Artificial intelligence-driven meta-analysis of brain gene expression data identifies novel gene candidates in Alzheimer's Disease

Caitlin A. Finney [1*], Fabien Delerue [1], & Artur Shvetcov [2*]

1. Dementia Research Centre, Department of Biomedical Science, Faculty of Medicine Health and Human Sciences, Macquarie University, Australia

2. Black Dog Institute, Australia

* Authors to whom correspondence should be addressed. E-mails: a.shvetcov@blackdog.org; caitlin_finney@hotmail.com

Abstract

Microarrays have identified thousands of dysregulated genes in the brains of patients with Alzheimer's disease (AD); yet identifying the best gene candidates to both model and treat AD remains a challenge. To this end, we performed a meta-analysis of microarray data from the frontal cortex and cerebellum of AD patients, followed by an artificial-intelligence driven approach to identify the top AD gene candidates. In the frontal cortex, gene candidates included mitochondrial complex V subunits: ATP5J, ATP5L and ATP5H. In the cerebellum, the top candidate was SC5D, which is involved in cholesterol biosynthesis and catabolism. An important finding was that there was no overlap in dysregulated pathways between the frontal cortex and cerebellum, suggesting that pathophysiological mechanisms in AD are brain-region specific. Combined, these results have significant implications for future models and therapeutic strategies in AD.

Expression profiling through microarray analysis is a valuable high-throughput tool that allows researchers to determine quantitative gene expression for a large number of mRNA transcripts [1]. Recent meta-analyses of publicly available microarray data have proven to be beneficial for identifying novel contributing genes for several diseases including influenza [2], atherosclerosis [3], chronic pain [4], cancer [5,6], and Parkinson's disease [7]. To date, however, few studies have used this approach to identify novel gene targets in the Alzheimer's disease (AD) brain. One such study took a comparative approach to find overlapping gene expression profiles in neurodegenerative disorders including AD, Lewy body disease, amyotrophic lateral sclerosis and frontotemporal dementia [8]. Another took a similar approach to compare cross-species transcriptional overlap between mouse models of AD and humans [9]. These meta-analyses, however, are limited in two ways. First, human transcriptomic data is high-dimensional and complex, making it difficult to unravel hidden patterns in the datasets [10]. Second, these studies identified hundreds to thousands of dysregulated genes in AD. Combined, these limitations leave researchers with the question of how to determine the best target(s) for either developing new mouse and cell models of AD or treating AD.

The use of artificial intelligence (AI) is a way to overcome these limitations. Importantly, not only is AI able to unravel patterns within complex data in an unbiased way [10], it also has the ability to reveal which gene target(s) should be investigated further. In fact, machine learning models have successfully been used to differentiate Parkinson's disease patients from healthy controls based on gene expression profiles [11]. Further, a recent study used AI to analyze transcriptomic data and identified a novel target that resulted in the successful treatment of inflammatory bowel disease [12]. This approach has also been attempted in the context of AD. Here, researchers used a combination of random forest and least absolute shrinkage and selection operator (LASSO) as a method of feature selection to identify "biologically relevant" differentially expressed genes in the brains of AD patients [13]. However, the genes identified had few, and often no, protein-protein interactions in STRING, thus raising questions about their true biological relevance [13]. Further, the study performed machine learning on a very small number of samples, likely leading to overfitting of the model and a high risk of bias [14]. As such, there remains a clear need to establish machine learning-driven methods for identifying biological relevant genes in AD.

To identify the best, biologically relevant gene targets in AD, we performed a meta-analysis of an unprecedented number of microarray data from the frontal cortex and cerebellum of patients with AD compared to healthy controls. Using unsupervised machine learning, via principal component analysis (PCA), we identified the genes that had the highest contribution to between-group separation between AD patients and healthy controls and thus were more likely to be dysregulated genes in AD. These genes were then analyzed using STRING to determine biologically relevant genes that have clearly established interaction networks. Supervised machine learning was then performed on the identified pathways and genes in each brain region to a) determine the pathways that were the best predictors of AD and b) rank the genes within the pathways based on their contribution to the machine learning model. We identified several dysregulated pathways in both the frontal cortex and cerebellum and, interestingly, no overlapping pathways between the two brain regions. The supervised machine learning model determined that in the frontal cortex there were several dysregulated gene candidates in AD including those involved in mitochondrial ATP and oxidative phosphorylation, signaling, cellular metabolic processes and mitochondrial ribosomal protein synthesis pathways. In the cerebellum, important candidate genes included those involved in development and differentiation, sterol and steroid metabolic processes, and intracellular transport. Overall, our findings indicate that these pathways, and their respective genes, represent ideal, biologically relevant candidates for developing both novel animal and cell models of, and treatments for, AD. Further, as there was no overlap between the frontal cortex and cerebellum, our findings point to regionally specific AD neuropathology and suggest that treatments and models may need to take a multifaceted approach to AD.

Results

Meta-analysis and unsupervised machine learning identifies novel dysregulated pathways and genes in AD

Two microarray datasets were re-analyzed and combined into a single frontal cortex dataset using a z-score approach (see *Methods*). For the cerebellum, only a single dataset was used and, similarly, gene expression was also represented by a z-score. Unsupervised machine learning was employed via a principal component analysis (PCA) of each dataset, respectively. This demonstrated that there was clear clustering between the AD patient and healthy control

groups (Figure 1 A and B). In each structure, the top 1000 genes that contributed the most to between-group variance, as indicated by principal component 1 (PC1), were then identified as the most dysregulated genes in AD (Supplementary Data). To determine the number of common pathways represented by each structure's 1000 genes, a second unsupervised machine learning approach was used: k-means clustering using STRING. Here, k-means clustering is a way to identify genes in the network that have similar interconnections and overlapping pathways [15].

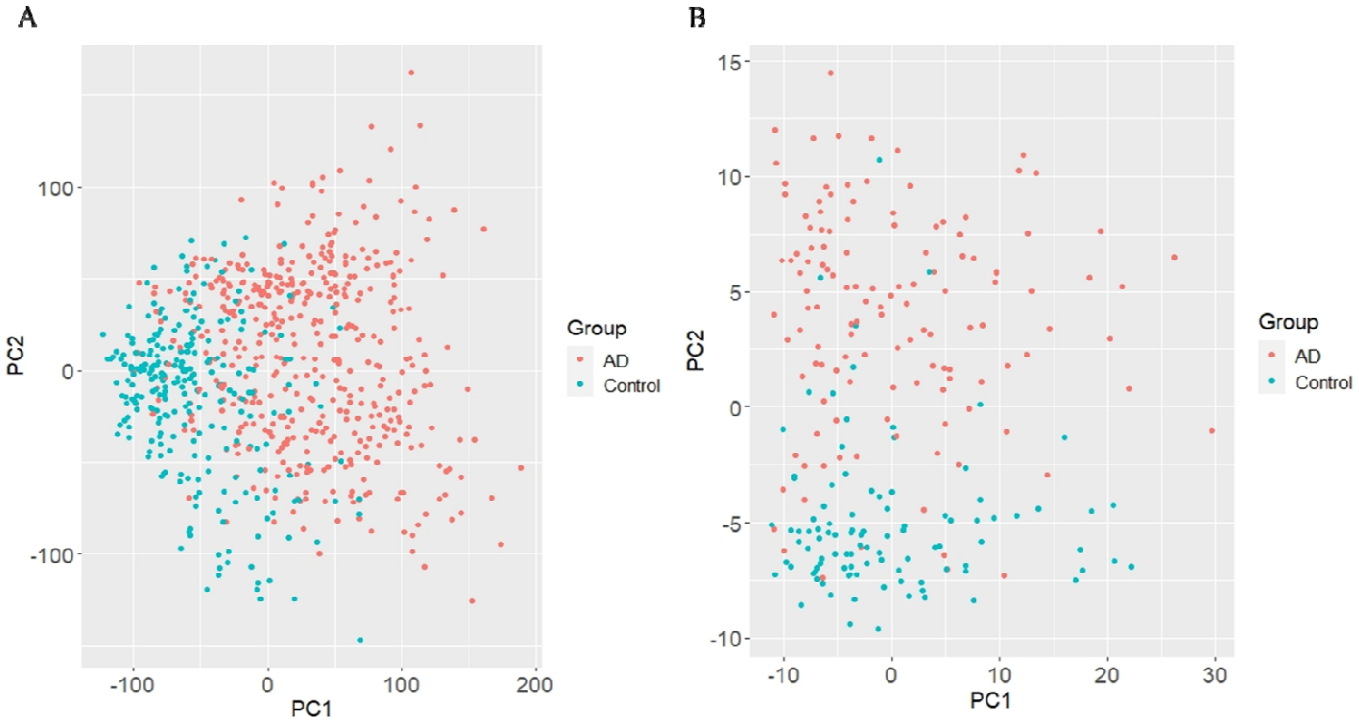


Figure 1. Principal component analysis of genes identified in the microarray meta-analysis revealsa clear between-group separation between the AD patients (red) and healthy controls (blue). **(a)** Frontal cortex. **(b)** Cerebellum.

In the frontal cortex, there were three distinct k-means clusters identified (Figure 2A) and we used Gene Ontology (GO) to characterize the biological functions of each. The first was characterized by signaling processes (Supplementary Figure 1). Within this cluster, 15 central gene nodes were identified that play important roles in voltage-dependent calcium channels, guanine nucleotide-binding protein (G protein) formation and activity, AMPA receptor, cAMP-dependent protein kinase A (PKA), and the SNARE complex (Supplementary Table 1). The second frontal cortical cluster was characterized by metabolic processes relating to

macromolecules, proteins, and DNA (Supplementary Figure 2). Here, there were 26 genes identified as being central nodes in the network with extensive roles in general cellular metabolic processes including DNA repair, transcriptional regulation, ubiquitin pathway regulation, and apoptosis (Supplementary Table 2). The third k-means cluster identified in the frontal cortex was related to mitochondrial processes. Within the overall cluster, there were two distinct sub-clusters identified based on biological function. The first was related to mitochondrial-specific energy, ATP, and oxidative phosphorylation and had 36 central nodes in the network (Supplementary Figure 3). These genes were all related to different mitochondrial subunits including: the F_0-F_1 ATP synthase (mitochondrial complex V, F-ATPase), V-ATPase, mitochondrial complex I, and mitochondrial complex III (Supplementary Table 3). The second sub-cluster within the third k-means cluster was defined by mitochondria-mediated cellular biosynthesis (Supplementary Figure 3). This was a smaller sub-network characterized by 10 central nodes and all genes were specific to the nuclear-encoded mitochondrial ribosomal 39S or 28S subunits, which play a central role in protein synthesis in the mitochondrion (Supplementary Table 4).

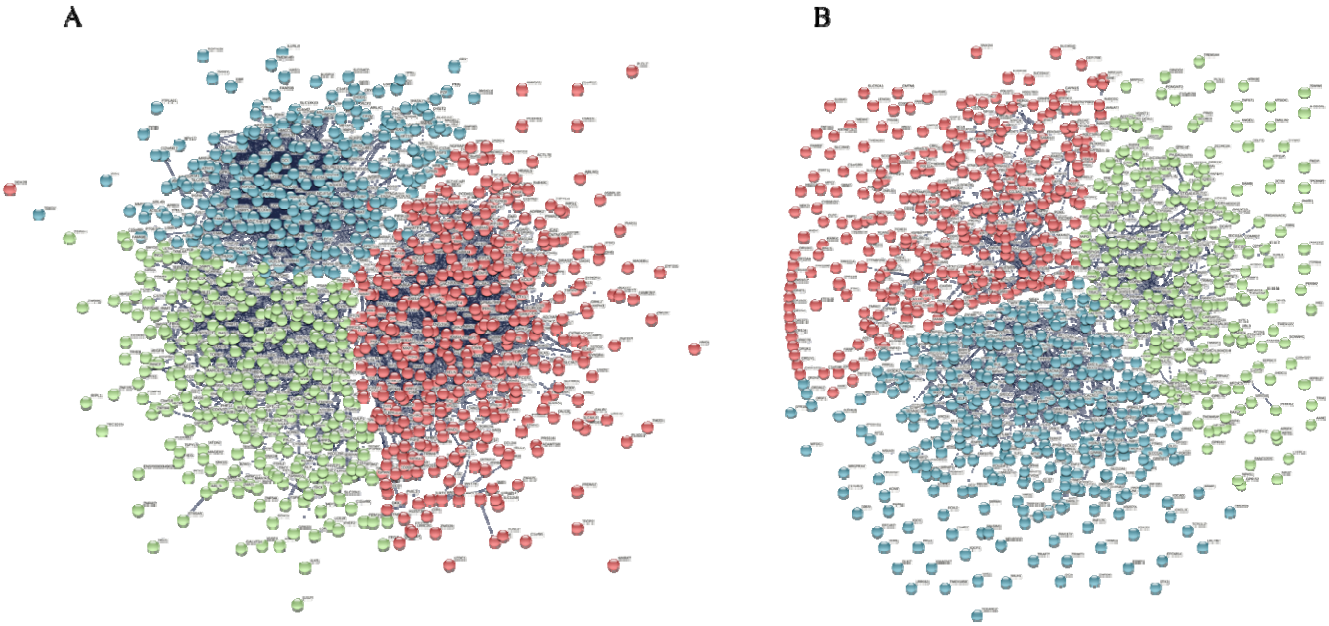


Figure 2. STRING k-means clustering of the top 1000 genes in AD identified by PC1 of the PCA. **(a)** Frontal cortex k-means clustering showed three distinct clusters of genes. Red n = 420, green n = 294, blue n = 286. **(b)** Cerebellum k-means clustering showed three distinct clusters of genes. Red n = 320, green n = 361, blue n = 319.

Similar to the frontal cortex, there were also three distinct k-means clusters identified in the cerebellum (Figure 2B). Unlike the frontal cortex, however, the clusters included different genes and represented different pathways that were unique to the cerebellum. The first cluster was specific to biological processes related to extracellular exosomes and vesicles (Supplementary Figure 4). It was small, with only three central nodes identified that were involved in the ESCRT-1 complex, a cytoplasmic protein tyrosine kinase and a GTPase belonging to the Ras superfamily, respectively (Supplementary Table 5). The second k-means cluster was characterized by genes involved in development and differentiation (Supplementary Figure 5). Seven central nodes in the network had a range of functions such as calcium signaling, transcription, cell cycle regulation, chromatin remodeling and transcriptional regulation (Supplementary Table 6). The third identified k-means cluster comprised four network sub-clusters. The first sub-cluster related to biological processes of protein glycosylation and glycoprotein biosynthesis (Supplementary Figure 6). The four central nodes identified within this sub-cluster were all involved in mucins (Supplementary Table 7). The second sub-cluster was specific to membrane lipid metabolic processes (Supplementary Figure 6). Only two central nodes were found, with genes involved in membrane-associated PAP activity and ceramide synthesis, respectively (Supplementary Table 8). The third sub-cluster was related to sterol and steroid metabolic processes (Supplementary Figure 6). This sub-cluster comprised five central nodes involved in cholesterol and sterol biosynthesis, and C-4 methylsterol oxidase activity (Supplementary Table 9). The fourth, and final, sub-cluster was defined by biological functions specific to intracellular transport (Supplementary Figure 6). There were six central gene nodes, playing roles in SNARE and BLOC-1 complexes, intra- and subcellular trafficking, and endosomal sorting (Supplementary Table 10).

Supervised machine learning reveals the top pathways and candidate genes for AD

The central gene nodes identified by the k-means clusters and sub-clusters within each brain structure were subsequently analyzed using supervised machine learning. A classification and regression tree (CART) algorithm was applied to each cluster, respectively, to compare the performance of the clusters in predicting AD. Performance indicators used included specificity (correctly identifies healthy control), sensitivity (correctly identifies AD patient) and accuracy (ratio of correct AD identifications to total number of samples).

Using these metrics, the top performing cluster for the frontal cortex was related to mitochondrial energy, ATP, and oxidative phosphorylation. This was followed very closely by the synaptic signaling, metabolic processes, and mitochondrial cellular biosynthesis clusters (Table 1). Within each cluster's model, we then identified the top three gene candidates by determining the correlation with output. For the mitochondrial energy, ATP, and oxidative phosphorylation cluster the top three genes were: ATP5J, ATP5L and ATP5H. All three genes encode different subunit of the F_0 complex of mitochondrial complex V including the F6, g, and d subunits. For the signaling cluster, GNG3 and GNG5 encode the γ and β subunits of heterotrimeric guanine nucleotide-binding (G-) proteins. SYT1 is a synaptic vesicle membrane that triggers neurotransmitter release at the synapse. The three top gene candidates for the metabolic processes cluster were COPS7A and COPS8, both of which play a regulatory role in ubiquitin pathways, and PSMB7, which encodes a subunit of proteosome 20S. The top three gene candidates in the mitochondrial cellular biosynthesis cluster were MRPS18A, MRPL15, and MRPS35, all of which encode mitochondrial ribosomal proteins (S18A, L15, and S35, respectively) (Table 1). The rankings and correlation with output calculations for the remaining genes in each k-means cluster and sub-cluster are listed in Supplementary Table 11.

Table 1. CART precision metrics for the k-means clusters and sub-clusters in the AD frontal cortex.

Cluster	**Sensitivity**	**Specificity**	**Accuracy**	**Top Three Gene Candidates**
Mitochondrial energy, ATP and oxidativephosphorylation	0.83	0.93	0.85	ATP5J, ATP5L, ATP5H
Signaling	0.83	0.92	0.86	GNG3, GNB5, SYT1
Metabolic processes	0.82	0.91	0.84	COPS7A, PSMB7, COPS8
Mitochondrial cellular biosynthesis	0.81	0.88	0.83	MRPS18A, MRPL15, MRPS35

In the cerebellum, CART indicated that only three k-means clusters and sub-clusters met the threshold (sensitivity $\geq$ 0.7) for precision modeling. The top performing model here involved the sub-cluster related to sterol and steroid metabolic processes. This was followed closely by the intracellular transport cluster and, to a lesser extent, the development and differentiation cluster (Table 2). Unlike in the frontal cortex, there were very few central nodes within each k-means cluster and sub-cluster in the cerebellum. As such, we then used correlation with output to

identify only a single top gene candidate in each cluster. For the sterol and steroid metabolic processes cluster, this was SC5D, which is involved in cholesterol biosynthesis and catalyzes the conversion of lathosterol into 7-dehydrocholesterol. In the intracellular transport cluster, the top gene candidate was SEC22B, a member of the SEC22 family of vesicle trafficking proteins that complexes with SNARE and plays a role in endoplasmic reticulum-Golgi protein trafficking. In the final cluster, development and differentiation, PAXIP1 was the top performing gene, playing a role in maintaining genome stability, mitosis progression and condensation of chromatin (Table 2). The rankings and correlation with output calculations for the remaining gene nodes in each of the clusters are listed in Supplementary Table 12.

Table 2. CART precision metrics for the k-means clusters and sub-clusters in the AD cerebellum.

Cluster	Sensitivity	Specificity	Accuracy	Top Gene Candidate
Sterol and steroid metabolic processes	0.76	0.84	0.81	SC5D
Intracellular transport	0.76	0.93	0.85	SEC22B
Development and differentiation	0.70	0.96	0.82	PAXIP1
Extracellular exosomes and vesicles	0.59	0.79	0.71	N/A – poor model
Protein glycosylation and glycoprotein biosynthesis	0.58	0.80	0.69	N/A – poor model
Membrane lipid metabolic processes	0.47	0.93	0.69	N/A – poor model

No overlap in pathways or gene candidates between the frontal cortex and cerebellum

Our findings clearly established a dichotomy between the frontal cortex and cerebellum in AD, with no common clusters, pathways, nodes, or gene candidates. Dysregulated clusters in the frontal cortex were related to the mitochondrion, synaptic signaling and general metabolic processes relating to macromolecules, protein and DNA. In the cerebellum, however, critical dysregulated clusters in AD centered around sterol and steroid metabolism, intracellular transport and development and differentiation. To further confirm that there was no overlap in dysregulated genes in the frontal cortex and cerebellum, we took the top 1000 genes previously identified in PC1 of the PCA for each region and compared them. We found that there were 32 overlapping genes between both datasets (Supplementary Table 13). Despite this, a STRING

network analysis showed that there was no interaction between the genes (Supplementary Figure 7).

Discussion

Alzheimer's disease is the most common form of dementia: 50 million people have dementia globally, with almost 10 million additional cases yearly [16]. There are currently no effective treatments for AD and over 99.6% of clinical trials of AD have failed thus far [17]. Further, most of what is currently known about AD has been established using familial AD (FAD) models, characterized by mutations in amyloid precursor protein (APP), presenilin 1 (PSEN1) and presenilin 2 (PSEN2), which account for less than 1% of AD cases [18-21]. Identifying novel gene candidates in AD is challenging. We approached this challenge by performing a meta-analysis of microarray data from the frontal cortex and cerebellum of patients with AD. We then used an artificial intelligence-driven method, specifically a combination of unsupervised and supervised machine learning, to identify novel gene candidates for future study.

Modelling of the frontal cortex found four dysregulated pathways in AD: 1) mitochondrial energy, ATP, and oxidative phosphorylation 2) signaling 3) metabolic processes and 4) mitochondrial cellular biosynthesis (Figure 3). The top performing model, however, was for mitochondrial energy, ATP, and oxidative phosphorylation. Here, the top three gene candidates included ATP5J, ATP5L and ATP5H, all of which encode subunits of the F_0 functional domain of mitochondrial complex V (also known as F_1F_0 ATP synthase). Complex V is the final complex in the electron transport chain and produces ATP through the phosphorylation of ADP [22-24]. Mitochondria are increasingly shown to contribute to the development and progression of AD, with evidence suggesting both primary and secondary dysfunctional mitochondrial cascades (for reviews see [25,26]). Specifically, mitochondrial dysfunction not only affects AD pathology, including APP activity and β amyloid (Aβ) accumulation, but AD pathology also leads to further mitochondrial dysfunction [26]. Little research has examined a role for complex V in AD, with most studies looking at it only as a consequence rather than a driver of AD pathology (for reviews see [23,24]). In support of the idea that complex V dysfunction may precede AD neuropathology, two recent genome-wide association study (GWAS) meta-analyses of approximately 25,00 and 50,000 people,

respectively, identified a shared ATP5H/KCTD2 locus for AD risk [27,28]. Another study using a C9ORF72-ALS/FTD mouse model found that mitochondrial dysfunction preceded neurodegeneration and this effect was driven by the complex V subunit ATP5A1 [29]. Interestingly, a recent study also demonstrated that disruption of mitochondrial complex I in dopaminergic neurons induced progressive parkinsonism in mice, for the first time showing that mitochondrial disruption is sufficient to induce neurodegeneration [30]. This evidence, combined with our findings, strongly suggest that future research should elucidate the role of mitochondrial complex V, and its subunits, in AD.

Although the mitochondrial energy, ATP, and oxidative phosphorylation pathway was the best predictor of AD, it is worthy to note that the other three pathways were close behind with respect to performance metrics. This suggests that the genes identified within the signaling, metabolic processes and mitochondrial cellular biosynthesis pathways may also be good candidate AD genes in the frontal cortex (Figure 6). Future research, therefore, would also benefit from examining the role of these genes further in AD pathogenesis and as possible treatment targets.

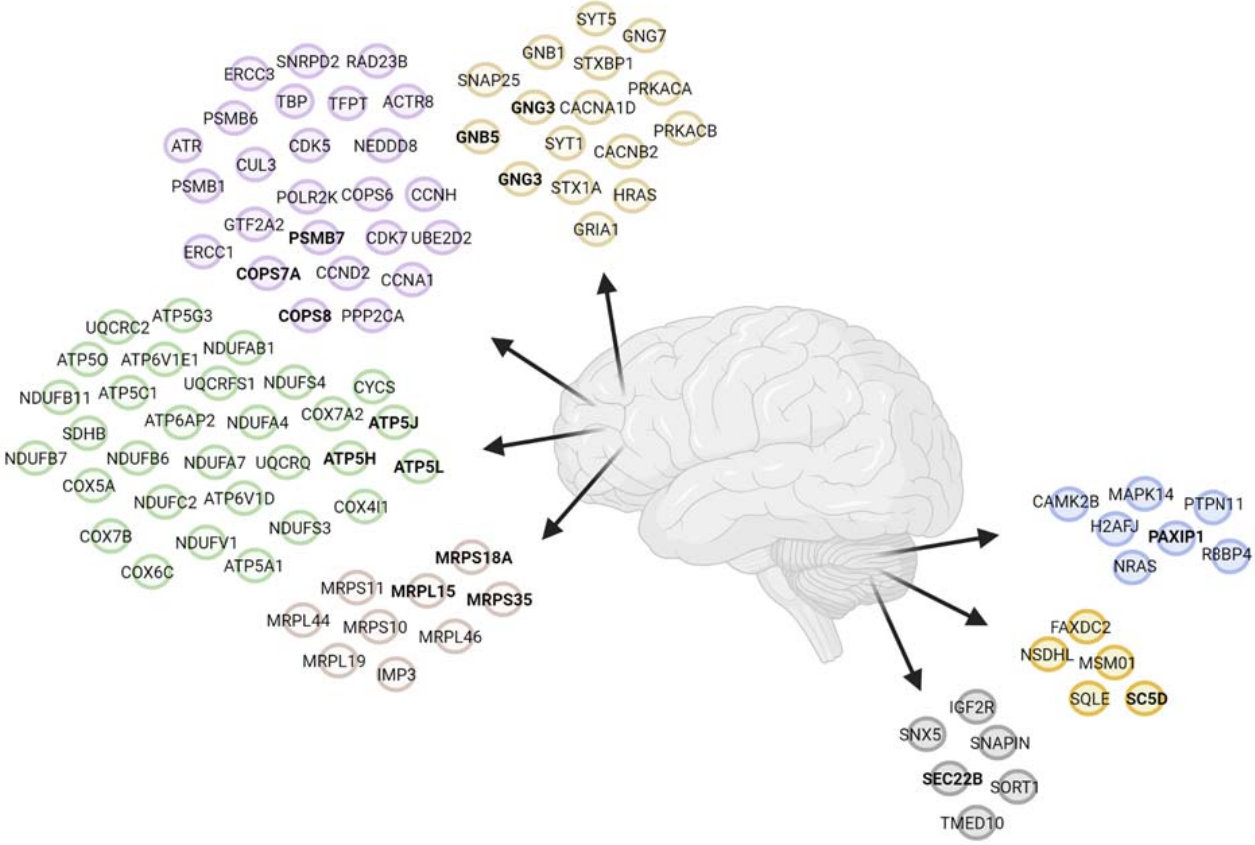


Figure 3. Summary of the identified novel gene candidates in AD in both the frontal cortex and cerebellum. In the frontal cortex, dysregulated pathways included: signaling (tan), metabolic processes (purple), mitochondrial energy, ATP, and oxidative phosphorylation (green), and mitochondrial cellular biosynthesis (brown). In the cerebellum, dysregulated pathways included: development and differentiation (blue), sterol and steroid metabolic processes (yellow), and intracellular transport (grey). Each circle represents a central gene node within each network. Gene nodes in bold font indicate the top performer(s) and candidate(s) based on modelling.

Cerebellar CART modelling found only three dysregulated pathways in AD: 1) sterol and steroid metabolic processes 2) intracellular transport and 3) development and differentiation (Figure 3). Importantly, despite genes in the other three pathways (extracellular exosomes and vesicles, protein glycosylation and glycoprotein biosynthesis and membrane lipid metabolic processes) being among the top 1000 dysregulated genes, they were poor predictors of AD and were unable to discriminate between a patient with AD and a healthy control. This highlights the idea that even if genes are identified as being dysregulated in AD, they may not be good target candidates. Despite having lower performance metrics than the frontal cortex, sterol and steroid

metabolic processes performed the best in the cerebellum. Here, the top gene candidate was SC5D, which is involved in cholesterol biosynthesis and specifically the conversion of lathosterol into 7-dehydrocholesterol. Few studies have examined a role for brain cholesterol biosynthesis and catabolism in AD. A recent study found that brain cholesterol homeostasis was altered in patients with AD, with SC5D levels shown to be negatively associated with neuritic plaque burden and neurofibrillary tangle pathology [31]. Another study using an SC5D knockout mouse model of Smith-Lemli-Opitz syndrome (SLOS), a malformation syndrome with cognitive deficits, found that the accumulation of sterol precursors like lathosterol had neurotoxic effects [32]. Future research would benefit from continuing to examine a role for SC5D in AD. The remaining two cerebellar CART models were relatively low performing in comparison and may represent additional gene targets for consideration in AD.

An important finding from our study was that there was no overlap in dysregulated pathways between the frontal cortex and cerebellum. Although there was an overlap of 32 genes, a STRING analysis indicated that they were disconnected nodes with no protein-protein interactions and networks, demonstrating that the overlap is likely noise. This finding suggests that AD pathophysiological mechanisms may be regionally specific, which has significant implications for how we both study and treat AD. First, it highlights the importance of comparing effects (e.g. treatments, genotypes, etc.) across multiple brain regions in both pre-clinical and clinical studies, rather than focusing only on a single region. A second implication of this finding is that a one-size-fits-all approach to treating AD neuropathology may be inappropriate. For example, treating dysregulated cholesterol homeostasis may address cerebellar but not frontal cortical AD pathology. It may be the case that this effect is due to varying levels of disease load across brain regions, such as amyloid plaque deposition spreading to the cerebellum only in the end stages of AD [33]. This, however, does not preclude the suggestion that AD treatments need to be multifaceted, with multiple targets, to adequately address widespread AD pathology across distinct regions.

In summary, our study reports an artificial intelligence-driven identification of novel gene candidates in the frontal cortex and cerebellum of patients with AD. Our findings highlight the importance of genes involved in mitochondrial complex V in the frontal cortex and sterol and steroid metabolism in the cerebellum in AD. Further, our findings also suggest that

pathophysiological mechanisms in AD are brain region specific. This has significant implications for future therapeutic strategies in AD.

Online Methods

Systematic Review of Publicly Available Data Repositories

To identify publicly available transcriptomic datasets, a systematic review of the Gene Expression Omnibus (GEO) database was performed. The key search terms used included "Alzheimer's disease" and "homo sapiens". Datasets were screened on the basis of the following inclusion criteria: (a) gene expression data generated using microarray platforms, (b) gene expression specific to the amygdala, hippocampus, entorhinal cortex, frontal cortex, temporal cortex or cerebellum, (c) clinically confirmed Alzheimer's disease patients and (d) inclusion of cognitively normal healthy controls. Datasets were excluded for the following: (a) use of other high throughput gene expression assays (e.g. RNA-sequencing), (b) gene expression in the periphery (e.g. blood), (c) not specifying confirmed AD diagnosis, and (d) not including cognitively healthy controls (e.g. controls with mild cognitive impairment).

Identification and Meta-Analysis of AD Microarray Datasets

A total number of 1010 datasets in GEO were screened. Based on the inclusion and exclusion criteria, 14 datasets were identified as being eligible for inclusion in the meta-analysis. The meta-data of each dataset was analyzed to determine if there was a sufficient sample size to undertake further analyses and machine learning. Of these datasets, 11 were excluded due to an insufficient sample size. Consequently, we were unable to include the following brain regions in the analyses: the amygdala, hippocampus, entorhinal cortex, and temporal cortex. Future work would benefit from focusing on expanding microarray expression data for these regions to enable large-scale analyses and the application of artificial intelligence methods.

Two brains regions were included in the current meta-analysis: the frontal cortex and cerebellum. Data from the frontal cortex originated from two GEO datasets: GSE44770 and GSE33000. Data for the cerebellum originated from the GEO dataset GSE44768. Samples from the studies consisted of AD patients, with confirmed antemortem clinical diagnosis and postmortem neuropathological assessment, and normal, non-demented, healthy controls [34,35].

After identifying the respective datasets for the frontal cortex and cerebellum, we then converted all normalized gene expression values into a z-score. For the frontal cortex only, genes common to both datasets were selected and then the two datasets were merged into one. Data processing and merging was done in R Studio v1.2.5033 (R v3.6.3) with packages GEOquery, and dplyr. The final, unified frontal cortex dataset had a total sample size of n = 697 (AD n = 439, healthy control n = 238). The cerebellum dataset had a total sample size of n = 230 (AD n = 129, healthy control n = 101).

Machine Learning and STRING Network Analysis

The conventional approach for identifying differentially expressed genes typically uses fold change, *p*-values, or a combination of the two. These methods, however, are limited and are unlikely to provide the information needed to make strong conclusions about dysregulated genes that may be important for AD. First, using standard fold change cutoffs to select differentially expressed genes is inherently problematic. Genes with either low, or high, absolute expression are more likely to either easily meet or miss the fold change threshold, respectively, regardless of whether or not the gene is truly differentially expressed [36,37]. Further, the *p*-value is well-known to not only provide limited information about the data at hand, but to also be easily misinterpreted, thus likely contributing to the replication crisis [38-40]. Specifically, calculating statistical significance using a *p*-value does not account for the degree to which the genes are, or are not, involved in differences between groups (in this case, between AD patients and healthy controls). One way to circumvent the inherent limitations of conventional approaches is to approach the problem of identifying *truly* important genes through the lens of artificial intelligence, such as machine learning. In other words, we treat the problem as a machine learning problem: we would like to determine which genes best predict the classification of a given sample as being from either an AD patient or healthy control.

In the present study, we used a two-stage machine learning approach (Figure 4). In the first stage, we used unsupervised machine learning to perform an initial feature selection: identifying the genes that are likely to be important candidates for distinguishing an AD patient from a healthy control. Feature selection is an important step to reduce the number of features and thus avoid the curse of dimensionality and reducing the computational cost for the final machine learning algorithm [41]. Unsupervised machine learning is unique in the sense that it does

not predefine any sample as being either an AD or healthy control sample. Instead it identifies similarities between the various samples that exist, irrespective of their group membership [42]. Samples with high similarity will cluster together and will inform us how the data is grouped and what the drivers of this differentiation are [42]. If true differences exist between AD patients and healthy controls, the unique clusters will be representative of this, and we will identify which genes are driving these differences. The unsupervised machine learning approach used here was a principal component analysis (PCA). This is an important technique that reduces dimensionality within a dataset while simultaneously minimizing any information loss [43]. PCA has an established use in the analysis of high throughput datasets, such as microarray, to reveal hidden patterns within the thousands of identified genes [44-46]. All the genes ($\geq$ 15,000 genes per dataset) that were identified within the frontal cortex and cerebellum datasets, respectively, were analyzed using PCA in R Studio v1.2.5033 (R v3.6.3) and visualized using ggplot (Figure 4). Principal component 1 (PC1) in a PCA represents all the genes that are the highest contributors to the clustering between groups. As such, the top 1000 genes that contributed to PC1 were used as potential gene candidates for AD (Figure 4). To further narrow down the list of gene candidates, we entered the list of 1000 genes into STRING v11 [47]. Importantly, STRING allows for the identification of interaction networks and gene-enrichment analysis [47]. We then identified possible distinct network clusters using k-means clustering in STRING (Figure 4). K-means clustering is an unsupervised machine learning approach which, here, is a way to identify genes in the network that have similar interconnections and overlapping pathways [15]. Subsequently, each distinct k-means cluster for the two brains regions were separately entered into STRING and a network analysis was performed using the following active interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion and a minimum required interaction score of 0.7 (high confidence) (Figure 4). Importantly, identifying pathways and interaction networks increases the likelihood of identifying strong, biologically relevant gene candidates for AD from a wealth of literature and experimental data. More specifically, this approach allows us to target biologically relevant pathways rather than stand-alone genes that may not have any evidence of interactions thus increasing the chance of network effects and achieving a response in AD. In STRING, each k-means cluster were characterized using biological processes identified by Gene Ontology [48]. We then selected the central node(s) in each k-means cluster based on their connections with other genes in the network, with identified

central node(s) having the highest number of connections (Figure 4). Critically, the use of central node(s) in the networks allows us to identify gene candidates that are fundamentally important to complex networks and metabolic pathways, further increasing the likelihood of identifying biologically relevant targets in AD.

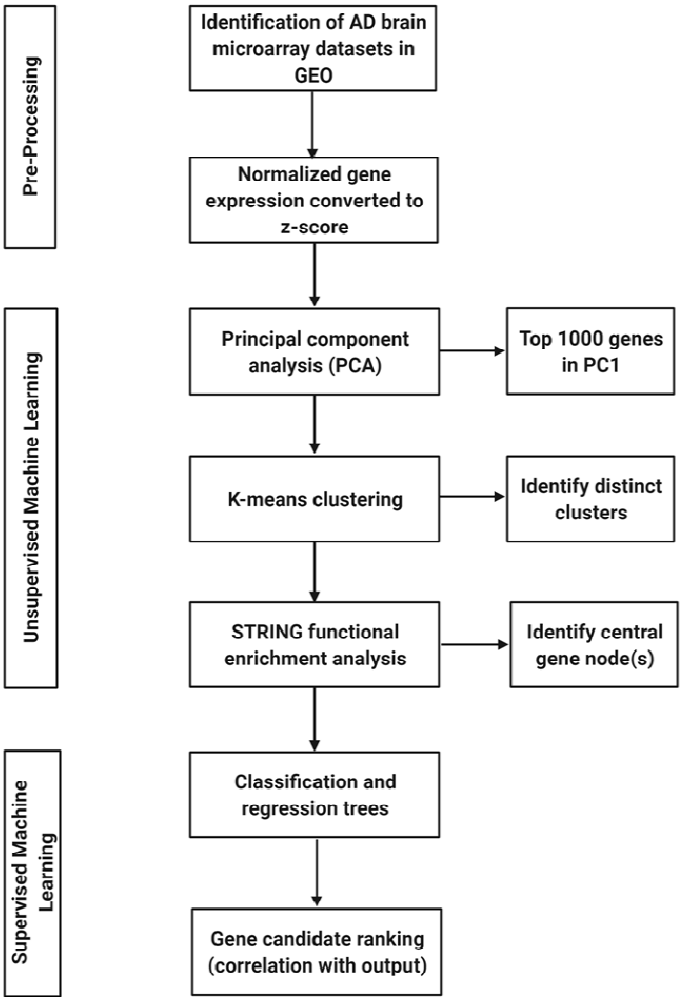


Figure 4. AI workflow used in the current study to identify new AD related genes.

After identifying the central node(s) in each k-means cluster for the frontal cortex and cerebellum, respectively, we then undertook the second stage of our two-stage machine learning approach: supervised machine learning using decision trees (classification and regression trees (CART)) (Figure 4). CART is a classification tool known as a white box model: it can identify which factors are important for the performance of the model [49]. In this case, it identifies which

genes are best able to separate AD patients from healthy controls within the machine learning model. The use of CART has been well-established in large clinical and public health projects characterized by high-dimensional, heterogeneous data [49,50]. In the current study, the central gene node(s) for each k-means cluster were used as features in the CART. The datasets for frontal cortex and cerebellum were each split into training (75%) and testing (25%) datasets. Training, tuning, and validating the model was done on the training dataset. Here, a 5-fold cross-validation was repeated three times to improve the accuracy estimates of the models [51]. Cross-validation was used also as a tool for identifying the top performing gene predictors. The final evaluation of the CART model performance was done on the previously withheld testing dataset. For further analysis, gene predictors from the model with the best fit were used and correlation with output was calculated. CART was performed in R Studio v1.2.5033 (R v3.6.3) with libraries rpart and caret.

Acknowledgements

This work was supported by a Macquarie University Research Acceleration Scheme Grant (MQRAS 173983854) awarded to C.A.F. and A.S.

Author Contributions

C.A.F. and A.S. conceived the experiment, performed the analyses and wrote the manuscript.

F.D. provided substantial input on both the experiments and manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Stears, R.L., Martinsky, T. & Schena, M. Trends in microarray analysis. *Nature Medicine* **9**, 140-145 (2003).
2. Rogers, L.R.K., de los Campos, G. & Mias, G.I. Micorarray gene expression dataset re-analysis reveals variability in influenza infection and vaccination. *Frontiers in Immunology* **10**(2019).
3. Xu, J. & Yang, Y. Potential genes and pathways along with immune cells infiltration in the progression of atherosclerosis identified via microarray gene expression dataset re-analysis. *Vascular* **28**, 643-654 (2020).
4. La-Croix-Fralish, M.L., Austin, S.-L., Zheng, F.Y., Levitin, D.J. & Mogil, J.S. Patterns of pain: Meta-analysis of microarray studies of pain. *Pain* **152**, 1888-1898 (2011).
5. Grutzmann, R. *et al.* Meta-anlaysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* **24**, 5079-5088 (2005).
6. Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. & Chinnaiyan, A.M. Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals dysregulation in prostate cancer. *Cancer Research* **62**, 4427-4433 (2002).
7. Kelly, J., Moyeed, R., Carroll, C., Albani, D. & Li, X. Gene expression meta-analysis of Parksinson's disease and its relationship with Alzheimer's disease. *Molecular Brain* **12**(2019).
8. Noori, A., Mezlini, A.M., Hyman, B.T., Serrano-Pozo, A. & Das, S. Systematic review and meta-analysis of human transcriptomics reveals neuroinflammation, deficient energy metabolism, and proteostasis failure across neurodegeneration. *Neurobiology of Disease* **149**(2021).
9. Wan, Y.-W. *et al.* Meta-analysis of the Alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell Reports* **32**(2020).
10. Signaevsky, M. *et al.* Artificial intelligence in neuropathology: Deep learning-based assessment of tauopathy. *Laboratory Investigation* **99**, 1019-1029 (2019).
11. Su, C., Tong, J. & Wang, F. Mining genetic and transcriptomic data using machine learning approaches in Parksinson's disease. *npj Parkinson's Disease* **6**(2020).
12. Sahoo, D. *et al.* Artificial intelligence guided discovery of a barrier-protective therapy in inflammatory bowel disease. *Nature Communications* **12**(2021).
13. Sharma, A. & Dey, P. A machine learning approach to unmask novel gene signatures and prediction of Alzheimer's disease with different brain regions. *Genomics* **113**, 1778-1789 (2021).
14. Navarro, C.L.A. *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: A systematic review. *BMJ* **375**(2021).
15. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605-D612 (2021).
16. World Health Organization. Global action plan on the public health response to dementia 2017-2025. (ed. Organization, W.H.) 44 (Geneva, 2017).
17. Cummings, J. Lessons learned from Alzheimer's disease: Clinical trials with negative outcomes. *Clinical and Translational Science* **11**, 147-152 (2018).
18. Elder, G.A., Gama Sosa, M.A. & De Gasperi, R. Transgenic mouse models of Alzheimer's disease. *Mount Sinai Journal of Medicine* **77**, 69-81 (2010).
19. Jankowsky, J.L. & Zheng, H. Practical considerations for choosing a mouse model of Alzheimer's disease. *Molecular Neurodegeneration* **12**(2017).
20. DeTure, M.A. & Dickson, D.W. The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration* **14**(2019).
21. Bali, J., Gheinani, A.H., Zurbriggen, S. & Rajendran, L. Role of genes linked to sporadic Alzheimer's disease risk in the production of β-amyloid peptides. *PNAS* **109**, 15307-15311 (2012).

22. Jonckheere, A.I., Smeitink, J.A.M. & Rodenbur, R.J.T. Mitochondrial ATP synthase: Architecture, function and pathology. *Journal of Inhereted Metabolic Disease* **35**, 211-225 (2011).
23. Patro, S. *et al.* ATP synthase and mitochondrial bioenergentics dysfunction in Alzheimer's disease. *International Journal of Molecular Sciences* **22**(2021).
24. Ebanks, B., Ingram, T.L. & Chakrabarti, L. ATP synthase and Alzheimer's disease: Putting a spin on the mitochondrial hypothesis. *Aging* **12**, 16647-16662 (2020).
25. Swerdlow, R.H. Mitochondria and mitochondrial cascades in Alzheimer's disease. *Journal of Alzheimer's Disease* **62**, 1403-1416 (2018).
26. Swerdlow, R.H., Burns, J.M. & S.M., K. The Alzheimer's disease mitochondrial cascade hypothesis: Progress and perspectives. *Biochimica et Biophysica Acta* **1842**, 1219-1231 (2014).
27. Boada, M. *et al.* ATP5H/KCTD2 locus is associated with Alzheimer's disease risk. *Molecular Psychiatry* **19**, 682-687 (2013).
28. Traylor, M. *et al.* Shared genetic contribution to ischaemic stroke and Alzheimer's disease. *Annals of Neurology* **79**, 739-747 (2016).
29. Choi, S.Y. *et al.* C9ORF72-ALS/FTD-associated poly(GR) binds Atp5a1 and compromises mitochondrial function in vivo. *Nature Neuroscience* **22**, 851-862 (2019).
30. Gonzalez-Rodriguez, P. *et al.* Disruption of mitochondrial complex I induces progressive parkinsonism. *Nature* **599**, 560-656 (2021).
31. Varma, V.R. *et al.* Abnormal brain cholesterol homeostasis in Alzheimer's disease - A targeted metabolomics and transcriptomic study. *npj Aging and Mechanisms of Disease* **7**(2021).
32. Jiang, X.-S., Backlund, P.S., Wassif, C.A., Yergey, A.L. & Porter, F.D. Quantitative proteomics analysis of inborn errors of cholesterol synthesis. *Molecular and Cellular Proteomics* **9**, 1461-1475 (2010).
33. Thal, D.R., Rub, U., Orantes, M. & Braak, H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD. *Neurology* **58**(2002).
34. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **25**, 707-720 (2013).
35. Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Molecular Systems Biology* **10**(2014).
36. Claverie, J.M. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8**, 1821-1832 (1999).
37. Mutch, D.M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.-A. The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics* **3**(2002).
38. Halsey, L.G. The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters* **15**(2019).
39. Altman, N. & Krzywinski, M. *P* values and the search for significance. *Nature Methods* **14**(2016).
40. Nuzzo, R. Scientific method: Statistical errors. *Nature* **506**, 150-152 (2014).
41. Verleysen, M. & Francois, D. *The curse of dimensionality in data mining and time series prediction*, (Springer, Berlin, 2005).
42. Tarca, A.L., Carey, V.J., Chen, X.-W., Romero, R. & Draghici, S. Machine learning and its application to biology. *PLOS Computational Biology* **3**(2007).
43. Jolliffe, I.T. & Cadima, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A* **374**(2016).
44. Ringer, M. What is principal component analysis? *Nature Biotechnology* **26**, 303-304 (2008).
45. Yeung, K.Y. & Ruzzo, W.L. Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763-774 (2001).
46. Reich, D., Price, A.L. & Patterson, N. Principal component analysis of genetic data. *Nature Genetics* **40**, 491-492 (2008).

47. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607-D613 (2018).
48. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330-338 (2019).
49. Morgan, J. *Classification and regression tree analysis*, (Boston University, Boston, 2014).
50. Lemon, S.C., Roy, J., Clark, M.A., Friedman, P.D. & Rakowski, W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine* **26**, 172-181 (2003).
51. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing* **21**, 137-146 (2011).

Artificial intelligence-driven meta-analysis of brain gene expression data identifies novel gene candidates in Alzheimer's Disease

Caitlin A. Finney [1*], Fabien Delerue [1], & Artur Shvetcov [2*]

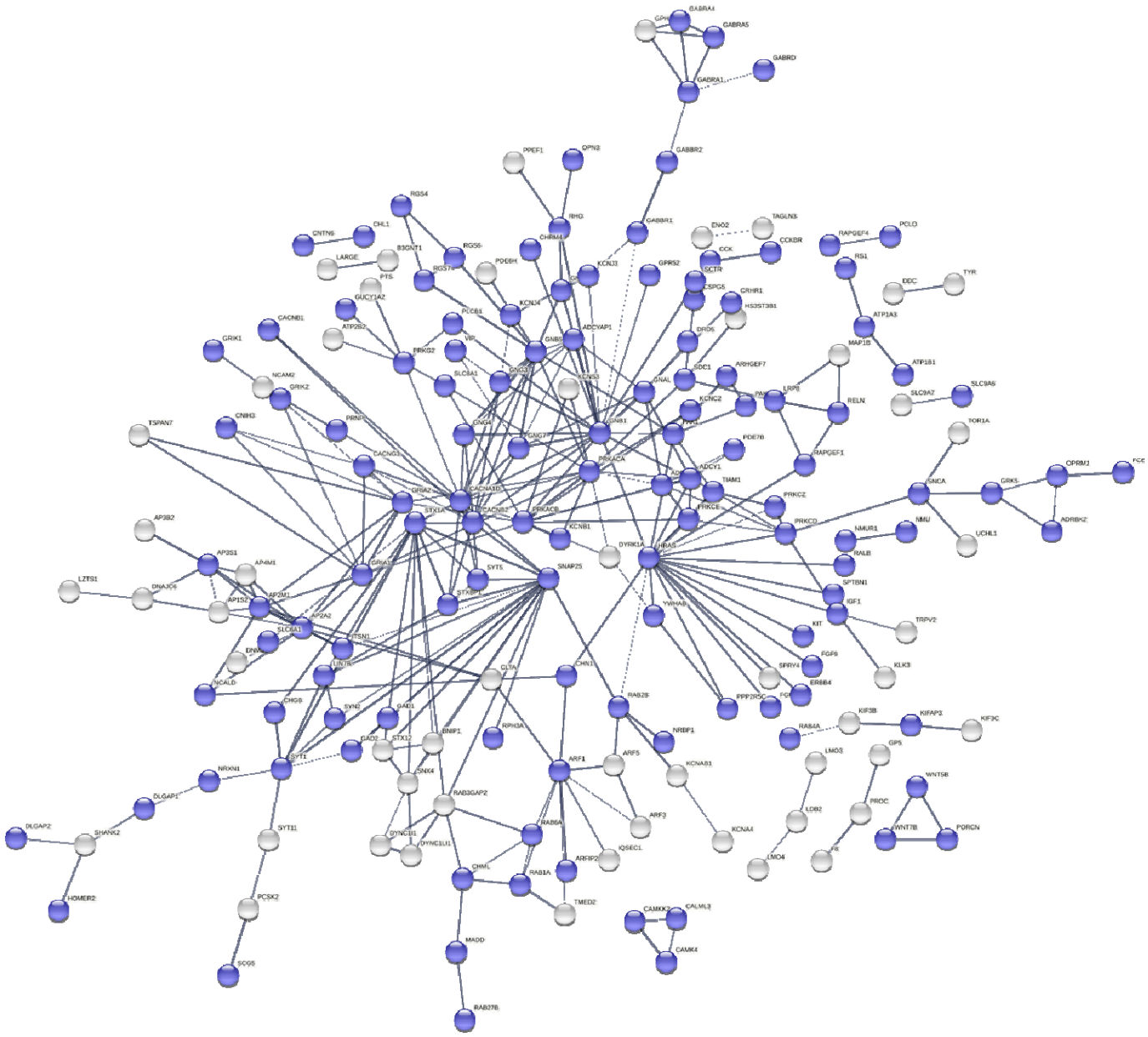
1. Dementia Research Centre, Department of Biomedical Science, Faculty of Medicine Health and Human Sciences, Macquarie University, Australia

2. Black Dog Institute, Australia

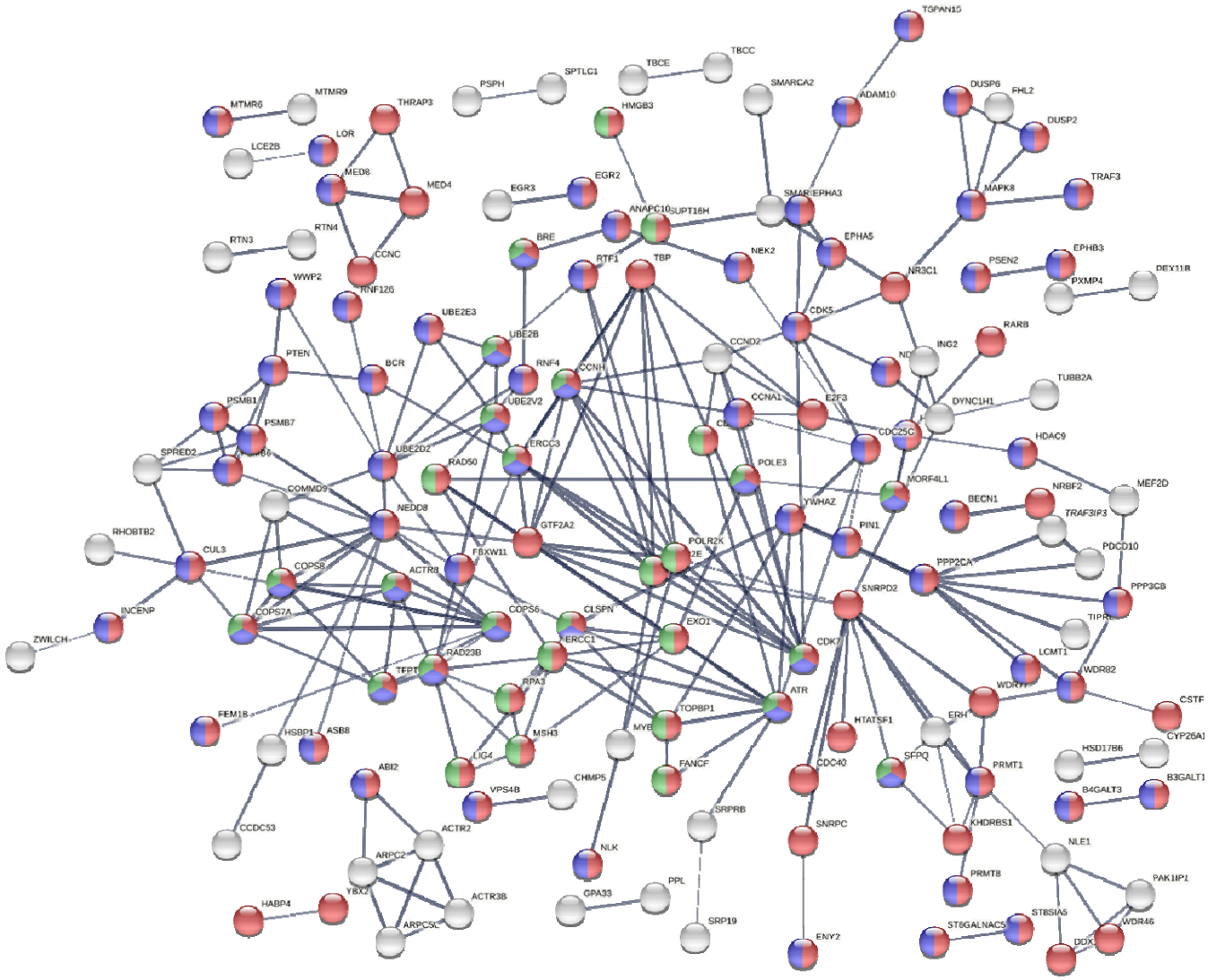
* Authors to whom correspondence should be addressed. E-mails: a.shvetcov@blackdog.org; caitlin_finney@hotmail.com

Supplementary Figures

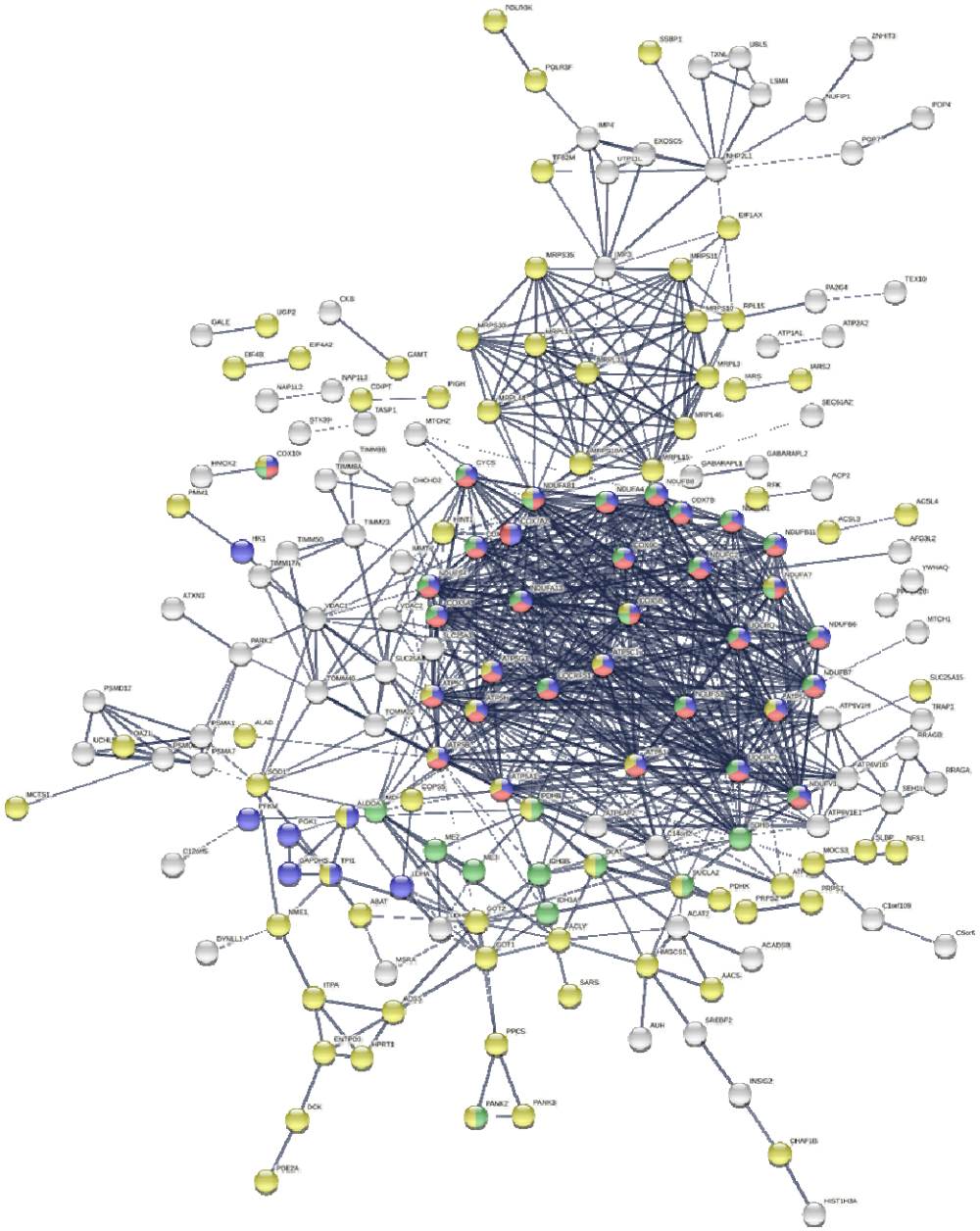
Supplementary Figure 1. Characterization of the biological processes of frontal cortex k-means cluster 1 (red) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. The blue nodes indicate genes involved in signaling (FDR = 1.68e-26).



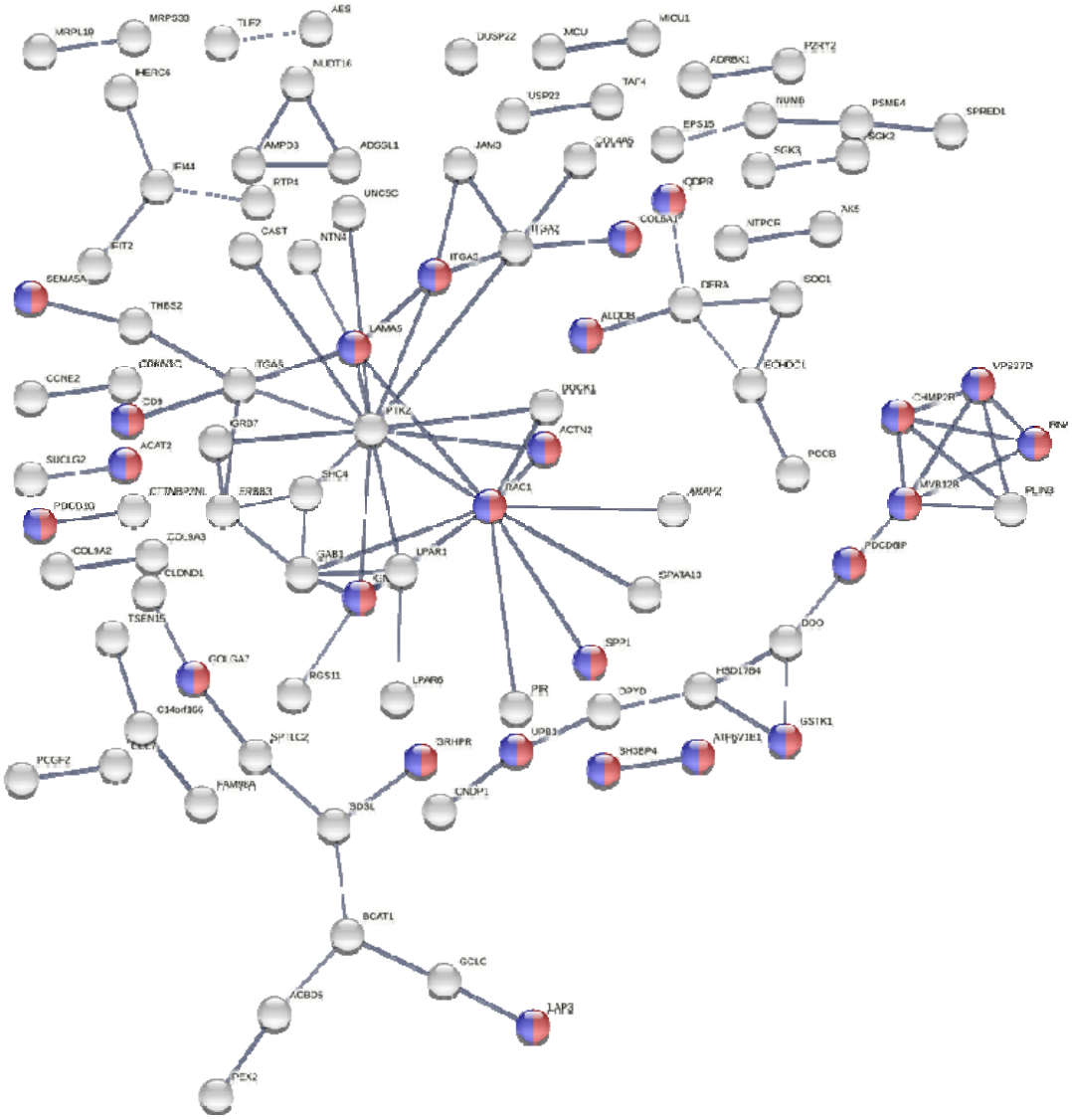
Supplementary Figure 2. Characterization of the biological processes of frontal cortex k-means cluster 2 (green) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. Red nodes indicate genes involved in macromolecule metabolic processes (FDR = 2.09e-13). Blue nodes indicate genes involved in protein metabolic processes (FDR = 5.22e-9). Green nodes indicate genes involved in DNA metabolic processes (FDR = 5.14e-6).



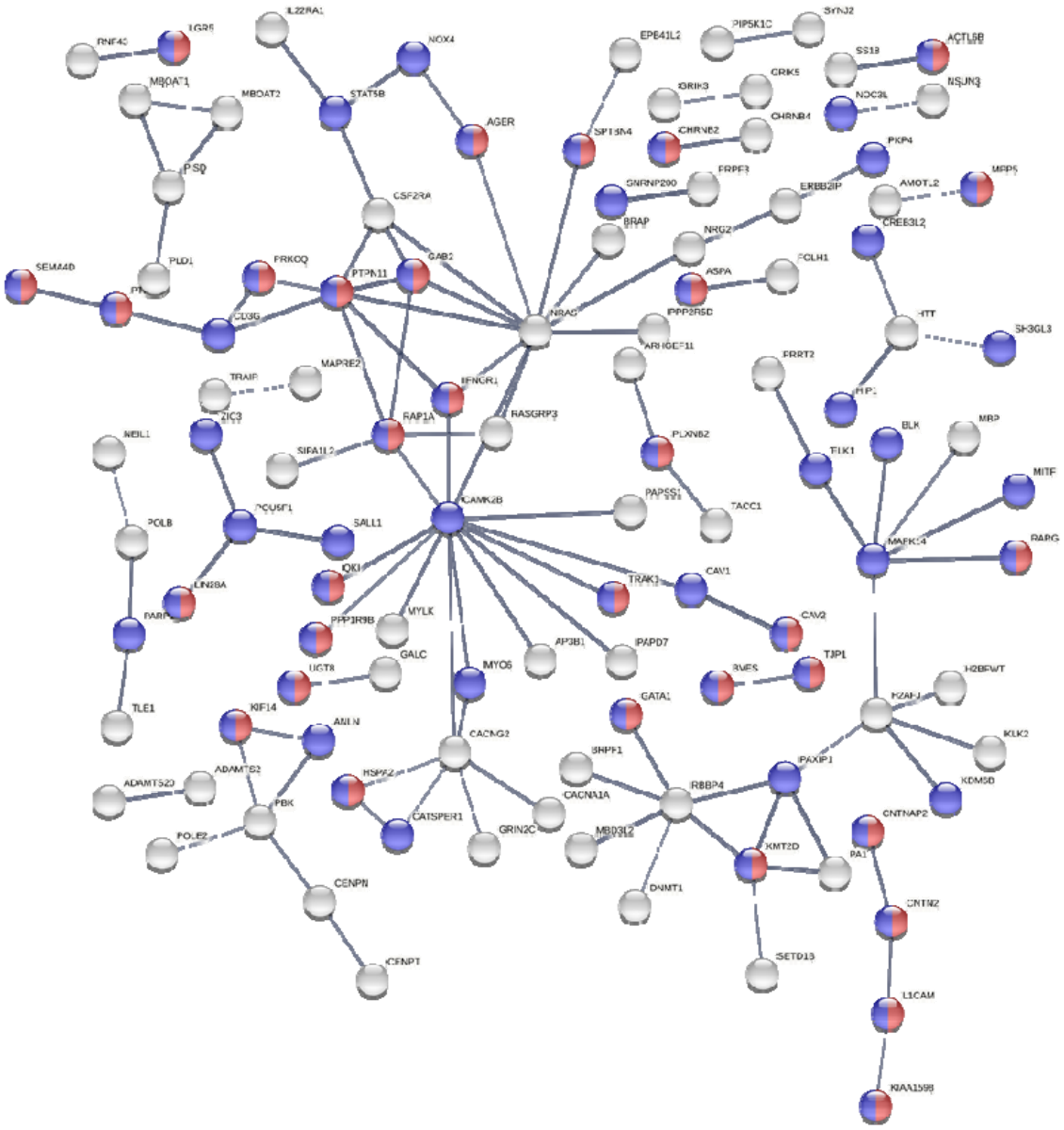
Supplementary Figure 3. Characterization of the biological processes of frontal cortex k-means cluster 3 (blue) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. Blue nodes indicate genes involved in ATP metabolic processes (FDR = 4.75e-26). Red nodes indicate genes involved in oxidative phosphorylation (FDR = 2.76e-24). Green nodes indicate genes involved in cellular respiration (FDR = 2.76e-24). Together, the blue, red, and green nodes form sub-cluster 3A: mitochondria (ATP, energy, and oxidative phosphorylation). Yellow nodes indicate genes involved in cellular biosynthetic processes (FDR = 2.01e-12) and form sub-cluster 3B: mitochondria (cellular biosynthesis).



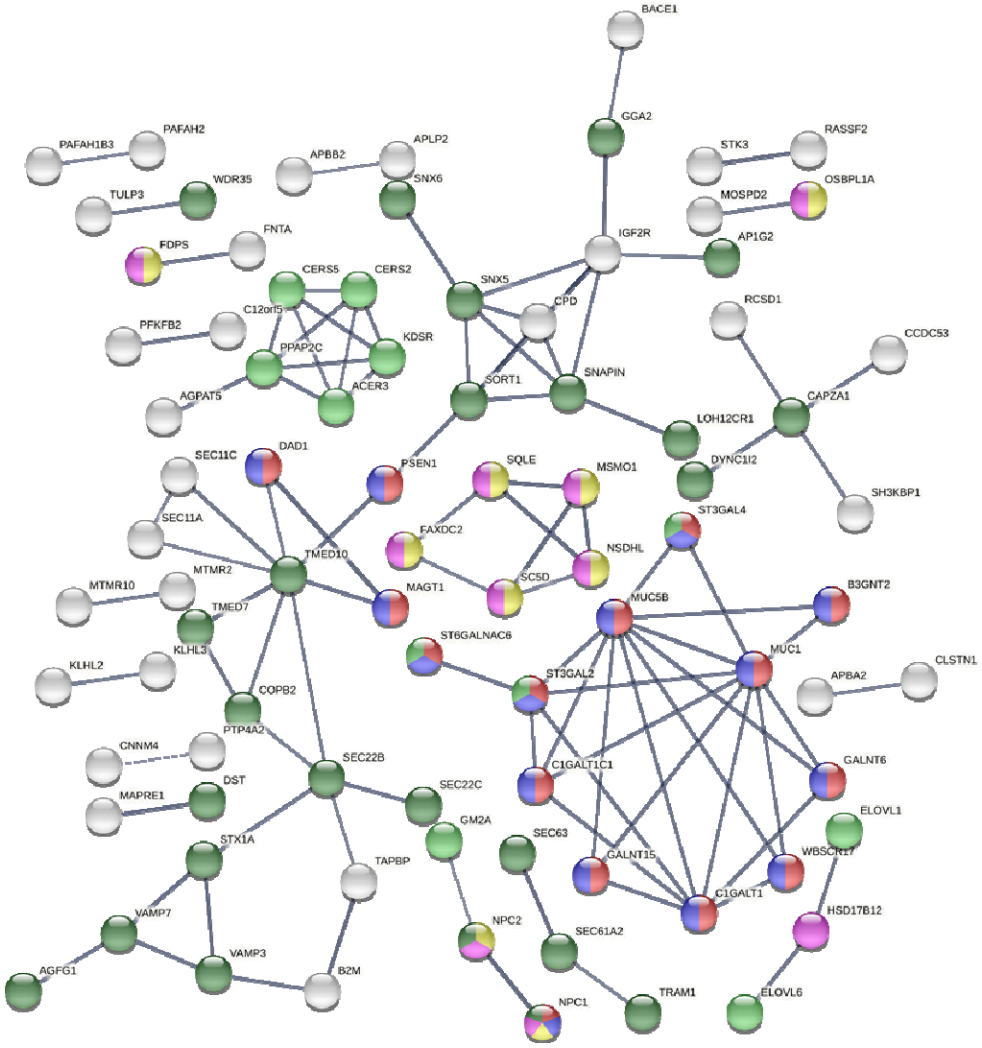
Supplementary Figure 4. Characterization of the biological processes of cerebellum k-means cluster 1 (red) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. Red indicates genes involved in extracellular vesicles (FDR = 0.0286). Blue indicates genes involved in extracellular exosomes (FDR = 0.0286).



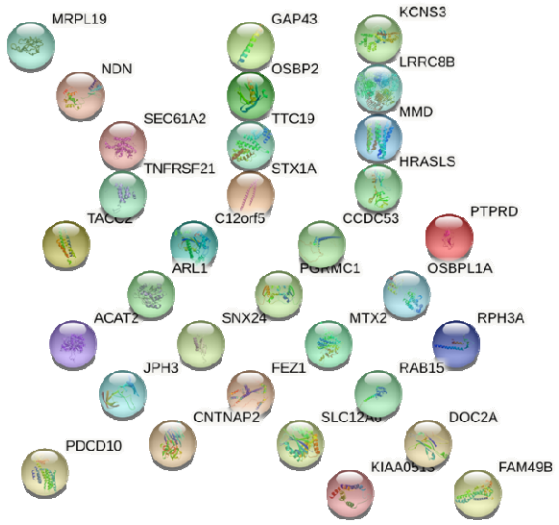
Supplementary Figure 5. Characterization of the biological processes of cerebellum k-means cluster 2 (green) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. Red nodes indicate genes involved in cell development (FDR = 8.53e-10). Blue nodes indicate genes involved in cell differentiation (FDR = 1.56e-9).



Supplementary Figure 6. Characterization of the biological processes of cerebellum k-means cluster 3 (blue) using Gene Ontology and STRING. The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence) and disconnected nodes in the network are hidden. Thickness of the line indicates the confidence in the interaction. Red nodes indicate genes involved in protein glycosylation (FDR = 1.35e-5). Blue nodes indicate genes involved in glycoprotein biosynthetic processes (FDR = 2.46e-5). Combined, red and blue nodes formed cluster 3A: protein glycosylation and glycoprotein biosynthesis. Light green nodes indicate genes involved in membrane lipid metabolic processes (FDR = 0.00064) and formed cluster 3B. Yellow nodes indicate genes involved in sterol metabolic processes (FDR = 8.86e-5). Pink nodes indicate genes involved in steroid metabolic processes (FDR = 0.0022). Combined, yellow and pink nodes formed cluster 3C. Dark green nodes indicate genes involved in intracellular transport (FDR = 0.0047) and formed cluster 3D.



Supplementary Figure 7. STRING network analysis of the overlap between the frontal cortex and cerebellum top 1000 dysregulated genes (n = 32 genes). The STRING network analysis includes the following interaction sources: experiments, databases, co-expression, neighborhood, and gene fusion. The minimum interaction score was set to 0.7 (high confidence). The overlapping genes were disconnected nodes and therefore did not have any interactions.



Artificial intelligence-driven meta-analysis of brain gene expression data identifies novel gene candidates in Alzheimer's Disease

Caitlin A. Finney [1*], Fabien Delerue [1], & Artur Shvetcov [2*]

1. Dementia Research Centre, Department of Biomedical Science, Faculty of Medicine Health and Human Sciences, Macquarie University, Australia

2. Black Dog Institute, Australia

* Authors to whom correspondence should be addressed. E-mails: a.shvetcov@blackdog.org; caitlin_finney@hotmail.com

Supplementary Tables

Supplementary Table 1. Frontal cortex k-means cluster 1 central node genes: signaling

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
CACNA1D	Calcium voltage-gated channel subunit α1 D	ENSG00000157388	α1 D subunit of voltage-dependent calcium channel	Downregulated
CACNB2	Calcium voltage-gated channel auxiliary subunit β2	ENSG00000165995	Subunit of voltage-dependent calcium channel protein	Downregulated
HRAS	HRas proto-oncogene, GTPase	ENSG00000174775	Bind GTP and GDP, have intrinsic GTPase activity, regulates exchange between plasma membrane and Golgi apparatus	Downregulated
GNB1	G protein subunit β1	ENSG00000078369	β subunit of heterotrimeric guanine nucleotide-binding proteins	Downregulated
GNB5	G protein subunit β5	ENSG00000069966	β subunit of heterotrimeric guanine nucleotide-binding proteins	Downregulated
GNG3	G protein subunit γ3	ENSG00000162188	γ subunit of heterotrimeric guanine nucleotide-binding proteins, γ subunit of both inhibitory and stimulatory complexes	Downregulated
GNG7	G protein subunit γ7	ENSG00000176533	Enable guanine nucleotide-binding protein β subunit activity	Downregulated
GRIA1	Glutamate ionotropic receptor AMPA type subunit 1	ENSG00000155511	AMPA receptor	Downregulated
PRKACA	Protein kinase cAMP-activated catalytic subunit α	ENSG00000072062	Catalytic subunit of protein kinase A, cAMP-dependent phosphorylation of proteins by PKA	Downregulated
PRKACB	Protein kinase cAMP-activated catalytic subunit β	ENSG00000142875	Catalytic subunit of cAMP-dependent protein kinase	Downregulated
SNAP25	Synaptosome associated protein 25	ENSG00000132639	Presynaptic plasma membrane protein involved in neurotransmitter release, part of SNARE complex	Downregulated
STX1A	Syntaxin 1A	ENSG00000106089	Regulation of ion channels and synaptic exocytosis	Downregulated
STXBP1	Syntaxin binding protein 1	ENSG00000136854	Syntaxin-binding protein, release of neurotransmitters via syntaxin regulation	Downregulated
SYT1	Synaptotagmin 1	ENSG00000067715	Synaptic vesicle membrane, triggers neurotransmitter release at synapse	Downregulated

SYT5	Synaptotagmin 5	ENSG00000129990	Negative regulator of vesicle fusion, calcium receptor or sensor	Downregulated

Supplementary Table 2. Frontal cortex k-means cluster 2 central node genes: metabolic processes (macromolecular, proteins and DNA)

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
ACTR8	Actin related protein 8	ENSG00000113812	Enable ATP binding activity, chromatin remodeling, double-strand break repair and transcription regulation	Downregulated
ATR	ATR serine/threonine kinase	ENSG00000175054	Serine/threonine kinase and DNA damage sensor, DNA stress signaling	Downregulated
CCNA1	Cyclin A1	ENSG00000133101	Control of germline meiotic cell cycle	Downregulated
CCND2	Cyclin D2	ENSG00000118971	Regulatory subunit of cyclin complex, required for cell cycle G1/S transition	Downregulated
CCNH	Cyclin H	ENSG00000134480	CDK-activating kinase, transcriptional regulation	Downregulated
CDK5	Cyclin dependent kinase 5	ENSG00000164885	Synaptic plasticity and neuronal migration via phosphorylation of cytoskeletal, endo- and exocytotic, and apoptotic proteins	Downregulated
CDK7	Cyclin dependent kinase 7	ENSG00000134058	Regulator of cell cycle progression, transcription initiation and DNA repair	Downregulated
COPS6	COP9 signalosome subunit 6	ENSG00000168090	Positive regulation of E3 ubiquitin ligases, regulation of cell cycle	Downregulated
COPS7A	COP9 signalosome subunit 7A	ENSG00000111652	Regulates activity of ubiquitin conjugation pathway	Downregulated
COPS8	COP9 signalosome subunit 8	ENSG00000198612	Positive regulation of E3 ubiquitin ligases	Downregulated
CUL3	Cullin 3	ENSG00000036257	Polyubiquitination and degradation, core component and scaffold protein of E3 ubiquitin ligase complex	Downregulated
ERCC1	ERCC excision repair 1, endonuclease non-catalytic subunit	ENSG00000012061	Nucleotide excision repair pathway, repair of DNA lesions	Downregulated
ERCC3	ERCC excision repair 3, TFIIH core complex helicase subunit	ENSG00000163161	ATP-dependent DNA helicase, nucleotide excision repair	Downregulated
GTF2A2	General transcription factor IIA subunit 2	ENSG00000140307	General transcription initiation factor	Downregulated
NEDD8	NEDD8 ubiquitin like	ENSG00000129559	Enables ubiquitin protein ligase binding activity	Downregulated

	modifier			
POLR2E	RNA polymerase II, I and II subunit E	ENSG00000099817	RNA polymerase II subunit, synthesis of mRNA	Downregulated
POLR2K	RNA polymerase II, I and II subunit K	ENSG00000147669	RNA polymerase II subunit, synthesis of mRNA	Downregulated
PPP2CA	Protein phosphatase 2 catalytic subunit α	ENSG00000113575	Phosphatase 2A catalytic subunit, negative control of cell growth and division	Downregulated
PSMB1	Proteasome 20S subunit β1	ENSG00000008018	Member of proteasome B-type family (T1B), core 20S β subunit	Downregulated
PSMB6	Proteasome 20S subunit β6	ENSG00000142507	Member of proteasome B-type family (T1B), core 20S β subunit	Downregulated
PSMB7	Proteasome 20S subunit β7	ENSG00000136930	Member of proteasome B-type family (T1B), core 20S β subunit	Downregulated
RAD23B	RAD23 homolog B, nucleotide excision repair protein	ENSG00000119318	DNA damage recognition in base excision repair, ubiquitin mediated proteolytic pathway	Downregulated
SNRPD2	Small nuclear ribonucleoprotein D2 polypeptide	ENSG00000125743	Involved in pre-mRNA splicing and small nuclear ribonucleoprotein biogenesis	Downregulated
TBP	TATA-box binding protein	ENSG00000112592	TATA-binding protein, transcription initiation factor	Downregulated
TFPT	TCF3 fusion partner	ENSG00000105619	DNA and protein kinase binding activity, apoptotic signaling pathway	Downregulated
UBE2D2	Ubiquitin conjugating enzyme E2 D2	ENSG00000131508	Ubiquitin proteosome system	Downregulated

Supplementary Table 3. Frontal cortex k-means cluster 3A central node genes: mitochondria (energy, ATP, and oxidative phosphorylation)

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
ATP5A1	ATP synthase F1 subunit α	ENSG00000152234	α subunit of F1 catalytic core of mitochondrial ATP synthase	Downregulated
ATP5B	ATP synthase F1 subunit β	ENSG00000110955	β subunit of F1 catalytic core of mitochondrial ATP synthase	Downregulated
ATP5C1	ATP synthase F1 subunit γ	ENSG00000165629	γ subunit of F1 catalytic core of mitochondrial ATP synthase	Downregulated
ATP5H	ATP synthase peripheral stalk subunit d	ENSG00000167863	d subunit of F0 complex of mitochondrial ATP synthase	Downregulated
ATP5G3	ATP synthase membrane subunit c locus 3	ENSG00000154518	c subunit of proton channel of mitochondrial ATP synthase	Downregulated
ATP5J	ATP synthase peripheral stalk subunit F6	ENSG00000154723	F6 subunit of F0 complex of mitochondrial ATP synthase	Downregulated
ATP5L	ATP synthase membrane subunit g	ENSG00000167283	g subunit of F0 complex of mitochondrial ATP synthase	Downregulated
ATP5MJ (C14orf2)	ATP synthase membrane subunit J	ENSG00000156411	Integral component of mitochondrial membrane, proton transporting ATP synthase complex	Downregulated
ATP5O	ATP synthase peripheral stalk subunit OSCP	ENSG00000241837	Component of the F-type ATPase found in mitochondrial matrix	Downregulated
ATP6AP2	ATPase H+ transporting accessory protein 2	ENSG00000182220	Transmembrane sector of V-ATPases	Downregulated
ATP6V1E1	ATPase H+ transporting V1 subunit E1	ENSG00000131100	V1 domain E subunit of V-ATPase	Downregulated
ATP6V1D	ATPase H+ transporting V1 subunit D	ENSG00000100554	V1 domain D subunit of V-ATPase	Downregulated
ATP6V1H	ATPase H+ transporting V1 subunit H	ENSG00000047249	V1 domain regulatory H subunit of V-ATPase, catalysis of ATP	Downregulated
COX4I1	Cytochrome c oxidase subunit 4I1	ENSG00000131143	Nuclear-encoded subunit IV isoform 1 of mitochondrial respiratory chain enzyme	Downregulated
COX5A	Cytochrome c oxidase subunit 5A	ENSG00000178741	Nuclear-encoded subunit Va of mitochondrial respiratory chain enzyme	Downregulated
COX5B	Cytochrome c oxidase subunit 5B	ENSG00000135940	Nuclear-encoded subunit Vb of mitochondrial respiratory chain enzyme	Downregulated
COX6C	Cytochrome c oxidase	ENSG00000164919	Nuclear-encoded subunit VIc of mitochondrial	Downregulated

	subunit 6C		respiratory chain enzyme	
COX7A2	Cytochrome c oxidase subunit 7A2	ENSG00000112695	Nuclear-encoded polypeptide 2 of subunit VIIa of mitochondrial respiratory chain enzyme	Downregulated
COX7B	Cytochrome c oxidase subunit 7B	ENSG00000131174	Nuclear-encoded subunit VIIb of mitochondrial respiratory chain enzyme	Downregulated
CYCS	Cytochrome c, somatic	ENSG00000172115	Central component of mitochondrial electron transport chain	Downregulated
NDUFA4	NDUFA4 mitochondrial complex associated	ENSG00000189043	Mitochondrial complex I, NADH dehydrogenase and oxidoreductase activity	Downregulated
NDUFAB1	NADH:ubiquinone oxidoreductase subunit AB1	ENSG00000004779	Mitochondrial complex I assembly and protein lipoylation	Downregulated
NDUFB1	NADH:ubiquinone oxidoreductase subunit B1	ENSG00000183648	Mitochondrial complex I assembly	Downregulated
NDUFB6	NADH:ubiquinone oxidoreductase subunit B6	ENSG00000165264	Mitochondrial complex I, NADH dehydrogenase and oxidoreductase activity	Downregulated
NDUFB7	NADH:ubiquinone oxidoreductase subunit B7	ENSG00000099795	Mitochondrial complex I, NADH dehydrogenase and oxidoreductase activity	Downregulated
NDUFB8	NADH:ubiquinone oxidoreductase subunit B8	ENSG00000166136	Mitochondrial complex I assembly	Downregulated
NDUFB11	NADH:ubiquinone oxidoreductase subunit B11	ENSG00000147123	Mitochondrial complex I, NADH dehydrogenase and oxidoreductase activity	Downregulated
NDUFC2	NADH:ubiquinone oxidoreductase subunit C11	ENSG00000151366	Mitochondrial complex I assembly	Downregulated
NDUFS3	NADH:ubiquinone oxidoreductase core subunit S3	ENSG00000213619	Iron-sulfur protein (IP) component of mitochondrial complex I	Downregulated
NDUFS4	NADH:ubiquinone oxidoreductase subunit S4	ENSG00000164258	Nuclear-encoded accessory subunit of mitochondrial complex I	Downregulated

NDUFV1	NADH:ubiquinone oxidoreductase core subunit V1	ENSG00000167792	Subunit mitochondrial complex I, carries NADH, flavin mononucleotide and iron-sulfur binding sites	Downregulated
SDHB	Succinate dehydrogenase complex iron sulfur subunit B	ENSG00000117118	Mitochondrial complex II, oxidation of succinate	Downregulated
UQCRC2	Ubiquinol-cytochrome c reductase core protein 2	ENSG00000140740	Mitochondrial complex III	Downregulated
UQCRFS1	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	ENSG00000169021	Mitochondrial complex III, enables oxidoreductase activity	Downregulated
UQCRQ	Ubiquinol-cytochrome c reductase complex III subunit VII	ENSG00000164405	Ubiquinone-binding protein, subunit of mitochondrial complex III	Downregulated

Supplementary Table 4. Frontal cortex k-means cluster 3B central node genes: mitochondria (cellular biosynthesis)

Gene	**Name**	**Ensembl ID**	**Function (NCBI)**	**Direction of Dysregulation**

IMP3	IMP U3 small nucleolar ribonucleoprotein 3	ENSG00000177971	Interacts with the U3 snoRNP complex	Downregulated
MRPL15	Mitochondrial ribosomal protein L15	ENSG00000137547	Nuclear-encoded mitochondrial ribosomal 39S subunit, protein synthesis in mitochondrion	Downregulated
MRPL19	Mitochondrial ribosomal protein L19	ENSG00000115364	Nuclear-encoded mitochondrial ribosomal 39S subunit, protein synthesis in mitochondrion	Downregulated
MRPL44	Mitochondrial ribosomal protein L44	ENSG00000135900	Nuclear-encoded mitochondrial ribosomal 39S subunit, protein synthesis in mitochondrion	Downregulated
MRPL46	Mitochondrial ribosomal protein L46	ENSG00000259494	Nuclear-encoded mitochondrial ribosomal 39S subunit, protein synthesis in mitochondrion	Downregulated
MRPS10	Mitochondrial ribosomal protein S10	ENSG00000048544	Nuclear-encoded mitochondrial ribosomal 28S subunit, protein synthesis in mitochondrion	Downregulated
MRPS11	Mitochondrial ribosomal protein S11	ENSG00000181991	Nuclear-encoded mitochondrial ribosomal 28S subunit, protein synthesis in mitochondrion	Downregulated
MRPS18A	Mitochondrial ribosomal protein S18A	ENSG00000096080	Nuclear-encoded mitochondrial ribosomal 28S subunit, protein synthesis in mitochondrion	Downregulated
MRPS30	Mitochondrial ribosomal protein S30	ENSG00000112996	Nuclear-encoded mitochondrial ribosomal 28S subunit, protein synthesis in mitochondrion	Downregulated
MRPS35	Mitochondrial ribosomal protein S35	ENSG00000061794	Nuclear-encoded mitochondrial ribosomal 28S subunit, protein synthesis in mitochondrion	Downregulated

Supplementary Table 5. Cerebellum k-means cluster 1 central node genes: extracellular exosomes and vesicles

Gene	**Name**	**Ensembl ID**	**Function (NCBI)**	**Direction of Dysregulation**

MVB12B	Multivesicular body subunit 12B	ENSG00000196814	Component of ESCRT-I complex, mediates sorting of ubiquitinated cargo protein from plasma membrane to endosomal vesicle	Downregulated
PTK2	Protein tyrosine kinase 2	ENSG00000169398	Cytoplasmic protein tyrosine kinase, involved in cell growth and intracellular signal transduction pathways	Downregulated
RAC1	Rac family small GTPase 1	ENSG00000136238	GTPase belonging to RAS superfamily of small GTP-binding proteins	Downregulated

Supplementary Table 6. Cerebellum k-means cluster 2 central node genes: development and differentiation

Gene	**Name**	**Ensembl ID**	**Function (NCBI)**	**Direction of**

				Dysregulation
CAMK2B	Calcium/calmodulin dependent protein kinase IIβ	ENSG00000058404	β chain of CaMKII enzyme, calcium signaling	Downregulated
H2AFJ	H2A.J histone	ENSG00000246705	Replication-independent histone that is variant of H2A histone	Upregulated
MAPK14	Mitogen-activated protein kinase 14	ENSG00000112062	Stress-related transcription, cell cycle regulation and genotoxic stress response	Downregulated
NRAS	NRAS proto-oncogene, GTPase	ENSG00000213281	N-ras oncogene encoding membrane protein that shuttles between Golgi apparatus and plasma membrane, intrinsic GTPase activity	Upregulated
PAXIP1	PAX interacting protein 1	ENSG00000157212	Maintains genome stability, condensation of chromatin and mitosis progression	Upregulated
PTPN11	Protein tyrosine phosphatase non-receptor type 11	ENSG00000179295	Regulation of cell signaling for mitogenic activation, metabolic control, transcription regulation and cell migration	Upregulated
RBBP4	RB binding protein 4, chromatin remodeling factor	ENSG00000162521	Nuclear protein part of Mi-2 complex involved in chromatin remodeling and transcription repression via histone deacetylation	Downregulated

Supplementary Table 7. Cerebellum k-means cluster 3A central node genes: protein glycosylation and glycoprotein biosynthesis

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
C1GALT1	Core 1 synthase, glycoprotein-N-acetylgalactosamine 3-β-galactosyltransferase 1	ENSG00000106392	Generates common core 1 O-glycan structure Gal-β-1-3GalNAc-R, precursor for extended mucin-type O-glycans on cell surface and secreted glycoproteins	Downregulated
MUC1	Mucin 1, cell surface associated	ENSG00000185499	Membrane-bound protein member of mucin family, intracellular signaling	Upregulated
MUC5B	Mucin 5B, oligomeric mucus/gel-forming	ENSG00000117983	Major gel-forming mucin in mucus	Upregulated
ST3GAL2	ST3 β-galactoside α-2,3-sialyltransferase 2	ENSG00000157350	Type II membrane protein catalyzes transfer of sialic acid from CMP-sialic acid to galactose-containing substrates	Upregulated

Supplementary Table 8. Cerebellum k-means cluster 3B central nodes genes: membrane lipid metabolic processes

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
PPAP2C	Phospholipid phosphatase 2	ENSG00000141934	Membrane-associated PAP activity	Upregulated
CERS5	Ceramide synthase 5	ENSG00000139624	Synthesis of ceramide	Upregulated

Supplementary Table 9. Cerebellum k-means cluster 3C central node genes: sterol and steroid metabolic processes

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
FAXDC2	Fatty acid hydroxylase domain containing 2	ENSG00000170271	Enables C-4 methylsterol oxidase activity, positive regulation of megakaryocyte differentiation and protein phosphorylation	Downregulated
MSMO1	Methylsterol monooxygenase 1	ENSG00000052802	Localized to endoplasmic reticulum membrane, cholesterol biosynthesis	Upregulated
NSDHL	NAD(P) dependent steroid dehydrogenase-like	ENSG00000147383	Localized to endoplasmic reticulum membrane, cholesterol biosynthesis	Downregulated
SC5D	Sterol-C5-desaturase	ENSG00000109929	Cholesterol biosynthesis, catalyzes conversion of lathosterol into 7-dehydrocholesterol	Downregulated
SQLE	Squalene epoxidase	ENSG00000104549	Catalyzes first oxygenation step in sterol biosynthesis, rate limiting enzyme in pathway	Downregulated

Supplementary Table 10. Cerebellum k-means cluster 3D central node genes: intracellular transport

Gene	Name	Ensembl ID	Function (NCBI)	Direction of Dysregulation
IGF2R	Insulin-like growth factor 2 receptor	ENSG00000197081	Receptor for insulin-like growth factor 2 and mannose 6-phosphate, intracellular trafficking of lysosomal enzymes, activation of transforming growth factor β and degradation of insulin-like growth factor 2	Upregulated
SEC22B	SEC22B homolog B, vesicle trafficking protein	ENSG00000265808	Member of SEC22 family of vesicle trafficking proteins, complex with SNARE and plays role in endoplasmic reticulum-Golgi protein trafficking	Downregulated
SNAPIN	SNAP-associated protein	ENSG00000143553	Coiled-coil-forming protein associated with SNARE and BLOC-1 complexes	Downregulated
SNX5	Sorting nexin 5	ENSG00000089006	Endosomal sorting, phosphoinositide-signaling pathway and macropinocytosis	Downregulated
SORT1	Sortilin 1	ENSG00000134243	Trafficking proteins to cell surface or subcellular compartments (lysosomes or endosomes)	Upregulated
TMED10	Transmembrane p24 trafficking protein 10	ENSG00000170348	Vesicular protein trafficking	Upregulated

Supplementary Table 11. Rankings and correlation with output calculations for frontal cortex genes

Rank	Gene	Correlation with Output
Mitochondrial Energy, ATP and Oxidative Phosphorylation Cluster		
1	ATP5J	-0.6659338
2	ATP5L	-0.6422822
3	ATP5H	-0.6375793
4	COX7A2	-0.6368574
5	CYCS	-0.6365838
6	NDUFS4	-0.6346278
7	NDUFA4	-0.6329362
8	UQCRFS1	-0.6202124
9	NDUFAB1	-0.6198662
10	ATP6AP2	-0.6196468
11	UQCRQ	-0.6141539
12	ATP6V1E1	-0.6140198
13	ATP5C	-0.6120485
14	NDUFA7	-0.6106448
15	NDUFB6	-0.6091033
16	COX4I1	-0.6086761
17	NDUFC2	-0.5978484
18	ATP6V1D	-0.5881721
19	NDUFS3	-0.584457
20	NDUFV1	-0.5723417
21	ATP5A1	-0.5709399
22	SDHB	-0.5633374
23	COX5A	-0.5575793
24	COX7B	-0.5537025
25	COX6C	-0.546658
26	NDUFB11	-0.5441992
27	ATP6V1H	-0.5422819
28	ATP5MJ (C14orf2)	-0.5200282

29	NDUFB1	-0.503262
30	ATP5O	-0.4448213
31	NDUFB7	-0.422224
32	UQCRC2	-0.3180723
33	ATP5G3	-0.1264683
Signaling Cluster		
1	GNG3	-0.6560367
2	GNB5	-0.6180685
3	SYT1	-0.6075133
4	STX1A	-0.6031572
5	CACNB2	-0.5709202
6	CACNA1D	-0.5705174
7	HRAS	-0.5683397
8	GRIA1	-0.5676012
9	SNAP25	-0.5601579
10	GNB1	-0.5420024
11	STXBP1	-0.4932177
12	PRKACA	-0.4919038
13	PRKACB	-0.4210342
14	SYT5	-0.3693286
15	GNG7	-0.04200474
Metabolic Processes Cluster		
1	COPS7A	-0.6606022
2	PSMB7	-0.6447864
3	COPS8	-0.6410712
4	POLR2K	-0.6148842
5	CCND2	-0.61435
6	PPP2CA	-0.5872804
7	CCNA1	-0.5736387
8	CDK7	-0.5648813

9	COPS6	-0.5590357
10	UBE2D2	-0.5504209
11	CCNH	-0.5502376
12	NEDD8	-0.5442927
13	CDK5	-0.5388578
14	TFPT	-0.4711317
15	TBP	-0.4630831
16	CUL3	-0.4555989
17	PSMB6	-0.3932825
18	GTF2A2	-0.3929735
19	PSMB1	-0.3728207
20	ERCC1	-0.3717774
21	ATR	-0.3309196
22	ERCC3	-0.3274254
23	SNRPD2	-0.3060215
24	RAD23B	-0.3024951
25	ACTR8	-0.2332241
Mitochondria Cellular Biosynthesis Cluster		
1	MRPS18A	-0.6245018
2	MRPL15	-0.5998443
3	MRPS25	-0.5965593
4	MRPS11	-0.5871252
5	MRPL46	-0.538171
6	MRPS10	-0.531931
7	MRPS30	-0.5295817
8	IMP3	-0.5127594
9	MRPL19	-0.3417131
10	MRPL44	-0.2388002

Supplementary Table 12. Rankings and correlation with output calculations for cerebellum genes

Rank	Gene	Correlation with Output
Sterol and Steroid Metabolic Processes Cluster		
1	SC5DL	-0.2871297
2	FAXDC2	-0.1674087
3	NSDHL	-0.1402028
4	MSMO1	0.0890526
5	SQLE	-0.01123897
Intracellular Transport Cluster		
1	SEC22B	-0.3435256
2	SNAPIN	-0.2606466
3	IGF2R	0.1803468
4	SORT1	0.1135118
5	MED10	0.08230732
6	SNX5	-0.07930106
Development and Differentiation Cluster		
1	PAXIP1	0.2523401
2	RBBP4	-0.219466
3	CAMK2B	-0.1495922
4	PTPN11	0.0713462
5	NRAS	0.04845482
6	H2AFJ	0.02250764
7	MAPK14	-0.01244674

Supplementary Table 13. Overlap between the frontal cortex and cerebellum top 1000 genes

OSBPL1A
LRRC8B
STX1A
FEZ1
PDCD10
PTPRD
NDN
KCNS3
OSBP2
SLC12A6
GAP43
SNX24
TNFRSF21
HRASLS
MMD
MRPL19
PGRMC1
RAB15
MTX2
KIAA0513
JPH3
CNTNAP2
CCDC53
FAM49B
RPH3A
C12orf5
SEC61A2
TTC19
TACC2
ACAT2

ARL1
DOC2A