at a younger age, which in turn permits characterisation of early AD pathology. Many such studies have been performed on various animal models of AD, particularly humanised transgenic models with multiple fAD-linked mutations such as the 5XFAD model [7–9] (reviewed in [10]) and knock-in models that introduce fAD-like mutations into endogenous genes [11–18]. However, detailed analyses comparing gene expression patterns between different animal models of AD to each other and to the human disease are rare. Largely, this is due to the presence of noise within datasets and batch effects (e.g. platform, tissue, library preparation) between datasets (reviewed in [19]). This has meant that a large amount of publicly available gene expression data from many different animal models of AD and human AD cohorts is not immediately usable for integrative and comparative analyses.

Previous research has addressed the difficulty behind integration and comparison of different gene expression datasets through several approaches, including meta-analysis (e.g. [20]), matrix factorisation (reviewed in [21]), as well as methods that attempt to compare overall systematic changes in the biology of a system at the

gene set or pathway level (e.g. GSEA [22] and PLAGE [23]). Recently, *pathwayPCA* was released which provides an R/Bioconductor interface to the Adaptive, Elastic-net Sparse Principal Component Analysis (AES-PCA) approach [24, 25]. AES-PCA is an unsupervised method for dimension reduction developed for preserving biologically correlated information in gene expression data while reducing noise. Similar to other dimension reduction methods (e.g. non-negative matrix factorisation, PCA) it is able to summarise a gene expression dataset into latent variables that can then be used to calculate significant associations with experimental conditions/groups, similar to the purposes of using conventional gene set enrichment analysis (GSEA) approaches. However, as AES-PCA is specifically designed for use with predefined gene sets with biological relevance, the latent variables themselves are also amenable to a variety of further analyses including clustering of samples and integration of different datasets at both the gene- and gene set levels.

Here, we present applications of AES-PCA to address currently unanswered research questions relevant to gene expression changes underlying Alzheimer's disease, including:

1. How do animal models of AD differ in the gene expression patterns underlying the earliest detectable changes?

2. How can we make preliminary comparisons between the overall biological activities that are altered in animal models vs. familial AD in humans?

3. For large, complex and heterogeneous datasets (e.g. sAD cohort), how can we find biologically meaningful and interpretable clusters

Importantly, we illustrate the utility of AES-PCA for the integration and comparison of different datasets by (a) reducing noise and emphasising relevant biological activities correlated to particular conditions within datasets, (b) enabling cross-species analysis and comparison of different animal models and human AD, and (c) improving integration of different datasets through enabling cross-platform analysis in a biologically interpretable manner.

Results

Overview of method

We consider a gene expression matrix, *M*, consisting of *n* samples arranged as columns and *p* genes arranged as rows, which we use to characterise transcriptional states of interest. When applying the unsupervised dimension reduction method known as AES-PCA [25], the first step is to map the *p* genes to pre-defined gene sets *s*, such as those provided by external databases such as MSigDB [26], KEGG [27] or WikiPathways [28]. This is followed by selecting a coherent subset of genes for each predefined gene set using the elastic-net estimator [29], which combines both L1 (absolute values) and L2 (quadratic) penalties into the model. The adaptive lasso component (L1) of the AES-PCA enables a consistent subset of genes to be chosen for each dataset via the shrinkage properties of the lasso. The latent variables (i.e. Principal Components) extracted by AES-PCA represent activities within each gene set (i.e. overall up-regulation or down-regulation of genes within these gene sets). As these latent variables are calculated for each sample, they represent sample-specific estimates of gene set activities, and their sparse loadings

allow for the extraction of relevant genes contributing to variation in gene set activity across samples. In addition, latent variables can be used to calculate significant associations with experimental conditions/groups similar to the purposes of using conventional gene set enrichment testing approaches (e.g. GSEA), through using a linear regression model or permutation test [24, 25].

In our analyses below, we use AES-PCA with the Hallmark gene set collection from the Molecular Signatures Database (MSigDB). The Hallmark gene set collection comprises 50 gene sets representing diverse and well-characterised biological activities (e.g. "oxidative phosphorylation", "inflammation"), each containing ~200 non-redundant genes, based on expert manual curation from multiple studies and existing gene sets [30]. The AES-PCA approach we use in our analyses is summarised in **Figure 1** and more detail can be found in [24, 25, 30].

Gene expression data can be summarised using the Hallmark gene set collection and AES-PCA to provide an interpretable representation of altered biological activities

The first step of the AES-PCA approach involves summarisation of gene expression data into gene set activities. Because we use the Hallmark gene set collection throughout our analyses, we first demonstrate that the biological activities represented within the Hallmark gene set collection are sufficient to distinguish subtle, early changes within young adult brains of two different animal models of AD

(transgenic 5XFAD mouse model and *psen1*[Q96_K97del/+] fAD-like zebrafish model) in a biologically-interpretable manner.

The transgenic 5XFAD mouse model is a popular animal model of AD, expressing human *APP* and *PSEN1* transgenes with a total of 5 fAD-linked mutations. It is characterised by extensive and rapid development of amyloid and tau brain pathology with age [7, 8, 31]. In contrast, the *psen1*[Q96_K97del/+] zebrafish model is a knock-in model that possesses a single fAD-like mutation in a heterozygous state [13]. Knock-in models with similar fAD-like genetic backgrounds typically do not show amyloid and tau brain pathology with aging [14, 32]. The clearly different genetic backgrounds of these two animal models would be expected to result in different changes to their brains at the molecular level. For example, a previous transcriptome comparison of different transgenic mouse models such as 5XFAD and 3xTg showed extensive differences between these models [20]. However, to our knowledge, analyses comparing the 5XFAD mouse model to any fAD-like model have not previously been done.

We first applied conventional gene set analysis methods (*fgsea* [33], *fry* [34], *camera* [35]) with the Hallmark gene set collection to these two datasets, resulting in the identification of significantly enriched gene sets in young adult (6-month-old) *psen1*[Q96_K97del/+] zebrafish brains and young adult (3-month-old) 5XFAD mouse brains when compared to their wild-type siblings (FDR-adjusted Wilkinson's $p <$ 0.05) (**Figure 2**). The gene sets which were significantly enriched in the *psen1*[Q96_K97del/+] zebrafish and 5XFAD mouse brains were largely non-overlapping,

indicating clear differences in the earliest detectable pathological changes in the brains of these models. Hallmark gene sets uniquely enriched in 5XFAD brains included "interferon alpha response", "interferon gamma response", "complement", and "inflammatory response" while those uniquely enriched in fAD mutation-like zebrafish brains included "glycolysis", "oxidative phosphorylation" and "MTORC1 signalling". Gene sets that were enriched in both mouse and zebrafish included "fatty acid metabolism", "heme metabolism", "E2F targets" and "cholesterol homeostasis". The enriched gene sets in young adult 5XFAD and fAD mutation-like brains are consistent with known information about these models: previous transcriptome analyses of young adult $psen1^{Q96\_K97del/+}$ brains demonstrated highly significant changes in gene expression corresponding to energy metabolism deficits [13, 36], while previous transcriptome analyses of young 5XFAD mouse brains indicated that 2-3 month-old brains were characterised by immune activation gene expression responses [9]. This gave us confidence that the subset of genes represented in the Hallmark gene set collection would likely be sufficient to capture relevant biological signals within our analysed datasets.

We repeated the above analysis using the AES-PCA method (see **Supplementary Figure 1** for a more detailed comparison of results from conventional gene set testing methods and AES-PCA). The sparse loadings of the latent variables (PCs) calculated with AES-PCA give additional information on relevant genes accounting for the variation in gene set activity between samples within each dataset, hence giving further insight on the differences between the $psen1^{Q96\_K97del/+}$ zebrafish and 5XFAD mouse models at the gene expression level (**Figure 3**). In the previous

result, our gene set analysis results indicated that several gene sets were significantly enriched in both the *psen1*^{Q96_K97del/+} zebrafish and 5XFAD mouse brains compared to their wild-type siblings. However, by inspecting the sparse loadings of these gene sets, we are able to see that there are extensive differences in the genes contributing to enrichment of these gene sets in most cases (**Figure 3**). Only two genes are in common: zebrafish *oxtr* and mouse *Oxtr* [oxytocin receptor] for the "epithelial mesenchymal transition" gene set; and, zebrafish *tsc2* and mouse *Tsc2* [tuberin] for the "PI3K AKT MTOR signalling" gene set. Notably, both of these genes are in opposite directions, supporting overall differences between these models at the gene expression level.

Overall, while there are some similarities in the biological activities being altered in the young adult brains of these two animal models at the gene set level, there are extensive and biologically relevant differences at the gene expression level. We hence show that the AES-PCA approach in conjunction with the Hallmark gene set collection is able to provide unique and complementary biological insights to conventional GSEA methods.

AES-PCA latent variables enable cross-species comparison of global transcriptional states of human familial AD and animal models of AD

An important area of AD research focuses on evaluating similarities and differences between animal models of AD and the human disease at the molecular level. These types of studies are important in the context of translational uses for animal AD

models (e.g. identification of relevant molecular pathways for therapeutic intervention). Currently, comparing animal and human gene expression datasets is not straightforward due to biological and technical sources of variation that cause batch effects and noise within and between datasets. Additionally, the lack of statistical power within low-replicate published datasets, and differences in scope between experimental design for studies in human and model organisms, can make meta-analytic approaches challenging. This motivated us to investigate whether the sparse latent variables calculated using AES-PCA could reduce noise and batch effects sufficiently to enable preliminary cross-platform and cross-species comparisons of different datasets.

Here, we consider a comparison of gene expression alterations in post-mortem brains of human early-onset familial AD (eofAD) patients compared to aged adult brains of the two animal models of AD introduced earlier (24-month-old $psen1^{Q96\_K97del/+}$ fAD-like zebrafish and 11-month-old 5XFAD mice). Typically, gene expression studies involving fAD patients are rare, and the only gene expression study of fAD that is available publicly to our knowledge is a microarray dataset by Antonell et al. [37]. Samples in this dataset are derived from posterior cingulate tissue from post-mortem brains of eofAD patients with *PSEN1* mutations (n=7), early-onset AD (eoAD) patients without mutations in fAD-causing genes (n=7), and age-matched healthy controls (n=7). Note that the $psen1^{Q96\_K97del/+}$ zebrafish and 5XFAD mouse datasets differ in the platform used to acquire gene expression data (RNA-seq) as well as on the brain tissue sampled (zebrafish: whole brain; mouse: cortex tissue). An initial Principal Component Analysis (PCA) plot of the gene

expression data from the three datasets supports the presence of batch effects that would normally make it difficult to analyse these datasets together (**Supplementary Figure 2**). In addition, applying a conventional GSEA approach using the Hallmark gene set collection indicates extensive disruption to many biological activities, which is difficult to interpret.

We applied AES-PCA separately to each dataset and applied dataset-specific standardisation (*z*-score) to transform the AES-PCA latent variable PC1 of each dataset to a comparable scale. The reason why we only consider PC1 is because this latent variable represents the major source of variation in each dataset, and in the AES-PCA method, the direction of PC1 corresponds directly to the overall direction of gene set activity. This gives the PC1 values a straightforward interpretation as the "pathway activity" of each gene set per sample. Interestingly, the PC1 latent variable representing Hallmark gene set activities for patients with early-onset fAD strongly resemble those with early-onset AD (without mutations in fAD-causing genes) (**Supplementary Figure 3**), indicating that the overall biological activities in eoAD in this particular dataset (regardless of whether they have an fAD-causing mutation or not) appear similar.

Finally, we used a mixed effects linear model on the AES-PCA latent variable PC1 to investigate whether similar biological changes were present in either aged 5XFAD mice or *psen1*$^{Q96\_K97del/+}$ zebrafish brains when compared to the eoAD human patients. The majority of Hallmark gene sets (48/50) in human eoAD were significantly captured by at least one of the zebrafish or mouse models (FDR-

adjusted *p*-value < 0.05) (**Figure 4**). Overall, aged fAD mutation-like zebrafish and

5XFAD mouse brains tended to capture complementary biological activities to each

other, suggesting that perhaps these models may have complementary utility when

studying molecular activities altered during human eoAD and eofAD. There were

also several gene sets showing similar changes in both the aged *psen1*[Q96_K97del/+]

zebrafish and 5XFAD mouse brains when compared to human eoAD (including

"xenobiotic metabolism", "heme metabolism"). We stress that the similarities and

differences shown in these comparisons should be considered preliminary results as

only one dataset was used for each animal model and for human eoAD, and the

brain tissue sampled in each dataset was different. This represents a common

scenario when relying on publicly available data, where it may not be possible to find

datasets that are comparable in terms of tissue, platform, and other factors. Our

results demonstrate that AES-PCA is capable of reducing noise within datasets and

batch effects between datasets to allow for preliminary cross-species and cross-

platform comparisons between different datasets.

AES-PCA reduces platform-specific batch effects to enable integrative

analysis of independent familial and sporadic AD datasets

We next investigated whether AES-PCA would enable us to analyse two human

Alzheimer's disease datasets differing in key aspects including tissue, platform, and

whether AD was early-onset vs. late-onset. One dataset comprises posterior

cingulate cortex-derived microarray data from early-onset AD (n=7), familial early-

onset AD (n=7) brains and healthy controls (n=7) (dataset referred to as **eoAD, microarray**) while the other comprises temporal cortex-derived RNA-seq data from sporadic AD brains (majority late-onset cases) (n=67) and healthy controls (n=65) (dataset referred to as **sAD, RNA-seq**).

After applying AES-PCA to both datasets separately, latent variables representing Hallmark gene set activity across all samples demonstrated reduced noise (e.g. dataset-, platform- and tissue-specific batch effects) and greater emphasis of the biologically meaningful signal in the data corresponding to AD diagnosis. **Figure 5** shows PCAs of both datasets before and after applying AES-PCA. While the **sAD, RNA-seq** samples still display a high level of heterogeneity compared to the **eoAD, microarray** samples, applying AES-PCA results in clearly improved separation of AD and control samples across Principal Component 2 in both datasets.

While separation between AD and control samples in the **eoAD, microarray** was pronounced based on latent variables captured by AES-PCA, samples in the **sAD, RNA-seq** dataset still displayed high heterogeneity, with some AD samples overlapping with control samples across PCs. We tested whether this variation might be explained through different age at death or ApoE genotype and found no significant correlation between age at death with AD diagnosis (Logistic regression $p = 0.82$) along with a strong association between ApoE genotype and AD diagnosis (Fisher's exact test $p = 7.7e\text{-}07$) consistent with increased AD risk through possessing at least one copy of ApoE4.

To explore other biologically-relevant signals within the data that could explain this variation in the **sAD, RNA-seq** samples besides ApoE genotype, we applied bootstrapped *k*-means clustering to the Hallmark AES-PCA latent variable PC1 of the **sAD, RNA-seq** samples, resulting in the identification of three clusters with high stability (Jaccard index > 0.9) (**Supplementary Text 1**). Interestingly, while each of the three clusters contains both AD and control samples, *Clusters 1* and *2* display distinct patterns of gene set activity (AES-PCA PC1 values) that are both remarkably homogenous for the samples within and are in the opposite direction to each other (**Figure 6**). It would seem that *Cluster 3* is comprised of samples with Hallmark gene set activity that does not clearly resemble either *Clusters 1* or *2*. The opposite Hallmark gene set activity shown by *Clusters 1* and *2* is biologically interesting as both clusters contain both AD and control samples showing very similar activities in Hallmark gene sets. This may imply that AD samples exhibit a large amount of diversity in the underlying molecular pathology at the gene expression level. Additionally, the idea that AD and control samples could show such similar gene set activity (AES PC1 values) within either *Clusters 1* or *2* may indicate that perhaps the Hallmark gene set collection may not be sufficient to capture all aspects of AD pathology that differentiate them from healthy controls.

To test whether choice of gene set used in the calculation of AES-PCA latent variables would be able to detect these differences between AD and control within clusters, we used a different gene set collection (IRE gene sets) previously found to have little overlap with the Hallmark gene set collection [36]. The IRE gene sets comprise four gene sets containing genes computationally predicted to contain Iron Responsive Elements in their 3' or 5' UTR. Without clustering, running AES-PCA

with the IRE gene sets from the ***sAD, RNA-seq*** data results in no significant differences detected between AD and control samples for any of the IRE gene sets (**Figure 7B**). Notably however, when samples are separated into the three clusters defined based on the Hallmark gene set AES-PCA PC1 values, each cluster shows distinct differences in AES-PCA PC1 activity for the IRE gene sets. Notably, there are clear significant differences between AD and control samples in AES-PCA PC1 activity for the "high-quality 3' IRE genes", "all predicted 5' IRE genes" and "high-quality 5' IRE genes" sets for Cluster 2 (**Figure 7A**, p-value from t-test < 0.05). The decreased activity of the "all predicted 5' IRE genes" set along with the increased activity of the "high-quality 5' IRE genes" set is similar to that observed in the early-onset AD dataset (**Supplementary Figure 5**). This result highlights how the reduced representation of gene expression data by AES-PCA is amenable to clustering, particularly in the identification of potentially biologically relevant subgroups.

In this example, we have shown that AES-PCA is an effective approach to comparing gene set activity across different platforms and sources of data and provides a noise-reduced representation of data that is amenable for identifying potentially biologically relevant subgroups or clusters of samples. However, we also note that effective use of AES-PCA to distinguish differences between conditions require careful consideration of gene sets that are representative of these gene expression differences.

Discussion

AD is a complex and heterogeneous disease that necessitates the use of animal models to explore early pathological states of the disease at the molecular level. In this paper we highlight an application of the AES-PCA approach and the utility in addressing problems related to the interpretation, integration, and comparison of different gene expression datasets across different platforms and species in the context of AD. Notably, we find that latent variables captured by AES-PCA and the Hallmark gene set collection effectively reduce noise within datasets and batch effects between datasets, thus enabling (a) interpretable visualisation of key differences underlying the early pathology of two different animal models of AD (*psen1*$^{Q96\_K97del/+}$ fAD mutation-like zebrafish and 5XFAD mice), (b) cross-species comparison of alterations in these animal models to human eofAD dataset, and (c) integrative analysis of human fAD and sAD in a biologically interpretable manner despite major platform-specific differences.

AES-PCA compared to conventional batch correction and meta-analytic approaches

Beyond finding enriched gene sets, a unique aspect of AES-PCA is the dimension reduction step where genes are selected to summarise overall gene set expression. In AES-PCA, the unsupervised nature of this step also shows effectiveness at implicitly reducing batch effects between datasets, which we demonstrated in this work. It is important to note that conventional batch effect removal methods such as RUVseq [38] and ComBat [39, 40] may also be used to explicitly remove batch

effects in gene expression datasets. However, the removal of batch effects using these methods works on the entire gene expression dataset, and typically require tuning of parameters (e.g. manually selecting the value of k in RUVseq, which is typically dataset-dependent) or specifying the experimental design (e.g. defining which samples belong to which groups in ComBat). RUVseq and ComBat estimate batch effects by using the entire set of genes for each dataset, and do not take gene set information into account by design. Instead, after batch correction, gene set tests are performed as a separate step using methods such as GSEA [22] or ROAST/FRY [34, 41]. This means that use of conventional batch correction methods is not directly comparable to the batch effect reduction that AES-PCA accomplishes. AES-PCA performs integrated analysis across all datasets, and appears to achieve batch effect reduction by prioritising the most representative genes of each gene set across all samples. In doing this, AES-PCA appears to reduce technical noise and batch effects between datasets in a reproducible manner when viewed at the gene set-level. This suggests that by reducing the complexity in the batch correction step compared to conventional methods such as ComBat or RUVseq, AES-PCA may be a useful alternative approach to these methods.

AES-PCA also differs fundamentally to traditional meta-analysis approaches like *p*-value combination methods and joint-analysis of datasets using a generalised linear model (GLM) (described in [42]). *p*-value combination methods (e.g. Fisher's method or Stouffer's method) occur after results have been independently obtained for each dataset separately. In contrast, AES-PCA analyses datasets simultaneously. GLMs are also a popular meta-analytic approach used to perform integrated analysis

between datasets, and a strength of this method is that known batch effects are able to be specified as random or fixed effects in the model. However, using a GLM requires that all batch effects must be known beforehand. In contrast, the unsupervised nature of the AES-PCA method means that knowledge of batch effects present within or between datasets is not required, meaning that use of AES-PCA may serve as a useful complementary approach to independently assess results from established methods like GLMs.

Applicability of AES-PCA in analysing Alzheimer's disease datasets

Our results reveal several preliminary insights of relevance to AD research that are consistent with existing literature and may act as starting points for further research.

As the first gene expression comparison of a knock-in fAD model and a transgenic model of AD, our results highlight clear, interpretable differences in the underlying molecular pathology in their young adult brains. Our results are consistent with previous work on these models suggesting that immune activation responses form the earliest pathological changes in the 5XFAD model [9], while young *psen1*$^{Q96\_K97del/+}$ zebrafish brains are characterised by early deficits in energy metabolism [13, 36]. In general, knock-in models have recently shown increasing popularity in AD research despite showing milder pathology compared to transgenic models as their genetic background is considered to resemble more closely that of human fAD [11, 12, 43]. Our results are in support of this idea (see **Figure 4**) and indicate that changes present in aged *psen1*$^{Q96\_K97del/+}$ zebrafish brains are better able to capture changes in human eoAD and eofAD related to energy metabolism

and other biological activities that are not captured in aged 5XFAD mouse brains. However, our results also suggest that 5XFAD mice are able to provide complementary information by capturing activity of other gene sets involving immune responses that are similar in human AD. Overall, our results highlight the importance of having access to younger age groups from animal AD models to investigate early changes in the brain at the molecular level, and emphasise the importance of comparative studies to assess the utility of different animal models of AD.

It has been well-established that the progression of symptoms and molecular pathology present in fAD resembles sAD [44, 45]. Our comparison of eofAD, eoAD (without known fAD mutation) and an independent cohort of sporadic AD brains supports this overall similarity at the gene expression level for the first time. The latent variables captured by AES-PCA suggest that overall changes in molecular activities in eoAD and eofAD are very similar to each other and largely consistent with sAD brains, despite the extensive heterogeneity observed in the sAD brains that remained even after AES-PCA. In our analyses, clustering of the sAD samples based on latent variables captured by AES-PCA indicated that some cases of sAD resembled fAD more than others. This may reflect the extensive diversity in sAD cases (e.g. environmental and genetic factors) in addition to diverse effects on the aging brain in both controls and sAD, and further highlights the need to better understand molecular differences underlying the heterogeneity of sAD.

Importantly, with the increasing popularity of data integration approaches and comparative analysis of different datasets across species and platforms (reviewed by

[46]), our results reveal the need for robust and representative gene sets across different biological activities of interest. In general, AES-PCA used in conjunction with the Hallmark gene set collection enabled interpretable, biologically relevant insights into the molecular activities underlying different conditions that could be straightforwardly compared between different species, platforms, and datasets. However, this approach was not able to distinguish all important differences between AD and control samples in the sAD dataset, which showed extensive heterogeneity. The use of IRE gene sets with AES-PCA however, was able to reveal significant differences in iron homeostasis activity between AD and control samples that were not captured by the Hallmark gene sets. This suggests that results can be influenced by the initial choice of gene sets chosen for inclusion, which is an important limitation to acknowledge in cases where the system under study (e.g. AD) is complex and not well-characterised at the molecular level. Despite this, we suggest that unsupervised techniques such as AES-PCA are a useful complementary approach alongside conventional gene expression analysis techniques as they have several unique advantages including ease of interpretation and reduction of biological and technical noise. These advantages make AES-PCA highly applicable to cross-platform and cross-species integration of multiple datasets, opening up possibilities for extracting further value from publicly available datasets which may otherwise have not been easily integrated.

Methods

Dataset availability, pre-processing and normalisation

The animal model RNA-seq datasets used in this work include: a **zebrafish fAD mutation-like model** derived from *psen1*$^{Q96\_K97del/+}$ and +/+ whole brains from 6- and 24-month-old fish (see [13, 36]; GEO accession GSE149149); and **5XFAD (C57BL6) mice** derived from 5XFAD and +/+ cortex tissue from 3-, 6-, 11- and 12-month-old mice (GEO accessions GSE140286 and GSE142633). Raw fastq files were downloaded from GEO and re-processed using *AdapterRemoval* to perform quality trimming with default parameters, followed by pseudoalignment with *Kallisto* (v.0.45) to either the zebrafish or mouse reference transcriptome using transcript descriptions from Ensembl Release 96 [47]. The "catchKallisto" function from the *edgeR* R/Bioconductor package was used to import count estimates [48]. Transcript-level count estimates were then summed to result in gene-level count estimates. Raw gene-level counts were normalised using the TMM method using the "calcNormFactors" function in *edgeR*. Normalised counts were transformed into log$_2$ cpm counts prior to analysis with AES-PCA.

The **human eoAD** dataset is a microarray dataset (Affymetrix Human Gene 1.1 ST Array) originally from a study by Antonell et al. [37]. We obtained raw data from GEO (accession GSE39420) and converted raw probe intensities to probeset expression while applying RMA normalisation using the "rma" function from the *oligo* R/Bioconductor package [49]. For probeset to gene annotations, we used the *hugene11sttranscriptcluster.db* R/Bioconductor package [50]. However, as there were many 1:many probeset:gene annotations, we performed analyses at the

probeset level. Normalised probeset expression was log$_2$ transformed prior to analysis.

The **human sAD** dataset is a subset of AD and control RNA-seq samples obtained from the Mayo Clinic RNA-seq study [51]. As raw data was not available for public download, we downloaded pre-processed, normalised count data from the Synapse database (https://www.synapse.org/#!Synapse:syn5550404) and filtered to only contain temporal cortex AD and control samples. Data was log$_2$ cpm transformed prior to analysis using the "cpm" function from *edgeR*.

Conversion of gene IDs in the Hallmark gene set collection

Gene sets used were Hallmark collection gene sets from MSigDB (v.7.0) [26, 30] which were downloaded from https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp as .gmt files and imported into R using the *GSEABase* R/Bioconductor package [52]. The imported gene sets contained human entrezgene identifiers, so we used the *biomaRt* R/Bioconductor package [53] to retrieve homologous Ensembl gene IDs for human, zebrafish, and mouse, in order to maintain compatibility with our processed RNA-seq datasets which used Ensembl gene identifiers for their respective species. For the human eoAD microarray dataset which used Affymetrix Human Gene 1.1 ST Array probesets, we used the *hugene11sttranscriptcluster.db* R/Bioconductor package to convert human entrezgenes in the gene sets into compatible probeset IDs.

Gene set enrichment analysis

Initially, gene set enrichment analysis was performed on the zebrafish and mouse normalised datasets using 3 gene set enrichment methods, *fgsea* [33], *camera* [35], and *fry* [34, 41] inspired by the EGSEA methodology [54]. Raw *p*-values were combined using Wilkinson's method and FDR-adjusted, with significantly enriched gene sets defined as having an FDR-adjusted *p*-value < 0.05. See Hin et al. [36] for further details on the gene set enrichment testing method.

AES-PCA

AES-PCA with the MSigDB Hallmark gene set collection was performed using the "AESPCA_pVals" function from the *pathwayPCA* R/Bioconductor package [24, 25] with default parameters, including FDR-adjustment for raw *p*-values calculated using permutation tests.

Testing for significant preservation of PC1 changes between different datasets

Prior to integrated analysis, PC1 latent variables calculated using AES-PCA for each dataset were standardised (z-score) by scaling by the dataset-specific mean and standard deviation. PC1 only was chosen as this variable captures the largest amount of variability, and was assumed to capture the majority of the biological signal.

Using the *lme4* [55] and *lmerTest* [56] R packages, we set up a mixed linear effects model (`lmer(value ~ group + (1|dataset)`) with dataset as a random effect (i.e. nesting term) in order to detect significant differences between AD (or AD-like

condition) and controls that were preserved across the following comparisons: fAD

mutation-like zebrafish brains and human eofAD; 5XFAD mouse brains and human

eofAD. Gene sets with AES-PCA PC1 values significantly preserved across

comparisons were defined as having FDR-adjusted p-value < 0.05.

Clustering of sAD human dataset

The "clusterboot" function from the *fpc* R package [57] was used to identify stable

clusters in the sAD human dataset including AD and control samples. The default

"bootmethod = boot" was used along with 1000 bootstraps and k-means clustering

as the clustering method. We repeated this for k = 3, 4, 5. The Jaccard index

(measure of cluster stability) was calculated for each cluster in each case. We used

k=3 for the final clusters as cluster stability of all 3 clusters had Jaccard index > 0.9

indicating highly stable clusters (Clusters with Jaccard index > 0.85 are considered

highly stable while those with Jaccard index < 0.6 indicate unstable clusters).

AES-PCA using IRE gene sets

For the human datasets, we ran AES-PCA again with default settings but with

numCores = 8 and human IRE gene sets obtained from [36] instead of Hallmark

gene sets, as the IRE gene sets have been previously shown to have little overlap

with Hallmark gene sets [36]. Differences in AES-PCA PC1 values between AD and

control groups were calculated using *t*-tests.

References

1. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12:e0190152. doi:10.1371/journal.pone.0190152.

2. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, et al. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annu Rev Biomed Data Sci. 2019;2:139–73. doi:10.1146/annurev-biodatasci-072018-021255.

3. Devi G, Scheltens P. Heterogeneity of Alzheimer's disease: Consequence for drug trials? Alzheimer's Res Ther. 2018;10:122. doi:10.1186/s13195-018-0455-y.

4. Ferreira D, Wahlund LO, Westman E. The heterogeneity within Alzheimer's disease. Aging (Albany NY). 2018;10:3058–60. doi:10.18632/aging.101638.

5. Au R, Piers RJ, Lancashire L. Back to the future: Alzheimer's disease heterogeneity revisited. Alzheimer's Dement Diagnosis, Assess Dis Monit. 2015;1:368–70. doi:10.1016/j.dadm.2015.05.006.

6. Badhwar AP, McFall GP, Sapkota S, Black SE, Chertkow H, Duchesne S, et al. A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. Brain. 2020;143:1315–31. doi:10.1093/brain/awz384.

7. Richard BC, Kurdakova A, Baches S, Bayer TA, Weggen S, Wirths O. Gene dosage dependent aggravation of the neurological phenotype in the 5XFAD mouse model of Alzheimer's disease. J Alzheimer's Dis. 2015;45:1223–36. doi:10.3233/JAD-143120.

8. Jawhar S, Trawicka A, Jenneckens C, Bayer TA, Wirths O. Motor deficits, neuron loss, and reduced anxiety coinciding with axonal degeneration and intraneuronal Aβ aggregation in the 5XFAD mouse model of Alzheimer's disease. Neurobiol Aging. 2012;33:196.e29-196.e40. doi:10.1016/j.neurobiolaging.2010.05.027.

9. Bundy JL, Vied C, Badger C, Nowakowski RS. Sex-biased hippocampal pathology in the 5XFAD mouse model of Alzheimer's disease: A multi-omic analysis. J Comp Neurol. 2019;527:462–75.

10. Myers A, McGonigle P. Overview of Transgenic Mouse Models for Alzheimer's Disease. Curr Protoc Neurosci. 2019;89. doi:10.1002/cpns.81.

11. Xia D, Watanabe H, Wu B, Lee SH, Li Y, Tsvetkov E, et al. Presenilin-1 knockin mice reveal loss-of-function mechanism for familial alzheimer's disease. Neuron. 2015;85:967–81.

12. Saito T, Matsuba Y, Mihira N, Takano J, Nilsson P, Itohara S, et al. Single App knock-in mouse models of Alzheimer's disease. Nat Neurosci. 2014;17:661–3. doi:10.1038/nn.3697.

13. Newman M, Hin N, Pederson S, Lardelli M. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production. Mol Brain. 2019;12:43. doi:10.1186/s13041-019-0467-y.

14. Hin N, Newman M, Kaslin J, Douek AM, Lumsden A, Nik SHM, et al. Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human PSEN2. PLoS One. 2020;15:e0227258.

15. Guo Q, Fu W, Sopher BL, Miller MW, Ware CB, Martin GM, et al. Increased vulnerability of hippocampal neurons to excitotoxic necrosis in presenilin-1 mutant knock-in mice. Nat Med. 1999;5:101–6.

16. Kawasumi M, Chiba T, Yamada M, Miyamae-Kaneko M, Matsuoka M, Nakahara J, et al. Targeted introduction of V642I mutation in amyloid precursor protein gene causes functional abnormality resembling early stage of Alzheimer's disease in aged mice. Eur J Neurosci. 2004;19:2826–38. doi:10.1111/j.0953-816X.2004.03397.x.

17. Siman R, Reaume AG, Savage MJ, Trusko S, Lin YG, Scott RW, et al. Presenilin-1 P264L knock-in mutation: Differential effects on Aβ production, amyloid deposition, and neuronal vulnerability. J Neurosci. 2000;20:8717–26.

18. Reaume AG, Howland DS, Trusko SP, Savage MJ, Lang DM, Greenberg BD, et al. Enhanced amyloidogenic processing of the β-amyloid precursor protein in gene-targeted mice bearing the Swedish familial Alzheimer's disease mutations and a "humanized" Aβ sequence. J Biol Chem. 1996;271:23380–8. doi:10.1074/jbc.271.38.23380.

19. Goh WW Bin, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. 2017;35:498–507.

20. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? Behav Brain Res. 2017;322 Pt B:311–28.

21. Stein-O'Brien GL, Arora R, Culhane AC, Favorov A V., Garmire LX, Greene CS, et al. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018;34:790–805.

22. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.

23. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. 2005;6:225. doi:10.1186/1471-2105-6-225.

24. Odom G, Ban Y, Liu L, Sun X, Pico A, Zhang B, et al. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection. bioRxiv. 2019;:615435. doi:10.1101/615435.

25. Chen X. Adaptive elastic-net sparse principal component analysis for pathway association testing. Stat Appl Genet Mol Biol. 2011;10:48. doi:10.2202/1544-6115.1697.

26. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP, et al. Databases and ontologies Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40. doi:10.1093/bioinformatics/btr260.

27. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30. doi:10.1093/nar/28.1.27.

28. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46:D661–7. doi:10.1093/nar/gkx1064.

29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20. doi:10.1111/j.1467-9868.2005.00503.x.

30. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1:417–25. doi:10.1016/j.cels.2015.12.004.

31. Giannoni P, Arango-Lievano M, Neves I Das, Rousset MC, Baranger K, Rivera S, et al. Cerebrovascular pathology during the progression of experimental Alzheimer's disease. Neurobiol Dis. 2016;88:107–17. doi:10.1016/j.nbd.2016.01.001.

32. Tiwari SS, Mizuno K, Ghosh A, Aziz W, Troakes C, Daoud J, et al. Alzheimer-related decrease in CYFIP2 links amyloid production to tau hyperphosphorylation and memory loss. Brain. 2016;139:2751–65. doi:10.1093/brain/aww205.

33. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative

statistic calculation. bioRxiv. 2016;:060012. doi:10.1101/060012.

34. Giner G, Smyth GK, Giner G, Smyth GK. <p>FRY: a fast approximation to ROAST gene set test with mean aggregated set statistics</p>. F1000Research. 2016;5.

35. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40:e133. doi:10.1093/nar/gks461.

36. Hin N, Newman M, Pederson SM, Lardelli MM. Iron Responsive Element (IRE)-mediated responses to iron dyshomeostasis in Alzheimer's disease. bioRxiv. 2020;:2020.05.01.071498.

37. Antonell A, Lladó A, Altirriba J, Botta-Orfila T, Balasa M, Fernández M, et al. A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. Neurobiol Aging. 2013;34:1772–8. doi:10.1016/j.neurobiolaging.2012.12.026.

38. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902. doi:10.1038/nbt.2931.

39. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27. doi:10.1093/biostatistics/kxj037.

40. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3. doi:10.1093/bioinformatics/bts034.

41. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26:2176–82. doi:10.1093/bioinformatics/btq401.

42. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014;15:91. doi:10.1186/1471-2105-15-91.

43. Sukoff Rizzo SJ, Masters A, Onos KD, Quinney S, Sasner M, Oblak A, et al. Improving preclinical to clinical translation in Alzheimer's disease research. Alzheimer's Dement Transl Res Clin Interv. 2020;6. doi:10.1002/trc2.12038.

44. Duara R, Lopez-Alberola RF, Barker WW, Loewenstein DA, Zatinsky M, Eisdorfer CE, et al. A comparison of familial and sporadic alzheimer's disease. Neurology. 1993;43:1377–84. doi:10.1212/wnl.43.7.1377.

45. Hellström-Lindahl E, Viitanen M, Marutle A. Comparison of Aβ levels in the brain of familial and

sporadic Alzheimer's disease. Neurochem Int. 2009;55:243–52.

46. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017;8 JUN:84. doi:10.3389/fgene.2017.00084.

47. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

49. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26:2363–7. doi:10.1093/bioinformatics/btq431.

50. MacDonald JW. hugene11sttranscriptcluster.db: Affymetrix hugene11 annotation data (chip hugene11sttranscriptcluster). 2017.

51. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016;3:1–10.

52. Morgan, M; Falcon, S; Gentleman R. GSEABase: Gene set enrichment data structures and methods. 2019.

53. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart - Biological queries made easy. BMC Genomics. 2009;10:22. doi:10.1186/1471-2164-10-22.

54. Alhamdoosh M, Law CW, Tian L, Sheridan JM, Ng M, Ritchie ME. Easy and efficient ensemble gene set testing with EGSEA. F1000Research. 2017;6:2010.

55. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48. doi:10.18637/jss.v067.i01.

56. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models . J Stat Softw. 2017;82:1–26. doi:10.18637/jss.v082.i13.

57. Hennig C. fpc: Flexible Procedures for Clustering. 2020.

58. Hennig C. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. J Multivar Anal. 2008;99:1154–76.
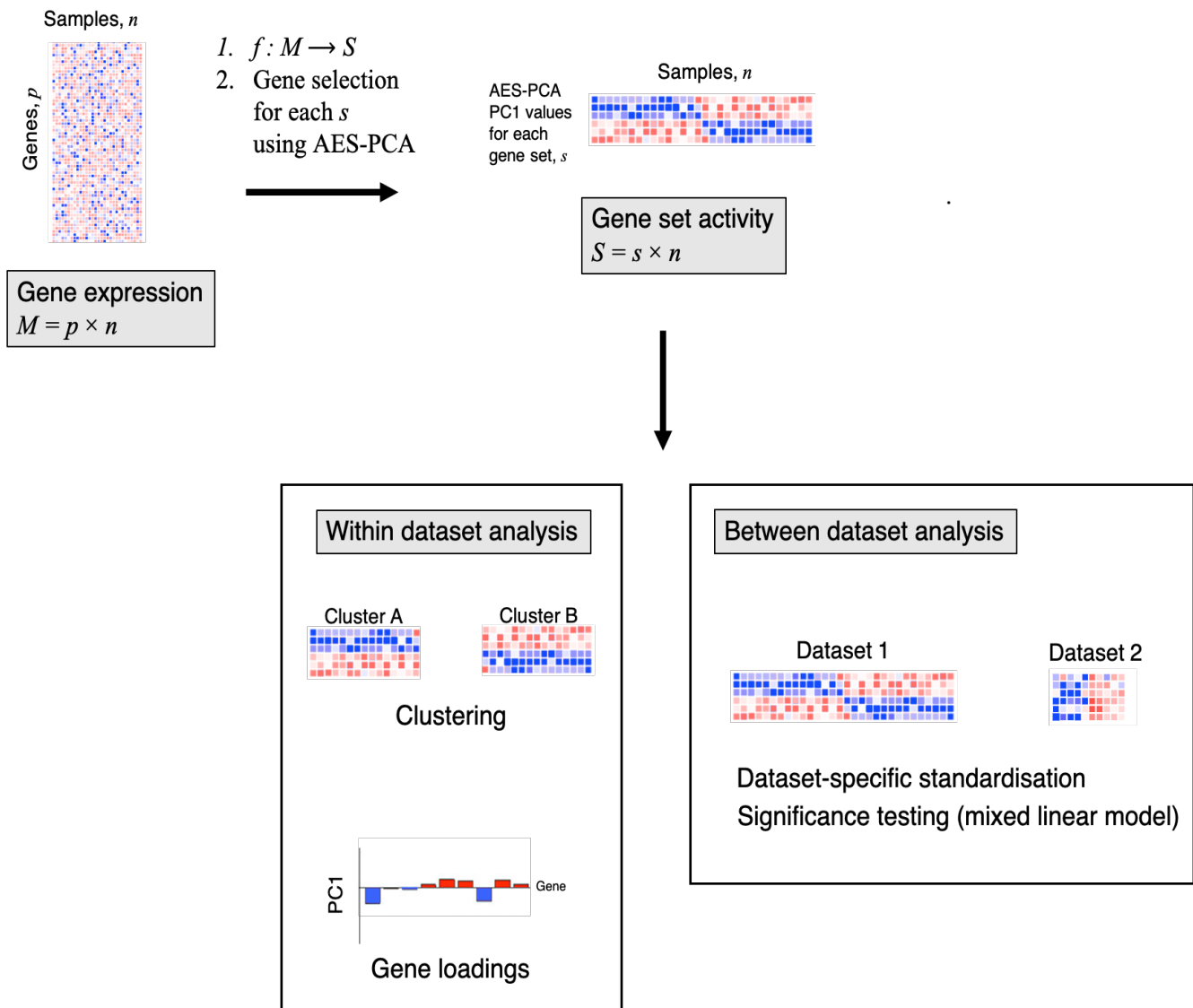
Figure 1. Schematic showing the application of AES-PCA method to gene expression analysis. Refer to Chen [25] for further details on the AES-PCA method.
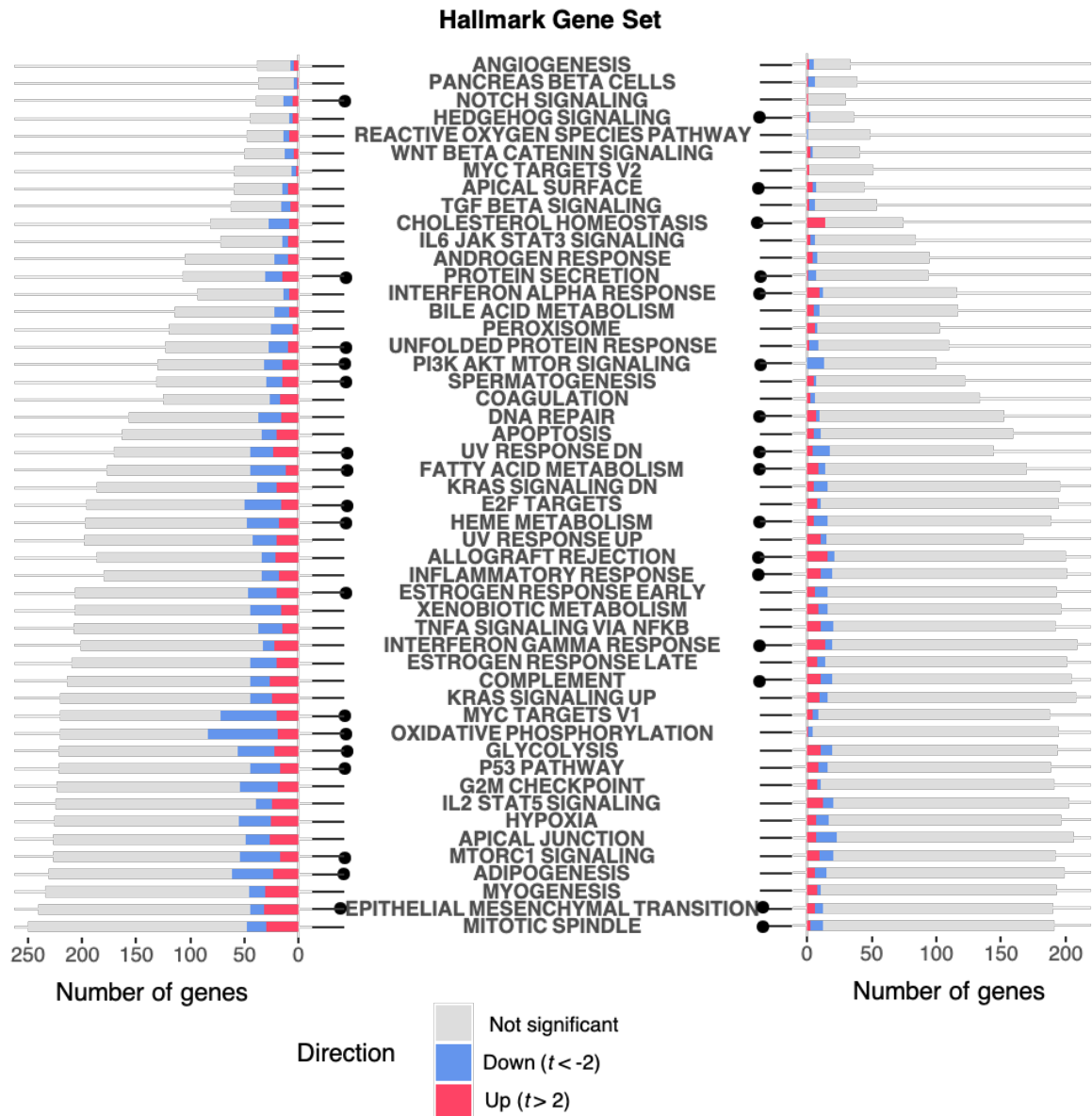
Figure 2. Comparison of molecular pathological changes in young adult brains of fAD mutation-like (*psen1*^{Q96_K97del/+}) zebrafish (6 months old) and 5XFAD mice (3 months old) compared to wild-type siblings using gene set enrichment testing of the MSigDB Hallmark sets. Gene sets were considered significantly enriched if the Wilkinson FDR-adjusted *p*-value (from *fgsea*, *fry*, and *camera*) was < 0.05 (indicated with dots to either left or right of gene set names). The proportion of genes within each gene set which were increased (*limma* t-statistic > 2) or decreased (*limma* t-statistic < -2) in expression are also indicated.
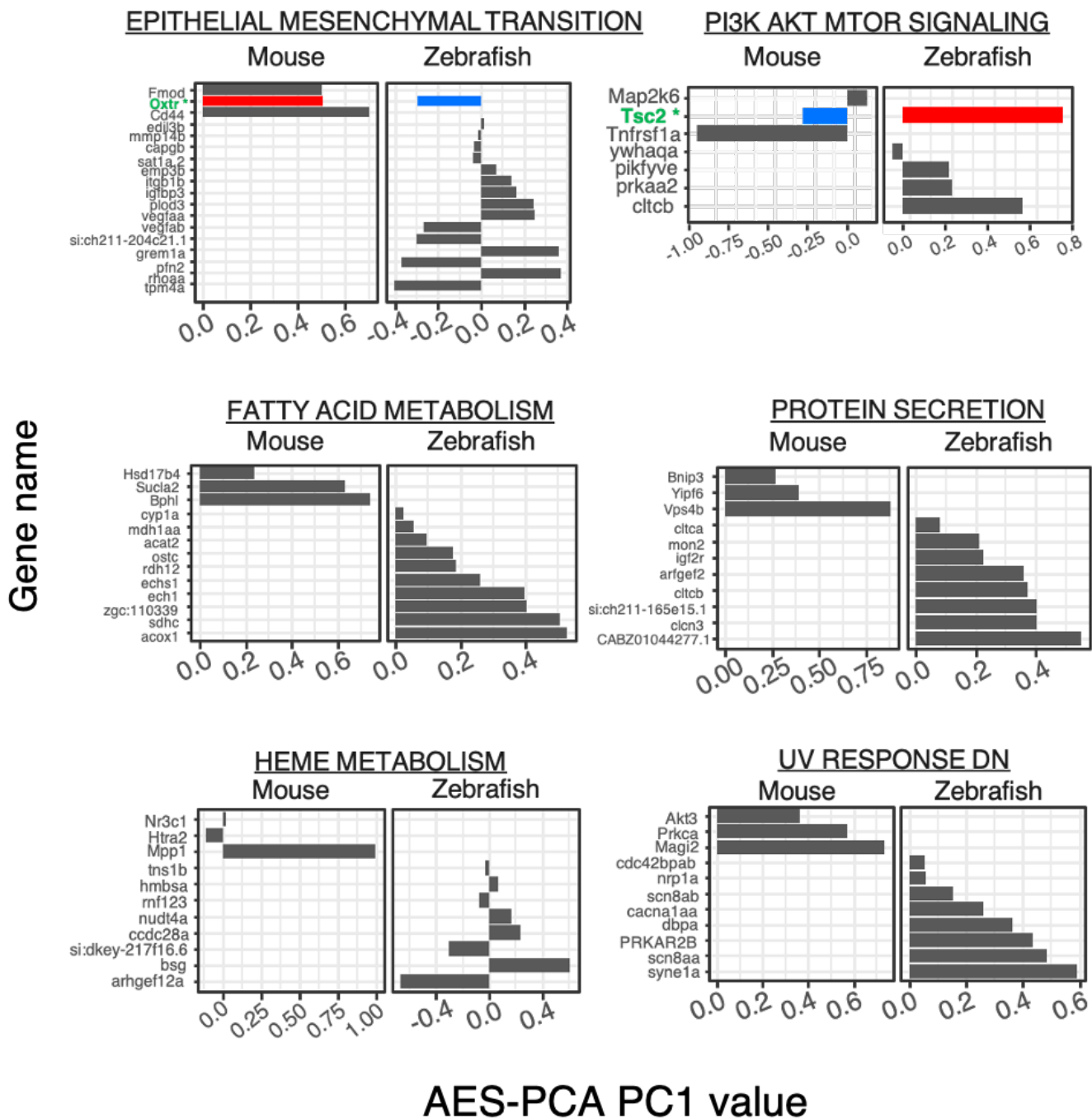
Figure 3. Genes contributing significantly to variation in gene set activity as determined from AES-PC1 in young adult fAD mutation-like zebrafish (6 months old) and 5XFAD mouse (3 months old) brains when compared to wild-type siblings. Gene sets shown here were found to be significant in both the young adult (6 month old) fAD mutation-like zebrafish (psen1Q96_K97del/+) and young adult (3 month old) 5XFAD mice using a combination of *fgsea*, *fry* and *camera*. Homologous genes which are present in both the mouse and zebrafish datasets are indicated in red colour.
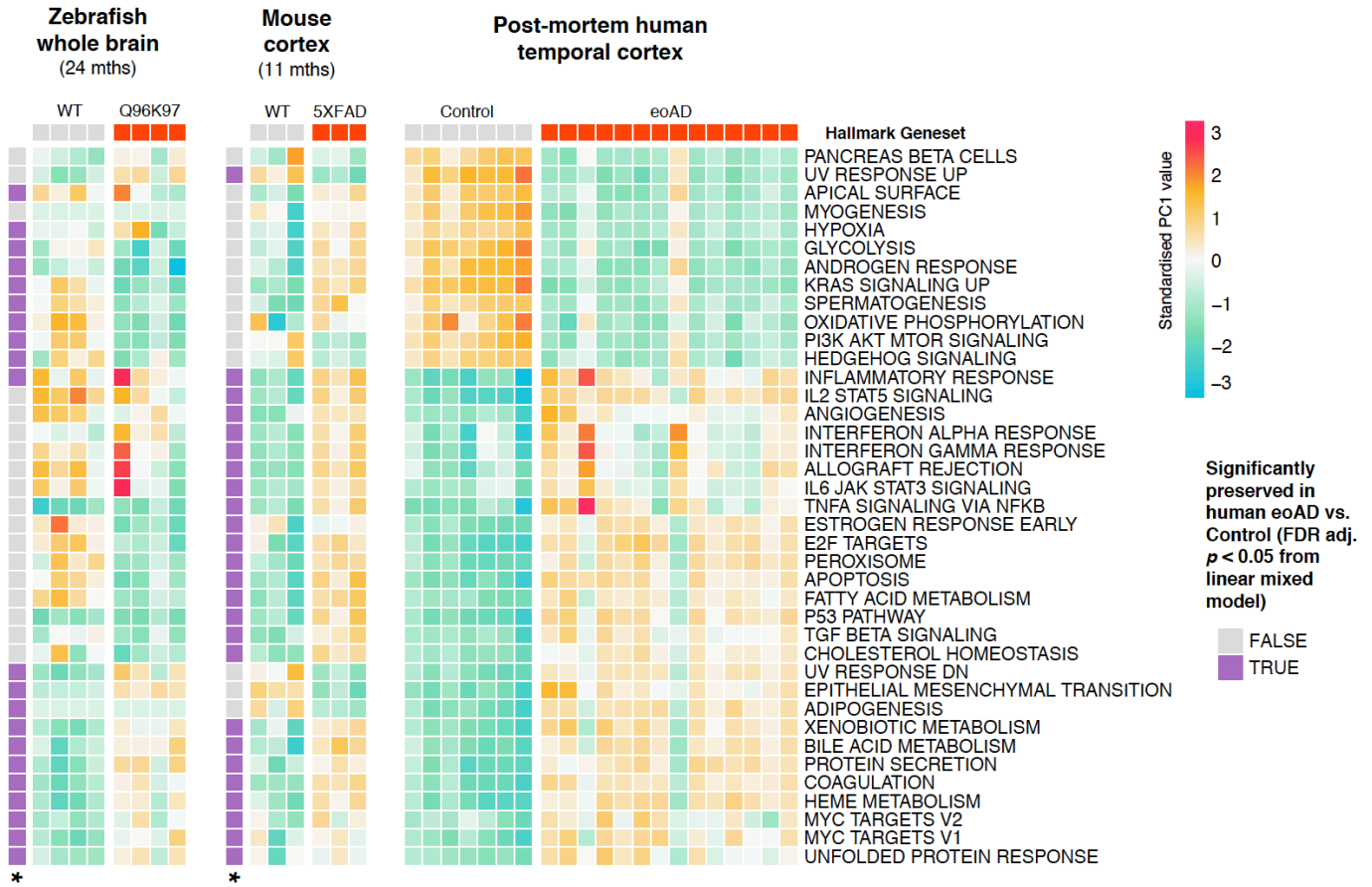
Figure 4. Preservation of brain gene expression changes between AD-like models and human early-onset fAD. Gene expression changes shown here are PC1 values calculated using AES-PCA and standardised through scaling by dataset-specific mean and standard deviation. Significant preservation was determined through linear mixed model (FDR-adjusted p-value < 0.05). Zebrafish values derived from whole-brain gene expression, mouse values derived from cortex gene expression, human values derived from posterior cingulate gene expression.
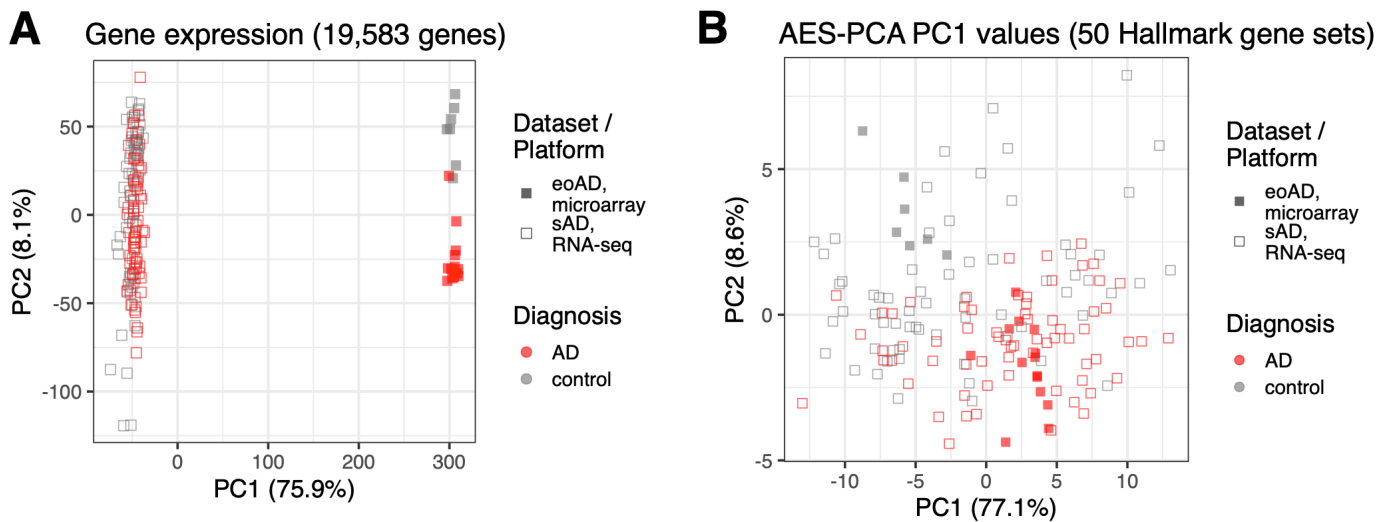
Figure 5. Principal Component Analysis (PCA) plots for two human AD gene expression datasets obtained from different platforms (*eoAD, microarray*; *sAD, RNA-seq*). A. PCA plot of genes expressed in both datasets. The *eoAD, microarray* dataset contained 23,076 genes identified by Ensembl gene identifiers (from 33,297 probesets), while the *sAD, RNA-seq* dataset contained 48,984 genes identified by Ensembl gene identifers. Only 19,583 genes which were present in both datasets were used for the PCA. Gene expression values were log2 transformed for each dataset prior to PCA. **B. PCA plot of PC1 values from AES-PCA from 50 Hallmark gene sets for both datasets**. AES-PCA PC1 values were standardised (*z*-score) for each dataset prior to PCA.

Figure 6. Heatmap of PC1 values calculated using AES-PCA for MSigDB Hallmark gene set activities in the *sAD, RNA-seq* dataset. PC1 values for the samples were clustered using bootstrapped *k*-means clustering (k=3). Annotations on top of the heatmap indicate ApoE genotype, age at death, diagnosis, and cluster assigned for each sample. There was no significant association between Cluster and ApoE genotype or age at death. The unclustered heatmap is provided in **Supplementary Figure 4**.
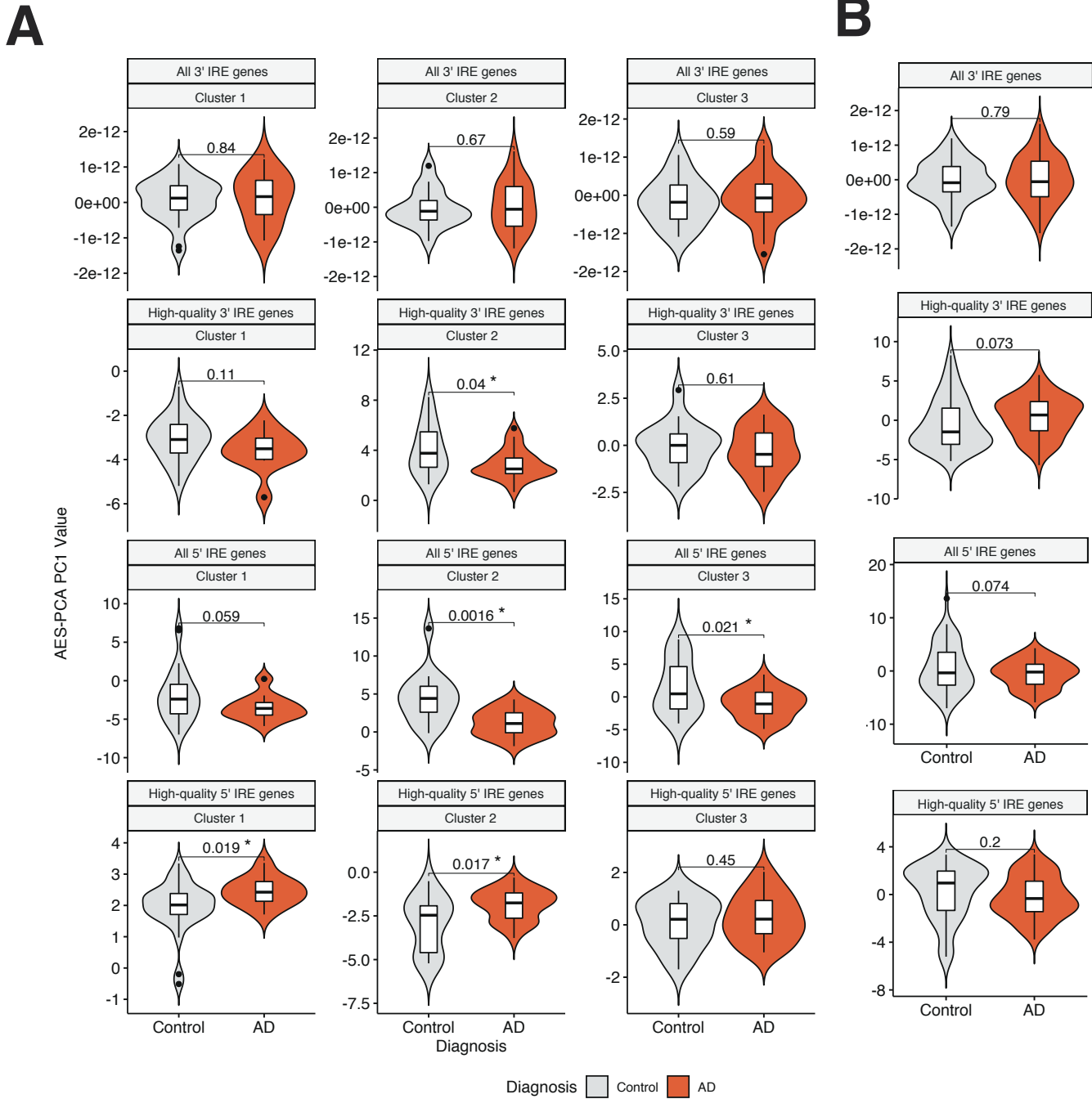
**Figure 7. AES-PCA PC1 values for IRE gene sets ("all 3' IRE genes", "high-quality 3'
IRE genes", "all 5' IRE genes", high-quality 5' IRE genes") calculated for the sAD,
RNA-seq samples. A. Dataset split into three clusters determined using bootstrapped
k-means with k=3. B. Full dataset.** Differences between AD and control groups were
tested using *t*-test (significance defined as $p < 0.05$). For further information on how IRE
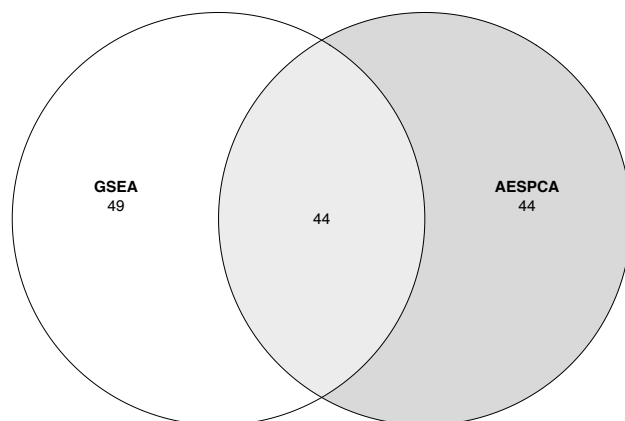gene sets were defined, see Hin et al. [36].

Supplementary Figures

Supplementary Figure 1

***psen1*^{Q96_K97del/+} vs. +/+ zebrafish (6-months-old)**
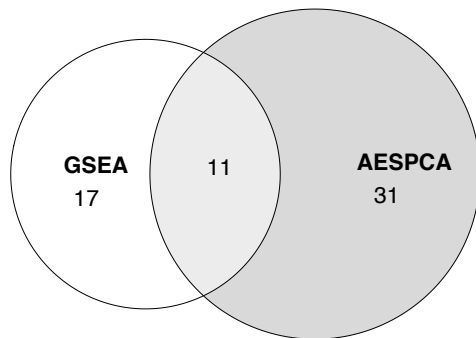
***5XFAD* vs. +/+ mice (6-months-old)**

Comparison of significantly enriched gene sets using conventional GSEA methods and AES-PCA in 6-month-old age groups of zebrafish and mouse datasets. The gene set enrichment analysis methods used (indicated as "GSEA" in the diagrams) were *fgsea*, *camera,* and *fry*, with raw *p*-values being combined using Wilkinson's method and FDR-adjusted (gene sets significantly enriched have FDR-adj. $p < 0.05$). For the zebrafish and mouse datasets shown, *fgsea*, *camera*, and *fry* were run on the complete dataset (including all age groups), using *limma* contrasts to test for associations between gene sets and specific comparisons (i.e. 6-month-old *psen1*^{Q96_K97del/+} vs. +/+ zebrafish and 6-month-old 5XFAD vs. +/+ mice). In contrast, AES-PCA was run on each age group for each dataset separately. AES-PCA is an unsupervised technique and hence requires additional permutation testing on the latent variables calculated in order to determine whether significant

associations with conditions exist. Currently, this permutation testing does not allow for "contrasts" to be specified in the same way as *limma*, and it is possible that running AES-PCA on age groups separately may result in decreased power to detect significant gene set associations compared to *limma* with conventional gene set enrichment testing methods. However, other properties of AES-PCA were proposed by Chen [25] to result in increased power for detecting significant gene set associations. The comparisons here suggest that neither AES-PCA or conventional gene set enrichment methods tested are clearly better than each other. The permutation testing for AES-PCA is implemented in the default functionality of the "AESPCA_pVals" function in *pathwayPCA*, and all *p*-values were FDR-adjusted (significantly enriched gene sets have FDR-adj. $p < 0.05$). Given that AES-PCA uses a different method to test for significant gene set associations compared to the way that *fgsea*, *fry*, and *camera* were used in a *limma* workflow, we would not expect results to be exactly the same, but somewhat similar. Here, the similarity between AES-PCA to the conventional gene set enrichment analysis methods tested (*fgsea*, *fry*, and *camera*) appears acceptable.
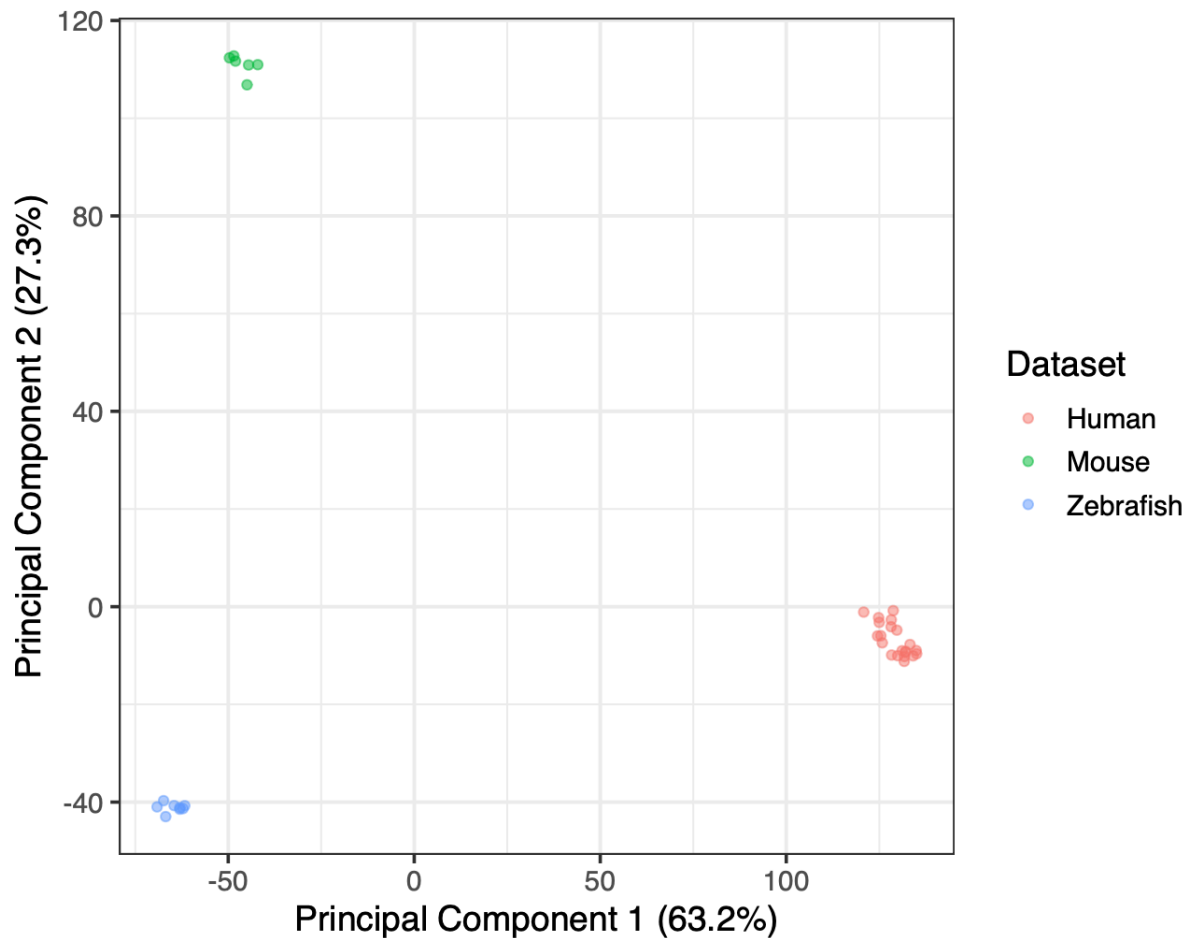
Principal Component Analysis (PCA) plot of log2 normalised brain gene expression for three datasets: post-mortem human eoAD and controls (posterior cingulate tissue), 24-month-old *psen1*^{Q96_K97del/+} zebrafish and wild-type siblings (whole brains), and 11-month-old 5XFAD and wild-type siblings (cortex tissue). The PCA indicates clear differences (batch effects) between the three datasets, with the largest source of variation (PC1) correlating with platform/technology (Mouse and Zebrafish datasets are RNA-seq while Human is microarray). The second largest source of variation (PC2) correlates with the differences between the Zebrafish and Mouse datasets. Only the 12,638 genes which were in common across all datasets were included in this PCA.

Supplementary Figure 3
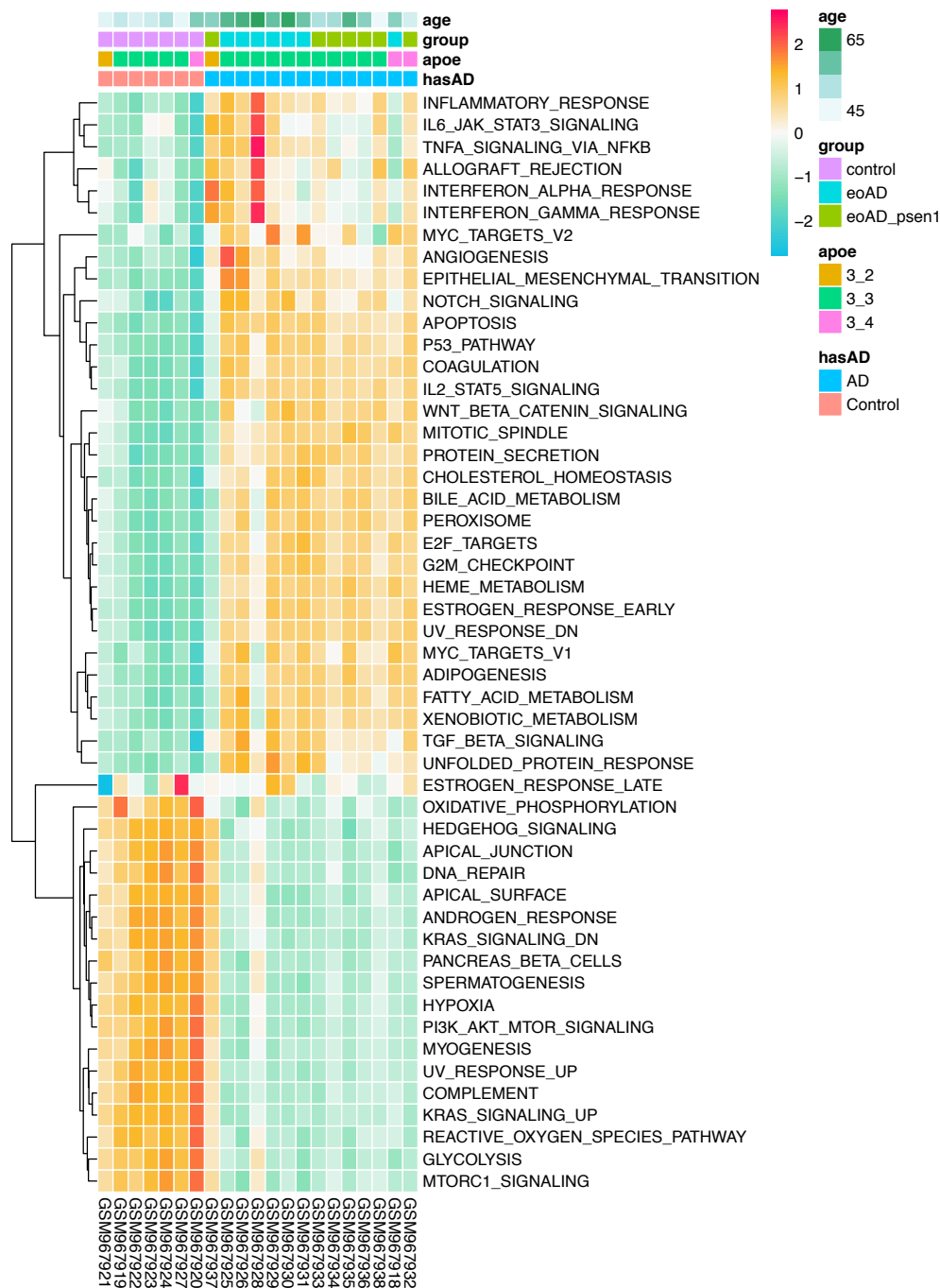
Heatmap of latent variables (PC1) representing Hallmark gene set activities calculated using AES-PCA for a microarray eoAD (early onset Alzheimer's disease) dataset. eoAD samples include tissue samples with known *psen1* fAD mutations ("eoAD_psen1" group) and without known fAD mutations ("eoAD" group). Overall, the latent variables representing Hallmark gene set biological activities are similar between eoAD_psen1 and eoAD samples, and clearly distinct from the age-matched controls.

Supplementary Figure 4
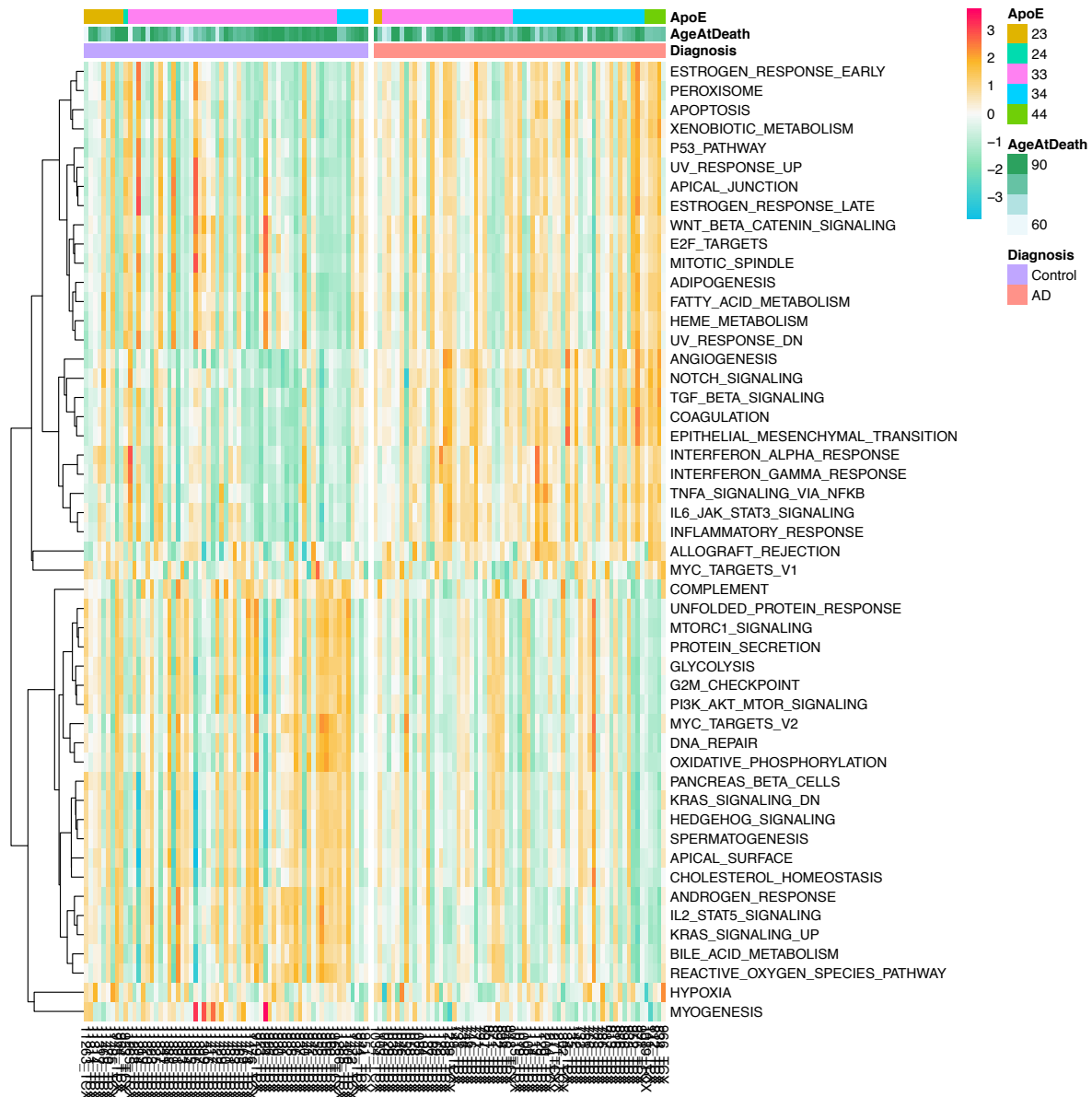
Heatmap of latent variables (PC1) representing Hallmark gene set activities calculated using AES-PCA for a sAD (sporadic Alzheimer's disease) dataset. This visualisation shows a clear trend of differences between AD and control samples, although a large amount of heterogeneity is still present. This heterogeneity was not fully explained by ApoE genotype and was not correlated to age (see main manuscript text), which formed the motivation to perform additional clustering on this dataset.
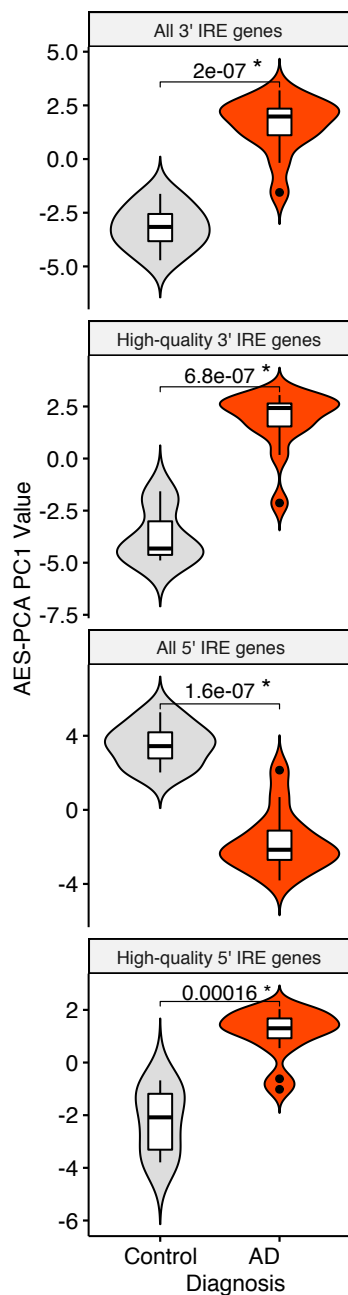
AES-PCA PC1 values for IRE gene sets ("all 3' IRE genes", "high-quality 3' IRE genes", "all 5' IRE genes", high-quality 5' IRE genes") calculated for the microarray eoAD dataset. Differences between AD and control groups were tested using *t*-test (significance defined as $p < 0.05$). For further information on how IRE gene sets were defined, see Hin et al. [1].

Supplementary Text 1. Determination of stable clusters for the RNA-seq, sAD dataset.

Prior to clustering, the PC1 values were standardised by scaling the data to have zero mean and unit variance. Bootstrapped k-means clustering was performed using the *fpc* R package [1] and the "clusterboot" function, using the clustering method "kmeansCBI" and 1000 bootstraps, for k = 3, 4, and 5. Running the "clusterboot" function gives cluster assignments (i.e. which genes are assigned to which clusters) as well as cluster stability assessments which are shown below. As a general guide, Hennig [2] suggests that a valid, stable cluster should have a mean Jaccard similarity value ≥ 0.75, while clusters considered highly stable should have a mean Jaccard similarities ≥ 0.85. Clusters with mean Jaccard similarity values between 0.6 and 0.75 are not considered stable, although may indicate patterns in the data. Clusters with mean Jaccard similarity values < 0.6 are considered unstable [2]. For the purposes of our analysis, we used k = 3 as all three clusters had Jaccard similarity values > 0.85 indicating highly stable clusters.

k = 3

* Cluster stability assessment *

Cluster method: kmeans

Full clustering results are given as parameter result

of the clusterboot object, which also provides further statistics

of the resampling results.

Number of resampling runs: 1000

Number of clusters found in data: 3

Clusterwise Jaccard bootstrap (omitting multiple points) mean:

[1] 0.9391498 0.9486744 0.9019909

dissolved:

[1] 1 4 12

recovered:

[1] 985 964 913

k = 4

* Cluster stability assessment *

Cluster method: kmeans

Full clustering results are given as parameter result

of the clusterboot object, which also provides further statistics

of the resampling results.

Number of resampling runs: 1000

Number of clusters found in data: 4

 Clusterwise Jaccard bootstrap (omitting multiple points) mean:

[1] 0.8238979 0.5824745 0.6560716 0.4268291

dissolved:

[1] 52 361 161 781

recovered:

[1] 721 147 232 91

k = 5

* Cluster stability assessment *

Cluster method: kmeans

Full clustering results are given as parameter result

of the clusterboot object, which also provides further statistics

of the resampling results.

Number of resampling runs: 1000

Number of clusters found in data: 5

 Clusterwise Jaccard bootstrap (omitting multiple points) mean:

[1] 0.5658384 0.6252707 0.5852597 0.5082613 0.6743152

dissolved:

[1] 578 315 414 618 215

recovered:

[1] 345 244 193 182 364

Chapter 6

Conclusions and Future Directions

Conclusions and Future Directions

A significant gap in the field of AD research lies in our understanding of the molecular mechanisms underlying incipient stages of AD. This is particularly relevant in familial AD, where subtle changes in the brain may begin as early as young adulthood. To better understand these early stages of AD at the molecular level, transcriptome analyses of accurate animal models is essential. This strategy underpins the research within the manuscripts here, where transcriptome analysis and other bioinformatics-focused approaches are applied to knock-in zebrafish models to investigate molecular mechanisms relevant to human AD.

It is important to note that prior to these manuscripts, no transcriptome analysis of any knock-in animal model of a mutation causing AD had been performed. This is a major gap in the field considering that the biological effects of knock-in familial AD mutations are known to be subtle and potentially overlooked without appropriate methodologies to capture these effects [1–4]. To address this research gap, the manuscripts in Chapters 2, 3, and 4 contained the first such transcriptome analyses on two different knock-in zebrafish models capturing different aspects of AD pathology. Chapter 2 described the analysis of a unique zebrafish model of the human K115fs mutation (*psen1*^{K97fs} zebrafish) that modelled a particular aspect of sporadic AD (increased expression of the PS2V isoform). Importantly, the results of this chapter indicated the presence of biologically relevant stress and immune gene expression responses in young adult brains despite no visible amyloid pathology. Meanwhile, Chapters 3 and 4 highlighted a viewpoint more relevant to familial AD by

describing the first transcriptome analysis of a familial AD mutation-like model (*psen1*$^{Q96\_K97del}$). In particular, glucose metabolism deficits were shown to be amongst the most significant changes in young adult familial AD mutation-like brains (see Chapter 3). As the first transcriptome analysis of a knock-in model of a familial AD-like mutation, these results represent an important contribution to the field as they provide support for a potential role of decreased glucose metabolism in early stages of familial AD. Decreased glucose metabolism has been similarly observed in the brains of cognitively normal young adults predisposed to sporadic AD (carrying the ε4 *APOE* allele) [5], which highlights the importance of studying these early events in AD to determine if it is possible to slow or prevent the progression of AD. Existing evidence points towards a role for decreased glucose metabolism in both early-onset familial AD and later-onset sporadic AD, as FDG-PET scans show hypometabolism occurs in similar brain regions across both early-onset and late-onset AD patients (although early-onset brains show a greater magnitude of hypometabolism) [6]. Because of this, further studies elucidating the molecular mechanisms of glucose metabolism in familial AD mutation-like models may be useful for giving insights relevant to AD in general. In addition, there is evidence that the specific glucose hypometabolism patterns in AD may be important for distinguishing AD from other dementias, particularly in the parieto-temporal, frontal and posterior cingulate cortices [7], and this further underscores the ongoing importance of 'omics studies to characterise differences between AD and other dementias at the molecular level.

The results shown in Chapter 4 extend upon the findings of glucose metabolism deficits in the familial AD mutation-like model and showed that transcriptional changes indicative of iron dyshomeostasis were also likely to be important in the early stages of the disease. Importantly, this finding has not been previously shown. Given that iron dyshomeostasis pathways have been known to interact with diverse pathways implicated in AD (e.g. energy metabolism, immune responses and inflammation, hypoxia) [8], the results in Chapter 4 contribute to current knowledge regarding the early pathogenesis of AD, and generate new questions regarding the events surrounding iron dyshomeostasis in early stages of AD. Overall, as the first transcriptome analyses of zebrafish knock-in models of AD, the results in Chapters 2, 3, and 4 represent the first steps towards characterisation of the transcriptional states underlying early stages of AD.

Future steps continuing this work may involve integration of transcriptomes from different knock-in models to further define key molecular mechanisms that are common or different between familial AD and sporadic AD. For example, the brain transcriptome of a *sorl1* zebrafish knock-in mutation model developed by the Alzheimer's Disease Genetics Laboratory has recently been characterised [9]. In humans, *SORL1* has been implicated in both familial and sporadic AD, so the analysis of a zebrafish *sorl1* mutation model could provide further insight into how these two subtypes of AD are similar at the molecular level. Similarly, knock-in models of *APP* and *PSEN1* in mice have been developed [10, 11], and comparative analysis with zebrafish knock-in mutation models would be useful for further elucidating conserved molecular mechanisms and how gene regulation and

expression may differ between these animals. Unfortunately, no transcriptome analysis has been performed for these mouse models yet, meaning that currently, a key gap still remains in the field.

Another important future step will involve the integration of different 'omics layers including proteomics and metabolomics from zebrafish knock-in mutation models. The majority of the work here has relied on transcriptome data. Findings in Chapter 2 indicated only mild correlation between genes in the transcriptome and encoded proteins in the proteome of the *psen1*^{K97fs} mutation model. This lack of correlation between genes and proteins has been well-documented in other species and studies [12, 13]. Recently, modelling the interactions between genes, proteins, and metabolites in interaction networks has emerged as a powerful method of integration of these 'omics types to give more biologically relevant results [14]. The incorporation of different 'omics data types will continue to play an important role in providing a greater understanding of the molecular changes and regulatory programs underlying familial and sporadic AD.

The rapidly improving availability of high-throughput omics technologies implies that bioinformatics will also continue to play an important role in elucidating molecular mechanisms of complex diseases such as AD. A recurring theme of the manuscripts included in this work involved the use of bioinformatics to provide systems-level insights into relationships between gene expression patterns, molecular pathways, and regulatory factors in human AD and animal models. For example, in Chapter 2, a network representation of gene co-expression patterns was instrumental in

summarising the expression patterns of over 20,000 expressed genes to less than

30 interpretable modules of co-expressed genes in a zebrafish dataset. Using the

same technique on an independent human sporadic AD dataset allowed for a

systems-level comparison of biological processes represented by these modules

between these datasets. This revealed an important molecular mechanism

(microglial responses regulated by ETS transcription factors) preserved between the

zebrafish and human datasets and also independently replicated by another lab

using human datasets [15] (see Chapter 2 for more details). This finding was not

evident in a straightforward differential gene expression analysis of the datasets,

hence highlighting the importance of bioinformatics strategies that incorporate

analysis at multiple levels (e.g. promoter motif analysis to identify potential

transcription factors regulating a common group of genes, identifying gene co-

expression modules to group genes meaningfully) across different datasets to

ensure that findings are replicable and relevant to the human disease.

Comparison of clusters of co-expressed genes was not the only effective method for

data integration found, with Chapter 4 focusing on the use of gene sets for

summarising overall changes in biological activities between biological groups and

datasets. Typically, gene sets are defined based on prior knowledge and manual

curation of genes involved in a particular pathway or biological activity. The unique

approach taken here to investigate iron homeostasis changes at the gene

expression level (by defining gene sets computationally based on genes predicted to

contain Iron Responsive Element motifs) underscores the importance of tailoring a

bioinformatics strategy to the biological question being asked. The success of this

approach was demonstrated by finding that iron dyshomeostasis is implicated at the transcriptional level across human sporadic AD and two animal models ($psen1^{Q96\_K97del}$ zebrafish and 5XFAD mice) which was not captured by gene sets from existing databases. Similar approaches of defining gene sets based on predicted regulatory motifs might be useful to explore further a range of different regulatory systems at the transcriptional level, beyond those currently available in existing gene set databases. For example, sterol regulatory elements are present on genes involved in cholesterol homeostasis, a biological activity known to be disrupted in AD (reviewed in [16]) and also shown to be significantly enriched in familial AD mutation-like ($psen1^{Q96\_K97del}$) zebrafish (see Chapter 4). There is some evidence that sterol regulatory element-binding proteins are dysregulated by tau alterations in AD [17], but no gene set exists that contains all genes with sterol regulatory elements. In addition, AD-linked genes including *APP* and *PSEN1* are under the control of several regulatory elements [18], yet comprehensive gene sets containing other genes with these elements have not been defined. The definition of gene sets based on regulatory elements still appears to be an under-explored strategy for investigating regulatory systems at the transcriptome level. This strategy would be particularly relevant for further delineating molecular mechanisms and regulatory systems of complex diseases such as AD.

Currently, many different animal models of AD are available to researchers, with different models potentially recapitulating different aspects of AD pathology. Understanding the key differences between these models at the molecular level is an ongoing area of research. Through the manuscripts included here, it is evident that

bioinformatics will continue to play an important role in accomplishing this goal. The

integration of different datasets is a key area of current research in bioinformatics,

and Chapters 2 and 4 investigated two targeted approaches to exploration and

comparison of particular biological activities between zebrafish and human datasets

(using gene co-expression modules and gene sets respectively). The manuscript in

Chapter 5 extended this work and used a combination of advanced dimension

reduction (the AES-PCA approach) and predefined gene sets to compare broadly

the global transcriptional states of zebrafish, mouse, and familial and sporadic

human AD datasets while allowing for cross-species and cross-platform differences.

Overall, the approach was successful at performing preliminary integration of

different datasets in a straightforward and interpretable manner, whilst also providing

biological insights to generate further hypotheses.

Currently, data integration studies between animal model and human AD datasets

are not widely described in the literature despite their importance for assessing the

relevance of different animal models to AD. One of the only studies directly

comparing the brain transcriptomes between animal models of AD and human AD

was performed by Hargis and Blalock [19], where meta-analysis was used to assess

concordance of brain gene expression patterns between several transgenic mouse

models and human AD. The results indicated that different transgenic mouse models

of AD did not show concordance to each other, and no transgenic mouse model was

able to fully recapitulate the transcriptional changes characteristic of human AD. In

Chapters 4 and 5, it was further demonstrated that 5XFAD mouse brain

transcriptomes also differed extensively compared to familial AD mutation-like

zebrafish. Considering that many drugs developed to treat human AD have showed success when tested in transgenic mouse models but not in human AD (reviewed in [20]), it will be important in coming years to further compare a number of different transgenic and knock-in models of AD to each other and determine which models accurately capture important aspects of human AD.

Another limitation of the work shown here is the reliance on bulk tissue samples where results give insight into global transcriptional states while obscuring potential differences in transcriptional programs between different cell types. Single-cell RNA-seq technology is emerging as an important approach for overcoming this limitation. For the first time, a large-scale study in 2019 used single-cell RNA-seq to capture 80,660 cell transcriptomes over 48 individuals with varying amounts of AD pathology [21]. Importantly, the study showed that when AD pathology was not severe, many disease-associated gene expression changes involved in myelination and inflammation were highly cell-type specific. However, when AD pathology was severe (i.e. late stages of the disease), many of the gene expression changes that occurred were common across cell types and indicative of a global stress response [21]. Given that the results in Chapters 2, 3, 4, and 5 indicated important events occurring in young adult brains with relevance to human AD, it will be important in future work to delineate these global transcriptional states observed by understanding the cell types responsible for specific gene expression changes.

Overall, these manuscripts have described the first steps to characterisation of the early AD transcriptome through the use of zebrafish models. The results further our

understanding of familial and sporadic AD at the molecular level and are a valuable

source of hypothesis generation for future work. In addition, bioinformatics-focused

strategies were instrumental in extracting biologically relevant insights from these

datasets and aiding in their integration. Collectively, the work here represents an

initial contribution towards the ultimate goal of being able to develop therapeutic

interventions capable of slowing or preventing the molecular progression of AD using

a data-driven, bioinformatics-led approach.

References

1. Guo Q, Fu W, Sopher BL, Miller MW, Ware CB, Martin GM, et al. Increased vulnerability of hippocampal neurons to excitotoxic necrosis in presenilin-1 mutant knock-in mice. Nat Med. 1999;5:101–6. doi:10.1038/4789.

2. Kawasumi M, Chiba T, Yamada M, Miyamae-Kaneko M, Matsuoka M, Nakahara J, et al. Targeted introduction of V642I mutation in amyloid precursor protein gene causes functional abnormality resembling early stage of Alzheimer's disease in aged mice. Eur J Neurosci. 2004;19:2826–38. doi:10.1111/j.0953-816X.2004.03397.x.

3. Siman R, Reaume AG, Savage MJ, Trusko S, Lin YG, Scott RW, et al. Presenilin-1 P264L knock-in mutation: Differential effects on Aβ production, amyloid deposition, and neuronal vulnerability. J Neurosci. 2000;20:8717–26.

4. Reaume AG, Howland DS, Trusko SP, Savage MJ, Lang DM, Greenberg BD, et al. Enhanced amyloidogenic processing of the β-amyloid precursor protein in gene-targeted mice bearing the Swedish familial Alzheimer's disease mutations and a "humanized" Aβ sequence. J Biol Chem. 1996;271:23380–8. doi:10.1074/jbc.271.38.23380.

5. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, et al. Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia. Proc Natl Acad Sci U S A. 2004;101:284–9. doi:10.1073/pnas.2635903100.

6. Kim EJ, Cho SS, Jeong Y, Park KC, Kang SJ, Kang E, et al. Glucose metabolism in early onset versus late onset Alzheimer's disease: An SPM analysis of 120 patients. Brain. 2005;128:1790–801. doi:10.1093/brain/awh539.

7. Mosconi L. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease: FDG-PET studies in MCI and AD. Eur J Nucl Med Mol Imaging. 2005;32:486–510. doi:10.1007/s00259-005-1762-7.

8. Lumsden AL, Rogers JT, Majd S, Newman M, Sutherland GT, Verdile G, et al. Dysregulation of neuronal iron homeostasis as an alternative unifying effect of mutations causing familial alzheimer's disease. Front Neurosci. 2018;12 AUG:533.

9. Barthelson K, Pederson S, Newman M, Lardelli M. Brain transcriptome analysis reveals subtle effects on mitochondrial function and iron homeostasis of mutations in the SORL1 gene implicated in early onset familial Alzheimers disease. bioRxiv. 2020;:2020.07.17.207787. doi:10.1101/2020.07.17.207787.

10. Xia D, Watanabe H, Wu B, Lee SH, Li Y, Tsvetkov E, et al. Presenilin-1 knockin mice reveal loss-of-function mechanism for familial alzheimer's disease. Neuron. 2015;85:967–81. doi:10.1016/j.neuron.2015.02.010.

11. Saito T, Matsuba Y, Mihira N, Takano J, Nilsson P, Itohara S, et al. Single App knock-in

mouse models of Alzheimer's disease. Nat Neurosci. 2014;17:661–3. doi:10.1038/nn.3697.

12. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci Rep. 2015;5. doi:10.1038/srep10775.

13. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016;165:535–50. doi:10.1016/j.cell.2016.03.014.

14. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet. 2019;10 JUN:535. doi:10.3389/fgene.2019.00535.

15. Zhang B, Gaiteri C, Bodea L-G, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2013;153:707–20. doi:10.1016/j.cell.2013.03.030.

16. Chang TY, Yamauchi Y, Hasan MT, Chang C. Cellular cholesterol homeostasis and Alzheimer's disease. J Lipid Res. 2017;58:2239–54. doi:10.1194/jlr.R075630.

17. Wang C, Zhao F, Shen K, Wang W, Siedlak SL, Lee H gon, et al. The sterol regulatory element-binding protein 2 is dysregulated by tau alterations in Alzheimer disease. Brain Pathol. 2019;29:530–43. doi:10.1111/bpa.12691.

18. Theuns J, Van Broeckhoven C. Transcriptional regulation of Alzheimer's disease genes: Implications for susceptibility. Human Molecular Genetics. 2000;9 16 REV. ISS.:2383–94. doi:10.1093/hmg/9.16.2383.

19. Hargis KE, Blalock EM. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? Behav Brain Res. 2017;322 Pt B:311–28.

20. Drummond E, Wisniewski T. Alzheimer's disease: experimental models and reality. Acta Neuropathologica. 2017;133:155–75.

21. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570:332–7. doi:10.1038/s41586-019-1195-2.