Universidade de Lisboa

Faculdade de Farmácia

Evaluating cortical transcriptomic differences between Alzheimer's disease PSEN1-mutated mouse models and human patients, and their implications in drug development

Sara Rodrigues Mendes

Dissertation supervised by Doctor Nuno Luís Barbosa Morais and co-supervised by Doctor Elsa Margarida Teixeira Rodrigues

Biopharmaceutical Sciences Master

2019

Universidade de Lisboa

Faculdade de Farmácia

Evaluating cortical transcriptomic differences between Alzheimer's disease PSEN1-mutated mouse models and human patients, and their implications in drug development

Sara Rodrigues Mendes

Dissertation supervised by Doctor Nuno Luís Barbosa Morais and co-supervised by Doctor Elsa Margarida Teixeira Rodrigues

Biopharmaceutical Sciences Master

2019

ACKNOWLEDGEMENTS

I want to thank my supervisor, Doctor Nuno Morais, for guiding me through this last year with enthusiasm and patience, and for sharing his technical knowledge with me. Above everything, I want to thank him the opportunity to start this project from ground 0 and experience the struggles and joys of the life of a true scientist. I am also deeply grateful for my supervisor from FFUL, Doctor Elsa Rodrigues, for her valuable suggestions and for being a constant presence and proactively engage and advise me throughout the entire process.

To my lab colleagues, which have grown into the most amazing friends and provided such a joyful working atmosphere, it was really a pleasure to work alongside you. Marie Bordone, Mariana Ferreira, Arthur Schneider, Nuno Agostinho, Marta Bica, Sofia Borges (and Amélia), thank you, it is truly the people that make you love what you do! Thank you for the late phone calls and the mornings talks (and laughs), for the guidance and the example, for killing my long R sessions and for carefully maintaining the lab sugar levels. Thank you for always going above and beyond to be there when needed, you were truly the best!

In special to Marie Bordone, my fellow neuroscientist and junior mentor, I am forever grateful for you, your knowledge, your guidance, and above all, your friendship. Thank you for laughing at my plots and have the patience of 300 monks combined, and for always believing in me and in this project. You are an inspiration.

A very special thank you to my boyfriend, João Calisto, for always being my best friend and best support, for distracting me when needed, for enduring my 5am thesis-writing wake ups and for never letting me give up. I am also deeply grateful to my sister and my parents, and remaining family, for watching me grow up and support me every second of every day. I am truly lucky, I love you all.

To the new friends and to the old ones that have always been here, thank you for caring and supporting me, and making life an amazing ride!

This work is a result of the GenomePT project (POCI-01-0145-FEDER-022184), supported by COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by Fundação para a Ciência e a Tecnologia (FCT).

The results presented in this thesis have been presented in one seminar of the Neuroscience seminar series in Instituto de Medicina Molecular João Lobo Antunes, and in the form of a poster in the iMed conference 11.0.

“We did not come to fear the future. We came here to shape it”

- **Barack Obama**

ABSTRACT

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disease of the central nervous system, being nowadays considered the most prevalent age-related dementia worldwide. AD pathology is characterized by the extracellular deposition of insoluble amyloid-beta plaques, and the intracellular accumulation of abnormally phosphorylated tau protein into neurofibrillary tangles. Other hallmarks include neuronal death, exacerbation of the immune system and chronic inflammation, synaptic loss, and brain atrophy.

The world population is rapidly aging, and an increase in the older population is foreseen, as well as in the prevalence of dementias such as AD. Currently, there is no effective treatment to neither decrease nor cease the damage of this disease, which, allied with the lack of new approved medicines since 2003, comprises a social, economic and health burden. Moreover, clinical trials have been exhibiting high failure rates, especially during toxicity and efficacy assessments, which implies a poor representation of the actual human disease in preclinical animal models. Thus, it is vital to evaluate what molecularly distinguishes them in terms of disease pathophysiology, and how can they be improved to better represent the human disease.

On this note, this project purposes to assess the dissimilarities between the AD-induced gene expression (i.e. transcriptomic) alterations between preclinical AD mouse models and human AD patients, both carrying mutations in the *PSEN1* gene. For this purpose, microarray data was used for both species, and gene expression differences between AD and non-AD conditions were assessed through linear modelling for each specie. To unveil the biological meaning behind this changes, gene set enrichment analyses (GSEA) were performed.

Mechanisms associated with the immune system, namely with the inflammatory response, appear up-regulated in both human AD patients and mouse models, whereas neurotransmitter trafficking processes appear down-regulated in both. The majority of the other most strikingly disrupted pathways varied between human and mouse, but were often in accordance with prior scientific knowledge on AD. However, a few of them appeared differently altered between species, such as diabetes mellitus associated pathways, that appeared down-regulated in human patients and up-regulated in AD mouse models.

The analysis of the joint dataset (resulting of merging the human and mouse datasets) unveiled synaptic and neuronal activity -related pathways as down-regulated in the disease common to both species, but less so in mouse AD compared to human patients. On the other hand, immune system genes and pathways were commonly up-regulated in the disease but more so in the human patients. These subtle variations between human and mouse transcriptomic information suggest that disease dynamics are potentially species-specific and reinforce the need to generate models that are able to more effectively replicate the human disease.

Additionally, we also identified compounds able to induce GE alterations opposite to those observed for the species-common component of the disease, as well as those capable of emulating human-specific AD-induced transcriptomic alterations. Those candidate compounds can be further explored as therapeutics to combat AD or as a vehicle to obtain novel and innovative mouse models that more effectively replicate the transcriptomic signature of the actual human disease. Two groups of compounds were considered: those with prescription information for neurology-related conditions and those prescribed for other conditions. Moreover, only compounds positioned at the phase III of clinical trials or already available in the market were considered.

For future work, it would be possible to perform a similar analysis but to assess genetic perturbations (i.e. knockdowns or overexpression) rather than compounds, which could, likewise, be able to induce an opposite transcriptomic profile to that of the species-common disease, and of those that could promote the development of a human AD signature in a mouse model.

Moreover, given the complexity of the brain in terms of cell type composition and interactions between cell types, and the consequences of a neurodegenerative disease upon these, it would be interesting to incorporate brain cell-type-specific signatures as explanatory variables in the linear model used to estimate GE changes, in order to decouple AD-associated cell-type-specific and systemic GE alterations from brain cellular composition changes (namely neuronal loss).

The end goal of the present project would be to evaluate the effects of carefully selected genetic perturbations and compounds in cell lines and mouse models, in order to obtain a model able to more accurately develop the human AD.

Keywords: Alzheimer’s disease; transcriptomics; PSEN1; mouse models; compounds

RESUMO

A doença de Alzheimer (AD) é uma doença progressiva e irreversível do sistema nervoso central, sendo atualmente a demência mais prevalente a nível mundial, cuja incidência aumenta com o avançar da idade. Esta patologia caracteriza-se pela acumulação extracelular de placas insolúveis de péptido amiloide-beta (Aβ), e pela acumulação intracelular de proteína tau irregularmente hiperfosforilada sob a forma de agregados fibrilares. Outras características patofisiológicas incluem morte neuronal e perda de sinapses, exacerbação do sistema imunitário e inflamação crónica, e atrofia cerebral.

O presente rápido envelhecimento da população mundial prevê, com o aumento da proporção de população envelhecida, um igual aumento da prevalência e incidência de doenças neurodegenerativas associadas à idade, como é o caso das demências, categoria em que se inclui a AD. Atualmente não existe um tratamento eficaz que abrande ou impeça a progressão desta doença, o que, simultaneamente com a escassez de aprovação de novos medicamentos que se tem sentido na última década, constitui uma preocupação social, económica e de saúde pública. Adicionalmente, a maioria dos ensaios clínicos em doenças neurodegenerativas, inclusive em AD, apresenta elevadas taxas de insucesso, especialmente a nível dos ensaios de toxicidade e eficácia. O insucesso nesta fase dos ensaios reflete as dificuldades de transpor os resultados obtidos através dos modelos animais durante os estudos pré-clínicos para a doença humana, sugerindo que esta não será bem representada por estes modelos. Neste sentido, é imperativo avaliar as diferenças moleculares que distinguem os modelos animais e os doentes com Alzheimer em termos da fisiopatologia da doença, e também desenvolver diferentes abordagens que possam auxiliar a descoberta de modelos animais mais representativos da AD.

Nesse sentido, este projeto propõe avaliar os perfis de alteração de expressão génica (também referenciadas como alterações transcritómicas) entre amostras de cérebro de controlos e doentes com AD, tanto para amostras humanas como para amostras obtidas a partir de modelos animais, sendo que em ambas as espécies as amostras relativas aos portadores de um fenótipo de doença apresentam mutações no gene da presenilina 1 (*PSEN1*). Com esta abordagem pretendeu-se comparar os perfis de alterações transcritómicas induzidos pela AD no cérebro obtidos para cada uma das espécies através de modelação linear de dados de *microarrays,* sendo que para essa análise foi considerada, para além da condição (controlo *versus* doente), outra informação sobre as amostras como a idade do dador. A interpretação biológica dessas alterações transcritómicas foi feita por análise de alguns genes encontrados diferencialmente expressos, e também com recurso a *Gene Set Enrichment Analysis* (GSEA),

um método que identifica as vias metabólicas mais desreguladas, tendo por base o perfil das alterações transcritómicas entre as condições em estudo.

As amostras de ratinhos com AD consideradas no presente estudo dividem-se em três categorias: (1) animais com mutações exclusivamente a nível do *PSEN1*, e animais que adicionalmente são portadores de mutações no gene *APP*, podendo estes ser (2) heterozigóticos ou (3) homozigóticos. Os três modelos mostraram inexistente ou fraca correlação com a doença humana, aquando da comparação dos perfis de alteração transcritómica. Adicionalmente, os dois primeiros modelos não apresentaram diferenças de expressão significativas entre as amostras controlo e as amostras doentes. Considerando que o modelo homozigótico com mutações em *PSEN1* e *APP* foi o único a apresentar alterações a nível do perfil transcritómico, todas as análises e comparações descritas consideraram apenas estes ratinhos.

Os resultados mostram que tanto em humano como em ratinho portadores de doença de Alzheimer existe uma sobre-expressão dos genes envolvidos nos mecanismos de regulação do sistema imunitário, nomeadamente a nível de inflamação crónica, e uma diminuição do transporte de neurotransmissores. As restantes vias mais alteradas com a AD diferem entre humano e ratinho, embora a maioria esteja alinhada com a bibliografia existente sobre a patologia. Algumas vias metabólicas também surgiram inversamente desreguladas entre as duas espécies.

Genes envolvidos em vias metabólicas de diferenciação, proliferação e apoptose celular, processamento de DNA e RNA, e mecanismos relacionados com o sistema cardiovascular surgiram sobre-expressos na doença humana; enquanto genes envolvidos em vias associadas com atividade sináptica e neuronal, canais de transporte de membrana, e com a diabetes surgiram sub-expressos. Contrariamente, no caso do ratinho, verificou-se um exacerbar da diabetes, juntamente com o de vias metabólicas relacionadas com a colesterol e interações celulares; e sub-expressão de genes envolvidos na atividade mitocondrial e respiração celular, e em mecanismos de expressão génica, nomeadamente a nível do spliceossoma.

Realizou-se também uma análise conjunta dos dados de humano e ratinho, com a qual se observou uma maior variância de expressão génica entre os controlos humanos e os indivíduos doentes humanos, comparativamente com os ratinhos controlo e doentes, reforçando a possibilidade de que o desenvolvimento e a progressão da doença em ratinho não sejam demarcados o suficiente para que, a nível do transcritoma, exista uma explícita diferenciação entre as condições de doença e de não-doença.

As diferenças de expressão génica para os dados conjuntos foram modeladas linearmente, incorporando nos modelos, como variáveis, informação não só sobre a idade e condição das amostras, mas também sobre a espécie a que pertencem. Deste modo, foi possível isolar o efeito doença do efeito espécie e obter as diferenças transcritómicas ocorridas mais preponderantemente em humano e em ratinho, bem como as diferenças comuns às duas espécies, ou seja, independentes da espécie.

Esta análise revelou que a diminuição da atividade neuronal e sináptica está associada à AD, mas que surge menos afetada nos modelos de ratinho comparativamente aos doentes humanos. Quanto aos genes envolvidos nos mecanismos de regulação do sistema imunitário que também se revelam sobre-expressos na doença geral, encontraram-se mais sobre-expressos na doença humana do que no modelo de ratinho considerado. Estas subtis diferenças entre a informação transcritómica do humano e do ratinho sugerem que as dinâmicas associadas à AD possam ser específicas da espécie, reforçando a necessidade de ajustar os modelos animais para que simulem mais eficientemente a patologia humana.

Este projeto teve ainda como objetivo encontrar compostos e perturbações genéticas (*knockdowns* ou sobre-expressões) com potencial de recapitular as diferenças de expressão mais específicas da AD humana, para que possam ser administrados/manipulados em modelos de ratinho com o intuito de melhorar modelos já existentes ou encontrar um novo e aperfeiçoado modelo animal que replique de forma mais fidedigna as alterações decorridas da doença humana. Adicionalmente, também serão de interesse perturbações químicas e genéticas com capacidade de replicar perfis transcritómicos antagónicos daquele encontrado para a generalidade da doença, que possam ser utilizados como novas terapêuticas ou como objetos de estudo dos mecanismos associados à AD.

Para o propósito mencionado acima, recorreu-se à base de dados do *Connectivity Map*, que inclui informação transcritómica para diversas linhas celulares, antes e após lhes serem administrados diferentes compostos ou alterações genéticas. Usando um *software* desenvolvido no nosso laboratório, *cTRAP*, podemos, a partir dos perfis de alteração de expressão genética encontrados no nosso estudo, obter as perturbações químicas e genéticas que recapitulem as alterações transcritómicas do nosso interesse.

No tempo do estudo, analisou-se apenas as perturbações químicas, fazendo uma separação entre dois grupos de compostos: (1) aqueles com indicações para doenças do foro neurológico e (2) os com indicação de foro não neurológico. Para cada uma das duas categorias foram

selecionados os 10 compostos mais relevantes, isto é, aqueles que estatisticamente estão mais correlacionados com as alterações transcritómicas de interesse. Apenas compostos em fase III de desenvolvimento clínico ou disponíveis no mercado foram considerados.

A análise similar das perturbações genéticas fica então referenciada para trabalho a desenvolver no futuro. Adicionalmente, dada a complexidade celular do sistema nervoso central em termos de heterogeneidade e proporção celular, a qual é afetada em estados de doença, e especificamente neste caso de doenças neurodegenerativas, seria de interesse acrescentar uma assinatura que distinga os vários tipos celulares à informação proporcionada ao modelo linear utilizado para derivar os perfis de alteração transcritómica. Desta forma, seria possível distinguir alterações de expressão genética associadas a mudanças na composição celular daquelas relacionadas com mecanismos específicos da AD.

O objetivo final do projeto será testar perturbações químicas e genéticas escolhidas cuidadosamente, em linhas celulares e em modelos de ratinho, e testar a sua capacidade em gerar um modelo animal cujo desenvolvimento e progressão da AD seja mais similar ao observado em condições de doença humana; bem como o potencial dos mesmos em reverter características desta patologia.

Palavras-chave: doença de Alzheimer; transcritómica; PSEN1; modelos de ratinho; compostos

CONTENT INDEX

ACKNOWLEDGEMENTS .... ***i***

ABSTRACT .... ***iii***

RESUMO .... ***v***

LIST OF TABLES .... ***xii***

LIST OF FIGURES .... ***xii***

LIST OF ABBREVIATIONS AND SYMBOLS .... ***xv***

CHAPTER I – INTRODUCTION ***1***

1. Alzheimer’s disease **1**

1.1. Relevance of Alzheimer’s disease 1

1.2. Pathophysiology of Alzheimer’s disease 2

1.3. Classification of Alzheimer’s disease forms and Risk Factors 6

1.4. Clinical diagnosis, symptomatology and disease stages 8

1.5. Current therapeutic options 9

2. Failure rate of clinical trials **9**

3. Animal models **12**

4. Transcriptomics and bioinformatics **14**

4.1. Value of gene expression data 14

4.2. Gene expression assessment 15

4.3. AD transcriptome comparison studies between human and mouse models 17

5. Motivation and project goals **19**

CHAPTER II – MATERIALS AND METHODS ***21***

6. Datasets **21**

6.1. Human dataset 21

6.2. Mouse dataset 22

6.3. Joint dataset 24

7. Software **24**

8. Methods **24**

8.1. Data pre-processing 24

8.1.1. Import data files 24

8.1.2. Data transformation and normalization 25

8.1.3. Probe set summarization and expression values 25
8.1.4. Quality control and outlier removal 26
8.2. Gene Annotation 27
8.2.1. Probe ID to gene symbol conversion 27
8.2.2. Mouse to human orthologue gene symbol conversion 28
8.3. Statistical hypothesis testing 28
8.3.1. Significance and p-value 28
8.3.2. Multiple testing correction 29
8.4. Statistical tests 29
8.4.1. Gaussian distribution assessment 29
8.4.2. Variance assessment 31
8.4.3. Correlation analysis 31
8.5. Principal Component Analysis 32
8.6. Clustering analysis 34
8.7. Multiple linear modelling 34
8.7.1. Interaction coefficient and centring variables 36
8.7.2. Differential gene expression analysis with a Bayesian approach 38
8.7.3. Visualizing differential gene expression 40
8.7.4. Choosing the most suitable linear model 42
8.8. Gene Set Enrichment Analysis 44
8.9. Identification of DE-recapitulating and DE-reverting genetic and chemical perturbations 45
8.9.1. DAVID software 46

CHAPTER III – RESULTS AND DISCUSSION* *47

9. Quality control and sample removal for human and mouse data 47
9.1. Human dataset 47
9.2. Mouse dataset 48

10. Human dataset analysis 49
10.1. Finding explanatory variables with an impact on gene expression 49
10.1.1. Age and PMI correlation 50
10.1.2. Assessing condition confounding with Age and PMI variables 51
10.2. Exploring AD-induced DEG in human patients 54
10.3. Disrupted biological pathways in human AD brains 57

11. Mouse dataset analysis 60
11.1. Finding explanatory variables with an impact on gene expression 60

11.2. Exploratory transcriptomic characterization of the three mouse datasets 61

11.3. Choosing the best mouse dataset for comparison with the human dataset 62

11.4. Exploring AD-induced DEG in the HO mouse model 65

11.5. Disrupted biological pathways in AD mouse models' brains 66

12. Summary of findings for the human and mouse independent datasets 69

13. Joint dataset 71

13.1. Choosing the explanatory variables for linear modelling 71

13.2. Decoupling species-specific and species-common AD-induced GE changes 72

13.3. Exploring AD-induced DEG in the joint dataset 75

13.3.1. AD-induced DEG common to mouse models and human patients 75

13.3.2. Species-specific AD-induced DEG for mouse models and human patients 76

13.4. AD-induced disrupted biological pathways in the joint dataset 79

13.4.1. AD-induced alterations common to mouse models and human patients 79

13.4.2. Species-specific AD-induced alterations for mouse models and human patients 80

13.5. Interaction coefficient vs PSEN from the independent datasets 82

13.6. Unveiling compounds that replicate human AD transcriptomic changes (*cTRAP*) 85

CHAPTER IV – CONCLUSIONS 90

1. Final remarks 90

2. Future perspectives 91

CHAPTER V – BIBLIOGRAPHY 94

CHAPTER VII – SUPPLEMENTARY INFORMATION .... A

A. Methods .... A

B. Quality control and sample removal .... B

a. Human dataset .... B

b. Mouse dataset .... D

C. Dataset analysis .... G

a. Human dataset .... G

b. Mouse dataset L

c. Joint dataset .... O

LIST OF TABLES

Table 1 | Dataset information summary 23
Table 2 | Example of a design matrix 35
Table 3 | Example of a design matrix with an *Interaction* coefficient 36
Table 4 | Linear Models 38
Table 5 | Statistics of differential GE 39
Table 6 | Up- and down-regulated mechanisms in the separate human and mouse datasets 70
Table 7 | Candidate compounds recapitulative of human-specific AD 86
Table 8 | Candidate compounds for reverting of non-specific AD 88

Table S1 | Summary of R functions, and respective packages, used for the analysis .... A

LIST OF FIGURES

Figure 1 | Estimated and projected global population by age group from 1950 to 2100 1
Figure 2 | Alzheimer's disease pathology 2
Figure 3 | Alzheimer's pathological hallmarks 4
Figure 4 | Subclasses of Alzheimer's disease 7
Figure 5 | Development of a new pharmaceutical drug 10
Figure 6 | mRNA quantification techniques: microarray and RNA sequencing 17
Figure 7 | Affymetrix microarray artefacts 26
Figure 8 | Normal or Gaussian distribution 30
Figure 9 | Principal Component Analysis 33
Figure 10 | Centring of the *Species* variable 37
Figure 11 | Correlation between t-statistic and logFC and B statistic values 40
Figure 12 | Density scatter plots 41
Figure 13 | Fitting data 42
Figure 14 | Comparison between statistics from different models 43

Figure 15 | Gene Set Enrichment Analysis 45

Figure 16 | PCA before sample exclusion coloured by condition (Human) 48

Figure 17 | PCA before outlier exclusion coloured by condition (HO Mouse) 48

Figure 18 | Principal Component Analysis (Human)...49

Figure 19 | Age and PMI comparison 50

Figure 20 | Age and PMI t-statistic comparison..51

Figure 21 | Age and PMI distributions for Controls, EOAD and PSEN samples...............51

Figure 22 | t-statistic comparison between explanatory variables of *PMI*-differing models' 52

Figure 23 | t-statistic comparison between Age and Condition variables...........................53

Figure 24 | t-statistic comparison of Age-differing models 54

Figure 25 | DEG in human AD...55

Figure 26 | Altered Reactome pathways for the *PSEN* coefficient (Human)......................57

Figure 27 | Genes that contribute to the Alzheimer's disease KEGG pathway 59

Figure 28 | Age and PSEN t-statistic comparison (Mouse) 60

Figure 29 | PCA of the mouse datasets 61

Figure 30 | Volcano plots of differential expression for the mouse datasets 62

Figure 31 | Human and mouse *PSEN* t-statistic comparison..63

Figure 32 | PCA of the HO mouse dataset – highlighting 2 month old mice......................64

Figure 33 | DEG in mouse model of AD...66

Figure 34 | Reactome pathways differentially expressed for the *PSEN* coefficient (Mouse) 68

Figure 35 | Gene expression distribution for the *SNRPB* gene 69

Figure 36 | Principal Component Analysis (Joint dataset)...71

Figure 37 | t-statistic comparison between explanatory variables of the same model 73

Figure 38 | t-statistic comparison of the same variable between centring-differing models 74

Figure 39 | AD-induced DEG that are common to mouse and human (Joint dataset)........76

Figure 40 | Interpretation of the *Interaction* coefficient based on different GE profiles 77

Figure 41 | AD-induced DEG that are species-specific – *PSEN*/*Species* interaction (Joint dataset)...77

Figure 42 | Reactome pathways differentially expressed for the *PSEN* coefficient (Joint dataset)...79

Figure 43 | Reactome pathways differentially expressed for the *Interaction* coefficient (Joint dataset) 80

Figure 44 | Gene expression distribution for the *SUMO1* gene 81

Figure 45 | *Interaction* from the joint dataset vs *PSEN* coefficients from the independent datasets 82

Figure S1 | Raw chip-images (Human) ......B

Figure S2 | NUSE and RLE plots (Human) C

Figure S3 | Intensity distribution boxplots before sample exclusion (Human) C

Figure S4 | Heatmap on normalized GE (A) before and (B) after sample exclusion (Human) ......D

Figure S5 | PCA after sample exclusion coloured by condition (Human) ......D

Figure S6 | Intensity distribution boxplots prior to outlier exclusion (HO Mouse) ......E

Figure S7 | Heatmap on normalized GE before outlier exclusion (HO Mouse) ......F

Figure S8 | PCA after outlier exclusion coloured by condition (HO Mouse) ......F

Figure S9 | Percentage of variance explained by each principal component (Human) ......G

Figure S10 | Normality assessment for age of human samples ......G

Figure S11 | Normality assessment for PMI of human samples ......H

Figure S12 | Volcano plots that supported PMI exclusion from linear models I

Figure S13 | *PSEN* disrupted Biological Processes and Kegg pathways (Human) ......J

Figure S14 | Alzheimer's disease KEGG pathway ......K

Figure S15 | PCA after outlier exclusion coloured by condition (single-mutant Mouse) L

Figure S16 | PCA after outlier exclusion coloured by age (single-mutant Mouse) L

Figure S17 | PCA after outlier exclusion coloured by condition (HET Mouse) ......M

Figure S18 | PCA after outlier exclusion coloured by age (HET Mouse) ......M

Figure S19 | PCA after outlier exclusion coloured by age (HO Mouse) ......N

Figure S20 | *PSEN* disrupted Biological Processes and Kegg pathways (Mouse) ......N

Figure S21 | *PSEN* disrupted Biological Processes and Kegg pathways (Joint dataset) ......O

Figure S22 | PSEN/Species *Interaction* disrupted Biological Processes and Kegg pathways (Joint dataset) ......P

LIST OF ABBREVIATIONS AND SYMBOLS

ACh	Acetylcholine	INF-γ	Interferon gamma
AChE	Acetylcholinesterase	KEGG	Kyoto Encyclopaedia of Genes and Genomes
AD	Alzheimer's disease	LOAD	Late-onset Alzheimer's disease
APOE	Apolipoprotein E	MAPT	Microtubule-associated protein tau
APP	Amyloid precursor protein	MCI	Mild Cognitive Impairment
$A\beta$	Amyloid-beta	mRNA	Messenger RNA
$A\beta_{40}$	$A\beta$ isomer with 40 residues	NCBI	National Center for Biotechnology
$A\beta_{42}$	$A\beta$ isomer with 42 residues	NES	Normalized ES
cDNA	Complementary DNA	NF-kB	Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B cells
ChAT	Choline acetyltransferase	NFT	Neurofibrillary tangles
CMap	Connectivity map	NMDA	N-methyl-D-aspartate
CNS	Central Nervous System	normex	Normal-exponential
CSF	Cerebrospinal fluid	NUSE	Normalized unscaled standard error
DE	Differential expressed	PC	Principal Component
DEG	Differential expressed genes	PCA	Principal Component Analysis
DGE	Differential gene expression	PET	Positron-Emission Tomography
DMT	Disease-modifying treatment	PLM	Probe-level model
DNA	Deoxyribonucleic acid	PMD/PMI	Post-mortem Delay/Interval
ds-cDNA	Double stranded cDNA	PS1	Presenilin 1
EMBL	European Molecular Biology Laboratory	PS2	Presenilin 2
EOAD	Early-onset Alzheimer's disease	PSEN	Presenilin gene symbol
ES	Enrichment Score	Rho	Spearman's correlation coefficient
FAD	Familial Alzheimer's disease	RLE	Relative log expression
FC	Fold-change	RMA	Robust multi-array analysis

FDR	False Discovery Rate	RNA-Seq	RNA sequencing
GE	Gene Expression	ROS	Reactive Oxygen Species
GEO	Gene Expression Omnibus	RT-qPCR	Reverse transcriptase quantitative polymerase chain reaction
GO	Gene Ontology	SAD	Sporadic Alzheimer’s disease
GSEA	Gene Set Enrichment Analysis	SSD	Sum of Squared Distances
HET	Heterozygous	TNF-α	Tumor necrosis factor alpha
HO	Homozygous	TREM2	Triggering Receptor Expressed on Myeloid cells 2
IL-1β	Interleukin 1 beta	WHO	World Health Organization

CHAPTER I – INTRODUCTION

1. Alzheimer's disease

1.1. Relevance of Alzheimer's disease

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disease of the central nervous system (CNS). It was first described in 1907 by the German psychiatrist and neuroanatomist Alois Alzheimer as "an unusual illness of the cerebral cortex" [1–3], and nowadays is the most prevalent age-related dementia worldwide [4–6].

According to the World Population Prospects 2019 Report, the worldwide population aged 65 years or older will more than double by 2050, reaching 1.5 billion people (Figure 1) [7]. Furthermore, the World Health Organization (WHO) assesses that 5-8% of individuals, above 60 years old, will have dementia, a syndrome that currently impacts 50 million people globally, with around 10 million new cases per year [8]. In total, WHO envisions that 152 million people will develop dementia by 2050, the majority of which will live in low and medium income countries [8].

The prevalence and incidence of AD increases with age, the onset being around 65 years old, and peaks in the range from 70 to 90 years old [6,7,9]. Hence, it is expected a parallel increase in the number of AD cases along with the population ageing.

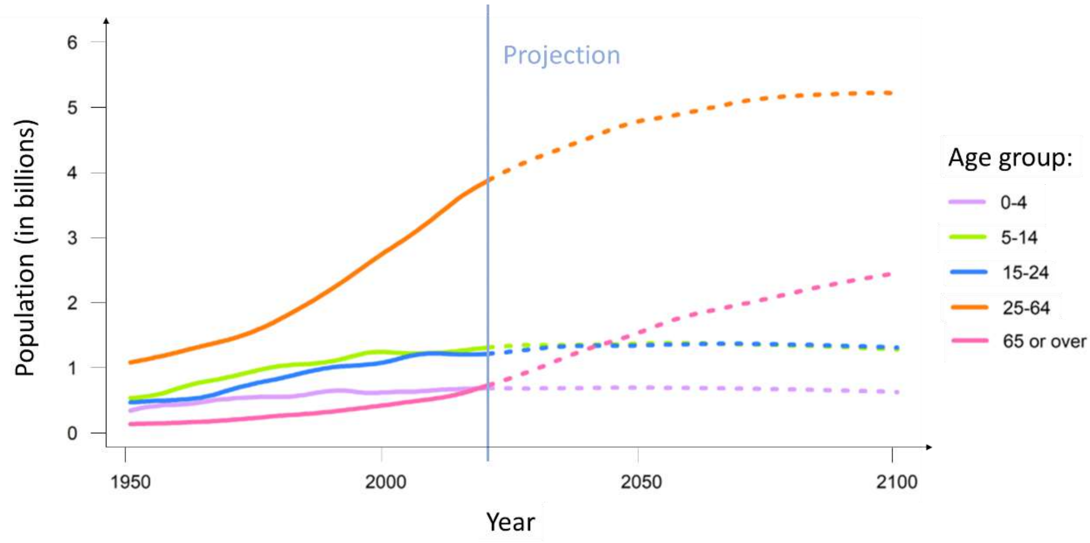


Figure 1 | Estimated and projected global population by age group from 1950 to 2100

World population (in billions) among five age groups, across time. Values were estimated since 1950 until 2018, and projected for future years until 2100. The most significant difference is among the 25-64 and 65+ years old groups, the latter being the fastest-growing and projected to more than double. Age groups below 24 years old are not expected to undergo significant changes. Based on *World Population Prospects 2019* report [7].

1.2. Pathophysiology of Alzheimer's disease

AD pathology comprises two major events: (A) the extracellular deposition of insoluble amyloid-beta (Aβ) plaques, and (B) the intracellular accumulation of abnormally phosphorylated tau protein into neurofibrillary tangles (NFT) in degenerating neurons [6,10,11]. In AD, Aβ plaques usually first affect the frontal and temporal lobes, hippocampus and limbic system, while neurofibrillary tangles originate in the temporal lobe and hippocampus [6]. Overall, both hallmarks spread to affect the entire neocortex and hippocampus [6], as illustrated in Figure 2.

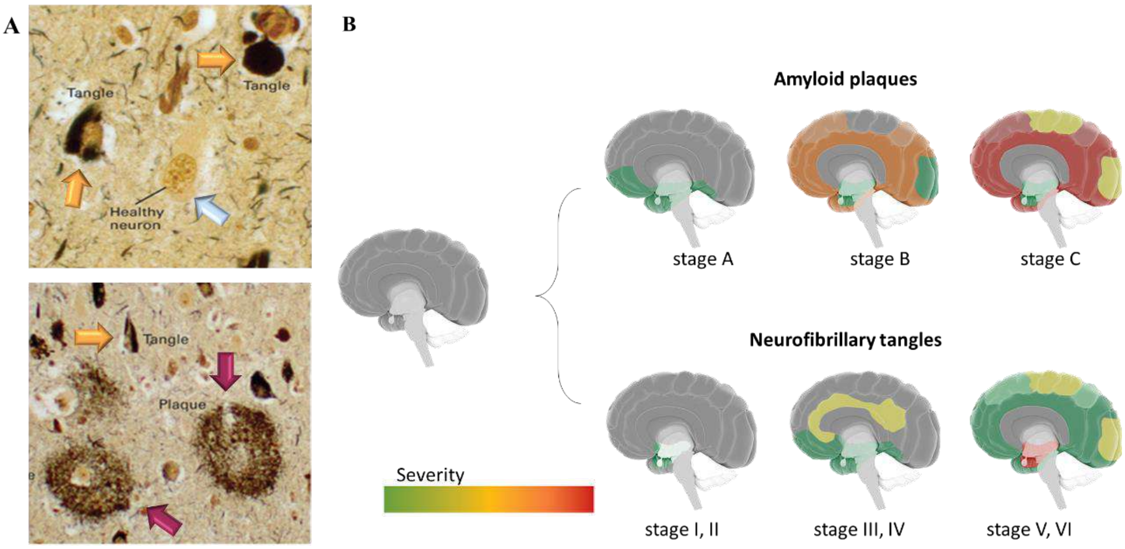


Figure 2 | Alzheimer's disease pathology

(A) Histopathology images of neurofibrillary tangles (orange arrows) and amyloid plaques (pink arrows) from a brain with AD, near a healthy neuron (light-blue arrow) – adapted from Kandel *et al* [12]. **(B)** Schematic representation of amyloid (top) and tau (bottom) pathologies progression through the brain – inspired by Masters *et al* [6].

In healthy neurons, phosphorylated tau protein, encoded by the microtubule-associated protein tau (*MAPT*) gene, regulates and stabilizes microtubules involved in neuronal development and in axonal transport [10,13]. In AD-damaged neurons, chemical alterations cause dissociation of tau and microtubules, with detached tau proteins eventually aggregating to form fibrillary tangles that block the neuronal transport system and hinder mechanisms such as metabolic and synaptic pathways [6,10]. Affected neurons can also excrete fibrillary tau into the intercellular space, which is then internalized by healthy neighbouring neurons and initiates tau pathology [6]. Hyperphosphorylated tau levels often appear augmented in the brain and cerebrospinal fluid (CSF) of AD patients [6] (Figure 3A).

On the other hand, amyloid plaques arise from amyloid precursor protein (APP), whose physiological function is not yet fully understood, although some studies hypothesize a direct involvement on synaptic maintenance and plasticity [6,14,15] and an indirect role on neuroprotection, neurite growth, signal transduction and apoptotic signalling, through its metabolic products [11,14]. APP can undergo processing by non-amyloidogenic pathway if cleaved by α-secretases, or the amyloidogenic pathway if sequentially cleaved by β-secretases and γ-secretases [16–19]. The γ-secretase complex includes presenilin 1 (PS1) and presenilin 2 (PS2) transmembrane proteins, respectively encoded by *PSEN1* and *PSEN2* genes [6,19,20]. These proteins can also be found intracellularly in endosomes and within the Golgi complex and endoplasmic reticulum [21].

Given that the knockout of PS1 leads to brain abnormalities and low longevity of animal models, this protein is postulated to have an important role in brain development and survival, contrarily to PS2, since the knockout mice are less affected [21]. The APP amyloidogenic metabolic pathway generates several length-differing Aβ peptides as products, of which the $Aβ_{40}$ and $Aβ_{42}$ oligomers, and the ratio of these two peptides, are the most relevant in AD [6,17,18]. In healthy individuals, $Aβ_{40}$ is usually benign and more abundant than the longest isomer [17]. Despite being neurotoxic, $Aβ_{42}$ is produced at a physiological rate that can be cleared in an healthy organism [17]. Regardless of mutations on *PSEN1* and *PSEN2* potentially resulting in genetic gain of a toxic function, biochemically they seem to lead to loss of function of the γ-secretase complex, resulting in an incomplete metabolization of APP protein and an increase in the longest Aβ peptide variant of 42 residues ($Aβ_{42}$) [22].

Notwithstanding the neuronal and astrocytic physiological production of Aβ, mutations in presenilins or APP genes lead to the overproduction of the toxic isomer, thus increasing the $Aβ_{42}$/ $Aβ_{40}$ ratio to levels beyond clearance capacity, which culminates with Aβ excretion into the extracellular space [6,17]. Neurotoxic $Aβ_{42}$ is more prone to aggregate into soluble oligomers that coalesce until reaching an insoluble fibril state, which ultimately deposits into senile plaques in the extracellular space between neurons [6,10,17,19]. Some studies suggest that the soluble Aβ oligomers that concentrate around the plaques are more toxic and better correlated with disease progression than the actual plaque deposition [19,23] (Figure 3B).

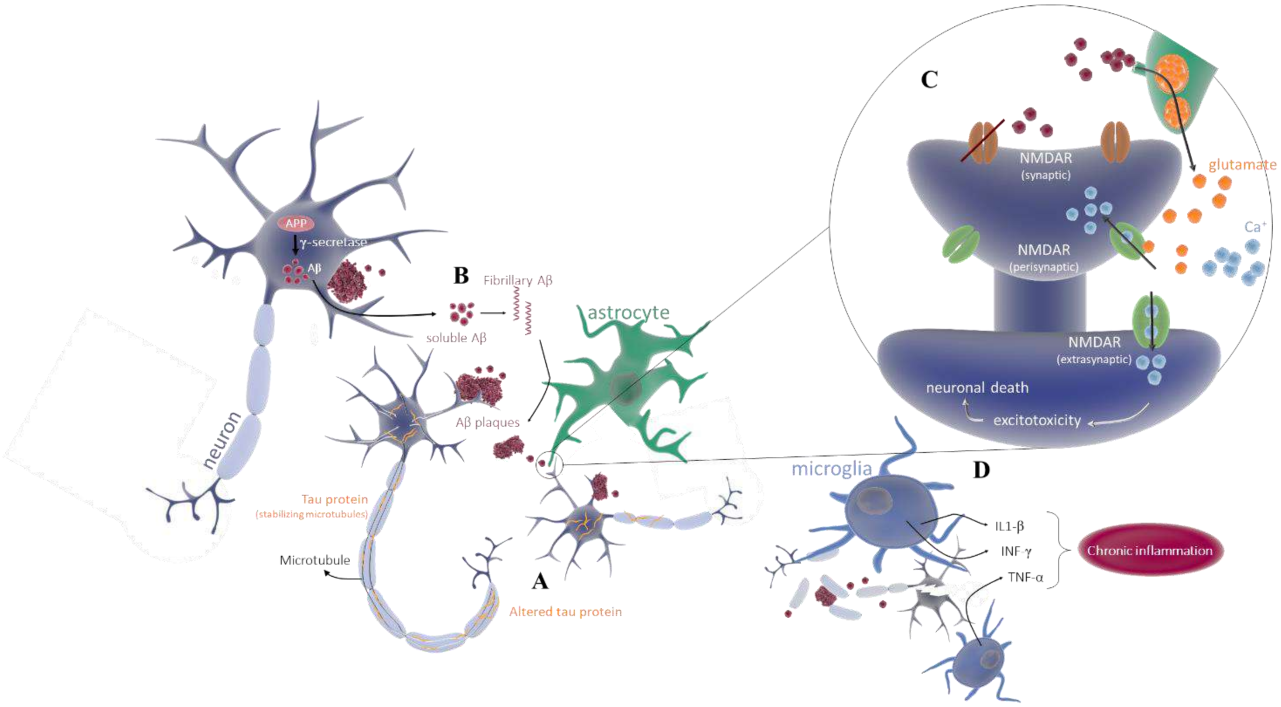


Figure 3 | Alzheimer's pathological hallmarks

AD is characterized by **(A)** tau pathology where tau protein decouples from microtubules and forms intracellular fibrillary tangles **(B)** amyloid pathology characterized by extracellular Aβ deposition into fibrillary plaques that are surrounded by soluble Aβ, **(C)** which is able to directly or indirectly affect NMDA receptors (NMDAR). **(D)** Both pathologies lead to neuronal death that, along with amyloid plaques, attract microglia, astrocytes and macrophages. This immune cell types release pro-inflammatory cytokines and other toxins that result in chronic inflammation and damage of the brain. Figure inspired by Master *et al* [6].

Glutamate is the most abundant neurotransmitter in the human brain and is an agonist ligand of glutamatergic receptors, which include the ionotropic glutamatergic N-methyl-D-aspartate (NMDA) receptors [24]. These ligand-gated ion channels have an important role during rapid neuronal communication and excitatory synapses [24], which are crucial during the learning process and memory formation [24,25]. Overall, the extracellular deposited Aβ is preponderantly located near excitatory synaptic clefts and NMDA receptors where, particularly the soluble and unrestricted Aβ form, can affect calcium influx and synaptic transmission [24,26]. The oligomers interact with neighbouring astrocytic receptors and induce glutamate exocytosis that, in turn, activates the extrasynaptic and perisynaptic NMDA receptors of neurons through the GluN2B subunit, thus increasing the influx of calcium ions (Ca^{2+}), consequently leading to a excitotoxicity state [17,24] (Figure 3C). As a consequence, of calcium overload, depolarisation of mitochondrial membrane potential is observed, together with an increase in reactive oxygen species (ROS), that not only promote phosphorylation of tau protein but also incite oxidative stress, leading to synaptic damage and neuronal death [17]. Furthermore, some studies have found evidence on oligomeric Aβ–induced loss of function of the synaptic NMDA receptors [25,26] (Figure 3C), possibly through degradation of EphB2 [26], a tyrosine kinase responsible for maintaining the structural integrity of these receptors, thus impairing long term synaptic potentiation and affecting cognition and memory [24,25].

Similarly to the glutamatergic system, cholinergic events are involved in both cognitive development, information processing and memory recall [27]. Physiological acetylcholine (ACh), the neurotransmitter associated with cholinergic neurons, is synthesized by the choline acetyltransferase (ChAT) enzyme, and is inactivated by the acetylcholinesterase (AChE) prior to its release in the synaptic cleft [27]. A decrease in neuronal ChAT expression, characteristic of AD, diminishes the amount of available ACh and hinders synaptic transmission and neuronal survival [27]. Some studies also suggest that the AChE enzyme may interact with Aβ peptide and exacerbate Aβ aggregation and deposition [28,29].

Additionally, Aβ plaques affect the vascular system, reducing the brain supply of oxygen and glucose, and damaging the blood-brain barrier [6,10,17]. Elimination of Aβ debris and dead cells often involves their uptake by microglia, which are considered the immune cells of the CNS, and astrocytes [17]. The elevated levels of Aβ debris occurring in AD exhaust cells' clearance capacity and lead to high concentration of activated microglia and astrocytes. These cells release pro-inflammatory cytokines, including interleukin 1β (IL-1β), tumour necrosis factor alpha (TNF-α) and interferon gamma (INF-γ), creating an imbalance in pro- and anti-inflammatory signalling that culminates in chronic inflammation of the diseased brain [10,17,30,31] (Figure 3D), which is known to exacerbate both tau and Aβ pathologies [31]. Besides cytokines, these activated glial cells also excrete toxic products including ROS [31].

TNF-α is crucial to initiate and regulate the inflammatory cascade of events and interplays with the transcription factor NF-kB (nuclear factor kappa-light-chain-enhancer of activated B cells), and β– and γ–secretases during neuronal apoptosis [31]. By interfering in the APP cleavage process, TNF-α is tightly related with Aβ production. In turn, Aβ stimulates the increase of TNF-α levels through activation of microglial NF-kB, feeding a cyclical loop of inflammatory exacerbation [31]. Another critical cytokine for Aβ plaque deposition is IL-1β, which has been implicated in APP synthesis, in addition to regulating other pro-inflammatory cytokines, namely TNF-α [31].

Physiologically, the triggering receptor expressed on myeloid cells 2 (TREM2) is an innate immune phagocytic receptor that acts as a bridge between the extracellular milieu and microglial intracellular signalling pathways. *In vivo*, TREM2 expression increases when in the presence of pro-inflammatory agents, with elevated TREM2 peripheral levels having been detected in AD patients [32]. Moreover, the R47H TREM2 variant is considered a risk factor for late-onset AD [31,32].

The accumulation of synaptic loss, decreased supply of oxygen and glucose to the brain, reduced capacity of glucose metabolization and neuronal loss, consequentially culminates in brain atrophy [10,30].

The amyloid cascade hypothesis assumes deposition of amyloid plaques as the trigger event for AD development and has been the most studied and accepted theory in the scientific community for decades. However, the high failure rate of clinical trials of compounds that target the amyloid process has been rising questions in the scientific community and led to the upsurge of other theories such as the tau hypothesis, which states formation of neurofibrillary tangles as the starting event of the disease [19,33].

1.3. Classification of Alzheimer’s disease forms and Risk Factors

AD can be classified as Early-Onset Alzheimer’s Disease (EOAD) or Late-Onset Alzheimer’s Disease (LOAD), depending if the patient is younger or older than 65 years of age, at the moment of the diagnosis, respectively [34–36] (Figure 4).

The disease can also be classified into Familial Alzheimer’s Disease (FAD), or Sporadic Alzheimer’s Disease (SAD) [6,34,35]. FAD, that only represents 1% of all AD cases, is linked to the inherited genetic background and is thought to be related with increased formation and aggregation of Aβ [6,34,35]. On the other hand, SAD is the most common form comprehending over 95% of diagnosed patients, and for which the causes are unknown [6,35] albeit being thought to stem from the inability to clear the Aβ depositions [6].

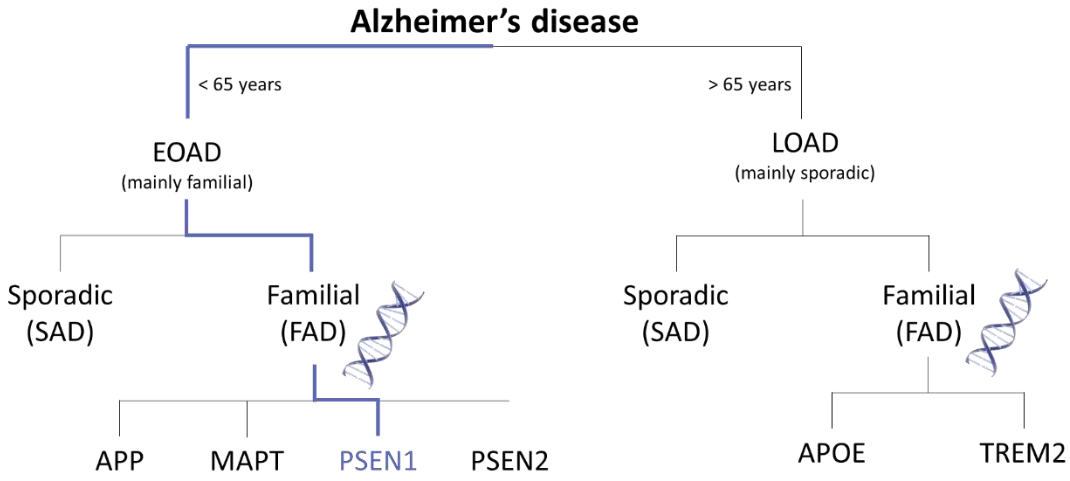


Figure 4 | Subclasses of Alzheimer's disease

AD can onset before (EOAD) or after (LOAD) the patient reaches 65 years of age, with its origin being genetic (familial AD) or sporadic. The current work focuses on AD caused by mutations in the presenilin-1 gene (blue path).

Although not exclusively, early-onset AD is usually of the FAD form while LOAD is commonly sporadic [6]. Despite these categorizations, all types of the disease are thought to have similar pathology and clinical symptoms [36].

FAD is associated with mutations in the *APP*, *PSEN1* and *PSEN2* genes, leading to alterations in APP cleavage and Aβ formation [6,34], and in *MAPT*, the gene coding for tau protein [6]. *PSEN1* is the most prevalently mutated gene [34]. Regarding SAD, despite its origin being unknown [35], genetic variations have been described in genes involved in the clearance of Aβ, affecting the organism ability to eliminate the plaques. Such genetic variants include apolipoprotein E (APOE) polymorphisms [6] and the R47H variant of *TREM2*, which has been associated with LOAD [31]. APOE is highly expressed in the brain, being the primary apolipoprotein involved in lipid metabolism in the CNS [6].

Besides genetics, several other risk factors have been identified, such as female gender and advanced age, with women being reportedly more affected by the disease [6,7,9,31] maybe due to, among other factors, their increased life expectancy [7] and decrease in the levels and/or action of sex hormones progesterone and estrogen [37]. Other conditions such as physical inactivity [38,39], low education levels [9,38,40], smoking [41], depression [38,42], mid-life obesity [9,38,43], metabolic disorders such as diabetes mellitus [9,38,44], mid-life hypertension and other cardiovascular pathologies [31,38,45–47], as well as traumatic brain injury [31,48] have also been linked to increased predisposition to develop AD. The latter three have been associated with immune system responses, such as inflammatory events that, as aforementioned, correspond to a hallmark of AD and prompt disease development [31].

1.4. Clinical diagnosis, symptomatology and disease stages

There are several methods that can be used for AD diagnosis, with the choice depending on the availability of the technique within the healthcare provider, its cost and the patient preference [49].

The evaluation can be non-invasive, which includes assessment of cognitive function through intellectual tests such as the clock test (where patients are asked to draw a clock figure [50]), as well as detection of amyloid plaque deposition through amyloid PET imaging [6,49]. Another diagnostic method, more invasive but less expensive than PET, is the lumbar puncture procedure, which measures the levels of $A\beta_{42}$, hyperphosphorylated tau protein and/or total tau protein content in the CSF [49].

AD symptoms start with memory loss, problem-solving difficulties, thinking impairment and confusion, followed by decline in oral and written communication and everyday tasks, such as getting dressed, behavioural and personality changes, agitation and depression [30,50,51]. Movement impairment is usually linked to later stages and, if severe, patients can become bedridden [30,51].

Diagnosed individuals fall into one of the three disease stages: preclinical AD, mild cognitive impairment (MCI) and dementia due to AD. In the first stage, pathophysiological hallmarks are detectable with diagnostic tools but clinical symptomatology is not yet present [30]. It is noteworthy that not all individuals diagnosed with preclinical AD further develop MCI or dementia [30]. Patients within the MCI stage show a stronger cognitive decline compared to regular aging, especially in thinking abilities, despite still being independent and able to perform everyday tasks [30]. In the last stage, dement individuals undergo a conspicuous impairment of memory, thinking and behavioural abilities, with symptoms deteriorating from mild to moderate to severe dementia, in parallel with neural damage [30].

1.5. Current therapeutic options

At the moment, there is no effective treatment neither to decrease nor cease the damage of AD [6,30]. Available therapies only temporarily improve symptomatology and can be classified into (1) pharmacological and (2) non-pharmacological treatments [6,30].

Pharmacological treatments operate either on increasing the amount of neurotransmitters or on blocking specific brain receptors to prevent excessive stimulation, acting on cholinergic, monoaminergic and glutamatergic systems [6,30,49]. Cholinergic inhibitors include donepezil, rivastigmine and galantamine, and prevent the inactivation of the acetylcholine neurotransmitter by the AChE [27], while memantine is a NMDA receptor inhibitor and acts on the glutamatergic system [6,49].

Regarding the non-pharmacological approaches, medical foods can be prescribed [6,52], mainly for patients that are intolerant or not respondent to the pharmacotherapy [6], which however has been shown more beneficial in pre-symptomatic or in early stages of AD [52]. Likewise, nutritional supplements aligned with physical exercise and lifestyle adjustments are also typical medical recommendations [6].

2. Failure rate of clinical trials

The drug development process starts with target identification and validation, followed by preclinical studies *in vitro* and *in vivo* that assess safety, toxicity, pharmacokinetics and efficacy [53]. After the preclinical stage, the third phase of drug development focuses on clinical trials, subdivided in four phases, conducted in human volunteers. The first phase (Phase I) only enrols healthy volunteers to determine safety, toxicity and the optimal drug dosage, while patients with the condition under study enter the trials in Phase II for evaluation of efficacy and secondary effects [53]. Phase III and IV further explore the previous studies, increasing the number of participants [53], with the fourth phase being posterior to market approval (Figure 5).

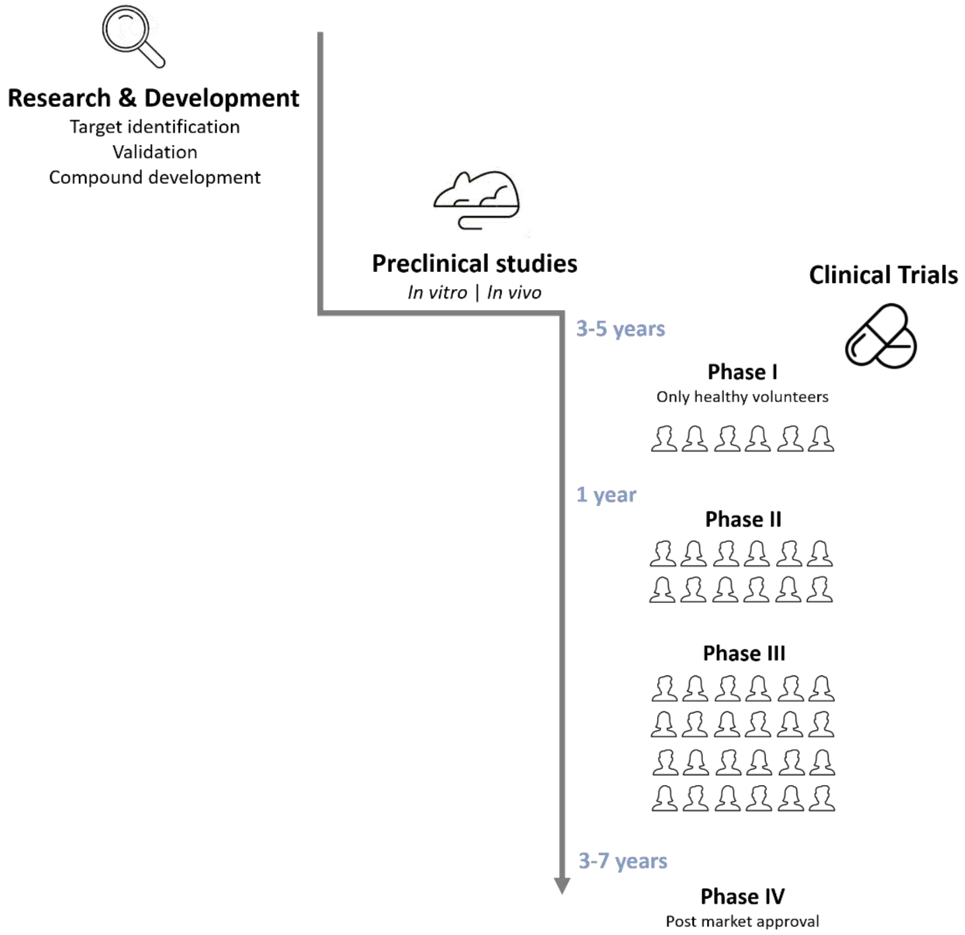


Figure 5 | Development of a new pharmaceutical drug

Drug development has three essential phases: research and development of new compounds, preclinical studies where the compound is tested *in vitro* and *in vivo*, and lastly clinical trials in human. The third and fourth sub-phases of clinical trials are the longest periods of drug development. Figure inspired by Orion Pharma's online available materials [54].

The most recently approved therapeutic drug for AD, Memantine, dates from 2003 [55,56], and was released by today's Actavis Generics that belongs to Teva Pharmaceutical Industries. Of the four currently approved therapeutic drugs, none constitutes a disease-modifying treatment (DMT) [49], being one considered as such if able to slow or completely interrupt the progress of disease [57].

In 2018, a total of 220 AD-related ongoing clinical studies were reported [57], encompassing not only therapeutic drugs, but also diagnostic techniques and other interventions, such as diet and exercise [57]. As of January 2018, an annual report stated 135 ongoing clinical trials testing 112 different agents, of which 63% were DMTs [58]. By July, five big-pharma compounds had already failed, four of which were in late-stage studies [55], belonging to Takeda Pharmaceuticals

[55,59], Merck & Co. [55,60], vTv Therapeutics [55,61], Eli Lilly and Astrazeneca [55,62], and Janssen [55,63]. Moreover, in the beginning of the year the pharma giant Pfizer announced that it is stepping out of the race of neuroscientific research and drug development after a series of failed clinical trials [64].

In February 2019, there were 156 AD-related ongoing clinical trials registered in ClinicalTrials.gov [65], with a total of 132 different therapeutic agents, of which thirty-two entered the trials in 2019 [56]. 73% of the overall therapeutic agents intended to modify the underlying molecular mechanisms of the disease, while the minority aimed for prevention or targeted cognitive and symptomatic enhancement [56]. Similar to last year, 2019 has been difficult for drug development targeting AD, given that, at the time of the present study, at least Roche [66], Biogen [67], Novartis and Amgen [68] have already abandoned some of their ongoing studies.

Defying all scrapped studies, the number of clinical trials seems to grow each year, with new approved trials entering the niche [56,58]. More than half of the trials are sponsored by the pharmaceutical industry, followed by 30-35% supported by academic centres [56,58]. A drug development program for AD costs around 5.07 billion euros, extends for an average of 13 years, since preclinical trials until market access approval, with the last Phase III being the most costly part of the trial [69].

Despite all efforts to convert biological breakthroughs into translational products for clinical application, none of the hundreds of drugs that underwent clinical trials until this point has been brought to the market since 2003 [70,71].

High failure rates of AD clinical trials have been associated to (1) the choice of the target pathological substrate [72,73], with the majority of DMTs aiming amyloid plaques [56], (2) insufficient comprehension of the underlying biochemical processes [72,73], (3) inadequate trial design, namely patient selection [72,73], and (4) incorrect drug dosage, high toxicity and/or low efficacy, which disclosures the limitations of extrapolating data obtained from preclinical animal models for human testing [71,73].

3. Animal models

Animal models have been, and currently are, vital for scientific advances regarding AD. They are extensively used to derive otherwise unreachable insights into the molecular pathways underlying the disease's pathophysiology. They also serve as preclinical models for drug target validation and testing of the therapeutic potential of innovative drugs, as well as other treatments, before clinical trials [70,74,75]. In summary, they have been useful for experiments that are overall impractical to execute in human patients.

Currently, there are 171 mouse models and 8 rat models available for research purposes [76], which struggle to effectively replicate the human disease, thus resulting in an elevated clinical trial failure rate of almost all pre-clinically approved therapies [70,71,74,75].

Mouse models do not naturally develop AD nor its pathological hallmarks possibly due to their low longevity [77] or the fact that their β-amyloid proteins do not tend to aggregate and form plaques [71]. Therefore, in order to study this disease, they have to be manipulated into carrying human mutations, making transgenic models the most commonly used ones [78]. Transgenic models are often genetically modified with human mutations, namely in *APP* and *PSEN* genes, resembling more the familial AD form, which represents only 1% of all cases, than the prevalent sporadic form [71,75]. This might be one of the reasons for therapies approved in preclinical trials to fail when tested in humans, especially because clinical studies often encompass patients with non-familial AD [71,75].

Furthermore, some of those mutations are introduced during the embryonic phase of the animal's development without any control of their genome location, quantity, transcription nor of the downstream RNA and protein processing [71,75]. All these mechanisms depend on the promoter used during insertion, which varies according to the animal model, contributing to discrepancies between models themselves and between the animal models and the human disease [71,75]. Since inserted transgenes are human, their expression in rodents is also highly unpredictable and can trigger mechanisms that are later confounded with disease pathology and lead to incorrectly interpreted results [71].

Regarding the pathophysiology, human AD is highly related to neuronal loss followed by brain atrophy, a hallmark that is not so evident in single-mutant mouse and rat models [70,71,75]. The low longevity of rodents might also prevent the full development of the disease, since AD is an age-related condition and more prone to occur in increased ages, as seen in humans [77].

Thus, in order to obtain a moderate or severe neuronal loss and a faster accumulation of Aβ plaques, double- or triple- mutated rodents are needed which, besides suffering brain atrophy before Aβ deposition, constitute a hazard regarding extrapolation of results due to excessive genetic manipulation leading to combinations and expression levels not found in human patients [70,71,75]. This accumulation of genetic mutations also results in an earlier disease onset, which differs from the human disease that appears in advanced ages [70]. This, allied to differences in the immune system that occur with aging, could possibly result in mouse models better representing early-onset AD, which are the minority of cases, or even preclinical AD [70]. The inflammatory pattern in rodents is also often less protuberant compared to humans, with chronic inflammation supporting AD development and progression [71].

Additionally, the pathological topography differs between humans and transgenic rodents, with Aβ plaques depositing first in the human neocortex, and, for the rodents, overall in hippocampus and cortex, despite its high dependence on the promoter used on the transgenes [71]. Beyond that, plaques are structurally different, with human plaques being insoluble while the rodent ones are fairly soluble, possibly due to dissimilar post-translational modifications [71].

The majority of rodent models fail to concomitantly develop Aβ plaques and neurofibrillary tangles, which has been a significant limitation for translational research [70,71,75,79]. The few that develop both hallmarks usually fail on neurodegeneration and loss of synapses, with all of them being genetically over-manipulated, which directly contributes to the extrapolation gap towards human disease [79]. The disease epigenetics and environmental contributions, of much importance to disease establishment and progression, are highly difficult to simulate in animal models, thus increasing the extrapolation gap [75]. Additionally, structurally similar homologue genes may have species-specific functions [71].

It has been suggest that adequate animal models should display the same genetic alterations as seen in the human disease and develop correspondent pathological features, while concordantly responding to therapeutics already approved to human patients [75]. The authors also note that a good model does not necessarily meet all known AD conditions, considering that not all AD cases are equal, being highly dependent on the stage desired to replicate [75].

4. Transcriptomics and bioinformatics

Transcriptomics is the scientific field that studies RNA transcripts of an organism and their functions, which has seen an exponential technological reform since the first published partial human transcriptome in 1991 [80,81], and nowadays is highly empowered by bioinformatic approaches [81]. Bioinformatics can be described as the application of computer software tools on data collected through research in the fields of, among others, life sciences, medical and biomedical fields, either in basic science or in applied/translational studies [82]. Those data can take a variety of forms, ranging from DNA, RNA or amino acid sequence, to protein structures and biological pathways, among others [82]. For the current project, mRNA datasets were the focus.

Computational techniques applied to life sciences research are emerging and revolutionizing how scientists perceive data by equipping them with powerful tools to handle and analyse scientific datasets from several different sizes and sources in effective and cost-efficient manners [82,83]. This analytical task becomes impractical for human end-users in case of high dimensionality data [82,83], being eased by high-tech evolution that allows to derive new knowledge that would possibly be unreachable by already existing laboratory techniques alone [82,83].

4.1. Value of gene expression data

Every individual has his/her own hereditary genetic information, with cells sharing the same genes. However, gene expression (GE) and post-transcriptional modifications are vital to provide cell variability and phenotypically differentiate an organism [81,84]. Post-translational changes play, as well, an important role in the case of protein-coding genes [81,84].

Genes are initially transcribed into mRNA, whose quantification can inform on gene activity [84]. GE is genetically and environmentally (including, for example, diet and physical performance, temperature or stress, and inflammatory status) influenced, being those differences important to outline cells' function, development stage and/or pathological stage [84]. On this note, a multitude of human diseases are associated with GE changes due to genetic mutations, polymorphisms or altered mRNA regulation [81,84]. Evaluating GE differences between conditions (for instance, healthy vs disease contexts), complements genomic studies by capturing a snapshot of the changes in gene regulation and transcription that may occur between them [81,84]. Moreover, comparing mRNA data of different tissues, conditions or longitudinal time points, assists in unveiling new insights on the biomolecular processes specifically occurring at dynamic states of our interest, such as a disease states [81], like AD.

Given the complexity of AD and all the aforementioned factors that can enter the equation of developing the phenotype, even when dealing with familial EOAD linked to the genetic background, it is imperative to consider transcriptomic data in its study. This type of data can not only potentially unveil biomolecular mechanisms and pathways related to the disease pathophysiology that would otherwise not be easily perceptible, but also be useful to unveil novel candidate targets that can be manipulated in order to elicit transcriptomic features of interest.

4.2. Gene expression assessment

At the moment, transcript quantification relies on two main techniques, microarrays and RNA sequencing (RNA-Seq) [81]. For the present work, we used data derived from microarray analysis. The microarray technique is advantageous compared to laboratory quantification methods, such as northern blotting and reverse transcriptase quantitative polymerase chain reaction (RT-qPCR), as it allows the simultaneous quantification of numerous transcripts [81].

Oligonucleotide microarrays or gene chips are solid slides printed with thousands of microscopic nano-wells (named spots), each containing a few copies of the same DNA oligonucleotide, which corresponds to a fragment of a gene of known sequence [81,84]. These oligomers, known as probes, are synthetized *in situ* and chemically attached to the slide in an orderly manner, with replicates of the same probe being strategically distributed in different spots throughout the chip [81,84] (Figure 6). The chips can either represent the whole known transcriptome of the organism in study or focus on genes of interest [84]. Affymetrix GeneChip® platforms are prominent amongst this type of microarrays.

Another contemporary technology is the BeadArray®, a bead-based microarray system, developed by Illumina, in which the probes constitute microscopic silica spheres coated with thousands of oligomers with the same sequence, which are randomly distributed within a chip micro-perforated with spots, similar to the one used by Affymetrix [85,86]. Probe identification is accomplished through a series of decoding hybridizations [85].

For GE assays, the mRNA under study is firstly isolated and reversely transcribed into cDNA, then fluorescently labelled and hybridized to the microarray. Lastly, a laser scanner is used to measure the intensity of the fluorophore [81,84]. The spot in which the hybridization occurs identifies the probe and, therefore, informs on the transcript sequence, while the measured intensity values relate to transcript abundance and are converted to GE values [81,84]. Samples from different conditions are assigned to different chips in order to assess GE information that can then be compared.

Notwithstanding all welfares related to the use of microarrays, transcriptome comparison is defied by the multitude of available methodologies for transcriptomic analysis, with even experiments with equal protocols generating discrepant results [74]. Inter-species assays have the additional challenge of being restricted by homologous genes, than can either be scarce or yield non-comparable species-specific functions [74].

RNA-Seq is a more recent technique where the isolated mRNA is fragmented, reversely transcribed to cDNA and then sequenced with high-throughput sequencing methods [81,87], such as those developed by Illumina IG, Applied Biosystems SOLiD and Roche 454 Life Sciences. The sequenced fragments are referred to as reads, being noteworthy that to one gene may correspond several reads. This process further allows the reconstruction of the original transcriptome through either computational *de novo* assembly or alignment of the reads against a reference genome (or transcriptome) (Figure 6) [81,87]. GE values are derived from the number of reads aligned to each gene [81], thus being a more accurate strategy than microarrays, as quantification is not dependent on hybridization success [87]. RNA-Seq can be performed on whole tissue or cell-specific (single-cell RNA-Seq) transcriptomes [81]. Despite RNA-Seq's robustness and accuracy, microarrays remain widely in use for monetary reasons, with RNA-Seq technology being often a more expensive process, as well as due to the higher availability of publicly available microarray databases [88]. Moreover, in the context of solely querying for GE, microarrays are much reliable, given the accuracy of their gene annotation already available for several species [88]. This was the main reason for choosing microarray data for the present project, given the scarcity of concordant mouse and human RNA-Seq datasets related with AD, at the time of the study.

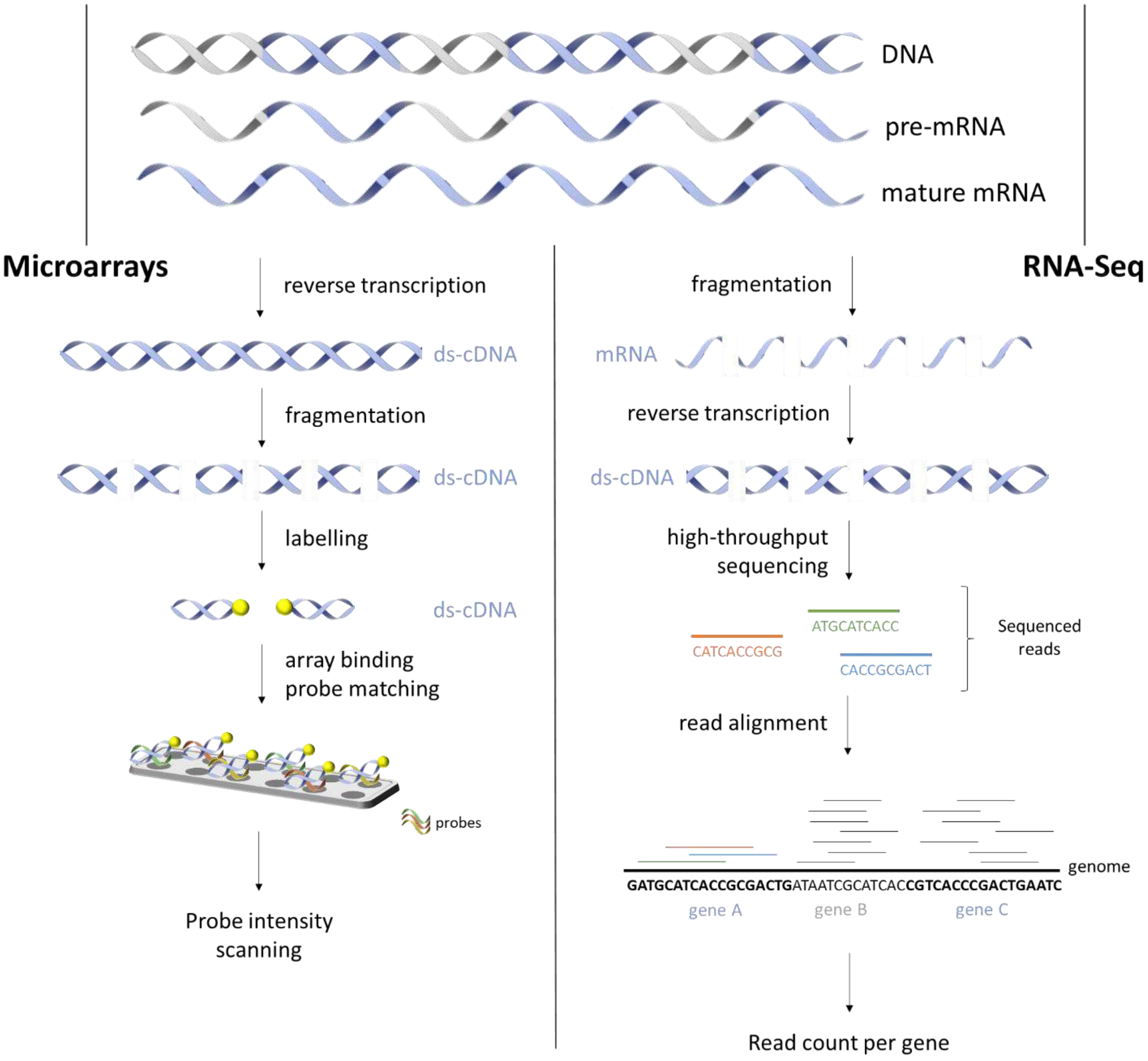


Figure 6 | mRNA quantification techniques: microarray and RNA sequencing

In both techniques, the first steps are mRNA isolation and its processing to fragments of double-stranded cDNA (ds-cDNA). Microarrays (left) are based on fluorescently labelled cDNA that hybridises with complementary pre-defined probes, while RNA-Seq (right) uses high-throughput sequencing of ds-cDNA fragments followed by genome alignment of the sequences reads. Figure inspired by Lowe *et al* [81].

4.3. AD transcriptome comparison studies between human and mouse models

At the time of writing, we found eight recent research papers, available in either PubMed (peer-reviewed) or BioRxiv (pre-prints), in which the authors compared human and mouse transcriptomic data relating to AD pathology [78,89–95].

The methodological approach chosen by these authors often relied on the overlap of differentially expressed genes (DEG) between species, or of enriched gene sets derived from databases [78,89–92,94,95] such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [96–98] or the Gene Ontology (GO) Resource [99,100]. It is noteworthy that by comparing *a priori* defined gene sets, one is limited by those known functional associations [101], which may lack gene specificity and disregard important details. Additionally, a couple of papers focused on specific cell types rather than whole-tissue transcriptomes [89,93]. However, at the beginning of the present study, to our knowledge, public single cell databases of concordant mouse and human data on AD were not available.

Moreover, none of the aforementioned papers explicitly considered using mouse models that matched human patients regarding the type of AD (such as if it is familial AD, or early- or late-onset) or the type of mutation, with even one of them [78] clearly admitting to compare transcriptomes of human idiopathic AD with those of several transgenic mouse and rat models [78]. As mentioned above, transgenic animals often carry human AD-related mutations, hence being more closely related to the human familial form of Alzheimer's [71,75]. Also, due to the low longevity of the animals, they more easily match patients with early-on-set AD than LOAD patients [77].

These studies were not concordant among themselves on whether mouse models correctly replicate the human disease, with some supporting that premise [91,92,94] and others being more judicious about it [78,89,90,93,95].

Regarding AD-associated biological pathways, there is a common trend along some of those studies that unveils immune responses, namely the inflammatory response, as an AD-induced enhanced mechanism, both in human patients and mouse models [78,90–93]. Oxidative stress [92] and the protein kinase cascade signalling [78] were also mentioned as up-regulated mechanisms, respectively in mouse models and human patients. Processes such as the tricarboxylic acid cycle [78] and mitochondrial activity [94] were described as affected in AD, as well as metabolic processes, protein transport and metabolism, transmembrane transport and vesicle trafficking [90,92]. Mechanisms linked to neuronal activity (such as neurodevelopment, long term potentiation and synaptic activity [91,92]) and cell cycle (such as cell differentiation, proliferation and death, and regulation of cell cycle [92,93]) were also found disrupted in AD.

5. Motivation and project goals

The world population is rapidly aging, and a swift of the social stratum is foreseen, with an increase in the population above 60 years old, as well as in the ratio between the elder and younger individuals [7]. The increase of the median population age is accompanied by age-related diseases such as dementia, of which AD is the most prevalent one. An increase of AD pervasiveness entails, besides health problems, social and economic burdens. The monetary cost of dementia's upsurge is estimated to be 32 000€ per patient in Europe, covering direct healthcare, non-medical care and indirect costs [102].

The scarcity of available treatments for AD, allied with the lack of new, more efficacious, approved medicines for the last 16 years, emphasizes the need for improved research on neurodegenerative diseases. Moreover, none of the existing therapeutics curtail nor halt disease progression, making it imperative to find new biochemical targets and drugs that can advance disease treatment. The difficulty in finding new medicines resides in the low success rates of clinical trials, with several new compounds passing preclinical tests but failing when tested in humans, especially during toxicity and efficacy assessments. This implies a poor representation of the actual human disease in preclinical animal models, thus being vital to evaluate what molecularly distinguishes them in terms of disease pathophysiology, and how can they be improved to better represent the human disease.

On this note, we purpose to assess the dissimilarities between AD preclinical mouse models and human patients at the transcriptome level. Contrasting to other studies with the same purpose, we carefully matched human and mouse samples regarding AD form and mutated gene. Therefore, the main goal of the present project is to address to what extent a *PSEN1*-mutated mouse model is able to recapitulate human FAD with mutations in the same gene, by analysing and comparing whole-transcriptome brain microarray data. This approach allows to find genes that are differentially expressed in AD samples compared to control non-diseased samples, and which of these genes vary, in the same or contrary direction, between species.

We also propose to pinpoint potential genetic alterations or compounds that can recapitulate the human transcriptomic signature of AD and that could be applied in a novel and more effective mouse model that better recapitulates the human disease.

Finally, the findings from this study are thought to provide insight not only into the biomolecular variances between AD *PSEN1* human patients and mouse models carrying human

PSEN1 mutations, but also on valuable and innovative targets that can henceforward be tested for a potential improved AD mouse model, as well as unveil new mechanisms that underlie disease progression.

CHAPTER II – MATERIALS AND METHODS

6. Datasets

For the present study, GE microarray data from *post-mortem* brain samples were used. For human samples, we resorted to the dataset published by Anna Antonell *et al*. [103], while mouse data were extracted from the dataset derived by Mar Matarin *et al*. [104]. Both datasets are publicly available in the National Center for Biotechnology (NCBI) Gene Expression Omnibus (GEO) repository [105]. Table 1 summarizes relevant information regarding the datasets and the samples considered for the current project.

6.1. Human dataset

Antonell and colleagues [103] derived a microarray GE dataset of *post-mortem* human brain samples from the **posterior cingulate area in the cortex**, with the intent of assessing transcriptomic differences between patients with sporadic EOAD, patients with FAD caused by mutations in the *PSEN1* gene and individuals without signs of neurodegenerative disease (considered as "healthy" individuals). This dataset and its annotation are publicly available under the NCBI GEO accession number **GSE39420** [103].

The authors considered 14 EOAD subjects in their experiment, out of which 7 held mutations in *PSEN1*, comprising the FAD-PSEN1 samples (4 with the M139T mutation, 2 with the V89L mutation and 1 with the E120G mutation). No known mutations were detected in *APP*, *PSEN1* or *PSEN2* for the remaining 7 samples, corresponding to the sporadic EOAD batch. Additionally, the authors included 7 samples from non-AD subjects, which showed no hallmarks of the disease.

For microarray analysis and differential GE assessment, the authors followed the Affymetrix microarray standard protocol for the **GeneChip Human Gene 1.1 ST Array Plate**. Information provided by the authors regarding the samples' neuropathological changes (Braak stage), *APOE* genotype, *PSEN1* mutation, age, gender and *post-mortem* delay/interval (PMD/PMI) [103] was contemplated in the present analysis.

6.2. Mouse dataset

The public mouse data used in the present study were generated by Matarin and colleagues [104] and obtained through the NCBI GEO accession number **GSE64398**. These data were derived from five different transgenic mice, carrying knock-in human mutations, four of which develop β-amyloid pathology at different rates and one that developed tau pathology and NFT. The first group includes *APP*-mutant mice carrying the double mutation K670N/M671L (TAS10), *PSEN1*-mutant mice with the M146V mutation (TPM) and a homozygous (HO-TASTPM) and a heterozygous (HET-TASTPM) crossbreeds of those two, all of which are regulated by a Thy1 promoter. The NFT transgenic mouse (TAU) had a knock-in mutation (P30IL) in *MAPT*, regulated by a CaMKII promoter. The authors also provided data for age-matched control mice, raised as littermates of the aforementioned transgenics except for the double-mutant mice. Thus, three subgroups of controls are present (littermates of *APP*-, *PSEN1*- and *MAPT*-mutated mice). Authors isolated brain samples at 2, 4, 8 and 18 months of age, from three different brain regions: hippocampus, cortex and cerebellum. GE was measured with **Illumina MouseRef-8 v2.0 BeadChip microarrays**.

Matarin and colleagues intended to assess discrepancies in GE regarding NFT-developing mice and those that developed β-amyloid pathology, towards non-AD control conditions. Additionally, they identified GE changes across the different rates of amyloidosis progression, i.e. across the different transgenic mice with different degrees of amyloidosis pathology (using the TPM, HET-TASTPM and HO-TASTPM mice, explained in more detail below). They evaluated GE differences across age and brain regions, as well.

In the context of this thesis, only **cortex** GE data from mice carrying mutations in *PSEN1* (i.e. the TPM, HO-TASTPM and HET-TASTPM animals) were used, in order to match the available human data. The three types of mice were divided into separate datasets and will be respectively referred to as single-mutant mice (PSEN), and homozygous (HO) and heterozygous (HET) double-mutant, for simplicity. Despite the latter combinations displaying three mutations, these mice are going to be referred to as “double-mutants” given that two genes carry mutations.

The authors characterized AD pathology development in the mouse models used in present study by their ability to develop amyloid plaques, which was assessed through immunohistochemistry. The single-mutant mice did not show any staining by 18 months, whereas the double-mutant mice developed the pathology at 8 months [104]. Contrarily to HET mice that displayed sporadic staining, reflecting low pathology development, HO mice were the most affected by plaque deposition, showing the highest level of Aβ staining, which could start to be slightly noticed even at 4 months [104].

Table 1 | Dataset information summary

Dataset/ Species	Human			Mouse					
Microarray technology	Affymetrix GeneChip			Illumina BeadArray					
Gene annotation package	GeneChip Human Gene 1.1 ST Array Plate			MouseRef-8 v2.0 BeadChip					
GEO accession number	GSE39420			GSE64398					
First Author	Anna Antonell			Mar Matarin					
Reference	[103]			[104]					
Samples (*used in the present study)	*PSEN1*-mutant EOAD	EOAD	Non-AD controls	*APP*-mutant	**PSEN1*-mutant	**APP*- and *PSEN1*-mutant	**APP*- and *PSEN1*-mutant	*MAPT*-mutant	*Controls
Name	PSEN	EOAD	Control	TAS10	TPM	HET-TASTPM	HO-TASTPM	TAU	Control
Number of individuals	7	7	7	-	14	16	15	-	36
Mutation	M139T V89L E120G	None	None	K670N M671L	M146V	K670N M671L M146V	K670N M671L M146V	P30IL	None
Available information	Condition (PSEN, EOAD, Controls) Age Gender *APOE* genotype PMI Braak stage *PSEN1* mutation			Condition (PSEN, EOAD, Controls) Age					
Information used in modelling	Condition (PSEN, EOAD, Controls) Age PMI			Condition (PSEN, EOAD, Controls) Age					

6.3. Joint dataset

From the junction of the aforementioned human and mouse datasets, a third dataset was compiled, that included all considered samples and only the orthologous genes between human and mouse. The methods to remove outliers and find the orthologous genes are summarized in sections 8.1 and 8.2 to follow.

Combination of datasets was performed on expression matrices (i.e. table-like data structures comprising the GE values) com with microarray probes as rows and samples as columns. The merger was implemented based on the gene symbols annotating the probes in both datasets, using the R programming language, more specifically the *merger* function [106] (software is explained in more detail in the next section).

7. Software

Data mining, manipulation and visualisation, as well as all statistical analyses and linear modelling were performed with the R programming language [106]. Specifically, the open-source and web-based R Studio [107], an integrated development environment for R, was used. In Table S1, the most relevant R packages and functions used in the present analysis are summarized.

8. Methods

8.1. Data pre-processing

8.1.1. Import data files

Probe intensity values were provided in CEL files for the human GeneChip microarrays, and in IDAT files for the mouse BeadArray dataset. These files were imported into the R environment using function *read.celfiles* from the *oligo* R package [108] for the human CEL dataset, and function *read.idat* from the *limma* R package [109] for the mouse IDAT one.

Sample metadata for both human and mouse were available as TXT files and were imported into R using the *read_delim* function from the *readr* R package [110].

8.1.2. Data transformation and normalization

The data pre-processing pipeline is identical for both used types of microarray datasets, comprising background correction, logarithmic transformation and quantile normalization [111,112]. Background correction is an important step to adjust the retrieved intensity in order to account for the background signal from non-specific binding, modelled by negative control probes specifically designed for non-specific binding [111,112]. For both the human and the mouse data, background correction was performed by the normal-exponential (normex) convolution model, which assumes observed intensities as the sum of an exponentially distributed foreground (specific binding) signal and normally distributed background (unspecific binding, i.e. noise) values [111].

Logarithmic transformation is performed to make the typically log-normal-like distributions of probe intensities and GE values more amenable to statistical analyses that assume distributions to be normal. With a logarithmization of base 2, one unit of logFC translates an increment or decrement of 100%, i.e. twice the GE value, between compared conditions (in the case of categorical variables), or by unit of the condition in study (in the case of continuous variables, such as age) [113]. Lastly, quantile normalization is performed in order to reduce the technical inter-sample data variance, by approximating the distributions of probe intensities across all arrays/samples, making them thereby comparable [112].

Function *rma* from the *oligo* R package that implements Robust Multi-Array Analysis (RMA), was used to pre-process the human Affymetrix GeneChip data [111,112], while function *neqc* from the *limma* R package was used on the mouse Illumina BeadArray data. Both functions implemented the normex convolution model for background correction, followed by quantile normalization and log2 transformation [111,114,115].

8.1.3. Probe set summarization and expression values

Replicates of the same probe appear in multiple positions within the same array, in order to encompass for technical variability and positional biases that may occur. Additionally, in GeneChip microarrays probes, different sequences target the same transcript. Each group of such probes is called a probeset.

Matrices with probe expression values for each sample were computed, with probes as rows and samples as columns. For the human data, function *exprs* from the *Biobase* R package [116] was used to compute the probe expression matrix from the output object of the *rma* function, whereas for the mouse dataset an expression matrix was obtained through the *EList* object resultant from the *neqc* function.

After transforming and normalizing the data, it is therefore important to summarize all values corresponding to the same probeset into a single value. The applied *rma* and *neqc* functions inherently perform probe summarization.

8.1.4. Quality control and outlier removal

An initial quality assessment of the human GeneChip data was performed by generating probe intensity grey-scaled images from the CEL files. These images show how probe intensities spatially distribute across each array and therefore allow the visual detection of potential spatial artefacts, such as surface scratches, particles and other contaminants [112], which are exemplified in Figure 7. Depending on the size and intensity of the artefact, an affected array may be considered as a technical outlier and discarded from further analysis. Given that Matarin *et al.* did not published individual intensity values (only intensities summarized by probe, i.e. probe replicates were absent), raw chip images could not be computed for the mouse data.

Function *image* from the *oligo* R package was provided with the output object of the *read.celfiles* function and used to display probe intensity images for the human data. The considered samples did not display any visible artefact besides the expected chip tag, located in the middle of the image (Figure S1).

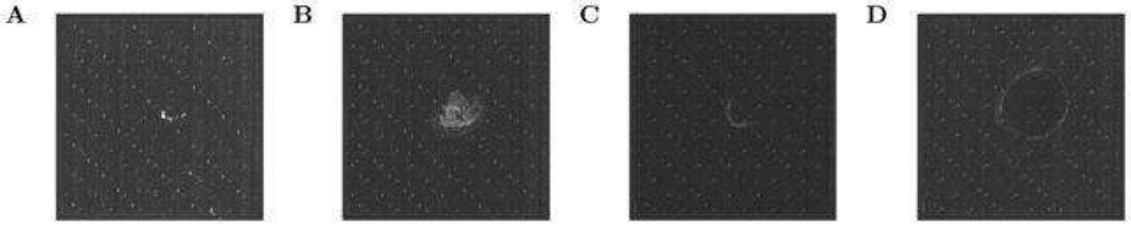


Figure 7 | Affymetrix microarray artefacts

Exemplifying images of possible technical artefacts detected by Petri *et al*. [117] in Affymetrix GeneChip microarrays.

For the same data, Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) plots were computed after probe-level model (PLM) fitting (Figure S2). A RLE plot is a representation of the relative expression distribution of all probesets across each sample, which, for each probeset, corresponds to the ratio between its expression values in the array, estimated through PLM, and its median expression value across all arrays [112]. The majority of probesets are assumed not to change much between samples, hence distributions are expected to fluctuate around zero [118].

NUSE is the individual probe error fitting the PLM. NUSE boxplots inform on the distributions of normalized standard error of probesets for each sample. Given that NUSE values are standardized at the probeset level across the arrays, the median of their distributions should be centered in 1.0. Samples with median NUSE values above 1.05 need further investigation since those deviations can be seen as a 5% average loss in precision [118]. RLE and NUSE were computed using functions *NUSE* and *RLE* from the *oligo* R package, using the object from PLM fitting as input. Similarly to what was observed for the raw images, no human sample was flagged as a potential outlier neither in RLE nor NUSE (Figure S2).

For both human and mouse microarrays, GE distribution boxplots were outlined for each sample, before and after data normalization. For that purpose, function *boxplot* from the *oligo* R package was used. After normalization, the distribution of expression values per sample should be similar across samples, with substantially deviating cases being flagged as outliers. No outliers were detected for neither the human nor the mouse data (Figure S3 and S6).

8.2. Gene Annotation

8.2.1. Probe ID to gene symbol conversion

The tables used for conversion between probe IDs (e.g. ILMN_2127842 for Illumina-derived data) and gene symbols (*HBA2* for human and *Hba2* for mouse) were directly imported from Bioconductor [116] through the *select* function from the *AnnotationDbi* R package [119]. Annotation packages *hugene11sttranscriptcluster.db* [120] and *illuminaMousev2.db* [121] for the human and mouse conversions were respectively used.

With microarrays, one gene is often profiled by different probes. Hence, when analysing differential GE, probes were ranked based on the corresponding t-statistic absolute value (derived after linear modelling, as explained in section 8.7), and, for each gene, the probe with the highest rank was kept.

8.2.2. Mouse to human orthologue gene symbol conversion

The table with the mouse-human orthology relations, that enabled the conversion from mouse to human gene symbols, was extracted from BioMart [122,123], in which the GRCm38.p6 annotation of mouse genes and respective human orthologues were selected within the Ensembl Genes 97 database [124]. Only genes with one-to-one orthology were considered orthologs.

8.3. Statistical hypothesis testing

When deducing information from data, it is important to have a clearly formulated question in the form of a testable hypothesis. Statistical inference relies on the formulation of the alternative or research hypothesis (H_1), that reflects our premises, and the null hypothesis (H_0), which includes all possibilities except the one we hypothesize [112]. These two hypotheses need therefore to be mutually exclusive and encompass all potential outcomes in their union.

8.3.1. Significance and p-value

Statistical tests evaluate if there is enough evidence to reject the null hypothesis based on the probability value (p-value) of erroneously rejecting a true null hypothesis in favour of the alternative hypothesis, i.e. of obtaining a false-positive call [112]. It is noteworthy that this value is computed by testing the assumption that the null hypothesis is true, hence it does not convey any probability of the alternative hypothesis being true [125].

The lower the p-value the lower the probability of inferring false positives [112]. Prior to testing, the researcher defines a significance level, α, based on the data prior to testing, which translates the amount of uncertainty one is willing to take when accepting an outcome [112,125]. The null hypothesis is rejected when the p-value, calculated based on the chosen statistic applied to the data, is lower than α [125].

8.3.2. Multiple testing correction

Considering GE data, when assessing if one gene is differentially expressed between two conditions, the commonly used significance level of 0.05 translates into a probability of correctly deciding not to reject the null hypothesis of $1 - \alpha$, that is, 95% [112]. When testing for N genes, the probability of being always correct becomes $(1 - \alpha)^N$, while the probability of finding at least one false positive within the N results is $1 - (1 - \alpha)^N$ [112]. Considering the aforementioned example, if 100 genes are tested, that would translate into a 99.4% chance of having at least one false positive.

There are several correction methods conceived for multiple testing. In the current work, the Benjamin and Hochberg approach [126] was used, where the false discovery rate (FDR), i.e. the proportion of false-positives towards the totality of positive results, is controlled. This method allows for some dependency between the variables in study which is important when dealing with expression data of genes that are often complexly related through biological processes [112].

For multiple testing correction, tests' p-values are ranked in ascending order. The last-in-rank FDR-adjusted p-value (q-value) equals its p-value, whereas the remaining are assessed by going backwards on the rank. The penultimate ranked p-value would be adjusted by choosing the lowest value of two options: (1) the last q-value or (2) a q-value obtained through equation 1 [126,127]. This procedure is completed by always considering the q-value immediately after; for instance, the 5th p-value would be adjusted by choosing between the q-value obtained for the p-value ranked 6th or through equation 1, whichever is the lowest.

$$qvalue = pvalue \times \frac{number\ of\ tests\ performed}{rank} \tag{1}$$

8.4. Statistical tests

8.4.1. Gaussian distribution assessment

Several statistical tests assume that data either follow a Gaussian/normal distribution or are sampled from a normally distributed population – these are called parametric tests. Contrarily, non-parametric methods do not make assumptions on data distributions nor analyse their actual values, but instead work with their ranks [125].

The theoretical Gaussian distribution has a bell shape and spreads symmetrically around its mean and infinitely to both the positive and negative directions (Figure 8), which usually does not apply to real data. Nonetheless, the majority of parametric tests is robust enough to perform well with distributions approximate to the Gaussian [125].

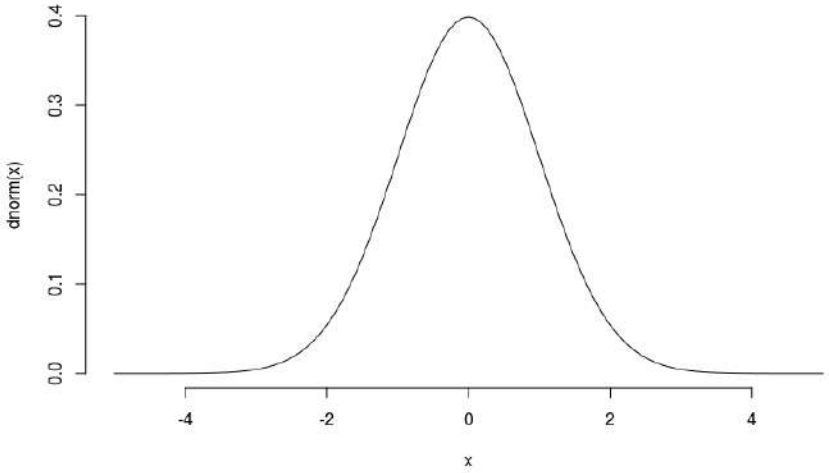


Figure 8 | Normal or Gaussian distribution

Representation of a theoretical normal distribution, with a mean of zero and a standard deviation of one. Data computed using the function *dnorm* from built-in R package *stats* [106], with the probability function on the y axis.

The Shapiro-Wilk test was used to assess data normality in our study, since sample sizes were lower than 50 (otherwise the Kolmogorov-Smirnov test would be used) [128]. In that case, the null hypothesis states that the sample in study follows a normal distribution, which has no evidence to be rejected if p-value is higher than the set significance level [128]. Function *shapiro.test* from built-in R package *stats* [106] was used to perform the Shapiro-Wilk test.

For the distributions for which normality was tested, Q-Q plots and probability distributions were also outlined (Figure S10-11). The first compares the quantiles of a theoretical normal distribution against the quantiles of the empirical data distribution, and the latter helps to visualize if data follows a normal distribution [128]. In the Q-Q plots, a linear relation is expected if normality is followed [128]. For this purpose, function *ggqqplot* from the *ggpubr* R package [129] was employed.

In the present study, for the two distributions (Age and PMI) evaluated for normality, a p-value associated with the Shapiro-Wilk test (displayed in the Q-Q plots in Figure S10-11) higher than a significance level of 0.05 was obtained. The assumption of normality was not rejected for the Age distribution (Figure S10), but it was for the PMI distribution (Figure S11). This decision was based on the non-Gaussian appearance of PMI probability distribution and the p-value which, despite being higher than 0.05, is small.

8.4.2. Variance assessment

Wilcoxon and t tests are, respectively, non-parametric and parametric hypothesis testing methods [125]. To assess differences between groups of samples, t-tests were conducted for variables that follow normal distributions, and Wilcoxon tests for those that do not [125]. Both tests were computed with function *stat_compare_means* from the *ggpubr* R package, used along *ggplot2* [130], differing in the *method* argument that discriminates between *t.test* and *wilcox.test*. The significance level was set at 0.05.

8.4.3. Correlation analysis

Correlation analysis infers the statistical association between two continuous variables, translating it into a value, the correlation coefficient [125]. The coefficient ranges from -1, cases in which variables are 100% inversely correlated, and +1, where they are 100% positively correlated, i.e. they simultaneously increase in the same direction. This analysis can either be performed in a non-parametrical or in a parametrical way by respectively employing Spearman’s or Pearson’s correlations [125].

For this project, Spearman’s rank correlation was chosen based on its ability to detect both linear and non-linear monotonic correlations, contrarily to Pearson’s correlation, that only tests for linear associations between variables [125]. Spearman’s method computes and compares the individual ranking of values for each variable, under the null hypothesis of absence of correlation (correlation coefficient = 0), i.e. that the ranked values of the variables in study are not covariant [125,131]. For the present study, the function *cor.test* from built-in R package *stats* was used to compute the correlation coefficient (*R* for Pearson’s or *Rho* for Spearman’s) and its respective p-value (p). To complement the analysis, a visual representation of the relationship between variables was also often plotted with function *smoothScatter* from the *graphics* built-in R package [106] or *ggscatter* from the *ggpubr* R package.

8.5. Principal Component Analysis

Due to the high dimensionality of data, as seen in microarray datasets where several thousands of genes are measured, techniques have been developed to assist researchers dealing with the thousands of variables under study and the high covariance between them [132,133]. Principal Component Analysis (PCA) is one of the most commonly used methods [132].

PCA acts as a descriptive statistical technique that reduces the dimensions of a dataset by summarizing its variability into new uncorrelated variables – known as principal components (PCs) –, while preserving the original information within the data [132,133]. In this sense, each PC is a linear combination of the original variables weighted by their individual contribution for the variance explained by that component [132,133].

PCA calculates new coordinates in the direction of most variance within the data (Figure 9), and then finds the projections of the data points in the new coordinate system [112]. At first, variables are centred by their average, sometimes scaled, and then a correlation matrix is derived [132,133], which is decomposed in eigenvectors and their respective eigenvalues [112]. Centring the variables is important in order to consider the variance within the data, instead of how much their averages stand apart from zero. Scaling is relevant when variables have different units or spread in very discrepant ranges [132] but, if applied to GE, it would equalize the contribution of each gene to separate the samples, across all genes, without distinguishing genes with highly and lowly variable expression. Therefore, scaling was not applied in the present analysis, such that genes with more variable expression contribute more to the separation of samples.

The obtained eigenvectors define the principal components, while the respective eigenvalues are scalars proportional to the variation explained by their components [112]. Principal components are ranked according to the percentage of variance that they explain, with the first (PC1) being the one with the highest variance [112,132]. Even though the number of components equals the number of variables, around 70% to 90% of the variance is typically explained by the first 10 or less components, with those lower in rank usually being dispensable [112,132,133]. That reduced number of variables is easier to explore and manipulate than the original amount [112,132,133], which for this work would translate into thousands of genes' expression values.

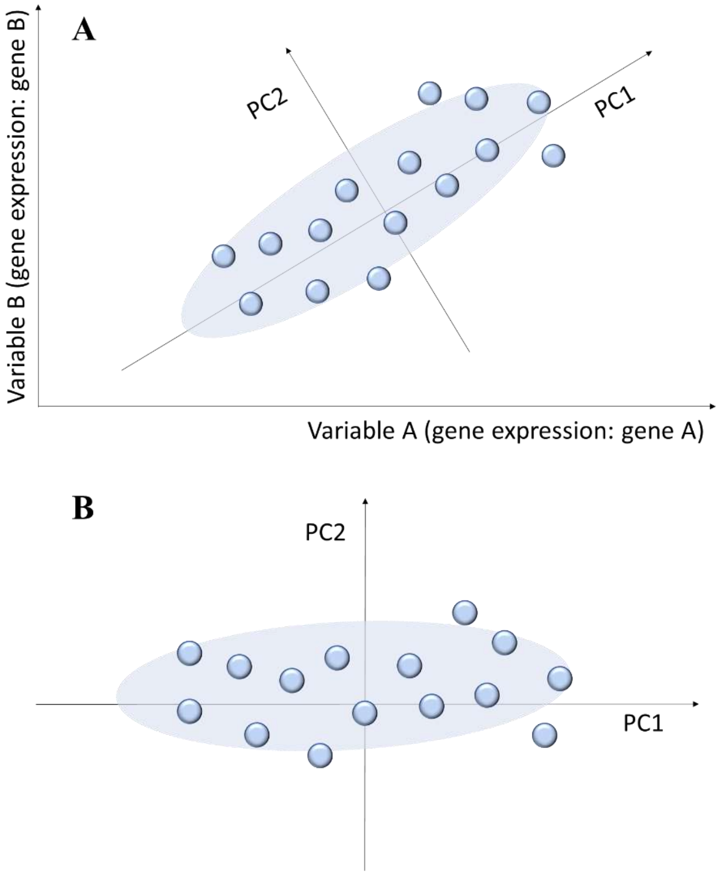

A
Variable B (gene expression: gene B)
Variable A (gene expression: gene A)
PC2
PC1
B
PC2
PC1

Figure 9 | Principal Component Analysis

(A) When applying PCA to GE, the genes' expressions correspond to the original variables, and the data points to the samples in study. For simplicity, the figure accounts for only two genes, and thus two dimensions. The first component follows the direction of the highest variance within the data.

(B) Graphically, PCA can be considered as a rotation of the data. Within this process, data are centered on the average and the origin is shifted towards the centroid of the data. The positions of the data points relative to each other are never affected during this analysis. Figure inspired by [109]

Comparing PCs in a plot allows to assess the heterogeneity of the data [133] and explore patterns, visualize relevant clusters [112], and roughly assess which features and variables could explain more the variability between samples. If the components with the highest variance are assumed to be the most important, PCA on GE data can also inform on relevant biological pathways [133]. This analysis becomes an important tool to highlight putatively important features to include in downstream differential expression analyses, being also practical to visually detect samples that might be considered outliers [132].

Computationally, the original matrix had samples as rows and genes as columns, that are centred using function *stdize* from the *pls* R package [134], with the scale argument turned FALSE. To compute the principal components, function *PCA* from the *FactoMineR* R package [135] was used, with the scaling argument also turned FALSE. A matrix with samples as rows and principal components as columns was retrieved, and components were plotted against each other using the *ggplot2* R package. Eigenvalues were also obtained using function *get_eigenvalue* from the *factoextra* R package [136], and the percentage of variance explained by each component was plotted resorting to the *fviz_eig* function, also from *factoextra*.

8.6. Clustering analysis

Clustering analysis consists on grouping similar data points together, such that they are closer to each other, by a given distance metric, than to data points of other clusters [112]. The similarity of data points can be assessed through several different distance functions, such as the Manhattan distance, correlation distance, the angle between vectors, among others [112]. For the present analysis, the Euclidean distance was used.

Heatmaps are graphical representations of data where the values are color-coded, being often used alongside cluster analysis. As part of such plots, hierarchical clustering can be represented through tree-like diagrams named dendograms, where samples are clustered together based on similarity until reaching one single cluster [112]. Functions *HeatmapAnnotation* and *Heatmap* from the *ComplexHeatmap* R package [137] were used to compute heatmaps and respective dendograms.

8.7. Multiple linear modelling

Linear modelling assumes a linear relationship between a dependent or response variable *y*, and the factors by which it is influenced, that are assumed to be independent predictor variables, also named explanatory variables (x_k) [112,125]. The current project deals with multiple linearity, as the expression of a given gene (dependent variable), is modulated by several different factors for which information is available, i.e. samples' metadata, which are treated as explanatory variables that are assumed to impact GE. Given k explanatory variables, a multiple linear regression model follows equation (2), where $y = GE$ of a given gene.

$$GE = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{2}$$

The β parameters are regression coefficients associated with each of the explanatory variables, except for β_0 which is an additive constant that graphically represents the intercept with the y axis [112,125] – referred to as *Baseline*. ε is the error term that works as an adjustment in regards of all factors that contribute to modulate GE (the dependent variable) that are not included in the explanatory variables [112,125].

Linear modelling finds the function that best fits the data based on the explanatory variables according to the least squares' method. Hence, the fittest model will be the one that minimizes the sum of squared distances (SSD) [112,125] between the data points and the regression line – these distances are called residuals.

Samples' description regarding these predictor variables was supplied to the modelling algorithm in the form of a design matrix. For instance, if a model had Age and Disease as predictor variables, the design matrix would be similar to Table 2.

Table 2 | Example of a design matrix

A design matrix has the samples as rows and explanatory variables as columns, where the Baseline is a vector of ones, since all samples contribute to it. In this example, the Age column corresponds to the donors' age in years and the Disease column is a binary categorical variable with 1 for the samples diagnosed with AD and 0 for the non-AD individuals.

Samples	Baseline	Age	Disease
AD1	1	20	1
AD2	1	17	1
AD3	1	16	1
Control1	1	35	0
Control2	1	27	0
Control3	1	32	0

Explanatory variables will also be referred to as predictors or coefficients, and their names will be in *italic*. When a variable is included in a linear model as the only explanatory variable, it may be denoted as a "single" variable. For instance, for a model translated by the equation 3A, the *Age* variable would be single-*Age*. Additionally, models differing exclusively in the presence of one explanatory variable will be referred to as *variable*-differing models, with *variable* replaced by its name (for instance, equations 3A and equations 3B represent *Age*-differing models).

$$GE = \beta_0 + \beta_{PSEN} PSEN \qquad (3A)$$

$$GE = \beta_0 + \beta_{PSEN} PSEN + \beta_{Age} Age \qquad (3B)$$

We used function *lmFit* from the *limma* R package to fit each model. For every model, the fold changes in expression associated with each contrast/variable are estimated for each gene.

8.7.1. Interaction coefficient and centring variables

In linear models, an *Interaction* coefficient can also be added to the model, if it is thought that two explanatory variables have a synergistic or offsetting relation, as such that impacts the dependent variable (y) differently than their additive effect alone. The *Interaction* consists in a multiplicative term between the two interacting explanatory variables [138]. An *Interaction* coefficient might also adjust the interpretation of the model. For instance, in the joint dataset, without *Interaction*, the *PSEN* coefficient would inform on the average AD-induced GE changes across samples, irrespectively of their species of origin. Adding the *Interaction* coefficient allows to distinguish between AD-induced GE changes specific of the human data and the ones specific of the mouse data (a more refined interpretation is discussed in results section 13.2). The design matrix of a model that integrates the *Interaction* effect is represented in Table 3.

Table 3 | Example of a design matrix with an *Interaction* coefficient

The *Interaction* effect is modelled by the multiplication of the two explanatory variables involved.

Samples	Baseline	Age	Disease	Interaction (Age * Disease)
AD1	1	20	1	20
AD2	1	17	1	17
AD3	1	16	1	16
Control1	1	35	0	0
Control2	1	27	0	0
Control3	1	32	0	0

When adding an *Interaction* coefficient to a model, centring of variables is also an important step, consisting in subtracting the mean of a variable across samples to each sample's individual value. This method is useful to diminish the correlation between the *Interaction* coefficient and its component variables, making the model more consistent with the purpose of estimating independent effects [138]. It can also ease interpretation of coefficient estimates. Centring shifts the "prediction centre" (i.e. the reference sample) to the centre of the available data (Figure 10), by turning the variable's mean to 0 [138]. In models where an *Interaction* coefficient is absent, the independence assumptions are already met and centring does not affect the coefficients' estimations, thus being irrelevant to apply [138]. Regarding the example in Figure 10, centring of the *Species* variable turns the linear model baseline to a conceptual human/mouse hybrid. It is noteworthy that the mathematical comparison between the GE in human and mouse samples remains unchanged, given that the distance between them in the model remains unchanged.

Depending on the effects that we hypothesise to interact and the meeting of independence assumptions, we can decide to centre all variables or some of our choice. For the present analysis, exclusive centring of the *Species* variable was tested – *Species*-centred model –, as well as centring of both *Species* and *PSEN* variables – fully-centred model. In these cases, the *Interaction* coefficient is obtained with the centred variables. Further comments on the choice of the centring approach are discussed in results section 13.2.

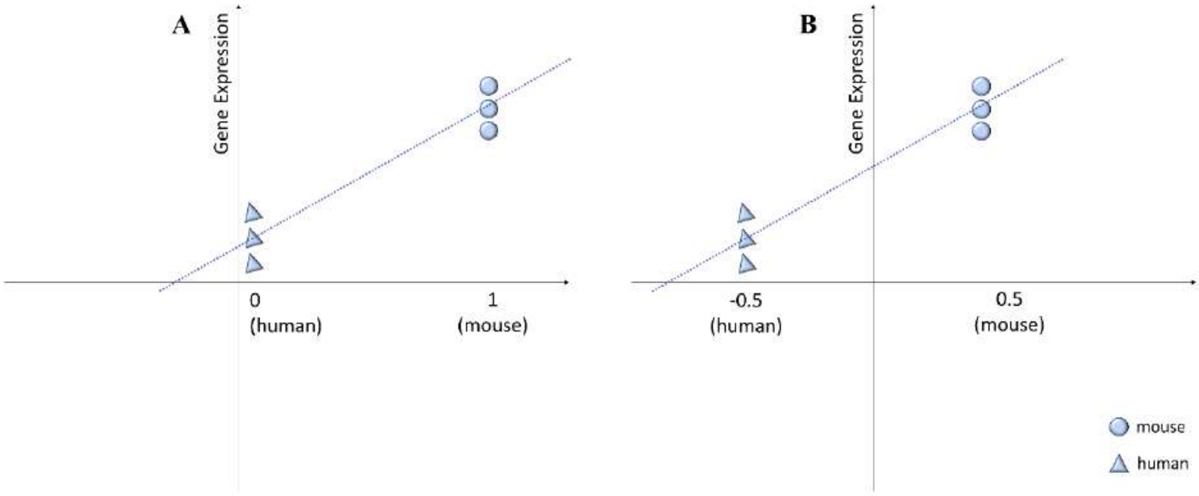


Figure 10 | Centring of the *Species* variable

In the case of modelling the joint dataset (which includes an *Interaction* coefficient), centring was attempted in order to obtain AD-induced GE changes common to the human patients and mouse models (in the *PSEN* coefficient), as well as those species-specific (in the *Interaction* coefficient). This shift in the data changes the baseline (i.e. the reference sample) from human controls to a conceptual human/mouse hybrid control (Figure 10). To centre the variables, the *scale* function from built-in R package *base* [106], with the scale argument turned FALSE, was employed.

For this project, several linear models of GE (Table 4) were tested in the human and mouse datasets. For the human dataset, different combinations of up to four variables were tested as explanatory variables, while for the mouse two were selected. Two predictors were also used for the joint dataset.

Table 4 | Linear Models

Summary of linear models for GE used in each dataset, and their respective explanatory variables.

Dataset	Model	Explanatory variables	Centring process
Human (with EOAD)	**1**	Disease	None
	2	PSEN + EOAD	
	3	Age	
	4	PMI	
	5	Disease + Age	
	6	Disease + PMI	
	7	Disease + Age + PMI	
	8	PSEN + EOAD + Age	
	9	PSEN + EOAD + PMI	
	10	PSEN + EOAD + Age + PMI	
Human (without EOAD)	**1**	PSEN	None
	2	Age	
	3	PSEN + Age	
Mouse	**1**	PSEN + Age	None
Joint	**1**	PSEN + Species + $\text{Int}_{\text{PSEN+Species}}$	None *Species* centred *PSEN* and *Species* centred

8.7.2. Differential gene expression analysis with a Bayesian approach

We used function *eBayes* from the *limma* R package to adjust the outcome from the linear models through an empirical Bayesian moderation, and function *topTable* from the same package to summarize the statistics of differential expression of each gene for the contrasts/effects in each given linear model. Those statistics are summarized in Table 5.

The Bayesian approach builds upon conditional probabilities where the estimated likelihood of an event occurring is based on prior knowledge, such as a previous event [139]. The Bayesian approach is inspired by the way human individuals gain knowledge and is useful when a researcher does not hold information on all variables that can affect a certain event, most commonly because of either data scarcity or a gap in the scientific knowledge [140].

Table 5 | Statistics of differential GE

Statistics obtained after fitting a linear model to the GE data, adjusting its errors through Bayesian moderation, and assessing differential GE, using the *limma* package in R [109]. These statistics can be retrieved for each coefficient in each linear model.

Statistic	Abbreviation	Meaning
Log fold-change	logFC	Average log2 of GE ratios between experimental conditions
Average expression	AveExpr	Average expression of each gene across samples
Moderated t-statistic	t	T-statistic of differential expression, with standard errors moderated through an empirical Bayesian model
Probability value	P.Value	P-value corresponding to t-statistic
Adjusted probability value	adj.P.Val	Adjusted p-value for multiple testing as described in section 8.3.2
B-statistic	B or log odds	Logarithmized empirical Bayes odds ratio of a gene being differentially expressed between conditions, assuming a default prior that 1% of genes are DE

The logFC represents the average magnitude of differential expression between experimental conditions. However, the logFC statistic alone provides no information on how precise is that estimate, and expression values can have a high variance within conditions. Moreover, with few replicates per condition, logFC estimates can be easily biased by outlier expression values [141].

The t-statistic is commonly used to deal with those issues by aiming at a compromise between magnitude and precision, being the ratio between the average logFC of a gene within a condition, and its associated standard error [141]. This statistic is therefore proportional to the magnitude of expression changes but penalizes genes whose expression variance within conditions is high when compared to that between conditions. However, given the high number (thousands) of genes tested and the small sample size (i.e. the low number of samples) associated with each test, there will be a few genes with an extremely low within-condition standard-error, just by chance, which are assigned with high t-statistic of differential expression, even if they have a small logFC [141]. To mitigate this source of "false positives", a common option is to "moderate" standard errors, i.e. to adjust randomly extreme values.

A Bayesian-moderated t-statistic encompasses variance information from all genes in the standard error. In this case, the gene-wide expression variance distribution constitutes the prior knowledge that is used to infer adjusted individual values of variance for each gene's expression [142,143]. Moderated t-statistics still encompass information on the magnitude of differential expression, being positively correlated with logFC (Figure 11A).

A new statistic, named B-statistic, can be computed from the moderated t-statistic as the logarithmized empirical (because priors are "empirically" derived from the data) Bayes odds ratio of a gene being differentially expressed. While t-statistics differentiate between up- and down-regulated genes, B-statistics inform on DEG but do not inform on the direction of that alteration. Absolute values of the moderated t-statistics are indeed perfectly positively correlated with B-statistic values (Figure 11B), as one is a surrogate of the other [144].

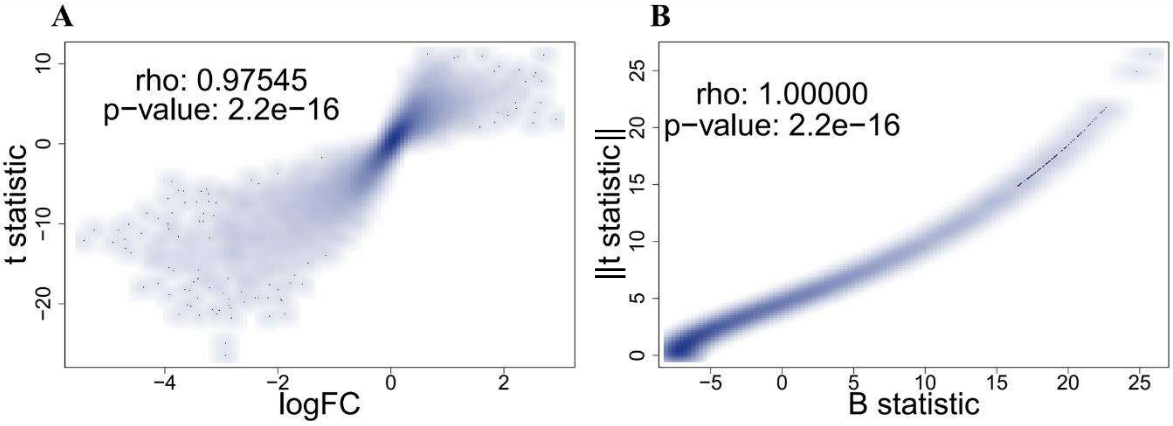


Figure 11 | Correlation between t-statistic and logFC and B statistic values

Example of t-statistics of differential expression that are highly correlated with (A) logFC values, as well as (B) in their absolute values with B-statistics. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

8.7.3. Visualizing differential gene expression

Plotting a smoothed scatter of the average expression of the genes against their logFC value (Figure 12A), for each contrast of the model, provides a preliminary visual quality control of the differential expression analysis. As most genes are not DE, a darker density cloud should appear around the zero logFC axis, through the entire range of average expression.

To visually identify DEG between the different conditions, B-statistic values were plotted against the logFC, for each contrast in each applied linear model. These V-shaped plots are called volcano plots (Figure S12) and allow to simultaneous visualize the significance (B-statistic in the y-axis) and the magnitude (logFC in the x-axis) of GE changes. Given that a B-statistic of 0 means that a given gene has 50% probability of being differentially expressed [109], only genes with positive B values were considered statistically significant. For categorical variables, a minimum logFC threshold of 1 was considered to define genes as differentially expressed, in order to obtain genes that are at least two times more up- or down-regulated in the condition of interest compared with the control condition. A threshold of 2 was also often used in order to refine the quantity of DEG and obtain the more extreme/interesting genes. In any case, the threshold was set after looking at the data.

Usually, the majority of the genes is not differentially expressed, thus volcano plots should present higher density of data points around a logFC of zero. To visualize shifts in data density that could compromise the interpretation of the plots, density volcano plots (Figure 12B) were simultaneously derived. On this note, function *stat_density_2d* function from the *ggplot2* R package was used.

All contrasts of all linear models, for the three datasets, demonstrated qualitatively similar patterns observed to those in Figure 12, both when plotting the logFC as a function of the average expression (Figure 12A), and in the volcano plots (Figure 12B).

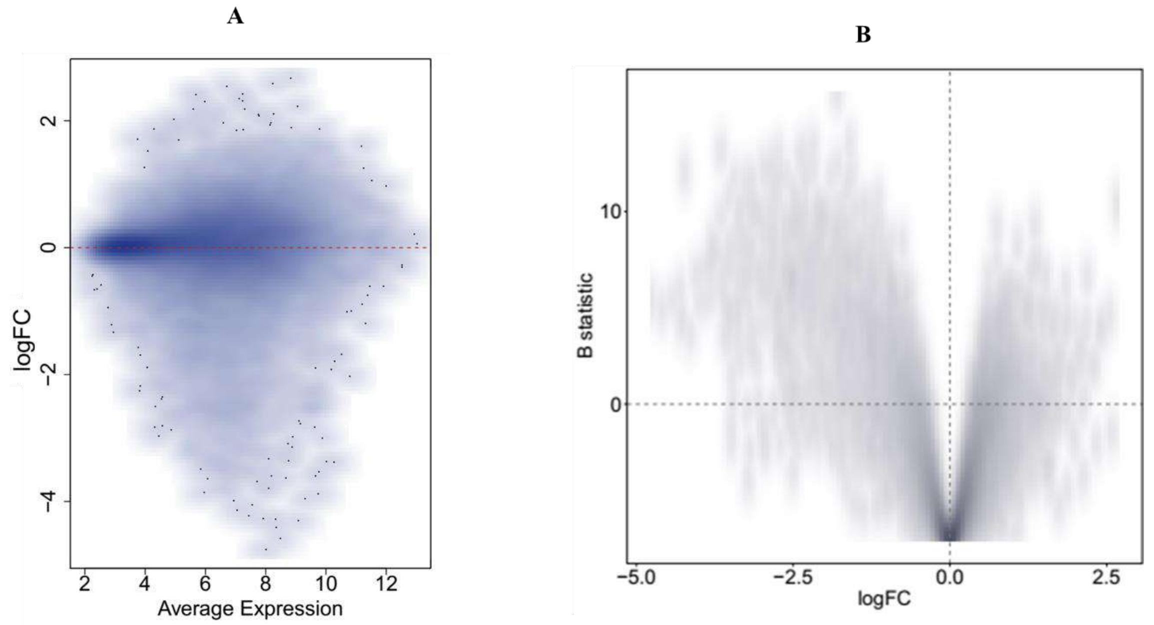


Figure 12 | Density scatter plots

Quality control of DEG patterns can be done through plotting genes' logFC values against (A) their average expression across all samples or (B) their statistical significance of differential expression. Given most genes are usually not differentially expressed, point density should be higher around a logFC of zero.

8.7.4. Choosing the most suitable linear model

Linear regression aims to correctly infer the linear relationship between the explanatory and dependent variables, based on sample data [125]. Nonetheless, it is easy to generate a model that **overfits** on the sample data by encompassing an excessive amount of explanatory variables which can also be redundant or too specific for that sample [125]. In the case of overfitting, the model explains the given sample data but is not applicable to other samples. In contrast, if the model does not encompass enough explanatory variables, it can **underfit** and lack relevant complexity, thus being unable to make accurate predictions [145] (Figure 12).

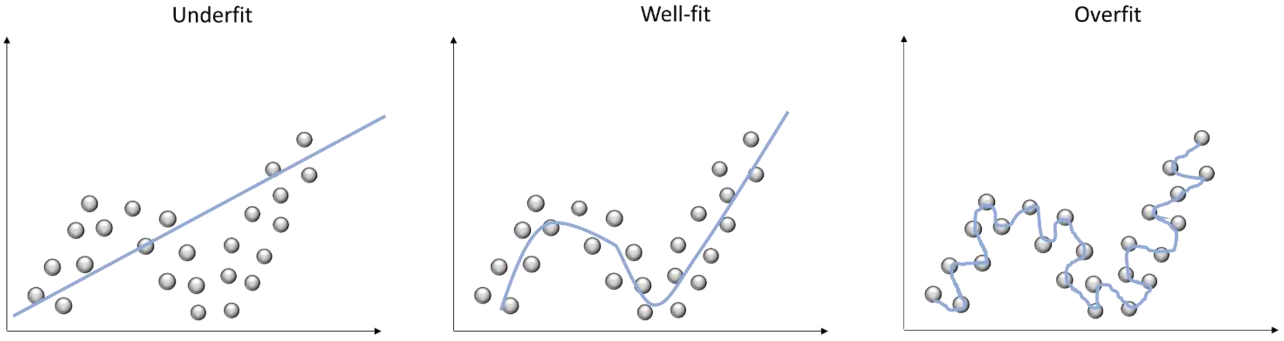


Figure 13 | Fitting data

Example of underfitting and overfitting a regression model. On the left panel, the model underfits the data by not adequately capturing their underlying structure; a linear model is fitted to clearly non-linear data. A good model (middle) approximates well the data's true underlying structure and provides a good representation of all points. An over complex model (right) adjusts itself to all available points and overfits the data, capturing their residual variation (i.e. noise) as if part of their underlying structure.

Models with different combinations of explanatory variables can be compared among themselves in order to infer their relative suitability for further analysis. On this note, scatter plots, regarding the coefficients associated with the same explanatory variables between linear models, can be computed for t-statistics, B-statistics and logFC, and these statistics can be tested for their correlation between variables and/or models.

For instance, let us consider model Z with explanatory variable α, and model Y with α and γ. A high correlation of logFC or B-statistic values of α between the two models (see example scatter plot in Figure 14, where each point is a gene) could mean that the γ explanatory variable is not confounded with α, i.e. it is independent from α (blue points in Figure 14).

Genes that, for a given explanatory variable, appear highly correlated between the two compared models, are similarly DE in both (i.e. have proportional magnitude and significance). Contrarily, genes deviating from correlation suggest that the α estimates are different between models (green and magenta points in Figure 14) and therefore that there is, at least, some interdependence between explanatory variables (between α and γ, for the considered example).

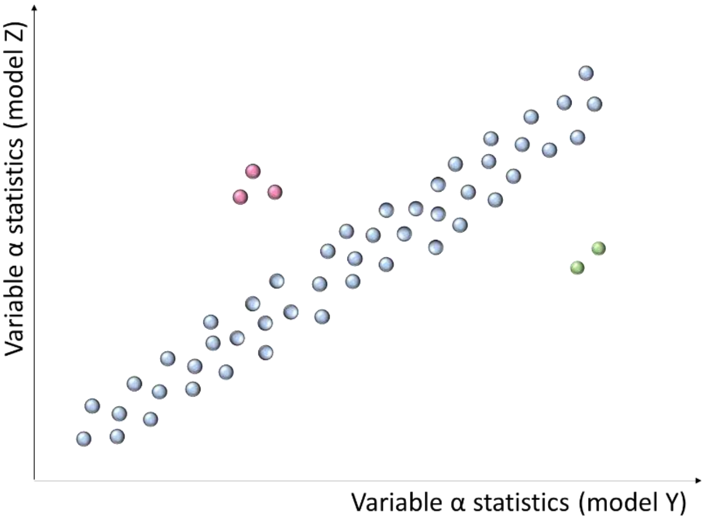


Figure 14 | Comparison between statistics from different models

Considering model Z with explanatory variable α, and model Y with α and γ, the plot represents the hypothetical comparison of a certain DE statistic (logFC, B, t statistic) associated with variable α between the two models. Blue points represent genes that are highly correlated between the two models (i.e. are equally differentially expressed), whereas green points and magenta those that are not.

Overall, the t-statistic was used to compare the different explanatory variables, as it encompasses information on the magnitude of differential GE, and correlates with its statistical significance (B-statistic).

By correlating different explanatory variables within the same model, one can also infer on redundancy. Highly correlated variables do not add valuable contributions to the model and indicate overfitting [125]. Collinearity or multicollinearity can be fixed either through inclusion of just one of the variables, or through the creation of a new variable that merges all the information [125].

The biological meaning and importance of the explanatory variables were also accounted for as exclusion or inclusion criteria, in the light of their relevance towards the main objective of the project of comparing the transcriptomes of AD human patients and mouse models. For instance, Age was always a considered variable given that AD is known to be an age-related dementia [6,7,9,31,37].

8.8. Gene Set Enrichment Analysis

To extract the biological meaning associated with GE differences (e.g. which pathways may be concomitantly dysregulated) genes need to be functionally annotated. Gene Set Enrichment Analysis (GSEA), a software distributed by the Broad Institute [101,146], was used to statistically assess the enrichment in predefined gene sets amongst genes altered between conditions [101,146]. These classes are defined *a priori* considering genes with chromosome proximity, common involvement in biological pathways and processes, shared association with disease pathways, among other criteria. The software receives a ranked list of genes based on a metric that quantifies their differences in GE between two conditions in study [101].

The geneset-defining databases selected for the present study were Reactome [147,148], established by the European Bioinformatics Institute and the European Molecular Biology Laboratory (EMBL), KEGG [96–98] and the Biological Processes from GO [99,100].

GSEA software was provided, from each dataset (human, mouse and joint), with a list of all considered genes ranked by their correspondent t-statistic of differential expression derived with linear modelling (ranging from positive to negative values). The program returns the pre-defined gene sets linked to already known pathways, from the aforementioned databases, that are significantly enriched in differentially expressed genes. For each pathway, while scrolling through the given ranked list of genes, GSEA calculates the Enrichment Score (ES) by incrementing or decrementing a score – called running ES – depending if the gene in the list is included in the gene set or not, respectively. The increment/decrement magnitude is proportional to the relevance of the gene to the gene set-associated pathway, which is defined within the program [101,146]. This process results in a running ES distribution of which the maximum absolute value corresponds to the overall ES associated with the gene set [101,146]. The software returns a list of up-regulated and a list of down-regulated pathways, which in the current study are distinguished based on the t-statistic signal associated with the genes that contribute to the enrichment of the pathway in question (Figure 15).

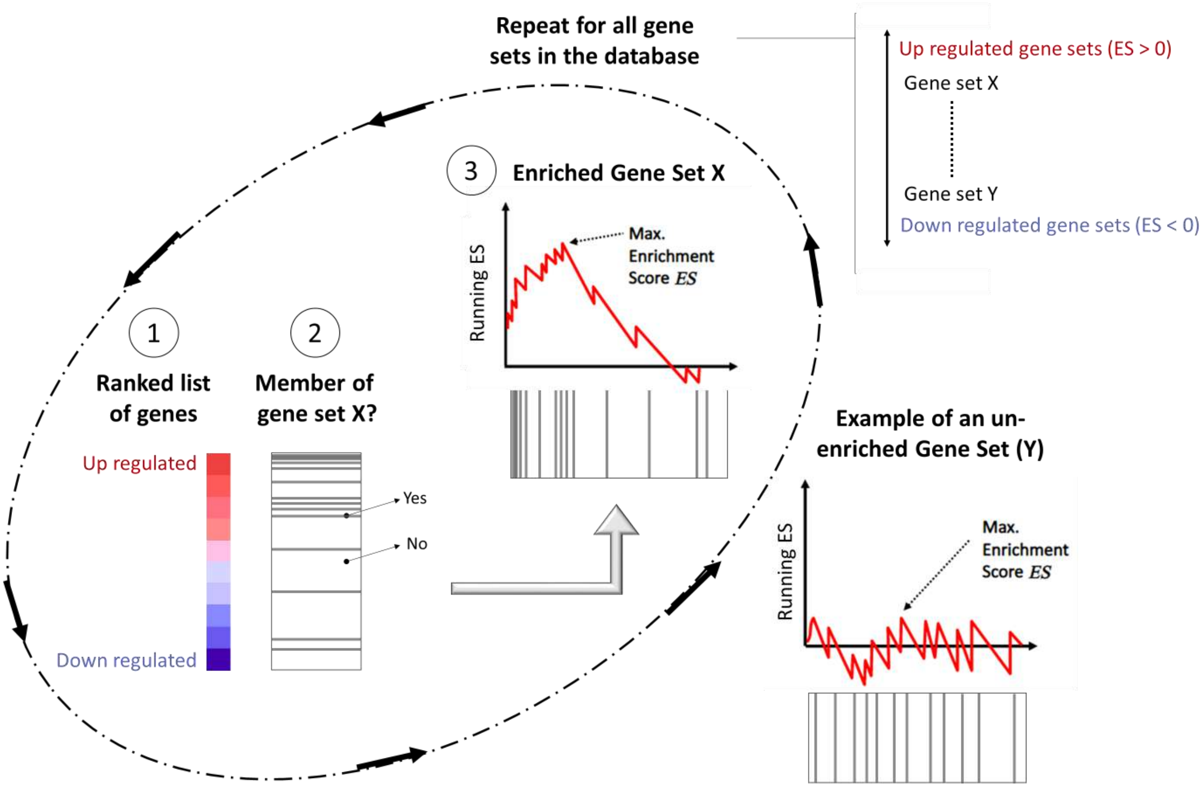


Figure 15 | Gene Set Enrichment Analysis

A ranked list of genes is given to GSEA software, which increases and decreases a running Enrichment Score (red) based on genes' contribution to the pathway, and returns a list of ranked pathways that are differentially expressed between the conditions in study, each associated with a final ES (Max Enrichment Score) value. Figure inspired by [146].

Since the number of genes present in the pathways can be very different from other pathways, a normalized ES (NES) that considers differences among gene set sizes is also computed, along with the p-value and FDR for statistical significance [101,146].

8.9. Identification of DE-recapitulating and DE-reverting genetic and chemical perturbations

To infer either genetic (gene over-expressions or knock-downs) or compound perturbations that recapitulate the changes in GE unveiled by applying linear modelling to the aforementioned datasets, *cTRAP*, an R tool developed in the host lab [149] was used. This package makes use of data from the Connectivity Map (CMap) of the Broad Institute [150,151]. The CMap database comprises GE data across several human cell lines, prior to and after being subjected to the different mentioned perturbations.

Similarly to GSEA, *cTRAP* accepts a list of ranked genes that, in the present analysis, was also based on t-statistics. The program then outputs a ranked list of perturbations among over-expressions, knockdowns and compound administrations, based on their ability to replicate similar or opposite transcriptomic changes. The correspondence between each returned alteration and the transcriptomic changes provided is evaluated through different statistical tests, of which Spearman's correlation was chosen due to the advantage of detecting non-linear relationships compared to Pearson's correlation, and thus enabling the identification of a broader range of relations. For each perturbation, a correlation coefficient value and correspondent p- and q-values were outputted. A positive test coefficient suggests perturbations that replicate the provided transcriptomic changes, while negative values reflect correlation with opposite transcriptomic changes.

Information on the genetic alterations and compounds is also provided, including the genetic target of the drug, its clinical development phase and the medical field and condition of prescription [150,151]. For the purpose of this thesis, only marketed compounds or those in phase III of clinical trials were considered of interest in order to ease future validation work.

8.9.1. DAVID software

The online DAVID software [152,153] was used to summarize gene targets associated with the compound perturbations obtained with *cTRAP* into categories with explicit functions. The list of target genes was inputted to the *Gene Functional Classification* tool and compared against the default *Homo sapiens* gene list provided by DAVID, as the transcriptomic background used to derivate all the drug gene targets was absent. DAVID outputs a list of the gene targets clustered by their functions. These gene families were then manually curated in order to reflect the core functions of its genes. The obtained information was appended to the *cTRAP* results.

CHAPTER III – RESULTS AND DISCUSSION

9. Quality control and sample removal for human and mouse data

9.1. Human dataset

To assess array quality and identify possible outliers within the human data, raw chip images (Figure S1) were generated along with NUSE and RLE plots (Figure S2). Given the absence of visual artefacts on those images, and the samples being within NUSE and RLE quality standard intervals (see section 8.1), no samples were considered to be outliers at this stage. However, the NUSE distributions of three controls (C1, C3 and C5), one EOAD (E2) and one PSEN (P4) were the most shifted towards values greater than 1 and apart from the remaining samples (Figure S2), and thus were flagged for further quality control steps.

Boxplots of normalized probe-intensity for human revealed E2 as the most deviant sample, but it aligns with the remaining samples after normalization (Figure S3). In the heatmap, C1 and P4 (deviant in NUSE) also did not cluster with the remaining samples of the same category. Additionally, samples E3 and P6 clustered with control samples (Figure S4A). Besides, C1, E3, E4 and P6 were also separated using PCA coloured by condition (Figure 16).

Despite the deviations observed for C1, E3, E4 and P6, they do not fall within the criteria to be considered as technical outliers, i.e. after normalization there are no deviant distributions in the boxplots, nor isolated samples in heatmaps' clusters or PCA distributions. Moreover, there can be variance attributable to biological factors not covered by the available metadata. Nonetheless, considering all the results from the exploratory analyses, C1, E3, E4 and P6 were discarded from posterior analyses of the human dataset, in order to maximize the distinction between non-diseased individuals and AD patients. After removing those samples, P4 was the only visually misplaced sample in the heatmap (Figure S4B). Nevertheless, since P4 clusters with diseased samples and is consistent with the remaining PSEN samples in PCA (Figure S5), P4 was kept in the dataset and therefore get a more robust GE signature for that distinction.

In other words, if the deviations observed for C1, E3, E4 and P6 had a biological meaning, they could provide extra information to the model and narrow down the DEG by considering the slight overlap in GE distributions between the controls and AD patients. However, the decision was to exclude those samples, sacrificing biological variability for robustness of the GE signatures and maximization of the separation between the controls and AD patients.

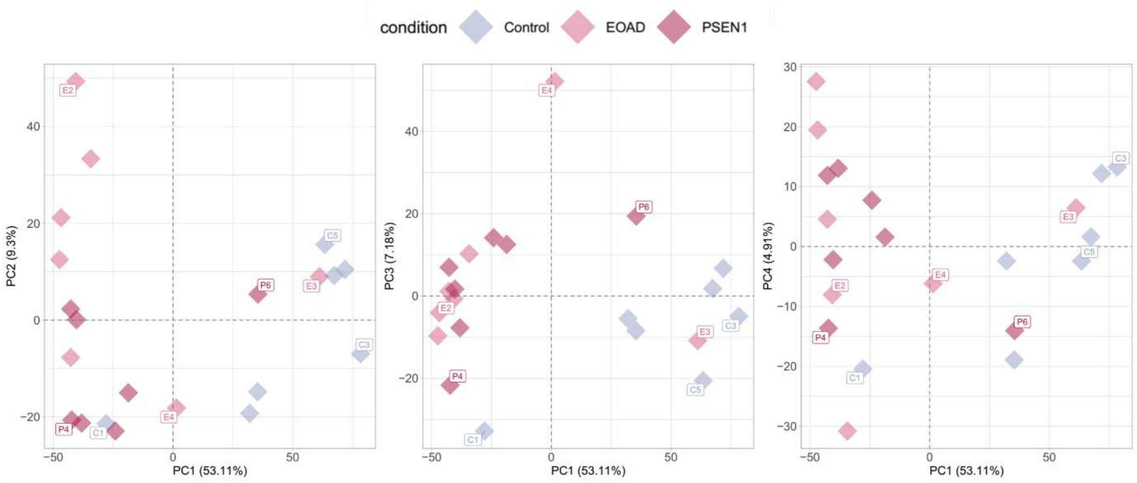


Figure 16 | PCA before sample exclusion coloured by condition (Human)

Principal components of normalized human GE data, with points coloured by AD condition. Flagged samples are identified with labels (C for Controls, E for EOAD and P for PSEN).

9.2. Mouse dataset

The boxplot for non-normalized GE mouse data (Figure S6) and the heatmap for normalized data (Figure S7) unveiled three deviant control samples (C22, C23 and C24). C22 and C23 substantiated as outliers in the heatmap and in PCA on normalized GE data (Figure 17), hence being discarded from further analyses. Considering that C24 clustered with the remaining samples in PCA (Figure 17 and Figure S8), this sample was kept. Results were concordant across the three murine datasets, with the results displayed below being only related the homozygous double-mutant mice.

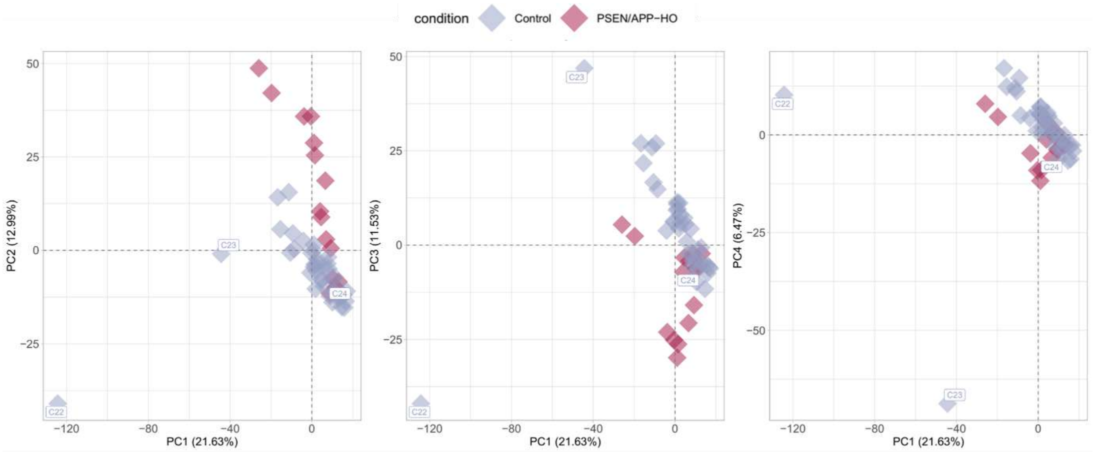


Figure 17 | PCA before outlier exclusion coloured by condition (HO Mouse)

Principal components of normalized mouse GE data, with points coloured by AD condition. Flagged outliers are identified with labels (C for Controls).

10. Human dataset analysis

10.1. Finding explanatory variables with an impact on gene expression

To evaluate the effect of potential explanatory variables on data variance, PCA on normalized GE data was plotted and coloured according to samples' available metadata, such as Condition (which includes the distinction between control, EOAD and PSEN samples), age (in years), PMI (in minutes), gender, *APOE* genotype, Braak stage and the type of *PSEN1* mutation. Only Condition, Age and PMI revealed observable trends in the data (Figure 18 and Figure S5).

The first principal component (PC1) explains 57.5% of the normalized GE variance (Figure S9), being explicit a separation of non-diseased subjects and AD patients, visible when coloured by Condition, while the second component seems to explain the variance between EOAD and *PSEN1*-mutated AD patients (Figure 18). Visible trends were absent along the other components, for the Condition variable (Figure S5).

However, samples' age and PMI also appeared associated with the pattern along PC1 (Figure 18), with samples from older individuals appearing to also present higher PMI values. The fact that these three variables (condition, age and PMI) vary along the same principal component suggests that they might be confounded to some extent. As such, *Condition*, *Age* and *PMI* were the explanatory variables used for linear modelling, and the decoupling of the three effects was attempted.

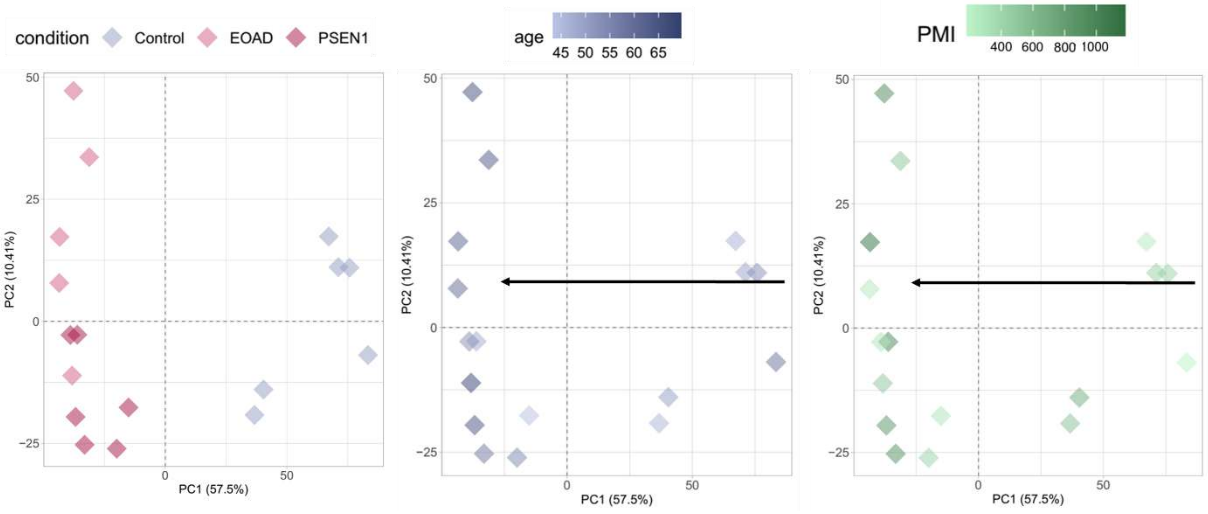


Figure 18 | Principal Component Analysis (Human)

Principal components of normalized human GE data, with points coloured by AD condition (left), age (middle) and PMI (right).

10.1.1. Age and PMI correlation

Given that age and PMI followed a similar pattern in PCA plots, in order to integrate information on both variables in the linear models, it is important to ensure their independence to avoid redundant information and overfitting. On this note, correlation between age and PMI original values was computed ($Spearman's\ rho\ =\ 0.26,\ p\ =\ 0.31$; Figure 19). Given the p-value of 0.31, the null hypothesis, which states that the variables are not correlated, could not be rejected at a significance level of 0.05. Thus, the variables were considered mostly independent of each other, also based on the low correlation coefficient, hence both could be integrated as predictors in the same linear models.

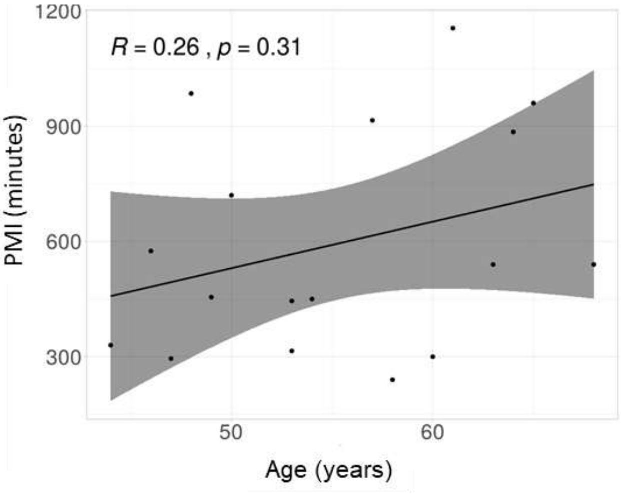


Figure 19 | Age and PMI comparison

Comparison between Age (in years) and PMI (in minutes) across human samples. Linear regression line in black, and its 95% confidence interval is represented by the grey area. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

To better unveil the importance of modelling age and PMI, t-statistic values for each coefficient were compared using different linear models. The comparisons were made between single-*Age* and single-*PMI* models (models 3 and 4, respectively – see Table 4), and between *Age* and *PMI* coefficients of models where both variables are used as predictors (models 7 and 10 – see Table 4). The variables are not much correlated when compared between different models where each variable is the only explanatory variable, nor when they are concomitantly present with *Disease* in a model (Figure 20A-B). Even though *Age*'s and *PMI*'s t-statistics were slightly correlated when the *Disease* variable is distinguished between *PSEN* and *EOAD* (Figure 20C), it was decided to still be feasible to use them simultaneously in linear models. Of note, low p-values indicate significance of the low correlations due to the very high number (thousands) of genes considered in the analysis, but the assessments above were based on the magnitude of association (*Rho*) between variables.

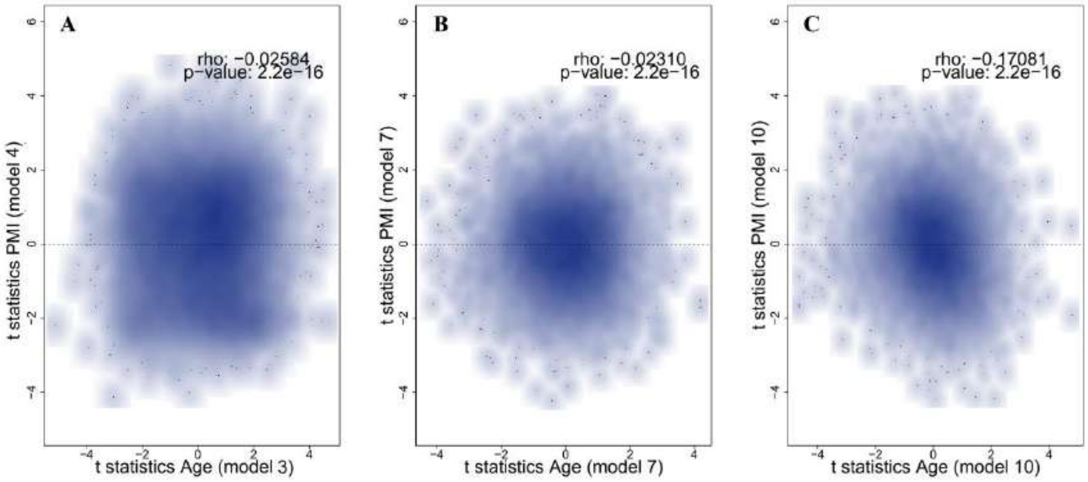


Figure 20 | Age and PMI t-statistic comparison

Comparison between t-statistics of differential expression retrieved with linear modelling for (A) single-*Age* and single-*PMI*, for (B) coadjutant *Age* and *PMI* coefficients concomitant with *Disease*, and (C) coadjutant *Age* and *PMI* coefficients concomitant with *PSEN+EOAD*. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

10.1.2. Assessing condition confounding with Age and PMI variables

The age and PMI trend along PC1 suggested a confounding with AD condition. In order to better understand the relation between the three variables, age and PMI distributions for non-AD, EOAD and PSEN1 samples were compared (Figure 21).

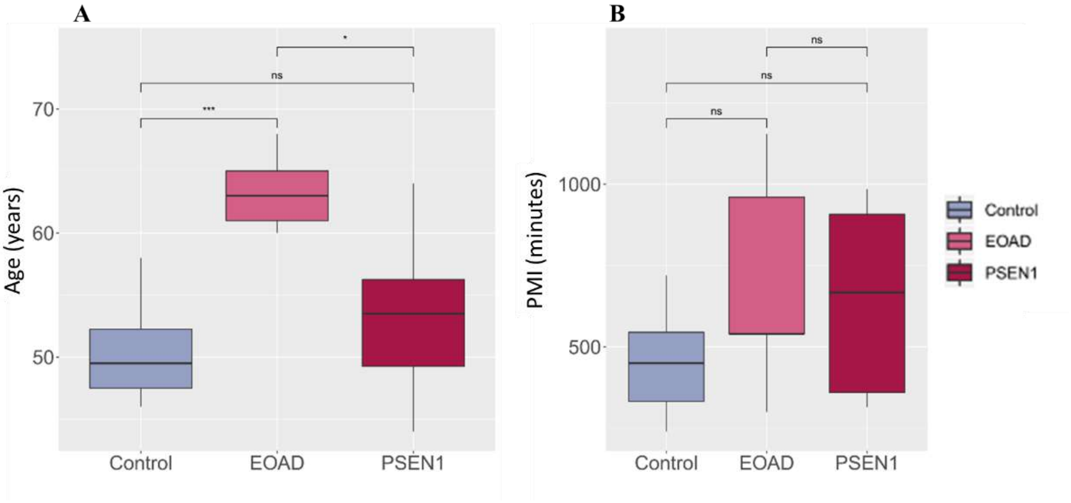


Figure 21 | Age and PMI distributions for Controls, EOAD and PSEN samples

(A) Age (in years) and (B) PMI (in minutes) distributions across Controls, EOAD and *PSEN1*-mutated samples. Between groups differences were respectively assessed with the t-test (A) and the Wilcoxon-test (B), in accordance with sections 8.4.1 and 8.4.2 and 3.4.2 ($* p < 0.05, ** p < 0.01, *** p < 0.001$, ns stands for non-significant, for $\alpha = 0.05$).

No statistically significant PMI differences were found between groups (Figure 21). Additionally, *PMI* should not be considered an explanatory variable, given that no genes are significantly differentially expressed based on PMI when considering this variable in any of the linear models (Figure S12). This suggests that PMI does not add relevant information to GE differences between samples.

When comparing t-statistic values of differential expression for all the *Condition* coefficients across *PMI*-differing models (models 1 and 6, 2 and 9, 5 and 7, and 8 and 10), correlation coefficients were close to 1 (Figure 22). These results corroborate the idea that using or not PMI as a predictor does not affect the relationship between the *Condition* variable and the response variable (gene expression) and, therefore, PMI does not add robustness to the model regarding the distinction between AD and non-AD samples. As such, we decided not to include *PMI* as an explanatory variable for linear modelling of GE.

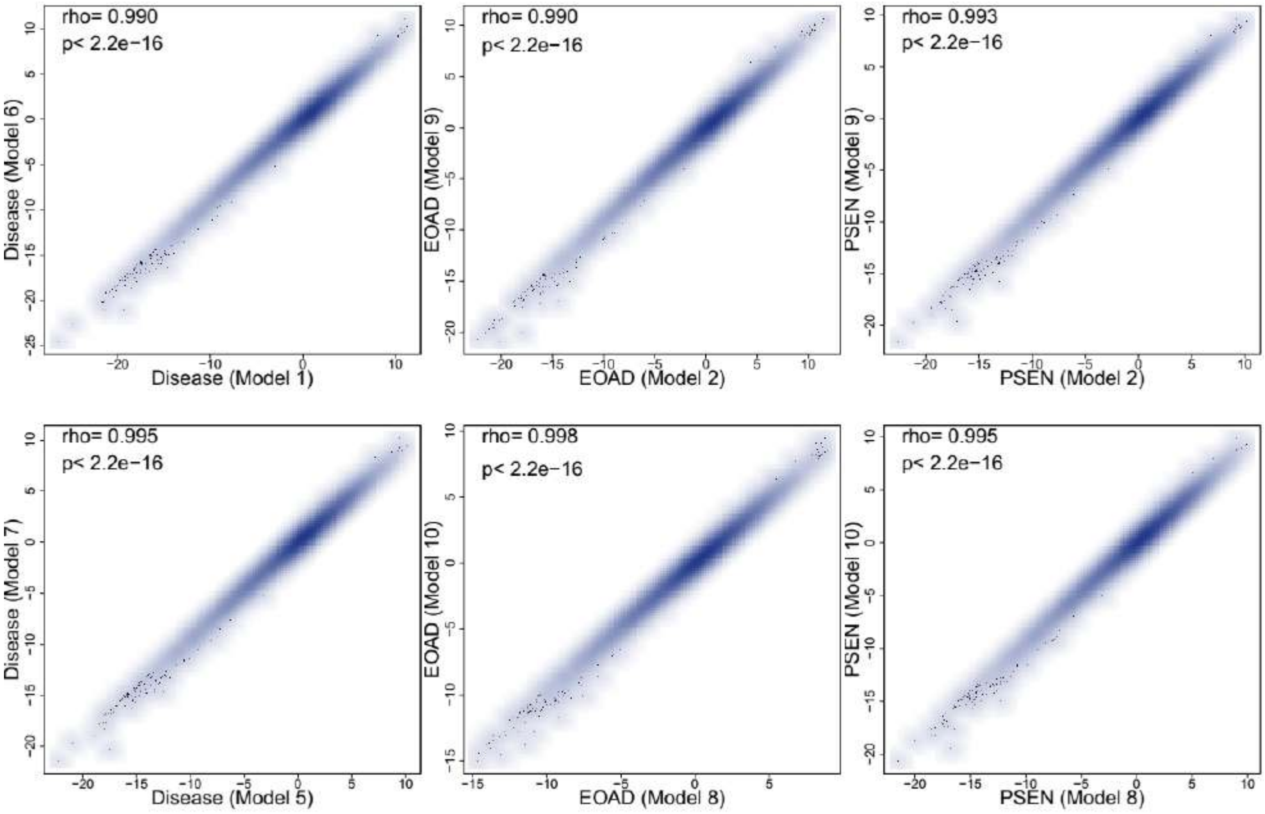


Figure 22 | t-statistic comparison between explanatory variables of *PMI*-differing models'

Comparison between t-statistics of differential expression retrieved with linear modelling for the same *Condition* variable across *PMI*-differing models. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

EOAD patients are significantly older than controls and PSEN patients (Figure 21), illustrating a possible confounding between AD condition and the age of the individual. As such, t-statistics were compared between single-*Age* (model 3) and single-*Condition* (Disease – which encompasses both EOAD and PSEN –, EOAD and PSEN; models 1 and 2) variables, and between coadjutant *Age* and *Condition* of models that include information exclusively on those variables (models 5 and 8) (Figure 23) – see Table 4 for linear models.

Single-*Age* and single-*Condition* appear to affect GE similarly, since these are highly correlated (Figure 23A). That effect is somewhat diluted and even inverted when the model considers both variables (Figure 23B), suggesting not only that both variables, to some extent, affect GE is the same way, but also that is difficult to fully decouple their independent effects, emphasizing the condition-age bias. Moreover, the *EOAD* coefficient is more correlated with *Age* than *PSEN* (Figure 23), which is in accordance with EOAD patients being the oldest (Figure 21).

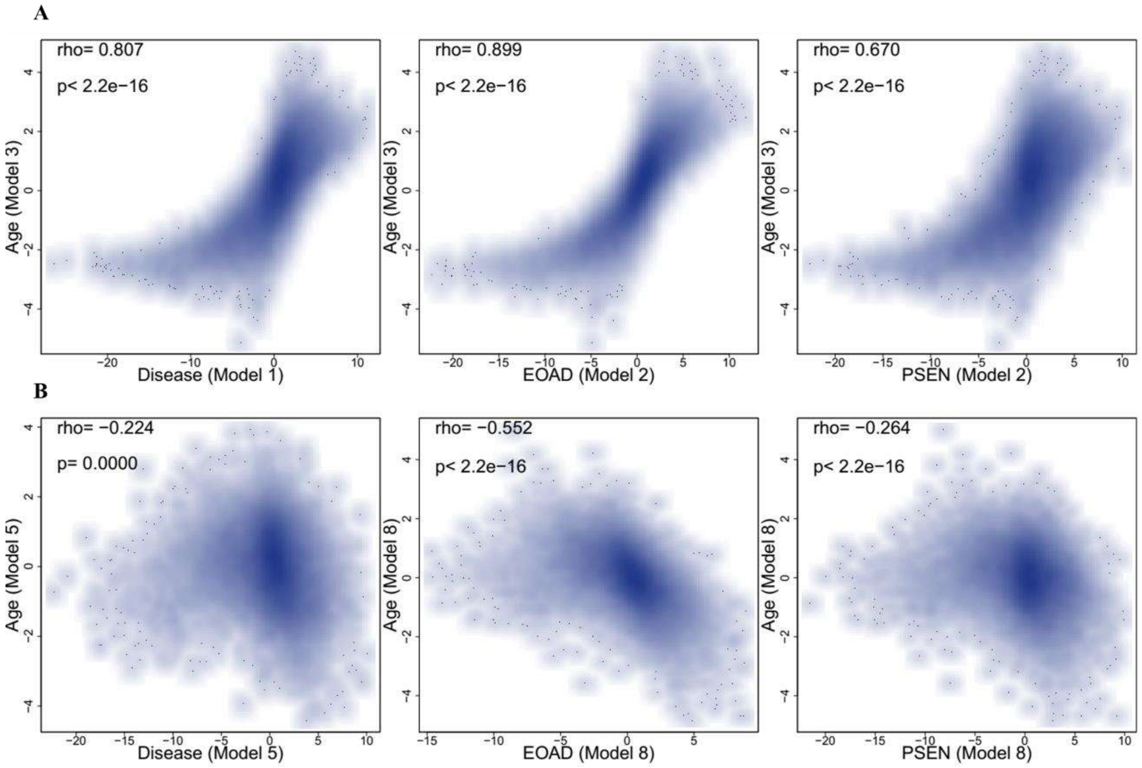


Figure 23 | t-statistic comparison between Age and Condition variables

Comparison between t-statistics of differential expression retrieved with linear modelling for (A) single-*Age* and single-*Condition* variables and for (B) coadjutant *Age* and *Condition* coefficients. Rho and *p* – Spearman’s rank correlation coefficient and p-value, respectively.

Condition variables appear highly correlated between *Age*-differing models, suggesting that *Age* as an explanatory variable does not have a big impact on the *Condition* effect (Figure 24), and thus could be removed from the models, similarly to what was observed for PMI. Nevertheless, considering its biological meaning and the fact that AD is a dementia with high prevalence in older individuals, *Age* was kept as a predictor. Instead, in order not to disrupt the linear models with a confounding effect on the association between *Age* and *Condition*, and given the project focus on comparing human and mouse data (the latter lacking EOAD samples), EOAD individuals were henceforth removed from the human dataset. If not for the confounding effect, it would be interesting to keep these samples in the dataset and gain insights on idiopathic AD. In conclusion, a model with *PSEN* and *Age* as explanatory variables (equation 4) was considered for further analyses.

$$GE = \beta_0 + \beta_{PSEN}PSEN + \beta_{Age}Age \tag{4}$$

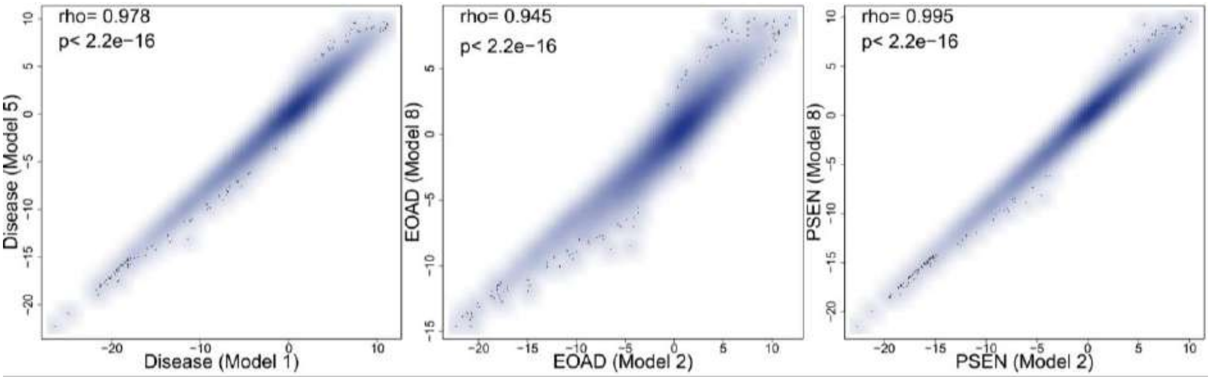


Figure 24 | t-statistic comparison of Age-differing models

Comparison between t-statistics of differential expression retrieved with linear modelling for the same *Condition* variable across *Age*-differing models. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

10.2. Exploring AD-induced DEG in human patients

The aforementioned linear model (equation 4) unveiled a few up-regulated genes in human AD compared to non-diseased individuals, but the majority of the DEG were down-regulated (Figure 25).

The most statistically significant up-regulated gene is *ARRDC4*, which participates in the internalization of activated G-protein coupled receptors [154] and, more recently, has been linked to the innate immune system, participating in inflammatory responses [155]. Other up-regulated genes include *FOXJ1*, *LPAR4* and *ST6GALNAC2*.

FOXJ1 encodes a transcription factor required for the formation of motile cilia, hair-like structures present in the surface of ependymal cells that outline the brain ventricles filled with CSF [156]. The motile cilia generates CSF movement that is vital for cleaning of waste products and transport of nutrients and signalling molecules [156].

LPAR4 encodes a lysophosphatidic acid (LPA) receptor. LPA participates in cellular survival, differentiation, proliferation and migration, and has also been related with neuronal and glial alterations in neuronal disorders [157]. As for *ST6GALNAC2*, it induces molecular modifications that might affect cellular communications [158].

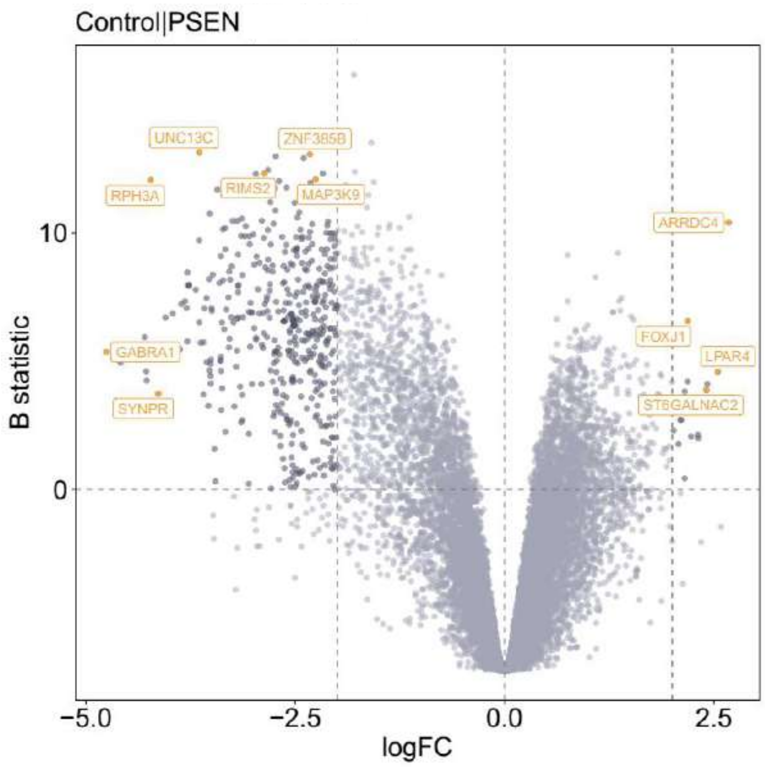


Figure 25 | DEG in human AD

Volcano plots of differentially expressed genes up- (positive logFC) and down-regulated (negative logFC) in human AD patients, compared with non-diseased individuals, for the *PSEN* coefficient. Genes in orange stood out from the most differentially expressed and/or statistically significant genes. Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed line) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 2; B > 0).

Some of the most down-regulated genes, such as *MAP3K9* [159] and *ZNF385B*, are associated with cellular responses, namely apoptosis. *MAP3K9* is involved in responses evoked by environmental triggers, which can culminate in neuronal apoptosis [159]. Intuitively, these processes were expected to be up-regulated in AD compared to non-disease conditions because of the neuronal death known to be associated with AD. However, we can also speculate that their down-regulation can be a feedback mechanism: given the increased and non-physiological neuronal death that occurs in a diseased state, the organism might impair natural apoptosis mechanisms in an attempt to control further damage. The ZNF385B protein contains zing-finger domains and its cognate gene was found to be highly expressed in B-cells and regulate the induction of their apoptosis when in ectopic conditions, by functioning as a p53-mediated DNA damage control checkpoint [160]. According to The Human Protein Atlas [161–163], *ZNF385B* is highly expressed in the brain, especially in the cerebellum and cerebral cortex [164], but a connection with its expression in B cells has not yet, to our knowledge, been proposed. Given the rise in immune responses seen in AD conditions, the need to reduce B-cell apoptosis might be expected, which would explain *ZNF385B* down-regulation. However, it is of note that *ZNF385B* expression was only detected, in the author's study, in B-cells of germinal centers, where B-cells proliferate and differentiate [160]. The gene may, of course, also have other function within the brain that have not yet been discovered.

Furthermore, synaptic functions were also apparently altered, with genes like *UNC13C* [165], *RPH3A* [166], *GABRA1* [167] and *SYNPR* [168], which are associated with synaptic vesicle cycle and trafficking, formation of synapses and neurotransmitter release, being down-regulated in AD. At a presynaptic level, rabphilin-3A, encoded by *RPH3A*, is recruited to synaptic vesicle membranes where it modulates synaptic vesicle trafficking. Reduction of *RPH3A* levels/activity has already been linked to AD pathology and increased concentrations of Aβ [166]. *GABRA1* acts as a GABA receptor, the main inhibitory neurotransmitter in mammals, and plays a role in synaptic inhibition [167]. The down-regulation of *GABRA1* suggests that, besides the aforementioned loss of synapses and decrease in the trafficking of neurotransmitters associated with AD, synaptic inhibition itself is also affected.

It is noteworthy that the down-regulation of those genes, especially the ones associated with synaptic activity and signal transmission, can also derive from neuronal loss, which is not detectable by the used linear model since it does not encompass for neural cell type proportion as an explanatory variable. Nevertheless, this would be an interesting and important factor to consider in future work.

10.3. Disrupted biological pathways in human AD brains

In order to extract the biological meaning associated with the AD-induced differentially expressed genes (given by the *PSEN* coefficient), Gene Set Enrichment Analysis (GSEA) was conducted to obtain pathways that might inform on gene function, which are pre-defined in databases such as Reactome, GO Biological Processes and KEGG. After converting probe annotation to gene annotation, genes were ranked by their t-statistics of AD-associated differential expression and ran in GSEA.

The most up-regulated Reactome pathways (Figure 26) and GO Biological Processes (Figure S13) in AD conditions are related with the immune system, DNA and cell cycle, and elastic fibres. KEGG pathways (Figure S13) follow the same pattern, with immune system-related conditions appearing up-regulated, such as viral infections, auto-immune diseases and transplants.

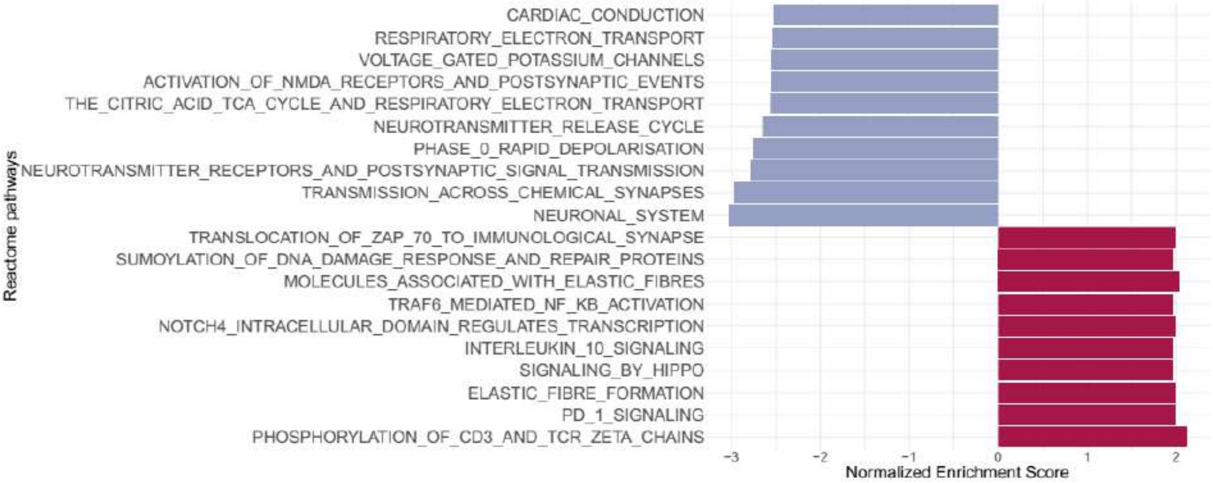


Figure 26 | Altered Reactome pathways for the *PSEN* coefficient (Human)

Representation of the 10 most significantly down- (blue) and up-regulated (magenta) Reactome pathways, in human *PSEN* brains. Pathways were considered significantly enriched if FDR < 0.05.

As aforementioned, immune responses are increased in AD, where there is an abnormal migration of glial cells towards brain regions with increased neural debris and Aβ deposition, resulting in the release of pro-inflammatory cytokines and ultimately triggering a chronic inflammatory state of the diseased brain [10,17,30,31].

Elastic fibres, of which elastin is the main component, are macromolecules present in the extracellular matrix of dynamic connective tissues, such as the lungs, skin and blood vessels [169]. These fibres are indeed vital for maintaining the flexibility and extensibility of arteries, and work along collagen fibrils [169]. Elastin has also been associated with deposition of amyloid-like structures in blood vessels, which disrupts the circulatory system [170]. Additionally, in AD, Aβ oligomers may migrate into the blood stream, blocking and disrupting brain arteries [170], and possibly the elastic fibres present in the cerebral vascular system, hence the activation of their synthesis mechanisms.

Cell cycle and the Hippo signalling pathways are up-regulated in PSEN conditions, as well. These pathways are involved in cellular differentiation, proliferation, and apoptosis [171], which aligns with glial activation and proliferation, immune cells proliferation and neuronal apoptosis, processes known to occur in an AD context. Moreover, Aβ oligomers and oxidative stress induce MST1-mediated (MST1 being an hypo kinase) phosphorylation of the Forkhead transcription factor *FOXO3*, triggering an apoptotic pathway that culminates in neuronal death [171–173].

The oxidative stress that underlies the pathogenic mechanisms of AD leads to oxidative DNA damage. The accumulation of disrupted DNA molecules triggers neuronal death and contributes to disease progression [174], being therefore predictable the activation of DNA repairing mechanisms to counter-balance these events (Figure S13).

Down-regulated Reactome pathways and Biological Processes in *PSEN* relate with synaptic and neuronal activity, as well as neurotransmitter trafficking. This is in accordance with AD-related mechanisms such as (1) microtubule disruption through tau hyperphosphorylation which destabilizes axonal transportation in affected neurons, thus disrupting vesicle and neurotransmitters trafficking [6,10], and (2) Aβ interference of transmembrane channels, such as NMDA receptors [24,26]. These mechanisms hinder neurotransmitters' release and culminate in synaptic loss and neuronal death, the major cellular hallmarks of AD. The down-regulation of NMDA receptors' activation suggests that Aβ oligomers are inducing their loss of function rather than their overstimulation, as discussed in section 1.2.

Moreover, Antonell *et al*. also found, in both EOAD and PSEN1 comparisons with non-AD samples, disruptions within pathways related to signal transmission, with great focus on calcium signalling, neuronal ligand-receptor interactions and long term potentiation, thus impacting neuronal plasticity [103]. The authors used DAVID software to identify these potentially dysregulated pathways from the DEG, not specifying if these pathways' activities are positively or negatively affected. However, they have clarified that the genes associated with the pathways were found down-regulated, i.e. pathways were enriched in down-regulated genes.

Unexpectedly, some KEGG pathways related with neurodegenerative diseases, such as AD, and with diabetes mellitus (considered as a risk factor of AD [175,176]), are downregulated in *PSEN*. Given the GSEA scoring system with ES (explained in section 8.8), it was hypothesized that, in the dataset used for the analysis, the number of genes that are usually down-regulated in AD is greater than of those usually up-regulated, thus having a major contribution to the negative ES scores of those pathways (Figure 27). In fact, GSEA unveils that the major contributors for the down-regulation of the AD pathway, i.e. a pathway that represents the unfolding of events that lead to the diseased state, are mitochondrial genes belonging to the *NDUF*, *SDH*, *UQCR*, *COX* and *ATP* families, whose altered levels have been associated with mitochondrial dysfunction in AD [177]. KEGG representation of the AD pathway (Figure S14) corroborates this hypothesis, showing that the expression of those genes (categorized in the Cx family) is inhibited in AD [178]. *GRIN2B* and *GRIN2D* genes encode proteins of NMDA receptors' subunits [179], whose loss of function was already suggested in the Reactome analysis above. Likewise, the *ITPR1* gene that encodes for a intracellular channel that mediates calcium release from the endoplasmic reticulum [180], and the *CALM3* gene that encodes a calcium-binding protein [181], were found down-regulated and contributing to the GSEA results regarding the AD pathway, which can be linked to alterations in membrane receptors, such as those seen in NMDA receptors, and altered Ca^{2+} homeostasis [181] in AD.

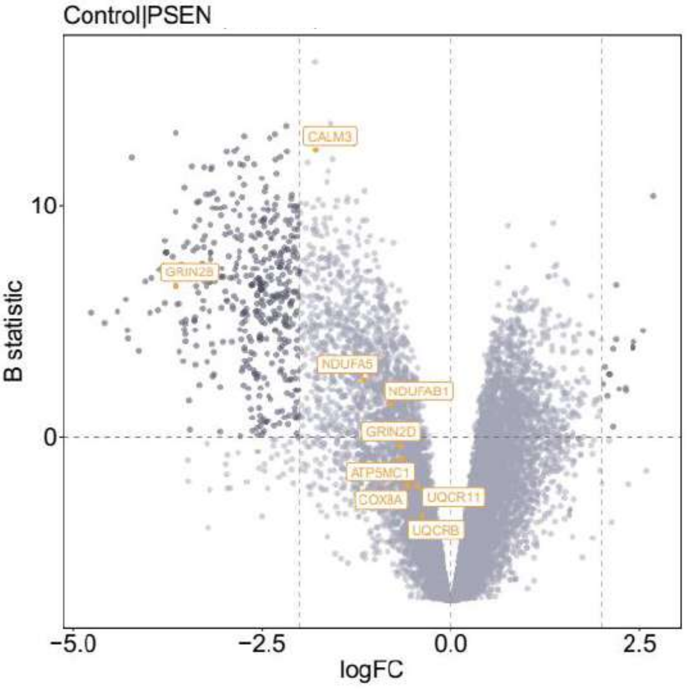


Figure 27 | Genes that contribute to the Alzheimer's disease KEGG pathway

Volcano plots of differentially expressed, i.e. up- (positive logFC) and down-regulated (negative logFC), genes in human AD patients, compared with non-diseased individuals, for the *PSEN* coefficient. Highlighted genes are a sample of those contributing to the down-regulation of Alzheimer's disease KEGG pathway. Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed line) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 2; B > 0).

11.Mouse dataset analysis

The murine dataset considered for the present analysis involved mice exclusively carrying a human *PSEN1* mutation (single-mutant mice, PSEN), and mice carrying heterozygous (HET) and homozygous (HO) combinations of human mutations in the *PSEN1* and *APP* genes – double-mutants. These three conditions were analysed as separate datasets.

11.1.Finding explanatory variables with an impact on gene expression

Biological features besides *Condition* and *Age* were not available for mouse data, thus these comprised the explanatory variables available for linear modelling. There was low correlation between the t-statistics obtained for these variables (Figure 28), showing that they are not substantially associated (the very high significance value can, in this case, be a consequence of a very large sample size), except perhaps mildly for the homozygous double-mutant mice. Henceforward, a model with *PSEN* and *Age* as explanatory variables was used for in the analysis of the three mouse datasets (equation 5).

$$GE = \beta_0 + \beta_{PSEN} PSEN + \beta_{Age} Age \tag{5}$$

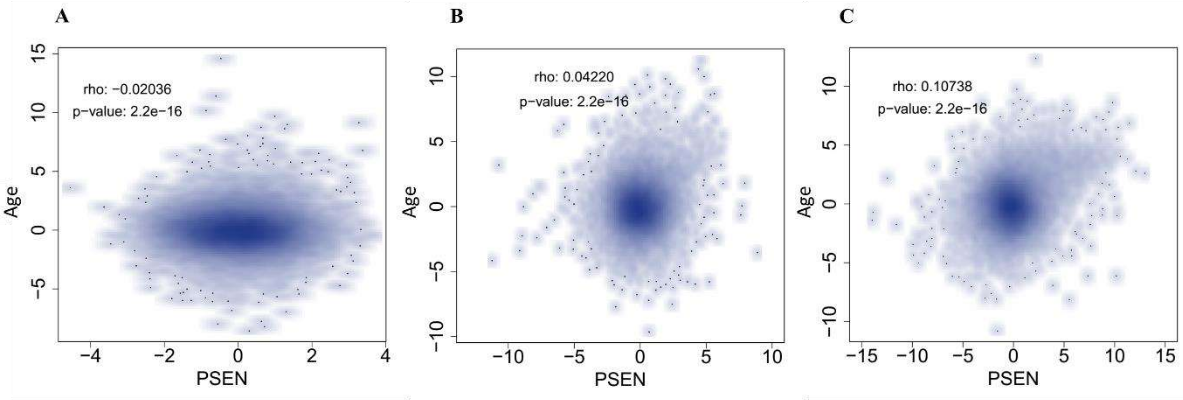


Figure 28 | Age and PSEN t-statistic comparison (Mouse)

Comparison between the t-statistics of differential expression for the *PSEN* and *Age* coefficients of (A) the single-mutant, (B) the heterozygous double-mutant and (C) the homozygous double-mutant mice. To ease visualization of the correlations, the limits on x-axis do not include *THY1* gene, which has a high t-statistic value on the *PSEN* coefficient (interpreted in section 11.2). Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

11.2. Exploratory transcriptomic characterization of the three mouse datasets

PCA plots for the single-mutant and HET mouse datasets displayed no pattern in the data, associated with either the AD condition or the age of the mice (Figure 29A-B). Moreover, no relevant separation between AD and non-AD samples was found. Regarding the homozygous double-mutant mice, PCA unveiled a relevant separation between non-AD and AD samples along PC1 and PC2 (Figure 29C).

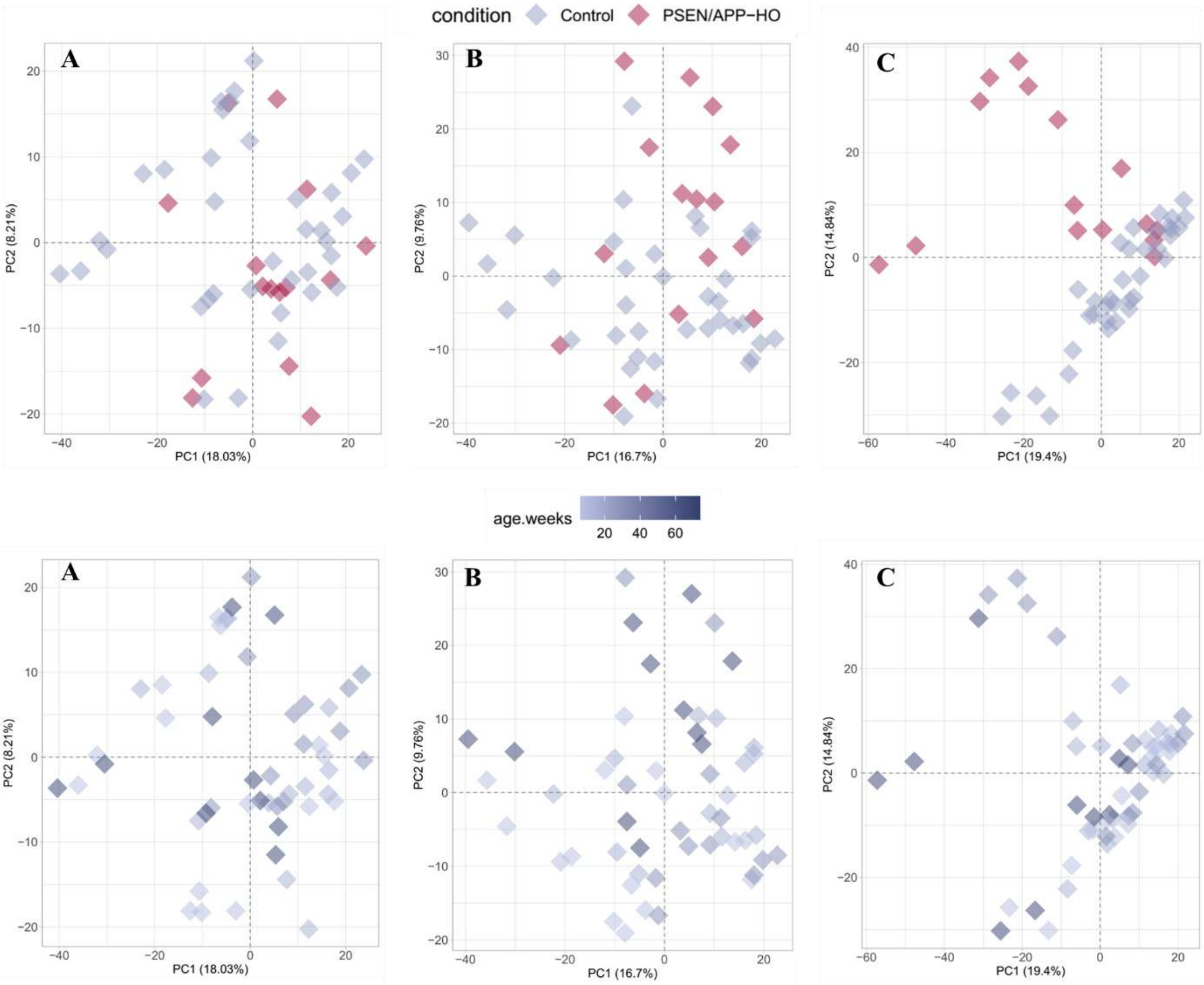


Figure 29 | PCA of the mouse datasets

Principal components of normalized GE, with points coloured by condition (top) and age in weeks (bottom), for (A) single-mutant mice, (B) heterozygous double-mutant mice and (C) homozygous double-mutant mice.

Moreover, linear modelling did not unveil statistically significant DEG regarding the *PSEN* coefficient for the single-mutant mice (Figure 30A). In the heterozygous double-mutant, the majority of the statistically significant DEG detected in *PSEN* have a small magnitude effect

($|logFC| < 1$) between AD and non-AD samples (Figure 30B), meaning that those genes were less than two-fold up- or down-regulated in AD. Contrarily, many statistically significant DEG ($|logFC| > 1$; $B > 0$) were detected for the homozygous double-mutant mice (Figure 30C).

THY1, the gene whose promoter is used to insert the mutated human genes in the transgenic mice, constantly appears more statistically significant (i.e. with higher B-statistic) than the others (Figure 30). In the mouse, this gene encodes a cell surface glycoprotein expressed in several cell types, including neurons and immune cells, such as T-cells [182,183], whereas in humans it is only expressed in neurons [182,184]. The encoded glycoprotein mediates cell-cell interactions and cell adhesion, promotes T-cell activation and inhibits neurite growth [182,183].

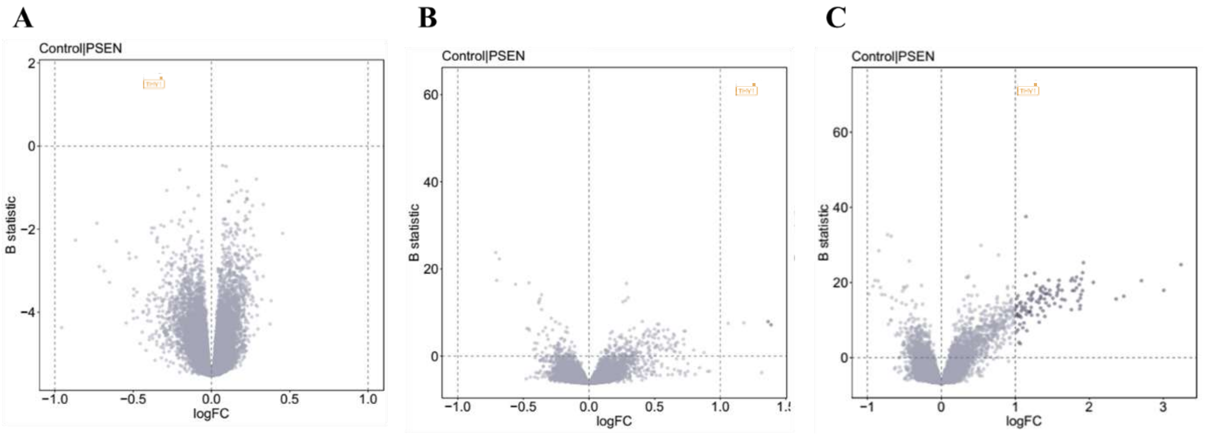


Figure 30 | Volcano plots of differential expression for the mouse datasets

Volcano plots of differential expression derived for *PSEN* and *Age* coefficients of (A) single-mutant, (B) heterozygous double-mutant, and (C) homozygous double-mutant mice. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively. Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed lines) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 1; B > 0).

11.3. Choosing the best mouse dataset for comparison with the human dataset

Under the assumption that mouse models recapitulate human AD, it would be expected the *PSEN* effects in both datasets (human and mouse) to be greatly correlated.

To address the similarities between human and mouse datasets, t-statistic values obtained for the *PSEN* coefficient (respectively using the models in equations 4 and 5) were compared between species, revealing negligible and low correlations of single-mutant and HET mice with human patients (Figure 31A-B). The comparison between the human dataset and the HO mice unveiled a higher correlation than the ones obtained for the other mouse datasets (Figure 31C), despite still being lower than expected for a *bona fide* mouse model.

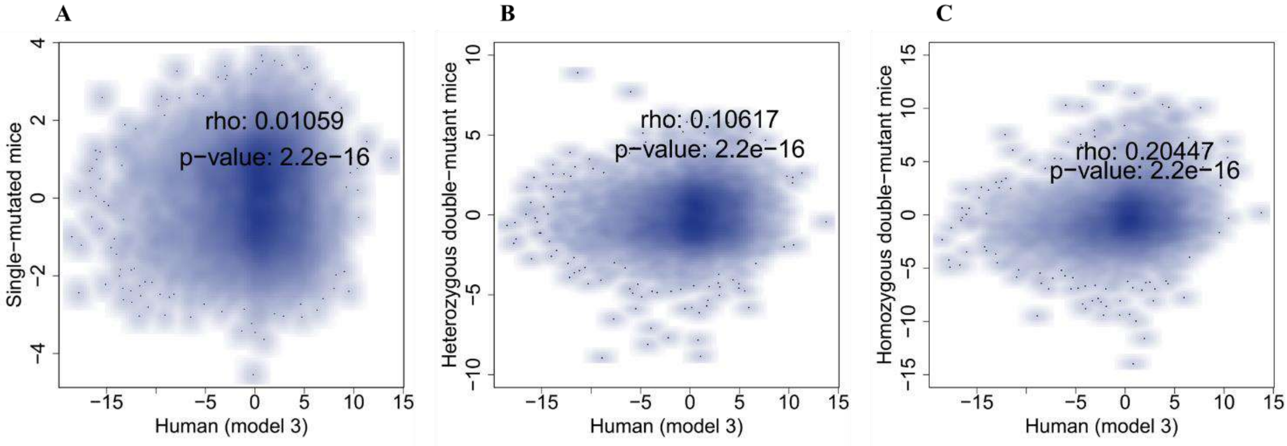


Figure 31 | Human and mouse *PSEN* t-statistic comparison

Comparison of t-statistics of differential expression derived for the *PSEN* coefficient between humans and (A) the single-mutant, (B) the heterozygous double-mutant and (C) the homozygous double-mutant mice. To ease visualization of the correlations, the limits on y-axis do not include *THY1* gene, which has a high t-statistic value on the *PSEN* coefficient (interpreted in section 11.2). Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

The lack of DEG and the low correlations obtained between human and single-mutant and HET mouse models, is concordant with the AD phenotype described by Matarin *et al* [104]. As they reported, single-mutant mice did not develop AD pathology, while a late development was observed for HET mice, with mild Aβ plaque deposition only visible when 8 months old or older (samples were collected at 2, 4, 8 and 18 months of age). The scarcity of AD pathology described by Matarin and co-workers [104], and of DEG, led to the decision of discarding the single-mutant and HET mouse datasets from further analysis.

Contrarily to the other two mouse model, HO mice were described, in the original study [104], as suffering from Aβ plaque deposition by the age of 4 months (strong signal is only seen at 8 months). The fact that these mice develop AD pathology earlier and more intensively compared to the other models, might relate with the observed stronger changes in GE reflected in the PCA pattern, corroborating the need to multi-mutate mouse models in order to have those develop a phenotype that resembles the human disease [70,71,75]. Multi-mutation is a common practice in AD research [70,71,75]. The low correlation between HO mouse and human t-statistics suggests that even multi-mutated mouse models that develop AD hallmarks, are not able to fully recapitulate the human disease [70,71,75].

The present analysis includes mice from all ages, with younger mice (2 and 4 month old) not expressing significant immunohistochemical marking of Aβ in the original study of Matarin and colleagues [104], thus being considered to have milder AD phenotype. Although the used linear model encompasses for age, if only mice with a stronger disease phenotype were included within the AD samples, the GE heterogeneity in those would be lower, considering GE is different between the mice without AD phenotype and those for which plaque deposition is detectable. These molecular variances can be illustrated with the PCA plot in Figure 32, where the samples from 2 month old mice are nearer the control samples compared to the remaining ones. We are left wondering if those phenotypic observations are reflected on GE in a way that is compatible with our models or if there are non-linear changes in GE that introduce noise in our models, so that young mutant mice should not be included. In future analyses, we should test if this would allow for a more robust separation of the AD group from the control group, potentially easing the discrimination of similarities and discrepancies between human patients and mouse models regarding AD-induced GE alterations.

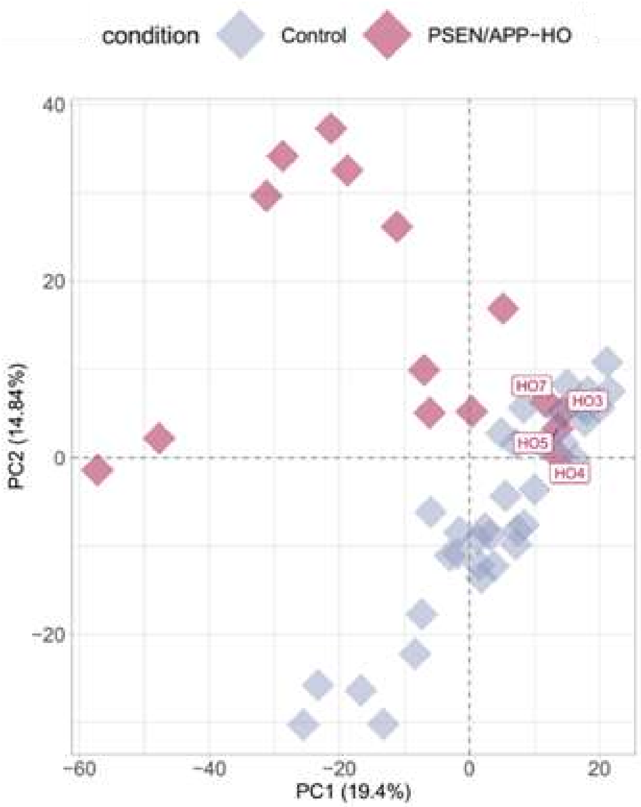


Figure 32 | PCA of the HO mouse dataset – highlighting 2 month old mice

Principal components of normalized GE, with points coloured by condition for the heterozygous double-mutant mice. 2 month old AD mice are identified with labels (HO).

11.4.Exploring AD-induced DEG in the HO mouse model

The linear model defined in equation 5 unveiled a few AD-induced up-regulated genes in the homozygous double-mutant mice, compared to control mice (Figure 33).

As seen before, the most significant up-regulated gene was *THY1*, which was already discussed above (see section 11.2). *LGALS3BP* is another example of an up-regulated gene whose protein controls cellular interactions, adhesion, migration and proliferation, participating in the migration of neuroblasts and differentiation of oligodendrocytes [185,186]. *LGALS3BP* also plays a role in innate immunity, namely by acting as a pro-inflammatory mediator [185,186].

Genes that act upon the immune system comprise the majority of the most up-regulated genes, including *CD59A*, *SLAMF9*, *TLR2*, *CD52*, *CST7* and *CCL3*. *CD59A* acts as an inhibitory agent towards complement system immunity and T-cell activation, thus regulating T-cell responses and protecting host cells from complement immune responses [187,188]. On a similar note, *LSP1* negatively regulates neutrophils adhesion, polarization, and migration [189]. *SLAMF9* plays a role in lymphocytic activation [190] and *TLR2* is vital for antigen recognition and activation of innate immune responses, with murine T cells constitutively expressing *TLR2,* while human T cells do not [191]. Amongst the most up-regulated genes are also *CST7,* which is expressed in microglia during demyelination, a white-matter pathology that has been described in AD in association with oligodendrocytes' function or quantity alterations [192,193]; and *CCL3,* whose coding protein is considered an inflammatory chemokine that promotes monocyte migration to affected regions [194].

Moreover, *CCL4* gene is, in fold-change, the most AD-induced up-regulated gene in the HO mouse model, which constitutes a pro-inflammatory chemokine as *CCL3,* and is produced from and secreted by glial cells and astrocytes in stress conditions, being involved in the migration of leukocytes to the affected region [195]. This prevalence of immune-related disrupted pathways is in accordance with the aforementioned immunological AD hallmarks.

Lastly, *CLEC7A* was also found to be up-regulated. This gene encodes a protein that is thought to participate in axonal regeneration of affected neurons, suggesting it is a defence mechanisms against the consequences of AD pathology, such as neuronal degeneration [196].

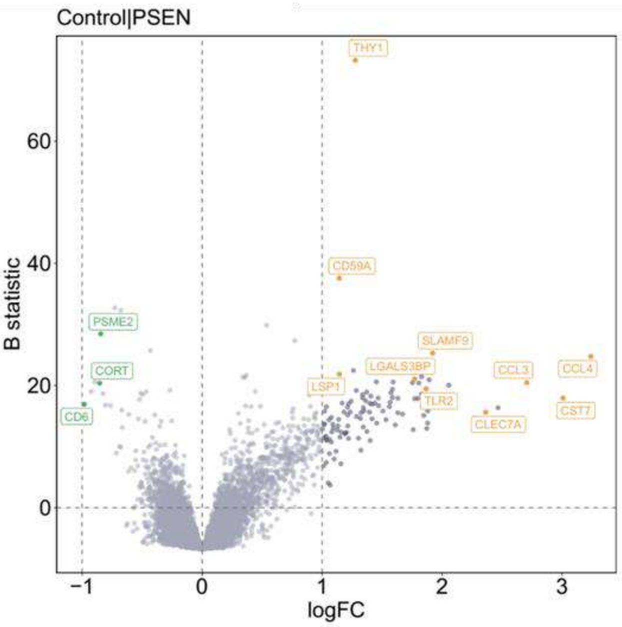


Figure 33 | DEG in mouse model of AD

Differentially expressed, i.e. up- (positive logFC) and down-regulated (negative logFC), gene in HO mouse model, compared with control mice, for the *PSEN* coefficient. Genes in orange are a sample of the most statistically significant differentially expressed genes, while those in green are a sample of down-regulated genes whose drop in expression is near two-fold. Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed line) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 1; B > 0).

Some genes whose drop in expression does not exceed but is near two-fold are *CD6*, *PSME2* and *CORT*. The first two genes are involved in the immune response, with *CD6* acting upon T-cell activation regulation [197]. *PSME2* encodes a subunit of the immunoproteasome, which seems to be induced by an inflammatory state and the presence of chemokines such as INF-γ and IL-1 [198]. The proteasome is vital for the assembling of antigenic peptides shown in MHC class I receptors, increasing the activation of T-cells, as well [198]. Lastly, *CORT* encodes a neuropeptide that mimics somatostatin [199], a growth hormone-inhibiting hormone that acts as a neurotransmitter and neuromodulator in the CNS [200]. Somatostatin interacts with G-protein coupled receptors and affects neurotransmission and cell proliferation, mainly holding an inhibitory function [200].

11.5. Disrupted biological pathways in AD mouse models' brains

Similarly to what was performed on the human data, GSEA was conducted to extract the biological meaning associated with the expression differences obtained for the *PSEN* coefficient through linear modelling. AD-induced disrupted pathways in homozygous double-mutant mice were obtained.

The majority of the unveiled up-regulated KEGG (Figure S20), Reactome (Figure 34) pathways and Biological Processes (Figure S20) in AD conditions relate with the immune system, similarly to the human results. Moreover, some AD-induced human down-regulated

KEGG pathways, such as those related with diabetes mellitus, are shown as up-regulated in mouse AD. These pathways' shift from down-regulated in human to up-regulated in mouse might reflect the important transcriptomic differences between mouse and human AD, therefore suggesting genes to be differently affected by the disease between both species. A more abrupt interpretation would be the mechanisms that underlie disease progression being completely different between human and mouse, justifying the poorly correlated transcriptomic changes and the inefficiency of mouse models in replicating the human disease. The hypothetical caveat of including younger mice that do not manifest evident AD phenotype, and the possibility of more pronounced neuronal death in humans compared to mice, can also play a role in the surge of these discrepancies.

Another group of up-regulated Reactome pathways was related with metabolic disruptions and cholesterol biosynthesis, with high levels of cholesterol having previously been linked to AD development [201]. For the human species, the brain is the organ with the highest levels of cholesterol, being crucial for maintenance of neuronal plasticity and function [201]. The brain-consumed cholesterol is synthetized within the CNS, majorly by glial cells, with special focus on astrocytes [201]. In a healthy state, the blood-brain barrier restricts the efflux of cerebral cholesterol into the peripheral circulatory system and prevents the influx of circulatory cholesterol into the brain, except for two oxidized forms (24S-hydroxycholesterol and 27-hydroxycholesterol) that are able to cross the blood-brain barrier [201]. In a disease state, where the blood-brain barrier is compromised, there is an influx of cholesterol-carrying lipoproteins into the brain that therefore disturbs the cholesterol levels within CNS [201]. Moreover, high levels of free cholesterol in neurons were found to favour the activity of β- and γ-secretases, thus increasing Aβ production [201]. In AD, the inflammatory state that affects the permeability of the blood-brain barrier [202] might lead to increased brain cholesterol levels, and therefore to higher production rates of Aβ.

An up-regulation of the Lysosome KEGG pathway is also observed (Figure S20). Deficits in lysosome axonal transport have been described in AD-diseased neurons, where lysosomes end up accumulating at amyloid plaque deposits, promoting disease progression [203].

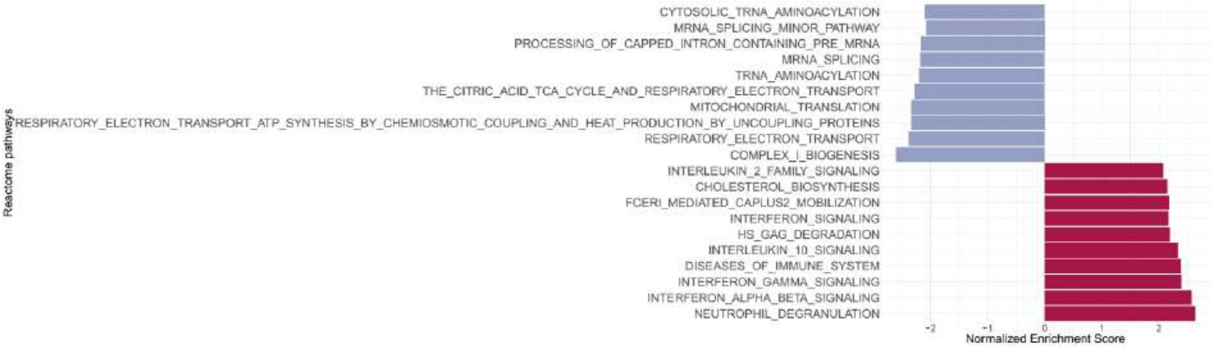


Figure 34 | Reactome pathways differentially expressed for the *PSEN* coefficient (Mouse)

Representation of the 10 most down- (blue) and up-regulated (magenta) Reactome pathways in mouse *PSEN* brains (homozygous combination dataset). Pathways were considered significantly enriched if FDR < 0.05.

Down-regulated Biological Processes (Figure S20), Reactome (Figure 34) and KEGG (Figure S20) pathways show an impairment at the level of RNA splicing. A recent study has unveiled Tau-mediated dysregulation of several spliceosome components and even loss of function of some vital proteins, such as the small nuclear ribonucleoprotein-associated protein B (SmB) [204]. Even though disrupted Tau is not detected in the mouse models used in the present study, the cognate gene of the SmB protein, *SNRPB*, shows decreased expression in AD condition for both the mouse and human datasets (Figure 35), suggesting that a different SmB-disruption mechanim is activated, at least in the mouse model. Some transfer RNA(tRNA)-related pathways are also down-regulated in Reactome, namely tRNA aminoacylation, which may relate with dysregulation of the spliceosome given that aminoacyl-tRNA synthetases also play a role in RNA splicing [205]. Alternative splicing is crucial for neuronal diversity and function, with disruptions of the spliceosome machinery being related to neurologic diseases, namely dementia [204].

Results also show the down-regulation of pathways associated with cellular respiration and mitochondrial processes. These alterations in the mitochondrial machinery might mirror its functional shift or loss, which might explain the aforementioned down-regulation of mechanisms related with GE regulation, as these are highly demanding in terms of energy requirements [206]. Moreover, mitochondrial impairment have been described to affect GE, alternative splicing and translational processes [206]. Prior to Aβ outburst and plaque formation, Aβ oligomers are found to accumulate in neuronal mitochondria in AD conditions, thus leading to the disability of respiratory functions that are crucial for cell maintenance [207]. Considering that neurons require a great amount of energy in order to properly function, dysfunction of the mitochondrial system has an immense effect in neuronal survival and AD development and progression [207]. Lastly, and in accordance to what was observed for the human data, KEGG pathways related with neurodegenerative diseases are also downregulated.

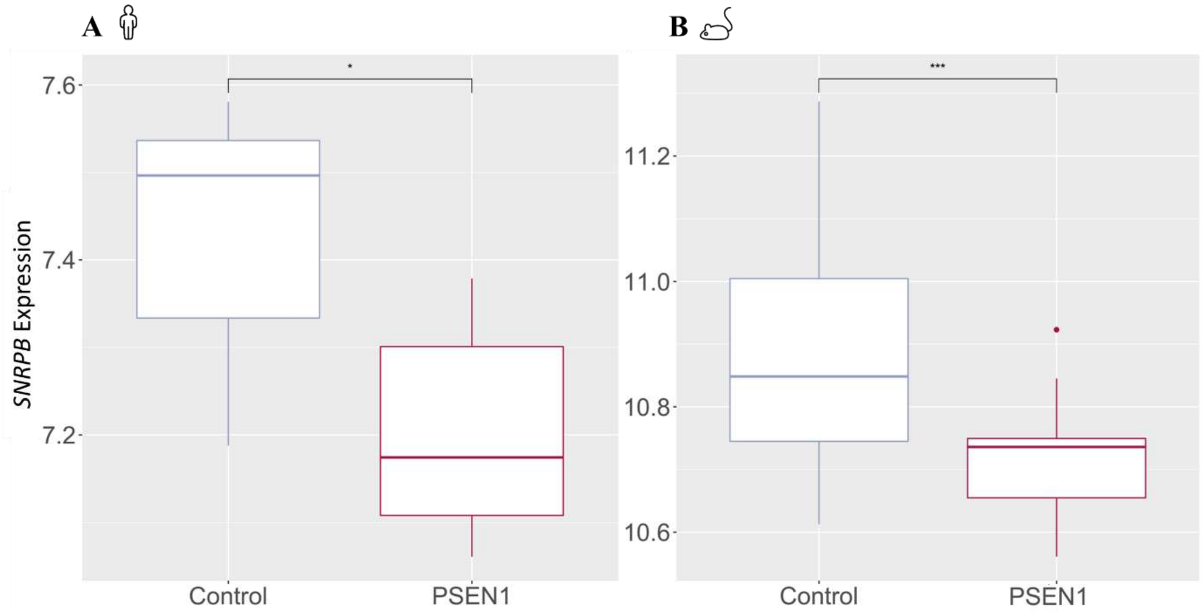


Figure 35 | Gene expression distribution for the *SNRPB* gene

SNRPB GE distribution across controls and PSEN1-mutated samples in the (A) human and (B) mouse datasets. T-test comparisons were conducted with a considered significance of $\alpha = 0.05$ ($* p < 0.05$,$* * p < 0.01$,$*** p < 0.001$, ns stands for non-significant).

12.Summary of findings for the human and mouse independent datasets

Our results show a prominence of AD-induced up-regulation of immune system mechanisms in both human and mouse datasets. Additionally, processes related with the cardiovascular system, cellular differentiation, proliferation and apoptosis, and nucleic acid processing were up-regulated in the human dataset. As for the mouse dataset, besides the immune system, cellular interactions, diabetes and cholesterol related pathways are up-regulated.

It is noteworthy that diabetes mellitus associated genes are down-regulated in human AD conditions, which is discrepant from mouse results. Overall, neural-related mechanisms, i.e. neuronal and synaptic activity, neurotransmitter trafficking and ionic channels, are also down-regulated in human AD; whereas spliceosome machinery and cellular respiration and mitochondrial processes appear down-regulated in the mouse dataset. A few genes, whose drop in expression does not exceed but is near two-fold, are associated with the immune system and neuronal mechanisms. Disrupted mechanisms are summarized in Table 6.

These discrepancies between the mouse and the human datasets suggest that different genes are being affected in AD or, more drastically, that the disease is developing through different mechanisms between species. This premise is reinforced by the lack of correlation between the human and mouse AD-associated transcriptomic changes. These discrepancies could be exacerbated by the inclusion of younger AD mice amongst the PSEN mouse samples, increasing heterogeneity in the disease development GE signature through the approximation of those samples to mouse controls and divergence from human AD patients.

Table 6 | Up- and down-regulated mechanisms in the separate human and mouse datasets

Datasets	Human	Mouse
Up-regulated mechanisms	Immune system Cardiovascular system Cellular differentiation/proliferation Apoptosis Nucleic acid processing	Immune system Cholesterol Diabetes mellitus Cell interactions
Down-regulated mechanisms	Ionic channels Neutransmitter trafficking Neuronal and synaptic activity Diabetes mellitus	Spliceosome/Gene expression Cellular respiration and mitochondrial processes Neutransmitter trafficking

However, the *PSEN* coefficients estimated by our linear modelling encompass, for each dataset, information on the mechanisms respectively underlying the human and mouse disease. To decouple the common and species-specific effects, the analysis described so far was repeated on the joint human and mouse dataset. Its results are explained in the following section.

13. Joint dataset

13.1. Choosing the explanatory variables for linear modelling

In order to obtain differential GE signatures that could be specifically linked to either mouse or human AD, as well as those common to both species, human and mouse data were merged into a joint dataset. PCA of the resulting normalized GE was used to unveil some patterns in the data, where points were respectively colour- or shape-distinguished by species and condition (Figure 36). Unsurprisingly, the plots unveiled species as the feature that explained most of the variance within the data (87.7%). This result was, to some extent, anticipated by the lack of correlation in the transcriptomic changes introduced by AD between human and mouse, seen in Figure 31C.

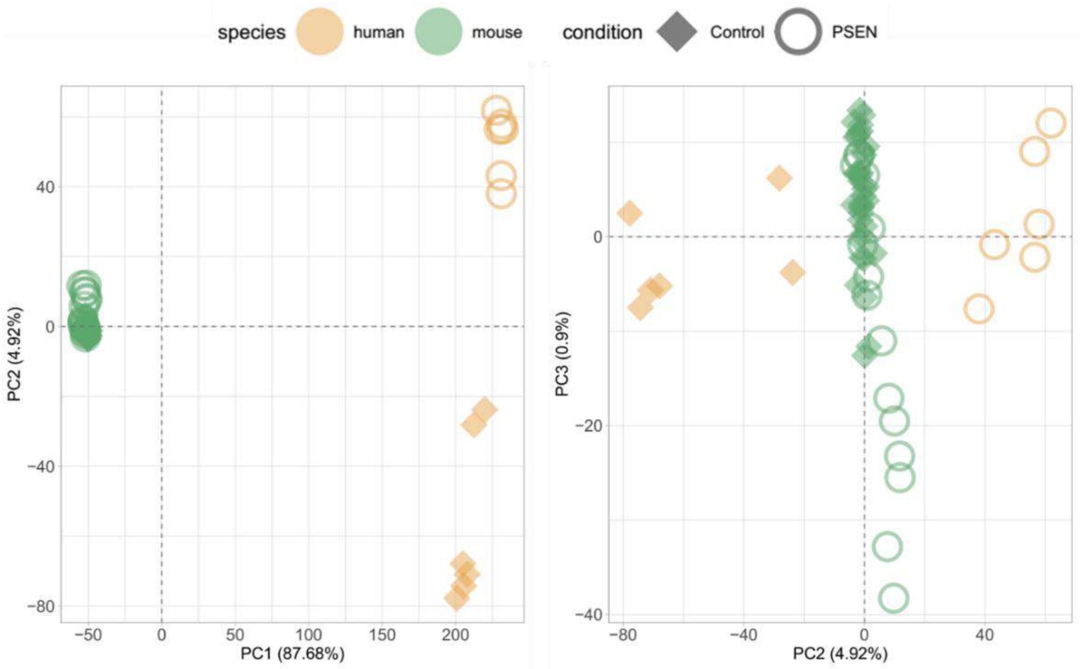


Figure 36 | Principal Component Analysis (Joint dataset)

Principal components of normalized GE data, with points shaped by AD condition and coloured by species.

Moreover, PC2 can decouple non-AD from AD human samples, as well as non-AD from AD mouse samples, albeit to a lesser extent. A separation of mouse controls and ADs is only obtained in PC3, which represents only 0.9% of the variance within the data. These results reinforce that differences between AD and non-AD samples are much subtler in mouse models than in humans, thus corroborating the premise of mouse models not effectively replicating the human disease, with pathology development not being strong enough to exhibit relevant molecular differences regarding non-diseased samples. As already discussed, the modest differences found between control and AD mice could also be accentuated by the heterogeneity within PSEN mouse samples, among which 2 and 4 month old mice, hardly showing any AD phenotype, are incorporated.

To fully decouple species-specific AD signatures from the ones common to the disease development in both species, GE of the joint dataset was modelled based not only on the *Species* and *PSEN*, but also on the interaction of those variables (equation 6). The *Interaction* coefficient allows to estimate species-specific AD transcriptomic changes.

$$GE = \beta_0 + \beta_{Species}Species + \beta_{PSEN}PSEN + \beta_{Interaction}Species \cdot PSEN \qquad (6)$$

13.2. Decoupling species-specific and species-common AD-induced GE changes

Given that an *Interaction* coefficient was added to the model, different data centring approaches were considered (see section 8.7.1). To choose which explanatory variables to centre, three linear models were compared, where (A) *Species* and *PSEN* were not centred, (B) only *Species* was centred, and (C) both predictors were centred. In all three models, *Species* and *PSEN* are independent from each other, as well as *Species* and *Interaction*, with the fully-centred model displaying the lowest correlation between *Species* and *Interaction* (Figure 37). *PSEN* is highly correlated with *Interaction* in the non-centred model, and less so in both centred models (Figure 37).

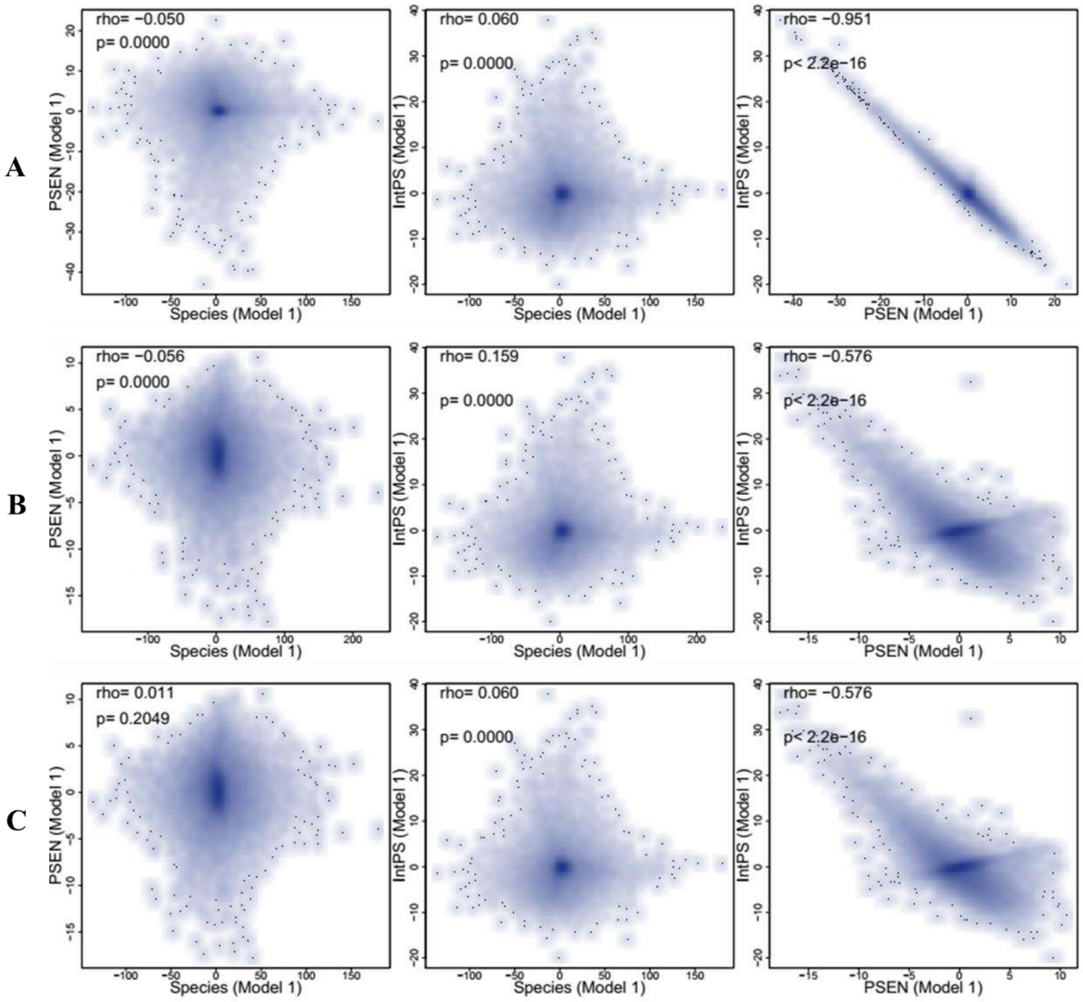


Figure 37 | t-statistic comparison between explanatory variables of the same model

Comparison of t-statistics of differential expression between the different explanatory variables of the linear model (*Species*, *PSEN* and *Interaction*) for (A) the non-centred model, (B) the *Species*-centred model and (C) the model with both *Species* and *PSEN* centred. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

When comparing the same variable between models that include an *Interaction* effect but have different centring approaches, it is expected the *Interaction* estimates to be constant irrespectively of the centring, whereas the remaining coefficients can vary with it when they are not the differently centred variable [138] (i.e. for instance, *Species* estimates are not affected by centring the *Species* variable), as it is seen in Figure 38.

In our data, the *Species* estimates correlate almost perfectly between the models where it is prone to change. The *PSEN* estimates vary equally between the *Species*-centred and fully-centred models compared with the non-centred one (Figure 38), which is expected since the coefficient is equally affected by both centring approaches.

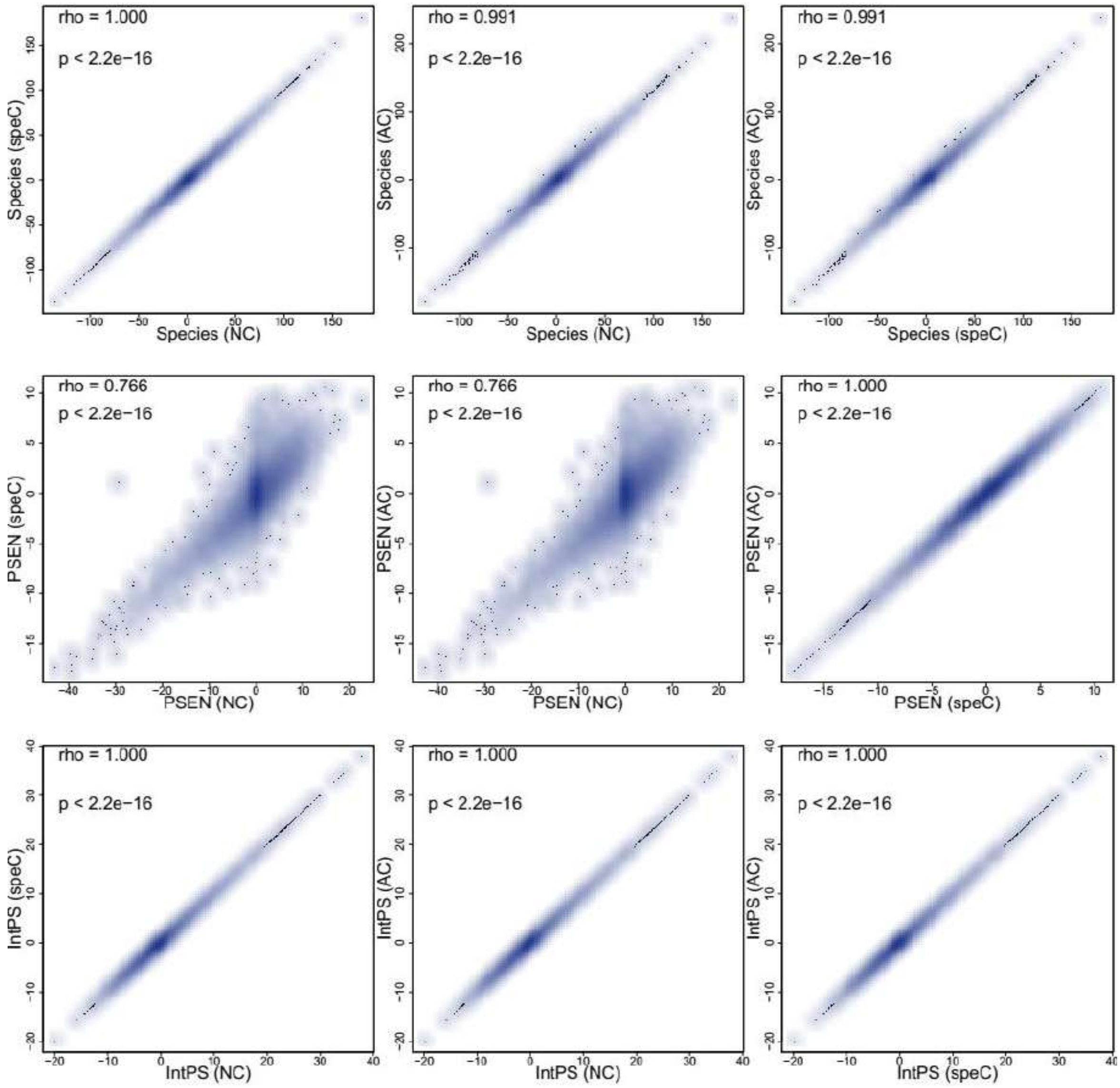


Figure 38 | t-statistic comparison of the same variable between centring-differing models

Comparison between t-statistics of differential expression retrieved with linear modelling across centring-differing models for the variables *Species* (top), *PSEN* (middle) and *Interaction* (bottom). Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively. NC stands for non-centred, speC for *Species*-centred and AC for all centred variables (*Species* and *PSEN*).

Recalling the explanation in section 8.7.1, the *Species*-centred model turns the baseline from a human individual (mouse is defined as the "positive" species) into a conceptual human/mouse hybrid. In a model with *Interaction* where both *Species* and *PSEN* are centred (referred to as "fully-centred" model), the baseline turns into a half-diseased hybrid, which is biologically less meaningful than the *Species*-centred baseline.

To favour the interpretability of the model, the *Species*-centred model was chosen for further analysis, also considering that the *Species-Interaction* correlation is not very high and the *PSEN* is as uncorrelated with the *Interaction* in that model as they are in the fully-centred one, since the *PSEN* estimates do not change between them.

13.3. Exploring AD-induced DEG in the joint dataset

13.3.1. AD-induced DEG common to mouse models and human patients

Using the linear model displayed in equation 6, differential expression regarding the *PSEN* coefficient unveils the AD-induced GE changes that occur in both human patients and mouse models – which will be referred to as "common AD changes" –, with up- and down-regulated genes respectively having positive and negative logFC values (Figure 39).

Common AD-induced up-regulated genes are majorly linked to the immune system, including genes such as *SLAMF9*, which regulates lymphocytic activity [190], and *CD52* that encodes surface proteins [208]. Other up-regulated immune genes found for the *PSEN* coefficient were *CD86* (whose expression enhances T cells' activation [209]), *GFAP* (encodes a protein highly expressed by reactive astrocytes in response to brain inflammation and injury, which suggests that both species have an increase of reactive astrocytes [210]) and *CST2* (encodes a protein abundant in immune cells, especially lymphocytes [211]). As seen for the mouse dataset, *LGALS3BP* is up-regulated in common AD, being involved in cellular interactions and cell cycle, namely of neuroblasts and oligodendrocytes, as well as in pro-inflammatory immune responses [185,186].

Genes *RPH3A*, *GABRA1*, *GABRG2* and *SULT4A1* appeared down-regulated for common AD. As explained above, *GABRA1* acts as a GABA receptor, the main inhibitory neurotransmitter in mammals, and plays a role in synaptic inhibition [167]. Additionally, the *GABRG2* gene, that plays a role in neuronal development and formation of inhibitory GABAergic synapses [167], also appeared down-regulated. GABAergic synapses are the main inhibitory synapses, with GABAergic neurons comprising 10-20% of the cortex brain region [212]. Lastly, *SULT4A1* gene encodes a conjugation enzyme entailed in the metabolization of hormones, drugs and neurotransmitters [213].

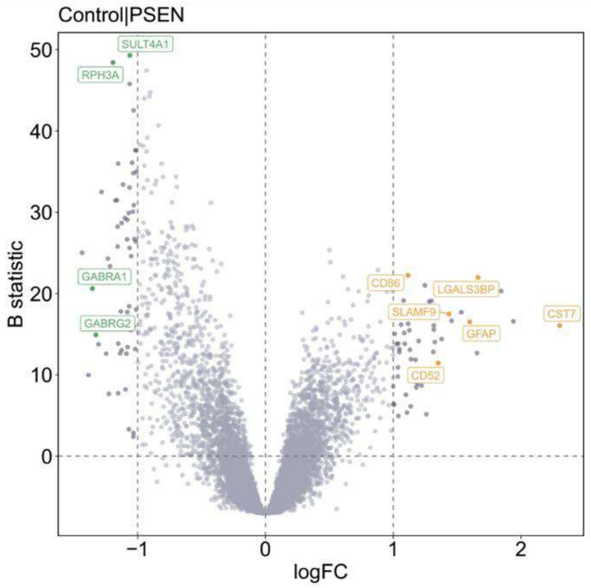


Figure 39 | AD-induced DEG that are common to mouse and human (Joint dataset)

Differentially expressed, i.e. up- (positive logFC) and down-regulated (negative logFC), genes in AD human and mice, for the *PSEN* coefficient. Genes in orange and green are a sample of the most differentially expressed genes, with green-coloured genes found DEG for the *Interaction* coefficient as well (section 13.3.2). Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed line) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 1; B > 0).

13.3.2. Species-specific AD-induced DEG for mouse models and human patients

The *Interaction* coefficient unveils species-specific AD-induced GE changes that go beyond the “common” changes, and that will be referred to as “species-specific AD changes”. Positive logFC values unveil genes that are more up-regulated or less down-regulated with the disease in mouse models compared to human patients (Figure 40A-C). To simplify, these genes will be referred to as “mouse-specific genes”.

On the other hand, negative logFC values unveil genes that are more up-regulated or less down-regulated with the disease in human patients compared to mouse models (Figure 40D-F). As before, these genes will be referred to as “human-specific genes”. The differentially expressed genes are displayed in Figure 41.

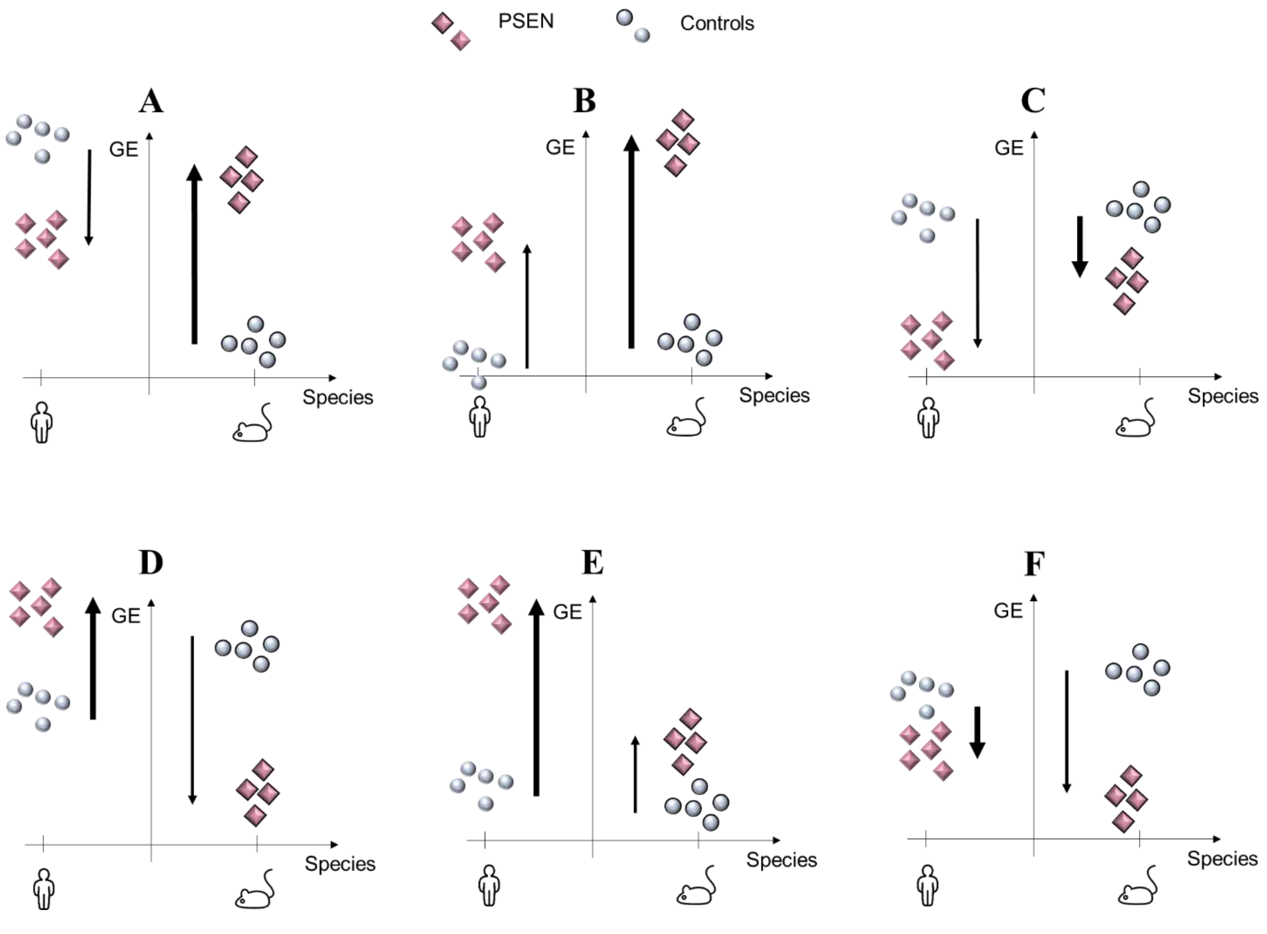
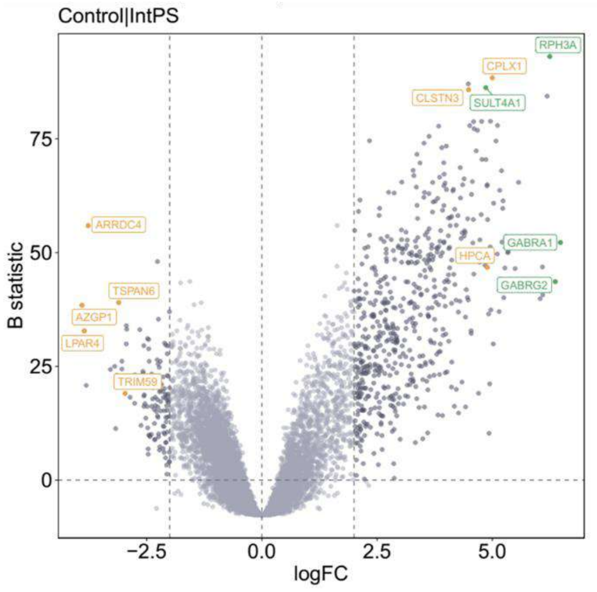


Figure 40 | Interpretation of the *Interaction* coefficient based on different GE profiles

Schematic images that represent several GE profiles that might appear in the *Interaction* coefficient with positive (A, B and C) and negative (D, E and F) logFC values. Up-sided arrows represent up-regulation and down-sided ones represent down-regulation. Thicker arrows signal the differential expression where the gene is most up-regulated or less down-regulated.

Figure 41 | AD-induced DEG that are species-specific – *PSEN*/*Species* interaction (Joint dataset)

Mouse-specific (positive logFC) and human-specific (negative logFC) AD-induced DEG in the joint dataset (*PSEN/Species Interaction* coefficient). Genes in orange and green are a sample of the most differentially expressed and/or statistically significant genes, with green-coloured genes having been found DE for the *PSEN* coefficient as well (section 13.3.1). Thresholds of magnitude (vertical dashed lines) and significance (horizontal dashed line) for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 2; B > 0).



Interestingly, genes commonly down-regulated in AD for both species (*RPH3A*, *SULT4A1*, *GABRA1* and *GABRG2*) appeared, in the *Interaction* coefficient, as mouse-specific, i.e. more up-regulated or less down-regulated in mouse than in human. Moreover, additional neural-related genes also appeared as mouse-specific, such as genes *CPLX1* (participates in synaptic vesicle exocytosis and neurotransmitter release [214]), *CLSTN3* (localizes in the postsynaptic membrane and assists in the presynaptic development and differentiation of inhibitory and excitatory synapses [215]) and *HPCA* (regulates calcium intracellular homeostasis, with disturbances in this equilibrium having been linked to the development of AD pathology [216,217]). This observation does not directly translate in an inconsistent disturbance of these mechanisms in AD mouse models, but rather suggests that these processes are likely more prominently down-regulated in human patients than in the murine models.

Consistently to the previous results for the human dataset, genes *ARRDC4* (that participates in the internalization of activated G-protein coupled receptors [154], as well as in the inflammatory response associated with the innate immune system [155]) and *LPAR4* (associated with cell differentiation, proliferation and migration, and described in neuronal and glial alteration in disease conditions [157]) also appeared as human-specific in the *Interaction* coefficient. Other human-specific genes include *TSPN6* (that encodes a transmembrane protein belonging to a family of proteins that have an immunoregulatory role and mediate cell interactions, development and migration [218]), *TRIM59* (is involved in cellular processes, similarly to *TPSN6*, and has been found overexpressed in several tumours [219]), and *AZGP1* (important for activation of immune responses, whose lower expression is associated with a poorer prognostic in oncologic patients [220]).

Considering that the immune-related pathways have been described as up-regulated in the individual human and mouse datasets, those observations suggest that such mechanisms are more activated in human patients than in mouse models. This kind of discrepancies might be behind the differences in disease development that distinguish the human disease from that induced in mouse models. Such differences can be induced by variances in the disease-induced immune responses between both species, which have been being identified throughout the years [221,222], namely relating to the innate neuroimmune system and neuroinflammation [221].

13.4. AD-induced disrupted biological pathways in the joint dataset

13.4.1. AD-induced alterations common to mouse models and human patients

To capture the biological alterations associated with the AD-induced GE changes common to or specific for human patients and mouse models, GSEA was respectively performed on the results obtained for the *PSEN* and *Interaction* coefficients through linear modelling. AD-induced disrupted gene-function pathways, concomitantly altered with the aforementioned expression differences, were obtained.

Similarly to both the human and mouse datasets, the most up-regulated Reactome, KEGG and BP pathways (in the joint dataset for the *PSEN* coefficient), associate with the immune system and elastic fibres' synthesis (Figure 42 and Figure S21). The most down-regulated pathways are linked with synaptic and neuronal activity, neurotransmitter trafficking and ion channels, resembling the human dataset, as well as with mitochondrial activity and cellular respiratory processes, like in the mouse dataset (Figure 42 and Figure S21).

KEGG pathways related with neuronal disease such as AD, Parkinson's or Huntington's keep appearing as down-regulated (Figure S21).

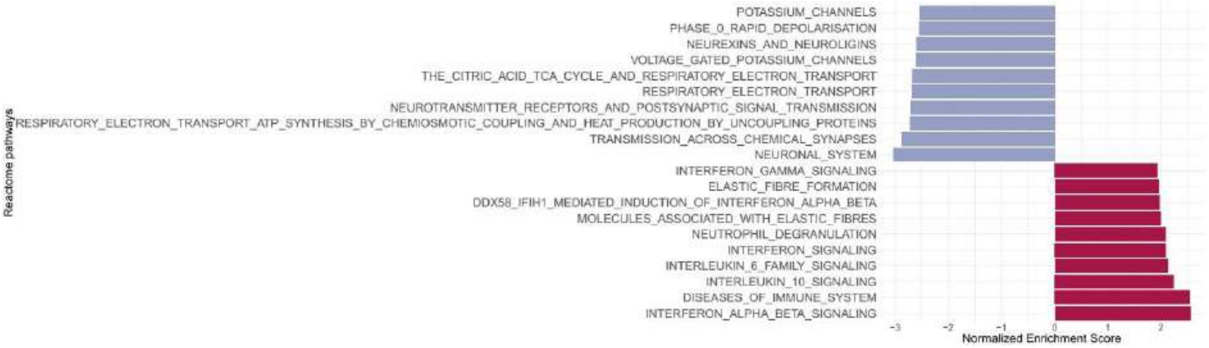


Figure 42 | Reactome pathways differentially expressed for the *PSEN* coefficient (Joint dataset)

Representation of the most down- (blue) and up-regulated (magenta) Reactome pathways commonly AD-induced in human patients and mouse models, based on DEG regarding the *PSEN* coefficient. Pathways were considered significantly enriched if FDR < 0.05.

13.4.2. Species-specific AD-induced alterations for mouse models and human patients

GSEA performed on the *Interaction* coefficient unveiled Reactome (Figure 43), KEGG and BP (Figure S22) pathways linked to the immune system and GE regulation as being human-specific, i.e. those down-regulated in the *Interaction* coefficient. Additionally, sumoylation of proteins is unveiled as a more enriched pathway in the human disease compared to mouse, corresponding to a post-translational process carried out by small ubiquitin-like modifier (SUMO) proteins 1, 2 and 3, that has been implicated in the pathophysiological development of neurodegenerative diseases, including AD [223]. Increased levels of SUMO1-modified proteins have been reported in brains of human AD patients and mouse models, with these proteins being known to target both APP and Tau proteins [223]. Despite the absence of significant differences between species in the expression alterations of the *SUMO1* gene induced by *PSEN1* mutations (Figure 44), it is worth noting that GE, i.e. mRNA levels, do not always have a direct correlation with protein levels or activity.

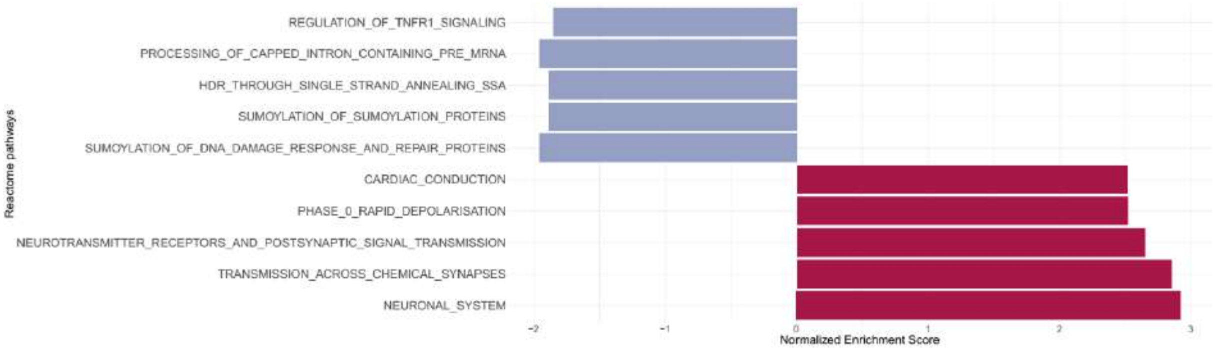


Figure 43 | Reactome pathways differentially expressed for the *Interaction* coefficient (Joint dataset)

Representation of the most up-regulated Reactome pathways specific for the human AD patients (blue) and for the AD mouse models (magenta), in addiction to an overall AD-induced disruption, based on DEG regarding the *Interaction* coefficient. Pathways were considered significantly enriched if FDR < 0.05.

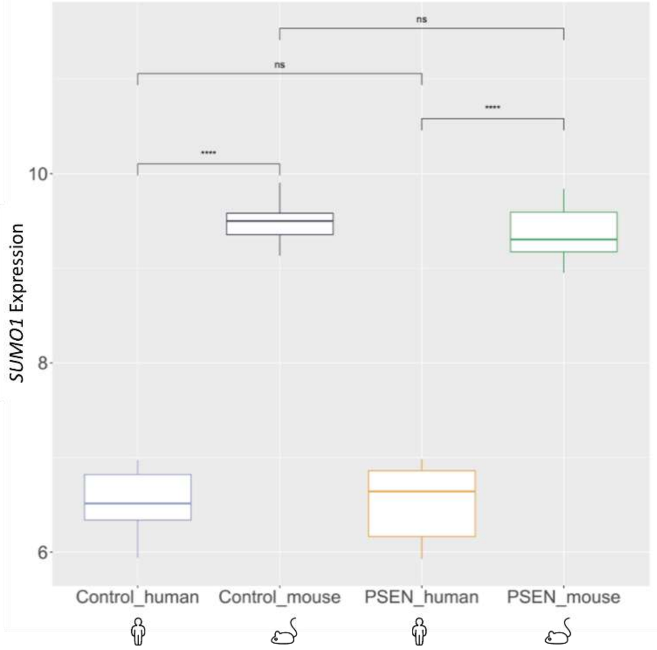


Figure 44 | Gene expression distribution for the *SUMO1* gene

SUMO1 expression distribution across controls and *PSEN1*-mutated samples, separated by species, in the joint dataset. T-test comparisons were conducted ($* p < 0.05, ** p < 0.01, *** p < 0.001$, ns stands for non-significant, for an $\alpha = 0.05$).

Pathways that appeared as more disrupted in AD mouse models, i.e. the up-regulated Reactome (Figure 43), KEGG and BP pathways (Figure S22) for the *Interaction* coefficient, majorly associate with signal transmission through neurotransmitters and synaptic functions. Moreover, pathways related with neurodegenerative diseases that appeared down-regulated in both isolated datasets of human and mouse, are shown as more up-regulated for the mouse models, suggesting they might be less affected in those than in human patients.

As explained for the DE mouse-specific genes, the relative up-regulation of neural pathways in mice does not directly translate in an inconsistent disturbance of these mechanisms in the murine AD models, but rather suggests that these processes are more prominently down-regulated in human patients than in the murine models, thus appearing as up-regulated for the latter. This explanation is aligned with mouse models showing mild AD pathology in the original study [104] that do not truthfully translate the human disease.

13.5. Interaction coefficient vs PSEN from the independent datasets

When comparing t-statistic values of differential expression obtained for the *Interaction* coefficient in the joint dataset and the *PSEN* coefficients in the independent human and mouse datasets, a high anti-correlation with the human *PSEN*, and a lower correlation with the mouse *PSEN*, are observed (Figure 45). This suggests that the *Interaction* coefficient of the joint dataset mainly translates the AD-induced human transcriptomic alterations.

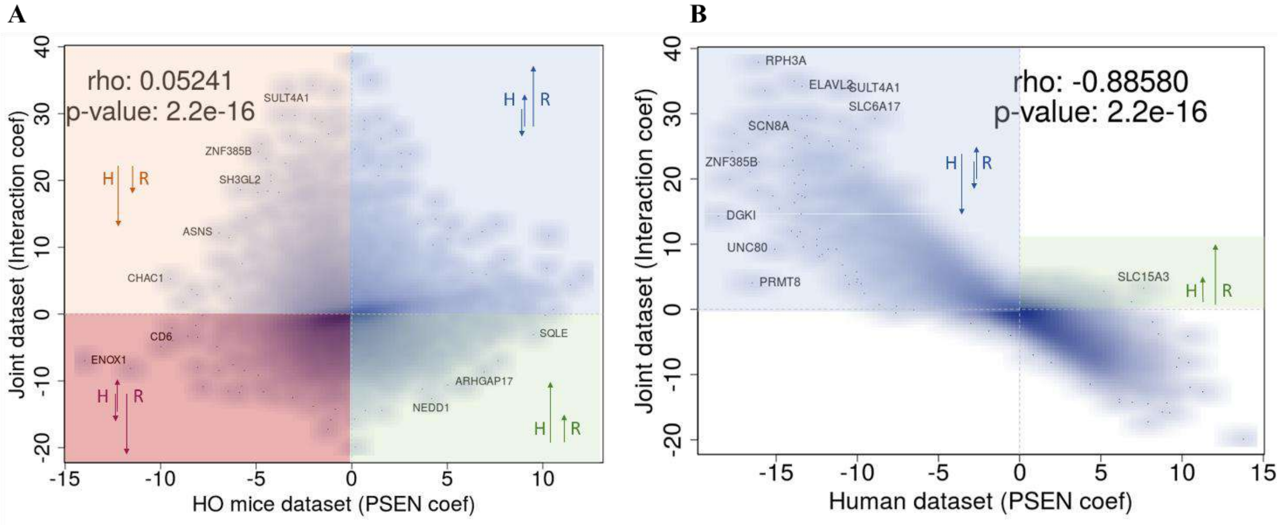


Figure 45 | *Interaction* from the joint dataset vs *PSEN* coefficients from the independent datasets

Comparison between the t-statistics of differential expression for the *Interaction* coefficient of the joint dataset and *PSEN* coefficients of (A) the mouse and (B) the human datasets. The colourful quadrants represent the possibilities of gene regulation associated with the outlier genes, i.e. genes that deviate from the correlation, while an up-sided arrow represents up-regulation and a down-sided arrow represents down-regulation (H for human and R for rodent/mouse). If two arrows are on the same direction, their relative size specifies in which dataset the gene is more altered. In the case of arrows pointing to different direction, that rule does not always apply, but the relevance lies on the opposite effects between the two species. Rho and *p* – Spearman's rank correlation coefficient and p-value, respectively.

In the mouse/joint comparison (Figure 45A), it is possible to find genes that are down- (orange quadrant) and up-regulated (green quadrant) in both human and mouse, but always more altered in the human patients than in the mouse models. This comparison also informs on genes that are more down-regulated in mice, being possible to be either up-regulated or less down-regulated in human patients (red quadrant). It also encompasses the opposite situation, where genes are more up-regulated in mice (blue quadrant), as already described. Genes in both the blue and red quadrants are therefore differently altered in AD conditions between mouse models and human patients and could be potential candidates to manipulate in mice to generate a novel and more accurate murine model for the human disease.

Genes in the orange quadrant include *SULT4A1*, *SH3GL2*, *ZNF385B*, *ASNS* and *CHAC1*. Both *SULT4A1* and *SH3GL2* are involved in synaptic and neurotransmitter activity [213,224], with *SH3GL2* acting upon neurotransmitter release and clearance from the synaptic cleft and having been found enriched in synaptic terminals [224]. As for *ZNF385B*, it encodes a protein that induces B-cell apoptosis [166], and its down-regulation can be a feedback mechanism to allow the action of the immune system in AD conditions (see section 10.2). The remaining genes are involved in brain and neuronal development and survival. *ASNS* encodes a protein vital for the making of asparagine synthetase, an enzyme which represents the main source of asparagine within the brain, given that asparagine does not physiologically accumulate in the brain when synthetized in other regions, nor it transposes brain barriers such as the blood-brain-barrier [225]. Asparagine is an important amino acid for brain development and neuronal myelination [225].

CHAC1 gene is up-regulated in stress conditions in an ATF4-dependent manner, hence indirectly informs on *ATF4* enrichment [226]. *ATF4* encodes a transcription factor that regulates neuronal survival and death through protein synthesis in the endoplasmic reticulum, and which is usually enriched in stress conditions [226,227]. Overexpression of *ATF4* results in ATP depletion, oxidative stress and cell death [226,227]. The cell death associated with AD can trigger a feedback mechanism that results in the down-regulation of ATF4 in order to decrease apoptosis rates, thus explaining the down-regulation of *CHAC1*.

Genes that are up-regulated in both human and mouse, but more in human, include *SQLE*, *ARHGAP17* and *NEDD1*. *SQLE* encodes an enzyme that participates in cholesterol biosynthesis, a condition that, as aforementioned, has been linked with AD pathology [228]. *ARHGAP17* translates into a Rho GTPase activating protein that activates the RhoA, Rac1 and Cdc42 proteins [229], which regulate actin configuration within the pre-synaptic terminal and stimulate Ca^{2+}-dependent release of neurotransmitters [230]. Lastly, *NEDD1* is part of protein complexes required for centrosomal microtubule nucleation [231].

Genes *ENOX1*, which positively regulates angiogenesis [232], and *CD6* appear in the red quadrant. *CD6* acts upon T-cell activation regulation [197] and is the most AD-induced down-regulated (in terms of logFC, i.e. magnitude of differential expression) gene in the mouse dataset (see section 11.4).

In the human/joint comparison (Figure 45B), we are interested in genes with high t-statistics for *Interaction* in the joint dataset and t-statistics close to zero for *PSEN* in the human dataset, which would comprise genes not DE in human but with alterations specific of the mouse dataset. Given the absence of such genes (somewhat expected, given that the AD-associated variance in the joint data is dominated by the human disease), we focused on genes down-regulated for *PSEN* in the human dataset but up-regulated for *Interaction* in the joint dataset, i.e. genes more up-regulated in mouse AD compared with human patients, and therefore contrarily regulated between the two species, or genes less down-regulated in mice compared to humans (blue quadrant in Figure 45B). *SLC15A3,* was identified in the green quadrant as up-regulated in both the human and joint datasets, suggesting it is enriched in both human and mouse datasets, however more in mouse models. Increased expression of the solute-carrier encoding gene *SLC15A3,* is associated with inflammatory immune responses [233].

As for the most DE genes in the blue quadrant in Figure 45B, two of them (*SULT4A1* and *ZNF385B*, described above) are down-regulated in AD in both species but less in the mouse model (orange quadrant of Figure 45A). *RPH3A*, as described above, is involved in synaptic vesicle cycle [166].

SLC6A17, *UNC80*, *SCN8A*, *ELAVL2*, *PRMT8* are also genes linked to synaptic and neurotransmitter activity, with *SLC6A17* participating in the pre-synaptic re-uptake of neurotransmitters after signal transmission [234], and *UNC80* and *SCN8A* playing important roles in the functioning of sodium channels, thus affecting electrical signal transmissions and synaptic activity [235,236]. Moreover, *ELAVL2* is associated with the transcription and splicing of RNA in human neurons, and participates in synaptic activity as well [237]; while *PRMT8* is important for function and plasticity of excitatory synapses as its dysfunction leads to altered ene and altered levels of multiple synaptic proteins [238]. Also on synaptic plasticity, *DGKI* encoded protein belongs to a family of kinases that regulate diacylglycerol levels, a lipid synthetized in synapses during signal transmission which seems to be involved in the recycling of presynaptic vesicles at excitatory synapses [239].

13.6. Unveiling compounds that replicate human AD transcriptomic changes (*cTRAP*)

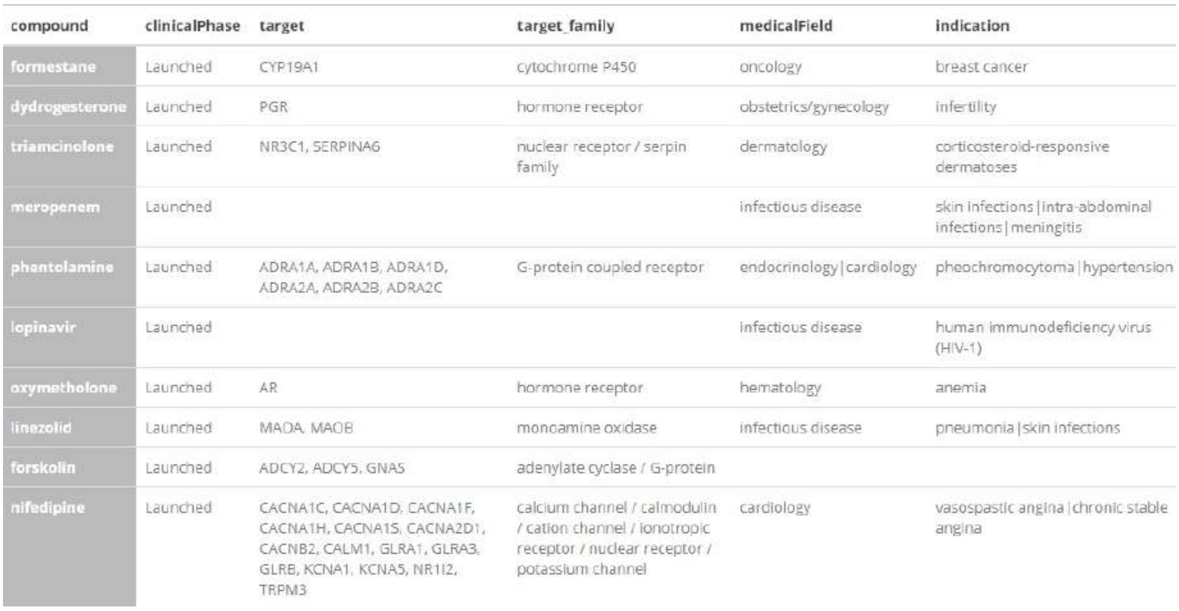
cTRAP was used in order to unveil compound perturbations with putative ability to induce, in a novel mouse model, GE changes recapitulative of those associated with human AD. The goal is then to find compounds that recapitulate opposite transcriptomic changes to those seen for the *Interaction* coefficient (Table 7), i.e. compounds that promote the up-regulation of the genes that, according to our model, are more disrupted in the human patients than in the mouse models (that correspond to the down-regulated genes in the *Interaction* coefficient).

Moreover, compound perturbations with the ability to induce GE changes opposite to the ones obtained for the disease common to both human and mouse models were also assessed, i.e. compounds that recapitulate opposite transcriptomic changes to the ones of the *PSEN* coefficient, with the intent to discover new potential therapies to combat the disease in human patients and that could be easily tested in mouse models (Table 8).

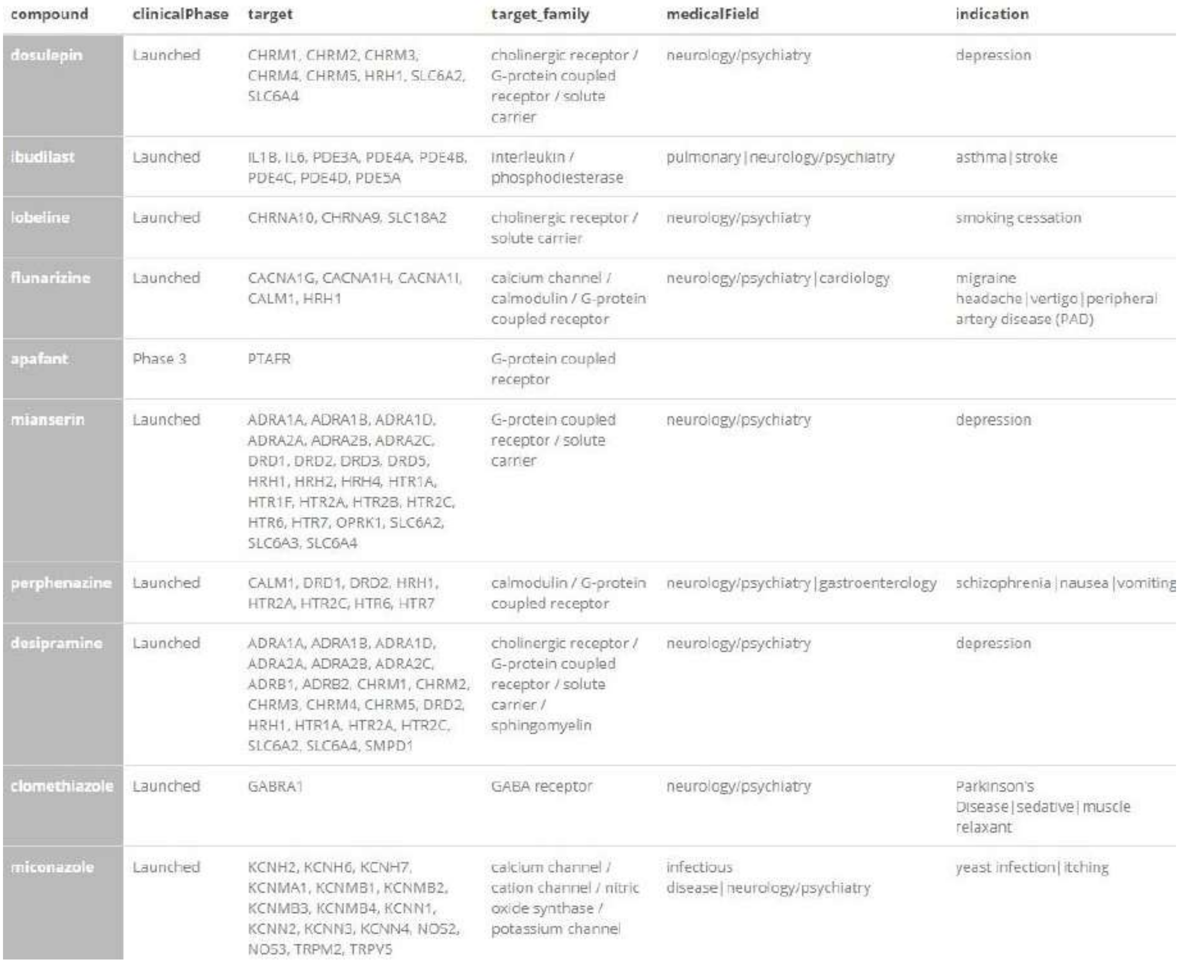
Overall, compounds were selected based on their clinical development phase, where those already undergoing phase III of clinical trials or having been launched into the market were favoured to facilitate access to the compounds in future work. The 10 most statistically significant compounds, with prescription information for (1) neurological conditions and for (2) other conditions, were chosen as potential candidates to be considered for future *in vitro* and *in vivo* testing.

Table 7 | Candidate compounds recapitulative of human-specific AD

Top 10 compounds prescribed for non-neurological (top) and neurological (bottom) conditions, inducing of GE alterations, and that most significantly ($q < 0.05$) anti-correlate with AD-induced GE changes obtained for the *Interaction* coefficient, i.e. compounds recapitulative of human-specific AD induced GE alterations.



compound	clinicalPhase	target	target_family	medicalField	indication
formestane	Launched	CYP19A1	cytochrome P450	oncology	breast cancer
dydrogesterone	Launched	PGR	hormone receptor	obstetrics/gynecology	infertility
triamcinolone	Launched	NR3C1, SERPINA6	nuclear receptor / serpin family	dermatology	corticosteroid-responsive dermatoses
meropenem	Launched			infectious disease	skin infections | intra-abdominal infections | meningitis
phentolamine	Launched	ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C	G-protein coupled receptor	endocrinology | cardiology	pheochromocytoma | hypertension
lopinavir	Launched			infectious disease	human immunodeficiency virus (HIV-1)
oxymetholone	Launched	AR	hormone receptor	hematology	anemia
linezolid	Launched	MAOA, MAOB	monoamine oxidase	infectious disease	pneumonia | skin infections
forskolin	Launched	ADCY2, ADCY5, GNAS	adenylate cyclase / G-protein		
nifedipine	Launched	CACNA1C, CACNA1D, CACNA1F, CACNA1H, CACNA1S, CACNA2D1, CACNB2, CALM1, GLRA1, GLRA3, GLRB, KCNA1, KCNA5, NR1I2, TRPM3	calcium channel / calmodulin / cation channel / ionotropic receptor / nuclear receptor / potassium channel	cardiology	vasospastic angina | chronic stable angina



compound	clinicalPhase	target	target_family	medicalField	indication
dosulepin	Launched	CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, HRH1, SLC6A2, SLC6A4	cholinergic receptor / G-protein coupled receptor / solute carrier	neurology/psychiatry	depression
ibudilast	Launched	IL1B, IL6, PDE3A, PDE4A, PDE4B, PDE4C, PDE4D, PDE5A	interleukin / phosphodiesterase	pulmonary | neurology/psychiatry	asthma | stroke
lobeline	Launched	CHRNA10, CHRNA9, SLC18A2	cholinergic receptor / solute carrier	neurology/psychiatry	smoking cessation
flunarizine	Launched	CACNA1G, CACNA1H, CACNA1I, CALM1, HRH1	calcium channel / calmodulin / G-protein coupled receptor	neurology/psychiatry | cardiology	migraine headache | vertigo | peripheral artery disease (PAD)
apafant	Phase 3	PTAFR	G-protein coupled receptor		
mianserin	Launched	ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C, DRD1, DRD2, DRD3, DRD5, HRH1, HRH2, HRH4, HTR1A, HTR1F, HTR2A, HTR2B, HTR2C, HTR6, HTR7, OPRK1, SLC6A2, SLC6A3, SLC6A4	G-protein coupled receptor / solute carrier	neurology/psychiatry	depression
perphenazine	Launched	CALM1, DRD1, DRD2, HRH1, HTR2A, HTR2C, HTR6, HTR7	calmodulin / G-protein coupled receptor	neurology/psychiatry | gastroenterology	schizophrenia | nausea | vomiting
desipramine	Launched	ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C, ADRB1, ADRB2, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, DRD2, HRH1, HTR1A, HTR2A, HTR2C, SLC6A2, SLC6A4, SMPD1	cholinergic receptor / G-protein coupled receptor / solute carrier / sphingomyelin	neurology/psychiatry	depression
clomethiazole	Launched	GABRA1	GABA receptor	neurology/psychiatry	Parkinson's Disease | sedative | muscle relaxant
miconazole	Launched	KCNH2, KCNH6, KCNH7, KCNMA1, KCNMB1, KCNMB2, KCNMB3, KCNMB4, KCNN1, KCNN2, KCNN3, KCNN4, NOS2, NOS3, TRPM2, TRPV5	calcium channel / cation channel / nitric oxide synthase / potassium channel	infectious disease | neurology/psychiatry	yeast infection | itching

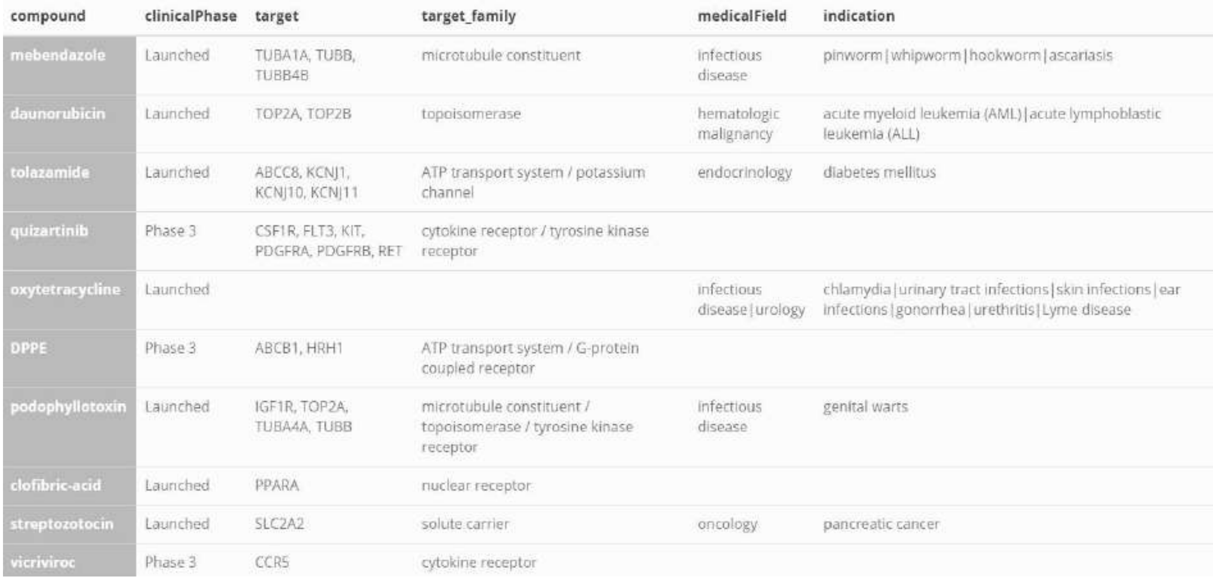
Considering the information on the compound's target family, most of the top 10 compounds act upon a membrane receptor or ionic channel, as the majority of pharmaceutical compounds do, which highlights the importance of those channels' alteration in AD.

The two compounds, targeting non-neurological conditions, whose induced perturbations more significantly anti-correlate with AD-induced human-specific transcriptomic changes are formestane [240] and dydrogesterone [241], respectively a steroidal aromatase inhibitor that supresses estrogen production [240], and a progesterone stereo-isomer that mimics progesterone and is highly selective for its receptors [241]. Sex steroid hormones interact with glial cells and neurons and physiologically promote neuronal plasticity and survival, thus having a positive impact in learning and memory processes [37]. Low levels of these hormones, such as those reached with menopause, have been considered a risk factor for neurodegenerative diseases, namely AD, with estrogen depletion promoting Aβ production and negatively regulating its degradation and clearance processes [37]. Regarding progesterone, despite its beneficial value towards brain health, the activation of one form of progesterone receptors (PR-A) might have a negative control effect not only upon the classical progesterone receptor isoform (PR-B) but also on estrogen receptors, and therefore antagonize estrogen beneficial effects and hinder the normal activity of the brain [242]. The relation between changes in the regulation of these female hormones and AD pathology development support the fact that women are more affected by the disease than men [6,7,9,31,37], and the upsurge of compounds such as formestane and dydrogesterone as potentially replicative of the human disease, given that they directly act upon the regulation of those female hormones. Interestingly, the human samples are mostly masculine, with only one female patient (out of 6) being included within the AD samples in the analysis.

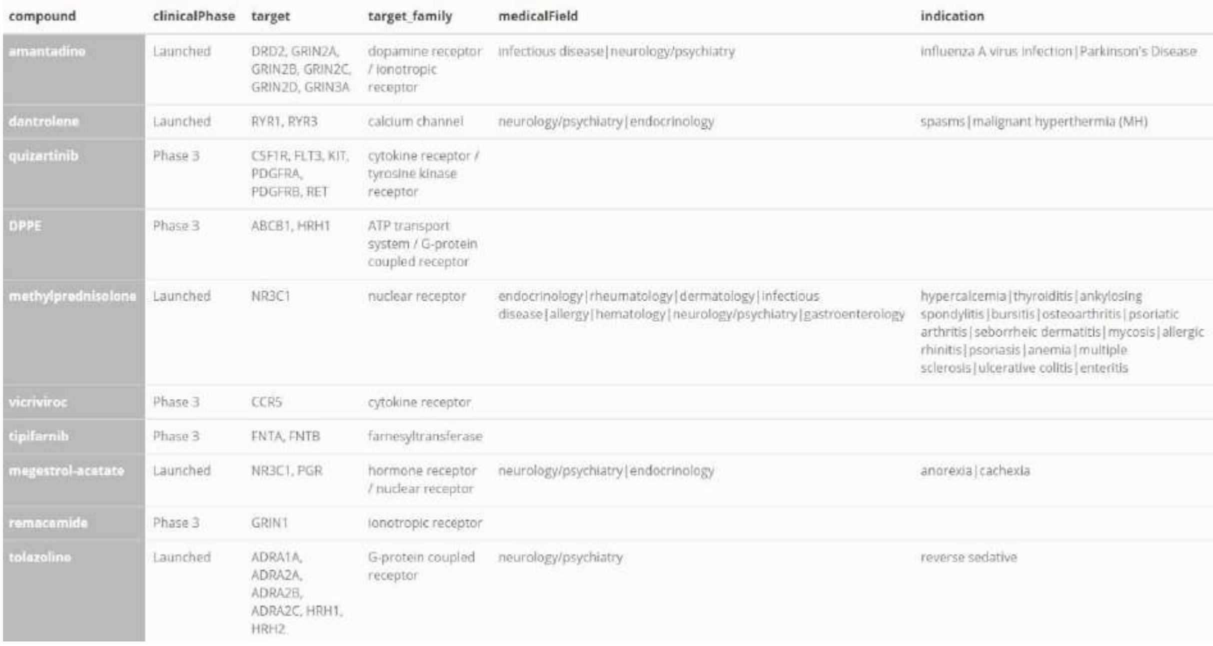
Of the compounds prescribed for neurological diseases, only clomethiazole is used to treat a neurodegenerative disease, which might highlight the fact that some of the different neurological diseases may display some common features with AD. The upsurge of compounds used in the treatment of those diseases as candidates to replicate human AD does not necessarily translate into a possible negative effect of the drug in its original therapeutic indication. Alternatively, it is possible for neurologic conditions to have a similar origin and deviate during pathological development in terms of GE.

Table 8 | Candidate compounds for reverting of non-specific AD

Top 10 compounds prescribed for non-neurological (top) and neurological (bottom) conditions, inducing of GE alterations, and that most significantly ($q < 0.05$) anti-correlate with AD-induced GE changes obtained for the *PSEN* coefficient, i.e. compounds that potentially revert species-common AD-induced GE alterations.



compound	clinicalPhase	target	target_family	medicalField	indication
mebendazole	Launched	TUBA1A, TUBB, TUBB4B	microtubule constituent	infectious disease	pinworm | whipworm | hookworm | ascariasis
daunorubicin	Launched	TOP2A, TOP2B	topoisomerase	hematologic malignancy	acute myeloid leukemia (AML) | acute lymphoblastic leukemia (ALL)
tolazamide	Launched	ABCC8, KCNJ1, KCNJ10, KCNJ11	ATP transport system / potassium channel	endocrinology	diabetes mellitus
quizartinib	Phase 3	CSF1R, FLT3, KIT, PDGFRA, PDGFRB, RET	cytokine receptor / tyrosine kinase receptor		
oxytetracycline	Launched			infectious disease | urology	chlamydia | urinary tract infections | skin infections | ear infections | gonorrhea | urethritis | Lyme disease
DPPE	Phase 3	ABCB1, HRH1	ATP transport system / G-protein coupled receptor		
podophyllotoxin	Launched	IGF1R, TOP2A, TUBA4A, TUBB	microtubule constituent / topoisomerase / tyrosine kinase receptor	infectious disease	genital warts
clofibric-acid	Launched	PPARA	nuclear receptor		
streptozotocin	Launched	SLC2A2	solute carrier	oncology	pancreatic cancer
vicriviroc	Phase 3	CCR5	cytokine receptor		



compound	clinicalPhase	target	target_family	medicalField	indication
amantadine	Launched	DRD2, GRIN2A, GRIN2B, GRIN2C, GRIN2D, GRIN3A	dopamine receptor / ionotropic receptor	infectious disease | neurology/psychiatry	influenza A virus infection | Parkinson's Disease
dantrolene	Launched	RYR1, RYR3	calcium channel	neurology/psychiatry | endocrinology	spasms | malignant hyperthermia (MH)
quizartinib	Phase 3	CSF1R, FLT3, KIT, PDGFRA, PDGFRB, RET	cytokine receptor / tyrosine kinase receptor		
DPPE	Phase 3	ABCB1, HRH1	ATP transport system / G-protein coupled receptor		
methylprednisolone	Launched	NR3C1	nuclear receptor	endocrinology | rheumatology | dermatology | infectious disease | allergy | hematology | neurology/psychiatry | gastroenterology	hypercalcemia | thyroiditis | ankylosing spondylitis | bursitis | osteoarthritis | psoriatic arthritis | seborrheic dermatitis | mycosis | allergic rhinitis | psoriasis | anemia | multiple sclerosis | ulcerative colitis | enteritis
vicriviroc	Phase 3	CCR5	cytokine receptor		
tipifarnib	Phase 3	FNTA, FNTB	farnesyltransferase		
megestrol-acetate	Launched	NR3C1, PGR	hormone receptor / nuclear receptor	neurology/psychiatry | endocrinology	anorexia | cachexia
remacemide	Phase 3	GRIN1	ionotropic receptor		
tolazoline	Launched	ADRA1A, ADRA2A, ADRA2B, ADRA2C, HRH1, HRH2	G-protein coupled receptor	neurology/psychiatry	reverse sedative

Similarly to the compounds putatively able to recapitulate human-specific AD-induced expression changes, the unveiled candidate compounds able to generate GE changes opposite to the overall disease in both species, predominantly target membrane receptors, namely dopamine and tyrosine kinase and cytokine receptors.

In the class of drugs that are not prescribed for neuronal disorders, the compound that was found to statistically anti-correlate the most with the overall AD-induced transcriptomic changes (common to both human patients and mouse models) is used as an anti-parasitic agent that acts upon tubulin polymerization disrupting microtubules. However, the repurposing of this compound to fight brain tumours has been described [243], which may justify its upsurge, given that the Connectivity Map database used to unveil these compound associations is based on tumorigenic cells lines [150,151].

The only compound targeting neurological conditions, amongst the 10 most significant selected, whose therapeutic indication is for a neurodegenerative disease, is an NMDA-antagonist prescribed for Parkinson's disease. A derivative from these compounds, memantine, is one of the approved therapeutics for AD, as mentioned in section 1.5 [6,49].

Overall, and considering that most compounds prescribed for brain diseases have already been studied for AD, as it is the case for dantrolene [244] and methylprednisolone [245], we believe the compounds with indication for non-neurological conditions hold the highest potential for innovation by being repurposed for AD pathology.

CHAPTER IV – CONCLUSIONS

1. Final remarks

The present study confirmed that GE differences between non-AD and AD human brain samples do not correlate with those from the commonly used mouse models. Nonetheless, the majority of altered pathways of both the human and mouse datasets are congruent with the literature available for AD, as discussed in the Results section.

The up-regulated mechanisms in human AD samples are associated with the immune and cardiovascular systems, cellular differentiation, proliferation and apoptosis, and nucleic acid processing; whereas genes involved in ion channels function and/or composition, neurotransmitters trafficking, and synaptic activity appeared down-regulated. Regarding the mouse dataset, genes linked to immune responses, cellular interactions, diabetes and cholesterol were unveiled as up-regulated, whereas spliceosome machinery, cellular respiration and mitochondrial processes appear down-regulated, as well as some neurotransmitter trafficking processes. It is noteworthy the prevalence of immune system-related mechanisms as up regulated in both species, which corroborates already established relevant AD hallmarks, such as chronic inflammation. The observed discrepancies in disrupted mechanisms between the two species might contribute to the differential AD development and progression.

Analysis of the joint dataset unveiled neural-related pathways as generally more downregulated in human patients than in mouse models. These results suggest that either (1) genes are not equally affected between species, thus leading to differences in the most disrupted pathways, or/and that (2) the disease dynamics are species-specific.

The aforementioned lack of correlation in GE alterations between human AD patients and AD mouse models, reinforces the need to obtain more trustworthy mouse models able to more effectively replicate the human disease. On this note, compounds that recapitulate AD-induced GE alterations specific of the human brain were assessed. The goal was to find potential candidate compounds to be applied into a novel mouse model that more accurately recapitulates human-specific AD-induced GE changes. The potential compounds found include formestane, dydrogesterone, triamcinolone and meropenem (prescribed for non-neurological conditions), and dosulepin, ibudilast, lobeline and flunarizine (prescribed for neurological conditions).

Compounds that recapitulate opposite GE alterations to those commonly induced by AD in both human patients and mouse models were also chosen as potential new disease-modifying therapeutics. Candidate compounds include mebendazole, daunorubicin, tolazamide and quizartinib (prescribed for non-neurological conditions), and amantadine, dantrolene, DPPE and methylprednisolone (prescribed for neurological conditions).

The present approach not only gives insights into the molecular mechanisms disrupted in AD and reinforces the differences that exist between the disease developed by animal models and humans, but also is refined enough to decouple AD-induced species-specific GE alterations from those common to both species. Moreover, candidate compounds able to replicate those transcriptomic alterations were also unveiled, which could therefore be used either to express the human disease in a novel and improved mouse model or as a therapeutic approach to reverse AD-induced GE alterations and disease development. The latter is, to our knowledge, the first study of its kind performed in AD. If successfully validated, this approach could revolutionize how research on AD is conducted.

2. Future perspectives

In addition to compound information, *cTRAP* can also inform on genetic perturbations (either overexpression or knockdown) that can recapitulate the human-specific and overall-AD GE changes. These genetic alterations could also potentially inform on molecular causes underlying the transcriptional differences and maybe unveil molecular mechanisms relevant for disease development and progression, being therefore a front worthy of future exploration.

Furthermore, brain tissue samples encompass a multitude of different cell types, namely neurons, astrocytes, endothelial cells, microglia and oligodendrocytes [246]. Studies have already attempted to find GE signatures for each cell type that can be applied to quantify relative cell-type proportions in brain samples [247,248], and even our lab has some ongoing projects with that same end goal. Cellular composition is indeed an important aspect when assessing differential GE between conditions, as it can be affected by the condition in study [247]. For instance, it is possible those AD-induced down-regulated genes associated with neuronal development and activity or with synaptic plasticity reflect, to some extent, the decrease in the number of neurons, and not necessarily a substantial down-regulation of those neural-specific genes in each cell. In future work, it would be interesting to incorporate cell-type composition in the linear models, to decouple cell-type-specific GE alterations from brain cell type composition changes.

As a validation step, the expression of some DEG could be confirmed by RT-PCR in AD and control brain samples. Additionally, it would be interesting to quantify the expression of those genes across different brain cell types, to gain a better insight on the cell-type specificity of the mechanisms underlying their differential expression in a diseased state.

If the present approach is validated, either in the "wet"-lab through the quantification of mRNA of selected genes, or through a bioinformatics analysis of an independent external dataset, the same pipeline could be applied to study other AD mouse models. Moreover, analysing a human idiopathic AD dataset could also give insight on how well studied mouse models emulate the idiopathic AD, therefore contributing to the conception of a novel idiopathic model. This would be an interesting approach given that the idiopathic disease is the most prevalent form of AD [6,35]. However, human controls and patients would need to be carefully matched regarding age, to minimize the effect of age-associated confounders.

After a careful selection of the final candidate compounds and gene perturbations, the goal is to test them *in vitro* and *in vivo*. Compound selection will not only consider their statistically significance, as it was used to select the top 10 most human AD correlated compounds, but also their genetic targets. A first approach would be to evaluate compounds with at least one of the targets considered as highly differentially expressed, which could be interesting if the target influences pathways involved in AD development. Compounds that target lowly DEG that are included in disease pathways and interesting gene networks, where other contributing genes also appear differentially expressed, could also be of interest.

The chosen compounds and genetic perturbations would be tested in non-diseased human and mouse neural cell lines, with the purpose of developing detectable AD hallmarks, such as high of levels of Aβ in the extracellular culture medium or increased presence of Tau protein within neurons. Other brain cell types could also be evaluated, especially, but not exclusively, if their estimated relative proportions are incorporated in the linear models and inferred to be relevant, as well as co-culture of different cell types. The outcomes of these analyses need to be interpreted with caution, given that isolated cell lines or even co-cultures of two neural cell lines do not effectively replicate the complex cell interactions ongoing in a human brain. To validate candidate genetic perturbations, a genomic editing tool like CRISPR [212] could be used.

Compound administration and gene editing of mouse models would also be tested, with thorough evaluation of AD development and progression. This study could be conducted both in healthy mice, to assess the development of the disease *de novo*, or in the same homozygous *APP/PSEN1* double-mutant mice used for the present analyses, in order to evaluate the improvement of an already existent mouse model. Compound administration in mice is a more intricate process compared with that in cell lines, as blood-brain barrier crossing needs to be considered, either by choosing drug candidates with characteristics that allow a physiological infiltration into the brain, or by adapting the administration method. Drug delivery across the blood-brain barrier can be achieved through a diversity of already described methods, including nanoparticle delivery (for instance, using the naturally-occurring, non-immunogenic brain-isolated exosomes [249]), intrathecal [250], intraventricular [250] or nasal [249] administration, or by chemical or physical disruption of the barrier [250]. Moreover, for the AD mouse model, if the blood-brain barrier is disrupted at the time of administration, the compound might be able to cross it if it is intravenously administered [249]. However, we can never discard the presence of other brain barriers such as the blood-CSF barrier, which can potentially become a hazard as well.

To sum up, future work should initially focus on selecting potential compound and genetic perturbations and validating the approach *in silico* or *in vitro*, so its use can be extended to testing the unveiled perturbations in what it could be an innovative mouse model able to further develop AD research.

CHAPTER V – BIBLIOGRAPHY

1. Stelzmann, R. A., Schnitzlein, H. N. & Murtagh, F. R. An english translation of alzheimer's 1907 paper, 'Uber eine eigenartige erkankung der hirnrinde'. *Clin. Anat.* **8**, 429–431 (1995).

2. Alzheimer, A. Über eine eigenartige Erkrankung der Hirnrinde. *Allgemeine Zeitschrift fur Psychiatrie und phychish-Gerichtliche Medizin* 146–148 (1907).

3. NIH-NIA, N. I. on A. Alzheimer's Disease and Related Dementias. Available at: https://www.nia.nih.gov/health/alzheimers. (Accessed: 3rd September 2018)

4. Wang, J., Gu, B. J., Masters, C. L. & Wang, Y.-J. A systemic view of Alzheimer disease — insights from amyloid-β metabolism beyond the brain. *Nat. Rev. Neurol.* **13**, 612 (2017).

5. Swerdlow, R. H. Pathogenesis of Alzheimer's disease. *Clin. Interv. Aging* **2**, 347–59 (2007).

6. Masters, C. L. *et al.* Alzheimer's disease. *Nat. Rev. Dis. Prim.* **1**, 15056 (2015).

7. United Nations Population Division. *World Population Prospects 2019 Report*. (2019).

8. World Health Organization. Dementia. (2019). Available at: https://www.who.int/news-room/fact-sheets/detail/dementia. (Accessed: 12th August 2019)

9. Niu, H., Álvarez-Álvarez, I., Guillén-Grima, F. & Aguinaga-Ontoso, I. Prevalence and incidence of Alzheimer's disease in Europe: A meta-analysis. *Neurol. (English Ed.* **32**, 523–532 (2017).

10. National Institute on Aging, N. What Happens to the Brain in Alzheimer's Disease? (2017). Available at: https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease. (Accessed: 14th August 2019)

11. Zheng, H. & Koo, E. H. Biology and pathophysiology of the amyloid precursor protein. *Mol. Neurodegener.* **6**, 27 (2011).

12. Kandel, E. R. *Principles of neural science*. (2013).

13. Caillet-Boudin, M.-L., Buée, L., Sergeant, N. & Lefebvre, B. Regulation of human

MAPT gene expression. *Mol. Neurodegener.* **10**, 28 (2015).

14. Norstrom, E. Metabolic processing of the amyloid precursor protein -- new pieces of the Alzheimer's puzzle. *Discov. Med.* **23**, 269–276 (2017).

15. Kohli, B. M. *et al.* Interactome of the Amyloid Precursor Protein APP in Brain Reveals a Protein Network Involved in Synaptic Vesicle Turnover and a Close Association with Synaptotagmin-1. *J. Proteome Res.* **11**, 4075–4090 (2012).

16. Pearson, H. A. & Peers, C. Physiological roles for amyloid β peptides. *J. Physiol.* **575**, 5 (2006).

17. Kumar, A., Singh, A. & Ekavali. A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacol. Reports* **67**, 195–203 (2015).

18. Chen, W. *et al.* Familial Alzheimer's mutations within APPTM increase Aβ42 production by enhancing accessibility of ε-cleavage site. *Nat. Commun.* **5**, 3037 (2014).

19. Makin, S. The amyloid hypothesis on trial. *Nature* **559**, S4–S7 (2018).

20. Wang, S., Mims, P. N., Roman, R. J. & Fan, F. Is Beta-Amyloid Accumulation a Cause or Consequence of Alzheimer's Disease? *J. Alzheimer's Park. Dement.* **1**, (2016).

21. ALZFORUM. Presenilin-1 (PSEN1). Available at: https://www.alzforum.org/alzpedia/presenilin-1-psen1. (Accessed: 19th September 2019)

22. De Strooper, B. Loss-of-function presenilin mutations in Alzheimer disease. Talking Point on the role of presenilin mutations in Alzheimer disease. *EMBO Rep.* **8**, 141–6 (2007).

23. Esparza, T. J. *et al.* Amyloid-β oligomerization in Alzheimer dementia versus high-pathology controls. *Ann. Neurol.* **73**, 104–19 (2013).

24. Liu, J., Chang, L., Song, Y., Li, H. & Wu, Y. The Role of NMDA Receptors in Alzheimer's Disease. *Front. Neurosci.* **13**, 43 (2019).

25. Snyder, E. M. *et al.* Regulation of NMDA receptor trafficking by amyloid-β. *Nat. Neurosci.* **8**, 1051–1058 (2005).

26. Shi, X.-D. *et al.* Blocking the Interaction between EphB2 and ADDLs by a Small Peptide Rescues Impaired Synaptic Plasticity and Memory Deficits in a Mouse Model

of Alzheimer's Disease. *J. Neurosci.* **36**, 11959–11973 (2016).

27. Ferreira-Vieira, T. H., Guimaraes, I. M., Silva, F. R. & Ribeiro, F. M. Alzheimer's disease: Targeting the Cholinergic System. *Curr. Neuropharmacol.* **14**, 101–15 (2016).

28. De Ferrari, G. V. *et al.* A Structural Motif of Acetylcholinesterase That Promotes Amyloid β-Peptide Fibril Formation [†]. *Biochemistry* **40**, 10447–10457 (2001).

29. Carvajal, F. J. & Inestrosa, N. C. Interactions of AChE with Aβ Aggregates in Alzheimer's Brain: Therapeutic Relevance of IDN 5706. *Front. Mol. Neurosci.* **4**, 19 (2011).

30. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **15**, 321–387 (2019).

31. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. *Alzheimer's Dement. (New York, N. Y.)* **4**, 575–590 (2018).

32. Gratuze, M., Leyns, C. E. G. & Holtzman, D. M. New insights into the role of TREM2 in Alzheimer's disease. *Mol. Neurodegener.* **13**, 66 (2018).

33. Du, X., Wang, X. & Geng, M. Alzheimer's disease hypothesis and related therapies. *Transl. Neurodegener.* **7**, 2 (2018).

34. Lanoiselée, H.-M. *et al.* APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. *PLOS Med.* **14**, e1002270 (2017).

35. Lava, N. Types of Alzheimer's Disease. *WebMD Medical Reference* (2018). Available at: https://www.webmd.com/alzheimers/guide/alzheimers-types. (Accessed: 3rd September 2018)

36. Chiaravalloti, A. *et al.* Comparison between Early-Onset and Late-Onset Alzheimer's Disease Patients with Amnestic Presentation: CSF and 18F-FDG PET Study. *Dement. Geriatr. Cogn. Dis. Extra* **6**, 108–119 (2016).

37. Pike, C. J. Sex and the development of Alzheimer's disease. *J. Neurosci. Res.* **95**, 671–680 (2017).

38. Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet. Neurol.*

13, 788–94 (2014).

39. Etnier, J. L. *et al.* The Physical Activity and Alzheimer's Disease (PAAD) Study: Cognitive outcomes. *Ann. Behav. Med.* **52**, 175–185 (2018).

40. Lee, J.-Y. *et al.* Illiteracy and the incidence of Alzheimer's disease in the Yonchon County survey, Korea. *Int. Psychogeriatrics* **20**, 976–85 (2008).

41. Cataldo, J. K., Prochaska, J. J. & Glantz, S. A. Cigarette Smoking is a Risk Factor for Alzheimer's Disease: An Analysis Controlling for Tobacco Industry Affiliation. *J. Alzheimer's Dis.* **19**, 465–480 (2010).

42. Johnson, L. A. *et al.* A depressive endophenotype of poorer cognition among cognitively healthy community-dwelling adults: results from the Western Australia memory study. *Int. J. Geriatr. Psychiatry* **30**, 881–886 (2015).

43. Kivipelto, M. *et al.* Obesity and Vascular Risk Factors at Midlife and the Risk of Dementia and Alzheimer Disease. *Arch. Neurol.* **62**, 1556–60 (2005).

44. Huang, C.-C. *et al.* Diabetes Mellitus and the Risk of Alzheimer's Disease: A Nationwide Population-Based Study. *PLoS One* **9**, e87095 (2014).

45. Launer, L. J. *et al.* Midlife blood pressure and dementia: the Honolulu-Asia aging study. *Neurobiol. Aging* **21**, 49–55 (2000).

46. Skoog, I. *et al.* 15-year longitudinal study of blood pressure and dementia. *Lancet (London, England)* **347**, 1141–5 (1996).

47. Kivipelto, M. *et al.* Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *BMJ* **322**, 1447–51 (2001).

48. Sivanandam, T. M. & Thakur, M. K. Traumatic brain injury: A risk factor for Alzheimer's disease. *Neurosci. Biobehav. Rev.* **36**, 1376–1381 (2012).

49. Weller, J. & Budson, A. Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Research* **7**, 1161 (2018).

50. Burns, A. & Iliffe, S. Alzheimer's disease. *BMJ* **338**, b158 (2009).

51. Bird, T. D. *Alzheimer Disease Overview*. *GeneReviews®* (University of Washington, Seattle, 1993).

52. Lange, K. W. *et al.* Medical foods in Alzheimer's disease. *Food Sci. Hum. Wellness* **8**, 1–7 (2019).

53. Horien, C. & Yuan, P. Focus: Drug Development: Drug Development. *Yale J. Biol. Med.* **90**, 1 (2017).

54. Orion Pharma. Drug development. Available at: https://www.orion.fi/en/rd/drug-development/. (Accessed: 21st September 2019)

55. Market Insider. Remember When Was The Last Time A New Alzheimer's Drug Was Approved? | Markets Insider. (2018).

56. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **5**, 272–293 (2019).

57. Alzheimer's Drug Discovery Foundation. *2018 Alzheimer's Clinical Trials Report*. (2018).

58. Cummings, J., Lee, G., Ritter, A. & Zhong, K. Alzheimer's disease drug development pipeline: 2018. *Alzheimer's Dement. (New York, N. Y.)* **4**, 195–214 (2018).

59. Adams, B. Takeda, Zinfandel ax Alzheimer's test for pioglitazone, chalking up another AD fail. (2018). Available at: https://www.fiercebiotech.com/biotech/takeda-zinfandel-ax-alzheimer-s-test-for-pioglitazone-chalking-up-another-ad-fail. (Accessed: 23rd August 2019)

60. Co., M. &. Merck Announces Discontinuation of APECS Study Evaluating Verubecestat (MK-8931) for the Treatment of People with Prodromal Alzheimer's Disease [Press Release]. (2018). Available at: https://investors.merck.com/news/press-release-details/2018/Merck-Announces-Discontinuation-of-APECS-Study-Evaluating-Verubecestat-MK-8931-for-the-Treatment-of-People-with-Prodromal-Alzheimers-Disease/default.aspx. (Accessed: 23rd August 2019)

61. Idrus, A. Al. vTv's azeliragon fails to improve Alzheimer's symptoms in phase 3. (2018). Available at: https://www.fiercebiotech.com/biotech/vtv-s-azeliragon-fails-to-improve-alzheimer-s-symptoms-phase-3. (Accessed: 23rd August 2019)

62. AstraZeneca. Update on Phase III clinical trials of lanabecestat for Alzheimer's disease [Press Release]. (2018). Available at: https://www.astrazeneca.com/media-centre/press-

releases/2018/update-on-phase-iii-clinical-trials-of-lanabecestat-for-alzheimers-disease-12062018.html. (Accessed: 23rd August 2019)

63. Janssen. Update on Janssen’s BACE Inhibitor Program [Press Release]. (2018). Available at: https://www.janssen.com/update-janssens-bace-inhibitor-program. (Accessed: 23rd August 2019)

64. Dilts, E. Pfizer ends research for new Alzheimer’s, Parkinson’s drugs. (2018).

65. U.S. National Library of Medicine. ClinicalTrials.gov. Available at: https://clinicaltrials.gov/ct2/home. (Accessed: 27th September 2019)

66. Roche. Roche to discontinue Phase III CREAD 1 and 2 clinical studies of crenezumab in early Alzheimer’s disease (AD) [Press Release]. (2019). Available at: https://www.roche.com/media/releases/med-cor-2019-01-30.htm. (Accessed: 23rd August 2019)

67. Biogen. Biogen and Eisai to Discontinue Phase 3 ENGAGE and EMERGE Trials of aducanumab in Alzheimer’s Disease [Press Release]. (2019). Available at: http://investors.biogen.com/news-releases/news-release-details/biogen-and-eisai-discontinue-phase-3-engage-and-emerge-trials. (Accessed: 23rd August 2019)

68. Novartis. Novartis, Amgen and Banner Alzheimer’s Institute discontinue clinical program with BACE inhibitor CNP520 for Alzheimer’s prevention [Press Release]. (2019). Available at: https://www.novartis.com/news/media-releases/novartis-amgen-and-banner-alzheimers-institute-discontinue-clinical-program-bace-inhibitor-cnp520-alzheimers-prevention. (Accessed: 23rd August 2019)

69. Cummings, J., Reiber, C. & Kumar, P. The price of progress: Funding and financing Alzheimer’s disease drug development. *Alzheimer’s Dement. (New York, N. Y.)* **4**, 330–343 (2018).

70. King, A. The search for better animal models of Alzheimer’s disease. (2018).

71. Cavanaugh, S. E., Pippin, J. J. & Barnard, N. D. Animal models of Alzheimer disease: historical pitfalls and a path forward. *ALTEX* **31**, 279–302 (2014).

72. Schott, J. M., Aisen, P. S., Cummings, J. L., Howard, R. J. & Fox, N. C. Unsuccessful trials of therapies for Alzheimer’s disease. *Lancet (London, England)* **393**, 29 (2019).

73. 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* **26**, 735–739 (2017).

74. Burns, T. C., Li, M. D., Mehta, S., Awad, A. J. & Morgan, A. A. Mouse models rarely mimic the transcriptome of human neurodegenerative diseases: A systematic bioinformatics-based critique of preclinical models. *Eur. J. Pharmacol.* **759**, 101–117 (2015).

75. Guo, B. & Zhou, Q. How efficient are rodent models for Alzheimer's disease drug discovery? *Expert Opin. Drug Discov.* **13**, 113–115 (2018).

76. ALZFORUM. Research Models. Available at: https://www.alzforum.org/research-models. (Accessed: 17th September 2019)

77. Gunn-Moore, D., Kaidanovich-Beilin, O., Gallego Iradi, M. C., Gunn-Moore, F. & Lovestone, S. Alzheimer's disease in humans and other animals: A consequence of postreproductive life span and longevity rather than aging. *Alzheimer's Dement.* **14**, 195–204 (2018).

78. Hargis, K. E. & Blalock, E. M. Transcriptional signatures of brain aging and Alzheimer's disease: What are our rodent models telling us? *Behav. Brain Res.* **322**, 311–328 (2017).

79. Drummond, E. & Wisniewski, T. Alzheimer's disease: experimental models and reality. *Acta Neuropathol.* **133**, 155–175 (2017).

80. Adams, M. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (80-.).* **252**, 1651–1656 (1991).

81. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **13**, e1005457 (2017).

82. Sousa, S. A. *et al.* Bioinformatics Applications in Life Sciences and Technologies. *Biomed Res. Int.* **2016**, 3603827 (2016).

83. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).

84. Sealfon, S. C. & Chu, T. T. RNA and DNA Microarrays. in *Methods in molecular biology (Clifton, N.J.)* **671**, 3–34 (2011).

85. Dunning, M. J., Smith, M. L., Ritchie, M. E. & Tavare, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **23**, 2183–2184 (2007).

86. Fan, J.-B., Hu, S. X., Craumer, W. C. & Barker, D. L. BeadArray^{TM}-based solutions for enabling the promise of pharmacogenomics. *Biotechniques* **39**, S583–S588 (2005).

87. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

88. Rao, M. S. *et al.* Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Front. Genet.* **9**, 636 (2018).

89. Masjosthusmann, S. *et al.* A transcriptome comparison of time-matched developing human, mouse and rat neural progenitor cells reveals human uniqueness. *Toxicol. Appl. Pharmacol.* **354**, 40–55 (2018).

90. Rothman, S. M. *et al.* Human Alzheimer's disease gene expression signatures and immune profile in APP mouse models: a discrete transcriptomic view of Aβ plaque pathology. *J. Neuroinflammation* **15**, 256 (2018).

91. Castillo, E. *et al.* Comparative profiling of cortical gene expression in Alzheimer's disease patients and mouse models demonstrates a link between amyloidosis and neuroinflammation. *Sci. Rep.* **7**, 17762 (2017).

92. Rojo, A. I. *et al.* NRF2 deficiency replicates transcriptomic changes in Alzheimer's patients and worsens APP and TAU pathology. *Redox Biol.* **13**, 444–451 (2017).

93. Galatro, T. F. *et al.* Transcriptomic analysis of purified human cortical microglia reveals age-associated changes. *Nat. Neurosci.* **20**, 1162–1171 (2017).

94. Altimiras, F. *et al.* Brain Transcriptome Sequencing of a Natural Model of Alzheimer's Disease. *Front. Aging Neurosci.* **9**, 64 (2017).

95. Wan, Y.-W. *et al.* Functional dissection of Alzheimer's disease brain gene expression signatures in humans and mouse models. *bioRxiv* 506873 (2019). doi:10.1101/506873

96. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic*

Acids Res. **28**, 27–30 (2000).

97. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

98. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* pro.3715 (2019). doi:10.1002/pro.3715

99. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

100. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

101. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

102. Cantarero-Prieto, D., Leon, P. L., Blazquez-Fernandez, C., Juan, P. S. & Cobo, C. S. The economic cost of dementia: A systematic review. *Dementia* 147130121983777 (2019). doi:10.1177/1471301219837776

103. Antonell, A. *et al.* A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. *Neurobiol. Aging* **34**, 1772–1778 (2013).

104. Matarin, M. *et al.* A Genome-wide Gene-Expression Analysis and Database in Transgenic Mice during Development of Amyloid or Tau Pathology. *Cell Rep.* **10**, 633–644 (2015).

105. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

106. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

107. RStudio Team. RStudio: Integrated Development Environment for R. (2015).

108. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).

109. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing

and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

110. Wickham, H., Hester, J. & Francois, R. readr: Read Rectangular Text Data. (2018).

111. Shi, W., Oshlack, A. & Smyth, G. K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* **38**, e204–e204 (2010).

112. Drăghici, S. *Statistics and data analysis for microarrays using R and Bioconductor*. (CRC Press, 2012).

113. Quackenbush, J. Microarray data normalization and transformation. *Nat. Genet.* **32**, 496–501 (2002).

114. Smyth, G. *et al.* Linear Models for Microarray Data User 's Guide (Now Including RNA-Seq Data Analysis). *Bioinformatics* (2011).

115. Ritchie, M. E., Dunning, M. J., Smith, M. L., Shi, W. & Lynch, A. G. BeadArray expression analysis using bioconductor. *PLoS Comput. Biol.* **7**, e1002276 (2011).

116. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

117. Petri, T., Berchtold, E., Zimmer, R. & Friedel, C. C. Detection and correction of probe-level artefacts on microarrays. *BMC Bioinformatics* **13**, 114 (2012).

118. Bolstad, B. M., Collin, F., Simpson, K. M., Irizarry, R. A. & Speed, T. P. Experimental Design and Low-Level Analysis of Microarray Data. *Int. Rev. Neurobiol.* **60**, 25–58 (2004).

119. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2019).

120. MacDonald, J. W. hugene11sttranscriptcluster.db: Affymetrix hugene11 annotation data (chip hugene11sttranscriptcluster). (2017).

121. Dunning, M., Lynch, A. & Eldridge, M. illuminaMousev2.db: Illumina MouseWG6v2 annotation data (chip illuminaMousev2). (2015).

122. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

123. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

124. Ensembl. Ensembl 97 and Ensembl Genomes 44 have been released! – Ensembl Blog. (2019). Available at: http://www.ensembl.info/2019/07/03/ensembl-97-and-ensembl-genomes-43-have-been-released/. (Accessed: 19th October 2019)

125. Motulsky, H. *Intuitive biostatistics : a nonmathematical guide to statistical thinking*. (Oxford University Press, 2014).

126. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

127. Benjamini, Y. & Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25**, 60–83 (2000).

128. Mishra, P. *et al.* Descriptive statistics and normality tests for statistical data. *Ann. Card. Anaesth.* **22**, 67–72 (2019).

129. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2019).

130. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

131. McDonald, J. H. *Handbook of Biological Statistics*. (Sparky House Publishing, 2014).

132. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).

133. Groth, D., Hartmann, S., Klie, S. & Selbig, J. Principal components analysis. *Methods Mol. Biol.* **930**, 527–47 (2013).

134. Mevik, B.-H. & Wehrens, R. The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Softw.* **18**, 1–23 (2007).

135. Lê, S., Josse, J. & Husson, F. FactoMineR : An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).

136. Kassambara, A. & Mundt, F. Package 'factoextra' for R: Extract and Visualize the Results of Multivariate Data Analyses. *R Packag. version* **1**, 1–76 (2017).

137. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

138. Afshartous, D. & Preston, R. A. Key Results of Interaction Models With Centering. *J. Stat. Educ.* **19**, 1–11 (2011).

139. Maltenfort, M. G. Understanding Bayesian Statistics. *J. Spinal Disord. Tech.* **28**, 294 (2015).

140. Kording, K. P. Bayesian statistics: relevant for the brain? *Curr. Opin. Neurobiol.* **25**, 130–133 (2014).

141. Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210 (2003).

142. Smyth, G. Limma moderated t-statistics and B-statistics. *Bioconductor* (2017). Available at: https://support.bioconductor.org/p/6124/. (Accessed: 25th October 2018)

143. Akalin, A. *Computational Genomics with R.* (CRC Pr I Llc, 2019).

144. Cui, X. & Churchill, G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210 (2003).

145. Nichols, J. A., Herbert Chan, H. W. & Baker, M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys. Rev.* **11**, 111–118 (2019).

146. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

147. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

148. Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).

149. de Almeida, B. P., Saraiva-Agostinho, N. & Barbosa-Morais, N. L. cTRAP: Identification of candidate causal perturbations from differential gene expression data. (2019).

150. Lamb, J. *et al.* The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-.).* **313**, 1929–1935 (2006).

151. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

152. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

153. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

154. Patwari, P. *et al.* Thioredoxin-independent regulation of metabolism by the alpha-arrestin proteins. *J. Biol. Chem.* **284**, 24996–5003 (2009).

155. Meng, J. *et al.* ARRDC4 regulates enterovirus 71-induced innate immune response by promoting K63 polyubiquitination of MDA5 through TRIM65. *Cell Death Dis.* **8**, e2866–e2866 (2017).

156. Olstad, E. W. *et al.* Ciliary Beating Compartmentalizes Cerebrospinal Fluid Flow in the Brain and Regulates Ventricular Development. *Curr. Biol.* **29**, 229-241.e6 (2019).

157. Lin, M.-E., Herr, D. R. & Chun, J. Lysophosphatidic acid (LPA) receptors: signaling properties and disease relevance. *Prostaglandins Other Lipid Mediat.* **91**, 130–8 (2010).

158. Samyn-Petit, B., Krzewinski-Recchi, M. A., Steelant, W. F., Delannoy, P. & Harduin-Lepers, A. Molecular cloning and functional expression of human ST6GalNAc II. Molecular expression in various human cultured cells. *Biochim. Biophys. Acta* **1474**, 201–11 (2000).

159. Xu, Z., Maroney, A. C., Dobrzanski, P., Kukekov, N. V & Greene, L. A. The MLK family mediates c-Jun N-terminal kinase activation in neuronal apoptosis. *Mol. Cell. Biol.* **21**, 4713–24 (2001).

160. Iijima, K. *et al.* ZNF385B is characteristically expressed in germinal center B cells and involved in B-cell apoptosis. *Eur. J. Immunol.* **42**, 3405–3415 (2012).

161. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, 1260419–1260419 (2015).

162. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science (80-.).* **357**, eaan2507 (2017).

163. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* **356**, eaal3321 (2017).

164. The Human Protein Atlas. ZNF385B. Available at: https://www.proteinatlas.org/ENSG00000144331-ZNF385B/tissue. (Accessed: 25th November 2019)

165. NCBI. UNC13C unc-13 homolog C [Homo sapiens (human)]. (2019). Available at: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=440279. (Accessed: 2nd November 2019)

166. Tan, M. G. K. *et al.* Decreased rabphilin 3A immunoreactivity in Alzheimer's disease is associated with Aβ burden. *Neurochem. Int.* **64**, 29–36 (2014).

167. Chen, X., Durisic, N., Lynch, J. W. & Keramidas, A. Inhibitory synapse deficits caused by familial α1 GABAA receptor mutations in epilepsy. *Neurobiol. Dis.* **108**, 213–224 (2017).

168. Singec, I. *et al.* Synaptic vesicle protein synaptoporin is differently expressed by subpopulations of mouse hippocampal neurons. *J. Comp. Neurol.* **452**, 139–153 (2002).

169. Kielty, C. M., Sherratt, M. J. & Shuttleworth, C. A. Elastic fibres. *J. Cell Sci.* **112**, 1093–1100 (2002).

170. Bochicchio, B., Pepe, A. & Tamburro, A. M. Elastic fibers and amyloid deposition in vascular tissue. *Future Neurol.* **2**, 523–536 (2007).

171. Wang, S.-P. & Wang, L.-H. Disease implication of hyper-Hippo signalling. *Open Biol.* **6**, (2016).

172. Yuan, Z. *et al.* Regulation of Neuronal Cell Death by MST1-FOXO1 Signaling. *J. Biol. Chem.* **284**, 11285–11292 (2009).

173. Sanphui, P. & Biswas, S. C. FoxO3a is activated and executes neuron death via Bim in response to β-amyloid. *Cell Death Dis.* **4**, e625–e625 (2013).

174. Coppede, F. & Migliore, L. DNA Damage and Repair in Alzheimers Disease. *Curr. Alzheimer Res.* **6**, 36–47 (2009).

175. Barbagallo, M. & Dominguez, L. J. Type 2 diabetes mellitus and Alzheimer's disease. *World J. Diabetes* **5**, 889–93 (2014).

176. Jia, J.-J., Zeng, X.-S., Song, X.-Q., Zhang, P.-P. & Chen, L. Diabetes Mellitus and Alzheimer's Disease: The Protection of Epigallocatechin-3-gallate in Streptozotocin Injection-Induced Models. *Front. Pharmacol.* **8**, 834 (2017).

177. Swerdlow, R. H. Mitochondria and Mitochondrial Cascades in Alzheimer's Disease. *J. Alzheimers. Dis.* **62**, 1403–1416 (2018).

178. KEGG PATHWAY: Alzheimer disease - Homo sapiens (human). Available at: https://www.genome.jp/kegg-bin/show_pathway?hsa05010. (Accessed: 22nd November 2019)

179. Fedele, L. *et al.* Disease-associated missense mutations in GluN2B subunit alter NMDA receptor ligand binding and ion channel properties. *Nat. Commun.* **9**, 957 (2018).

180. Gerber, S. *et al.* Recessive and Dominant De Novo ITPR1 Mutations Cause Gillespie Syndrome. *Am. J. Hum. Genet.* **98**, 971–980 (2016).

181. Ibarreta, D., Tao, J., Parrilla, R. & Ayuso, M. S. Mutation analysis of chromosome 19 calmodulin (CALM3) gene in Alzheimer's disease patients. *Neurosci. Lett.* **229**, 157–60 (1997).

182. Bradley, J. E., Ramirez, G. & Hagood, J. S. Roles and regulation of Thy-1, a context-dependent modulator of cell phenotype. *Biofactors* **35**, 258–65 (2009).

183. Jósvay, K. *et al.* Besides neuro-imaging, the Thy1-YFP mouse could serve for visualizing experimental tumours, inflammation and wound-healing. *Sci. Rep.* **4**, 6776 (2015).

184. Tokugawa, Y., Koyama, M. & Silver, J. A molecular basis for species differences in THY-1 expression patterns. *Mol. Immunol.* (1997). doi:10.1016/S0161-5890(98)00010-8

185. Nishikawa, H. & Suzuki, H. Possible Role of Inflammation and Galectin-3 in Brain Injury after Subarachnoid Hemorrhage. *Brain Sci.* **8**, (2018).

186. Shin, T. The pleiotropic effects of galectin-3 in neuroinflammation: A review. *Acta Histochem.* **115**, 407–411 (2013).

187. Sivasankar, B. *et al.* CD59 blockade enhances antigen-specific CD4+ T cell responses in humans: a new target for cancer immunotherapy? *J. Immunol.* **182**, 5203–7 (2009).

188. Du, Y. *et al.* NF-κB and enhancer-binding CREB protein scaffolded by CREB-binding protein (CBP)/p300 proteins regulate CD59 protein expression to protect cells from complement attack. *J. Biol. Chem.* **289**, 2711–24 (2014).

189. Wang, C. *et al.* Modulation of Mac-1 (CD11b/CD18)-mediated adhesion by the leukocyte-specific protein 1 is key to its role in neutrophil polarization and chemotaxis. *J. Immunol.* **169**, 415–23 (2002).

190. Dollt, C. *et al.* The novel immunoglobulin super family receptor SLAMF9 identified in TAM of murine and human melanoma influences pro-inflammatory cytokine secretion and migration. *Cell Death Dis.* **9**, 939 (2018).

191. Oliveira-Nascimento, L., Massari, P. & Wetzler, L. M. The Role of TLR2 in Infection and Immunity. *Front. Immunol.* **3**, 79 (2012).

192. Nasrabady, S. E., Rizvi, B., Goldman, J. E. & Brickman, A. M. White matter changes in Alzheimer’s disease: a focus on myelin and oligodendrocytes. *Acta Neuropathol. Commun.* **6**, 22 (2018).

193. Ma, J. *et al.* Microglial cystatin F expression is a sensitive indicator for ongoing demyelination with concurrent remyelination. *J. Neurosci. Res.* **89**, 639–649 (2011).

194. Baba, T. & Mukaida, N. Role of macrophage inflammatory protein (MIP)-1α/CCL3 in leukemogenesis. *Mol. Cell. Oncol.* **1**, e29899 (2014).

195. Takahashi, T. *et al.* Chemokine CCL4 Induced in Mouse Brain Has a Protective Role against Methylmercury Toxicity. *Toxics* **6**, 36 (2018).

196. Baldwin, K. T., Carbajal, K. S., Segal, B. M. & Giger, R. J. Neuroinflammation triggered by β-glucan/dectin-1 signaling enables CNS axon regeneration. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 2581–6 (2015).

197. Li, Y. *et al.* CD6 as a potential target for treating multiple sclerosis. *Proc. Natl. Acad. Sci.* **114**, 2687–2692 (2017).

198. McCarthy, M. K. & Weinberg, J. B. The immunoproteasome and viral infection: a complex regulator of inflammation. *Front. Microbiol.* **6**, 21 (2015).

199. NCBI. CORT cortistatin [Homo sapiens (human)]. (2019). Available at: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=1325.

(Accessed: 15th November 2019)

200. Liguz-Lecznar, M., Urban-Ciecko, J. & Kossut, M. Somatostatin and Somatostatin-Containing Neurons in Shaping Neuronal Activity and Plasticity. *Front. Neural Circuits* **10**, 48 (2016).

201. Shobab, L. A., Hsiung, G.-Y. R. & Feldman, H. H. Cholesterol in Alzheimer's disease. *Lancet Neurol.* **4**, 841–852 (2005).

202. Chiroma, S. M. *et al.* Inflammation in Alzheimer's disease: A friend or foe? *Biomed. Res. Ther.* **5**, 2552–2564 (2018).

203. Gowrishankar, S., Wu, Y. & Ferguson, S. M. Impaired JIP3-dependent axonal lysosome transport promotes amyloid plaque pathology. *J. Cell Biol.* **216**, 3291–3305 (2017).

204. Hsieh, Y.-C. *et al.* Tau-Mediated Disruption of the Spliceosome Triggers Cryptic RNA Splicing and Neurodegeneration in Alzheimer's Disease. *Cell Rep.* **29**, 301-316.e10 (2019).

205. Pang, Y. L. J., Poruri, K. & Martinis, S. A. tRNA synthetase: tRNA aminoacylation and beyond. *Wiley Interdiscip. Rev. RNA* **5**, 461–80 (2014).

206. Guantes, R. *et al.* Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res.* **25**, 633–44 (2015).

207. Hawking, Z. L. Alzheimer's disease: the role of mitochondrial dysfunction and potential new therapies. *Biosci. Horizons Int. J. Student Res.* **9**, (2016).

208. Zhao, Y. *et al.* The immunological function of CD52 and its targeting in organ transplantation. *Inflamm. Res.* **66**, 571–578 (2017).

209. Zheng, Y. *et al.* CD86 and CD80 differentially modulate the suppressive function of human regulatory T cells. *J. Immunol.* **172**, 2778–84 (2004).

210. Shigetomi, E., Saito, K., Sano, F. & Koizumi, S. Aberrant Calcium Signals in Reactive Astrocytes: A Key Process in Neurological Disorders. *Int. J. Mol. Sci.* **20**, (2019).

211. Nanut, M. P., Sabotič, J., Jewett, A. & Kos, J. Cysteine Cathepsins as Regulators of the Cytotoxicity of NK and T Cells. *Front. Immunol.* **5**, 616 (2014).

212. Caputi, A., Melzer, S., Michael, M. & Monyer, H. The long and short of GABAergic neurons. *Curr. Opin. Neurobiol.* **23**, 179–186 (2013).

213. Allali-Hassani, A. *et al.* Structural and Chemical Profiling of the Human Cytosolic Sulfotransferases. *PLoS Biol.* **5**, e97 (2007).

214. Mohrmann, R., Dhara, M. & Bruns, D. Complexins: small but capable. *Cell. Mol. Life Sci.* **72**, 4221–4235 (2015).

215. Lu, Z. *et al.* Calsyntenin-3 molecular architecture and interaction with neurexin 1α. *J. Biol. Chem.* **289**, 34530–42 (2014).

216. Wang, Y., Shi, Y. & Wei, H. Calcium Dysregulation in Alzheimer's Disease: A Target for New Drug Development. *J. Alzheimer's Dis. Park.* **7**, (2017).

217. Helassa, N., Antonyuk, S. V, Lian, L.-Y., Haynes, L. P. & Burgoyne, R. D. Biophysical and functional characterization of hippocalcin mutants responsible for human dystonia. *Hum. Mol. Genet.* **26**, 2426–2435 (2017).

218. Zou, F. *et al.* Expression and Function of Tetraspanins and Their Interacting Partners in B Cells. *Front. Immunol.* **9**, 1606 (2018).

219. Hao, L., Du, B. & Xi, X. TRIM59 is a novel potential prognostic biomarker in patients with non-small cell lung cancer: A research based on bioinformatics analysis. *Oncol. Lett.* **14**, 2153–2164 (2017).

220. Huang, Y. *et al.* Decreased expression of zinc-alpha2-glycoprotein in hepatocellular carcinoma associates with poor prognosis. *J. Transl. Med.* **10**, 106 (2012).

221. Franco Bocanegra, D. K., Nicoll, J. A. R. & Boche, D. Innate immunity in Alzheimer's disease: the relevance of animal models? *J. Neural Transm.* **125**, 827–846 (2018).

222. Zschaler, J., Schlorke, D. & Arnhold, J. Differences in innate immune response between man and mouse. *Crit. Rev. Immunol.* **34**, 433–54 (2014).

223. Maruyama, T., Wada, H., Abe, Y. & Niikura, T. Alteration of global protein SUMOylation in neurons and astrocytes in response to Alzheimer's disease-associated insults. *Biochem. Biophys. Res. Commun.* **500**, 470–475 (2018).

224. Yu, Q. *et al.* Overexpression of endophilin A1 exacerbates synaptic alterations in a mouse model of Alzheimer's disease. *Nat. Commun.* **9**, 2968 (2018).

225. Alfadhel, M. *et al.* Asparagine Synthetase Deficiency: New Inborn Errors of Metabolism. in *JIMD reports* **22**, 11–16 (Wiley-Blackwell, 2015).

226. Han, J. *et al.* ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat. Cell Biol.* **15**, 481–90 (2013).

227. Liu, J. *et al.* Brain-Derived Neurotrophic Factor Elevates Activating Transcription Factor 4 (ATF4) in Neurons and Promotes ATF4-Dependent Induction of Sesn2. *Front. Mol. Neurosci.* **11**, 62 (2018).

228. Brown, A. J., Chua, N. K. & Yan, N. The shape of human squalene epoxidase expands the arsenal against cancer. *Nat. Commun.* **10**, 888 (2019).

229. Harada, A. *et al.* Nadrin, a novel neuron-specific GTPase-activating protein involved in regulated exocytosis. *J. Biol. Chem.* **275**, 36885–91 (2000).

230. Doussau, F. The actin cytoskeleton and neurotransmitter release: An overview. *Biochimie* **82**, 353–363 (2000).

231. Haren, L. *et al.* NEDD1-dependent recruitment of the gamma-tubulin ring complex to the centrosome is necessary for centriole duplication and spindle assembly. *J. Cell Biol.* **172**, 505–15 (2006).

232. Venkateswaran, A. *et al.* The novel antiangiogenic VJ115 inhibits the NADH oxidase ENOX1 and cytoskeleton-remodeling proteins. *Invest. New Drugs* **31**, 535–544 (2013).

233. Song, F. *et al.* Regulation and biological role of the peptide/histidine transporter SLC15A3 in Toll-like receptor-mediated inflammatory responses in macrophage. *Cell Death Dis.* **9**, 770 (2018).

234. NIH. SLC6A17 gene. (2019). Available at: https://ghr.nlm.nih.gov/gene/SLC6A17. (Accessed: 16th November 2019)

235. NIH. UNC80 gene. (2019). Available at: https://ghr.nlm.nih.gov/gene/UNC80. (Accessed: 16th November 2019)

236. NIH. SCN8A gene. (2019). Available at: https://ghr.nlm.nih.gov/gene/SCN8A. (Accessed: 16th November 2019)

237. Berto, S., Usui, N., Konopka, G. & Fogel, B. L. ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism. *Hum. Mol. Genet.* **25**, 2451–2464 (2016).

238. Penney, J. *et al.* Loss of Protein Arginine Methyltransferase 8 Alters Synapse

Composition and Function, Resulting in Behavioral Defects. *J. Neurosci.* **37**, 8655–8666 (2017).

239. Goldschmidt, H. L. *et al.* DGKθ Catalytic Activity Is Required for Efficient Recycling of Presynaptic Vesicles at Excitatory Synapses. *Cell Rep.* **14**, 200–207 (2016).

240. Carlini, P. *et al.* Formestane, a steroidal aromatase inhibitor after failure of non-steroidal aromatase inhibitors (anastrozole and letrozole): Is a clinical benefit still achievable? *Ann. Oncol.* **12**, 1539–1543 (2001).

241. Schindler, A. E. *et al.* Classification and pharmacology of progestins. *Maturitas* **46**, 7–16 (2003).

242. Singh, M. & Su, C. Progesterone and neuroprotection. *Horm. Behav.* **63**, 284–90 (2013).

243. De Witt, M. *et al.* Repurposing Mebendazole as a Replacement for Vincristine for the Treatment of Brain Tumors. *Mol. Med.* **23**, 50–56 (2017).

244. Shi, Y., Wang, Y. & Wei, H. Dantrolene : From Malignant Hyperthermia to Alzheimer's Disease. *CNS Neurol. Disord. - Drug Targets* **17**, (2018).

245. Alisky, J. M. Intrathecal corticosteroids might slow Alzheimer's disease progression. *Neuropsychiatr. Dis. Treat.* **4**, 831–3 (2008).

246. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.* **8**, 8868 (2018).

247. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**, 785-801.e8 (2019).

248. Keller, D., Erö, C. & Markram, H. Cell Densities in the Mouse Brain: A Systematic Review. *Front. Neuroanat.* **12**, 83 (2018).

249. Dong, X. Current Strategies for Brain Drug Delivery. *Theranostics* **8**, 1481–1493 (2018).

250. Bellettato, C. M. & Scarpa, M. Possible strategies to cross the blood-brain barrier. *Ital. J. Pediatr.* **44**, 131 (2018).

251. Kanehisa Laboratories. KEGG PATHWAY: Alzheimer disease - Homo sapiens (human). (2018). Available at: https://www.genome.jp/kegg-bin/show_pathway?hsa05010. (Accessed: 21st December 2019)

CHAPTER VII – SUPPLEMENTARY INFORMATION

A. Methods

Table S1 | Summary of R functions, and respective packages, used for the analysis

Package	Functions	Reference
oligo	read.celfiles rma fitProbeLevelModel image NUSE RLE Boxplot	108
readr	read_delim	110
Biobase	exprs	116
AnnotationDbi	select	119
hugene11sttranscriptcluster.db	-	120
illuminaMousev2.db	-	121
pls	stdize	134
FactoMineR	PCA	135
factoextra	get_eigenvalue fviz_eig	136
ggplot/ggplot2	stat_density_2d ggplot	130
stats	shapiro.test cor.test dnorm	106
ggpubr	ggqqplot stat_compare_means ggscatter	129
graphics	smoothScatter plot	106
limma	read.idat neqc lmFit eBayes topTable	109

B. Quality control and sample removal

a. Human dataset

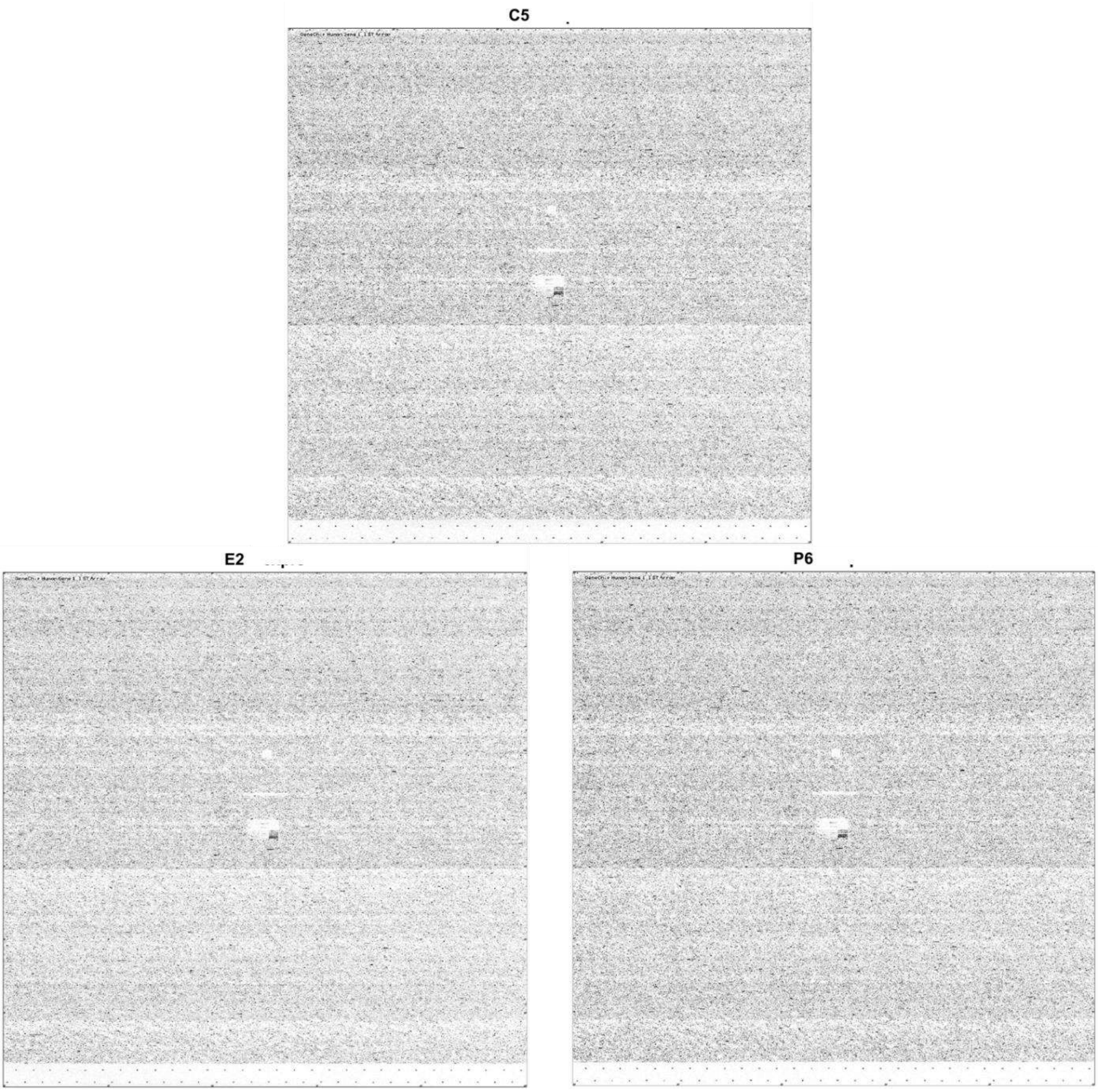


Figure S1 | Raw chip-images (Human)

Raw chip images generated for the human samples C3, E2 and P6

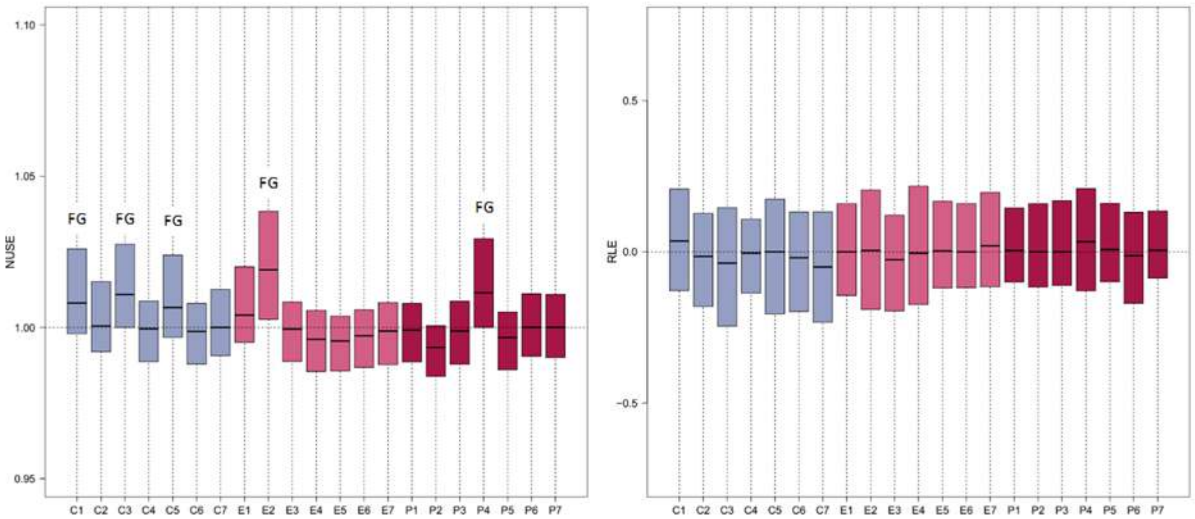


Figure S2 | NUSE and RLE plots (Human)

NUSE (left) and RLE (right) plots after probe level model fitting and before outlier exclusion. FG stands for flagged outlier, signalling samples whose distribution most deviates from 1 and from the remaining samples.

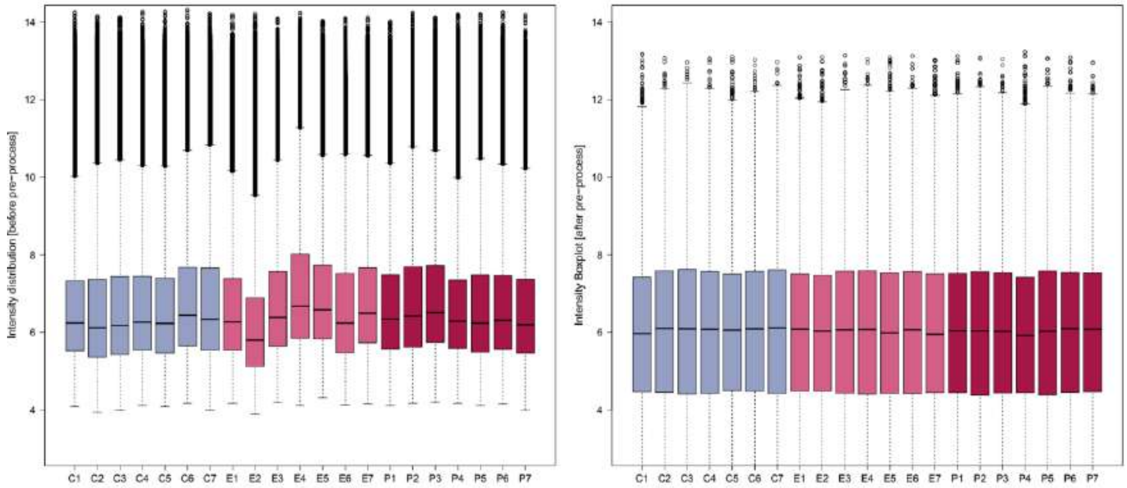


Figure S3 | Intensity distribution boxplots before sample exclusion (Human)

Intensity distribution boxplots prior to (left) and subsequent to (right) data normalization.

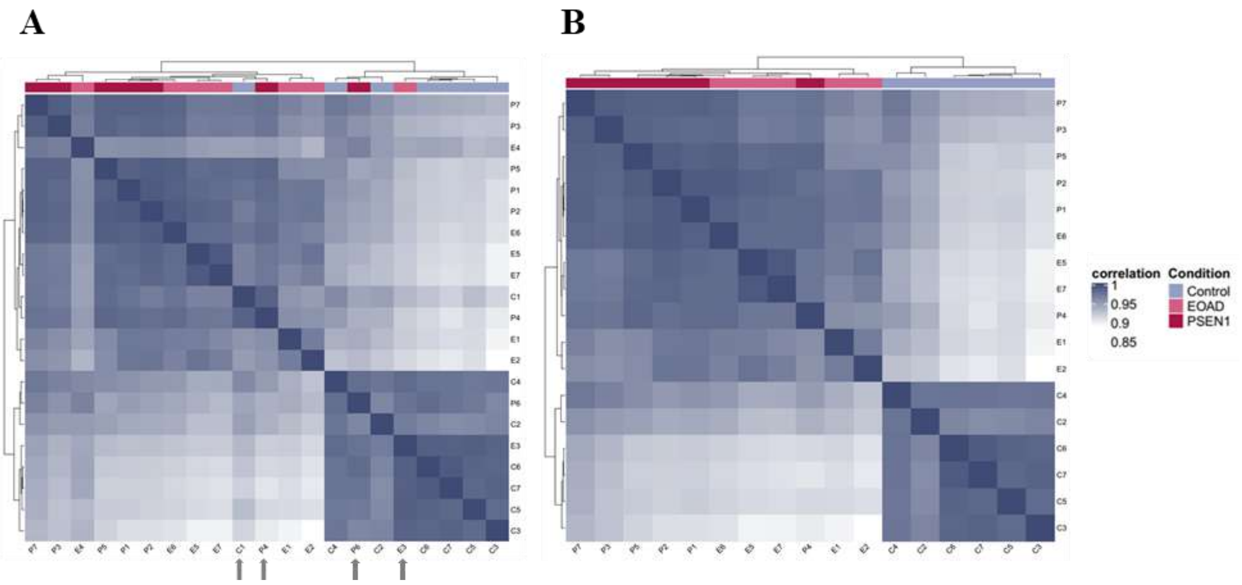


Figure S4 | Heatmap on normalized GE (A) before and (B) after sample exclusion (Human)

Flagged samples before sample removal are marked with an arrow.

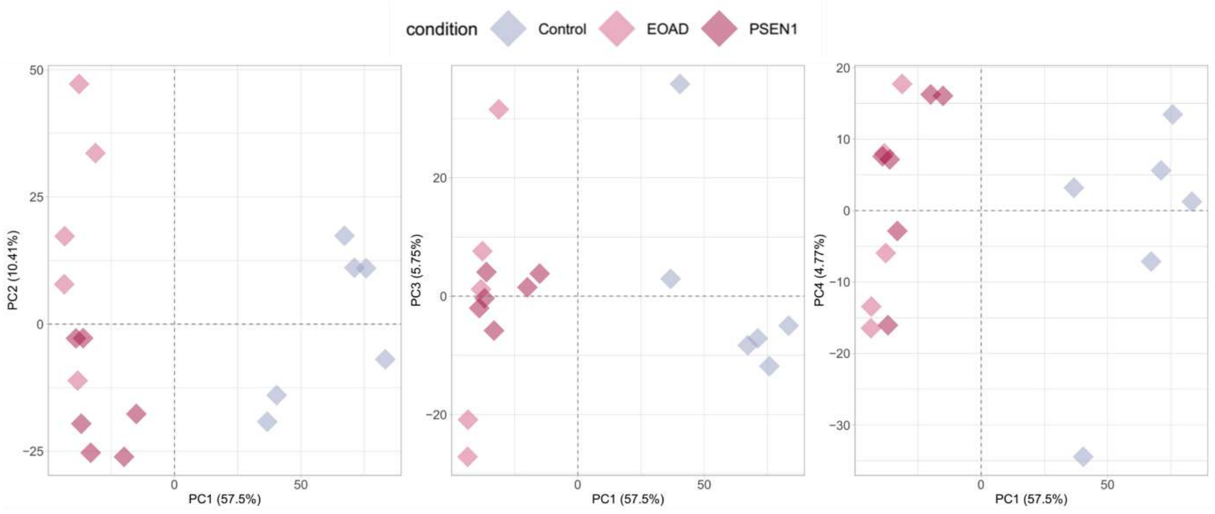


Figure S5 | PCA after sample exclusion coloured by condition (Human)

Principal components of normalized human GE data, with points coloured by AD condition.

b. Mouse dataset

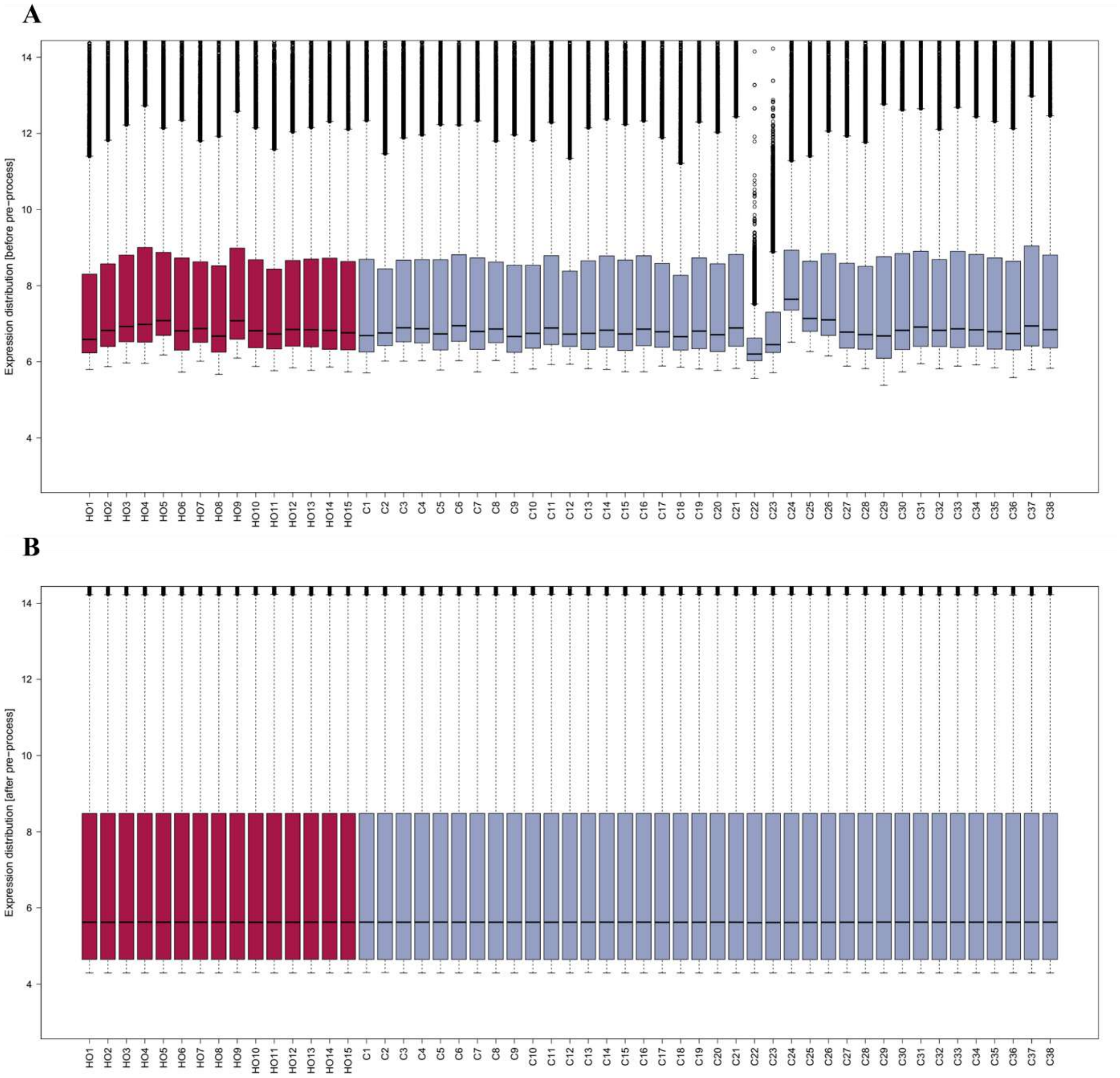


Figure S6 | Intensity distribution boxplots prior to outlier exclusion (HO Mouse)

Intensity distribution boxplots (A) before and (B) after data normalization, before outlier exclusion (HO Mouse).

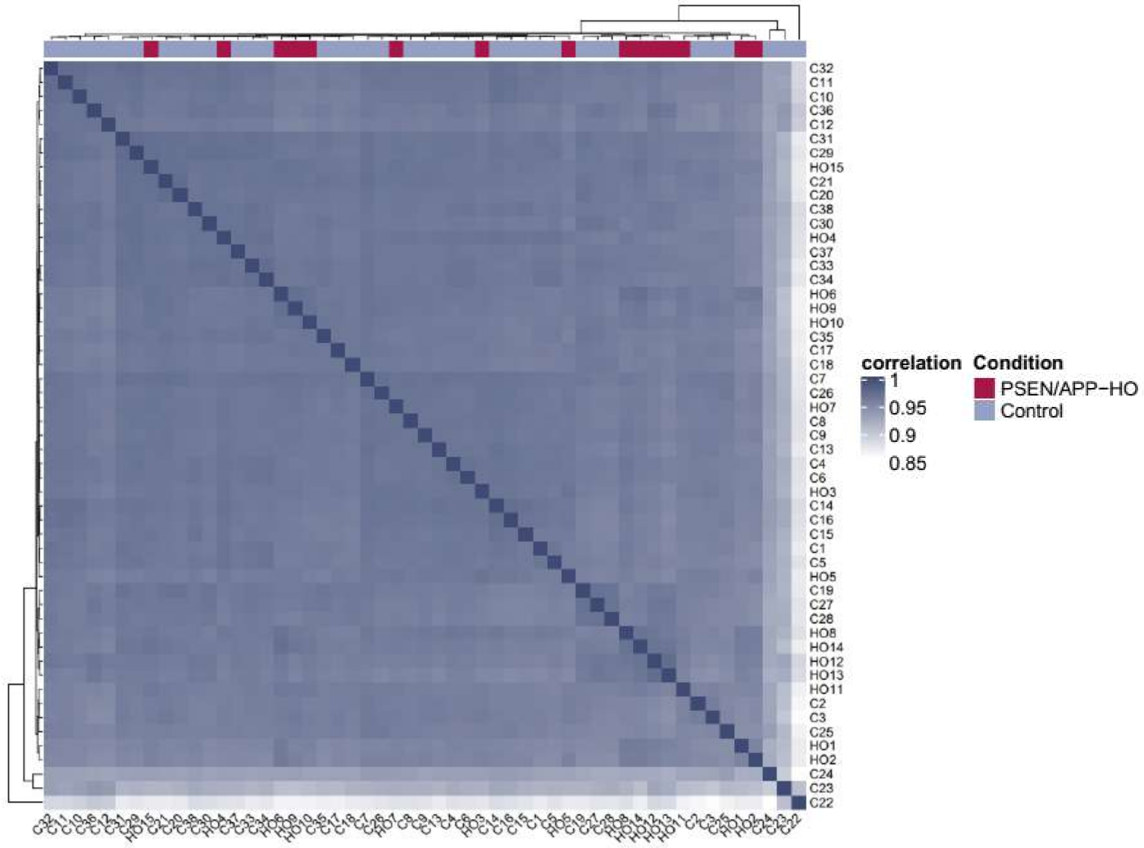


Figure S7 | Heatmap on normalized GE before outlier exclusion (HO Mouse)

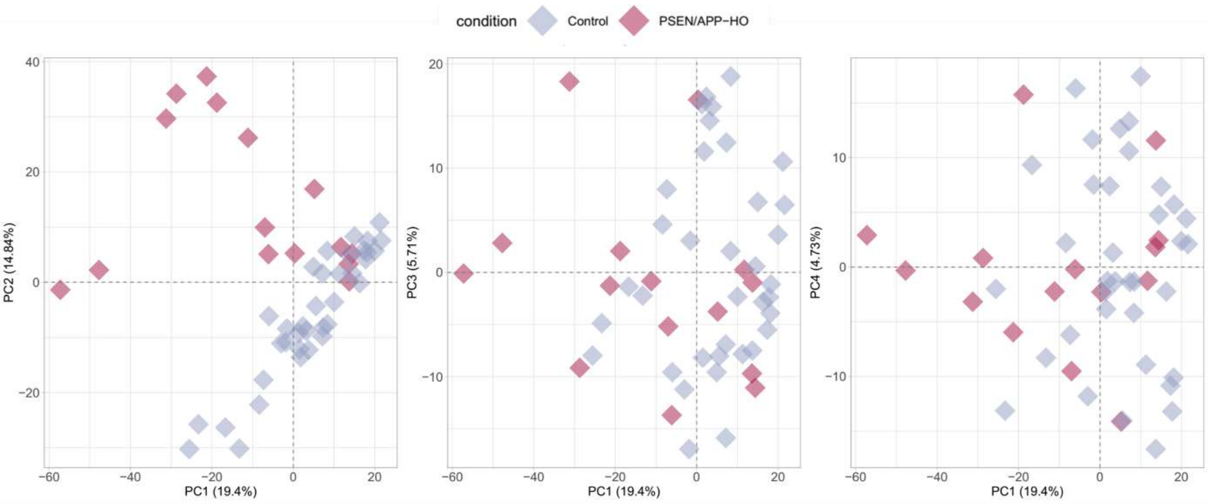


Figure S8 | PCA after outlier exclusion coloured by condition (HO Mouse)

Principal components of normalized mouse GE data, with points coloured by AD condition.

C. Dataset analysis

a. Human dataset

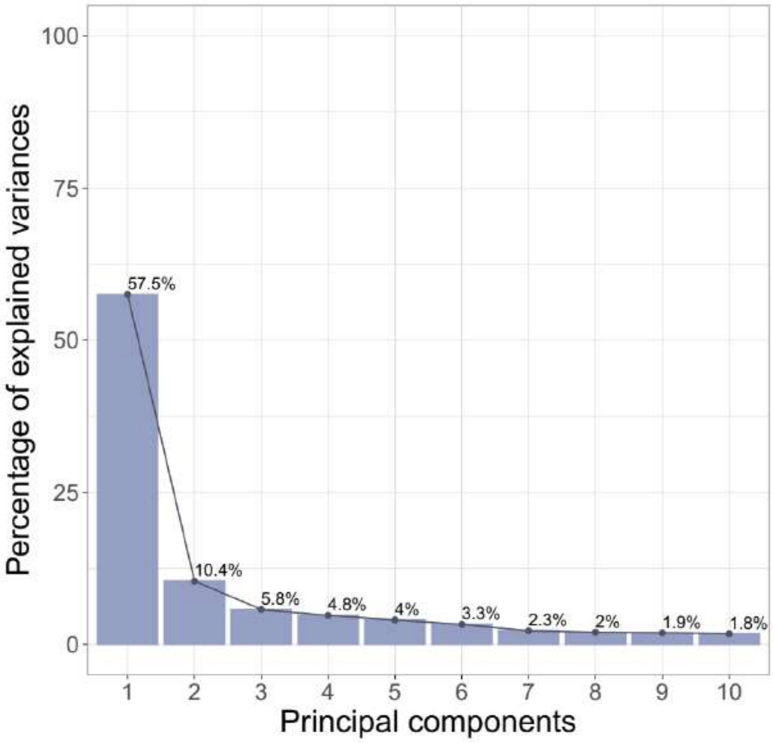


Figure S9 | Percentage of variance explained by each principal component (Human)

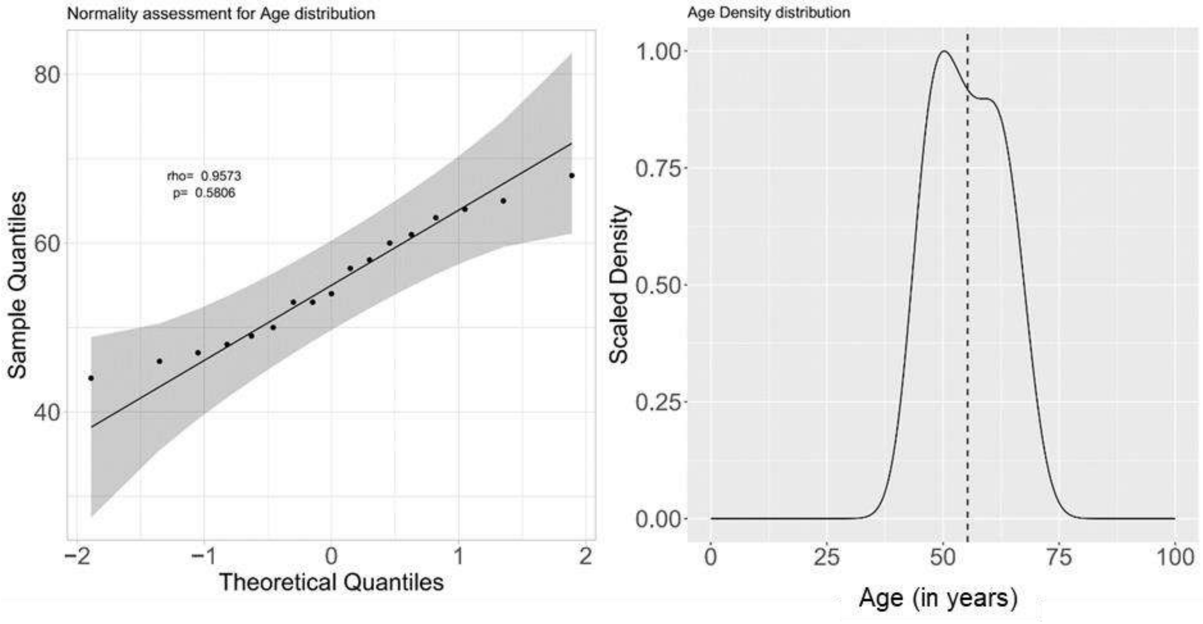


Figure S10 | Normality assessment for age of human samples

Q-Q plot (left) and Age probability distribution (right) plot to assess normality for human age. The dashed line in the distribution plot represents the mean.

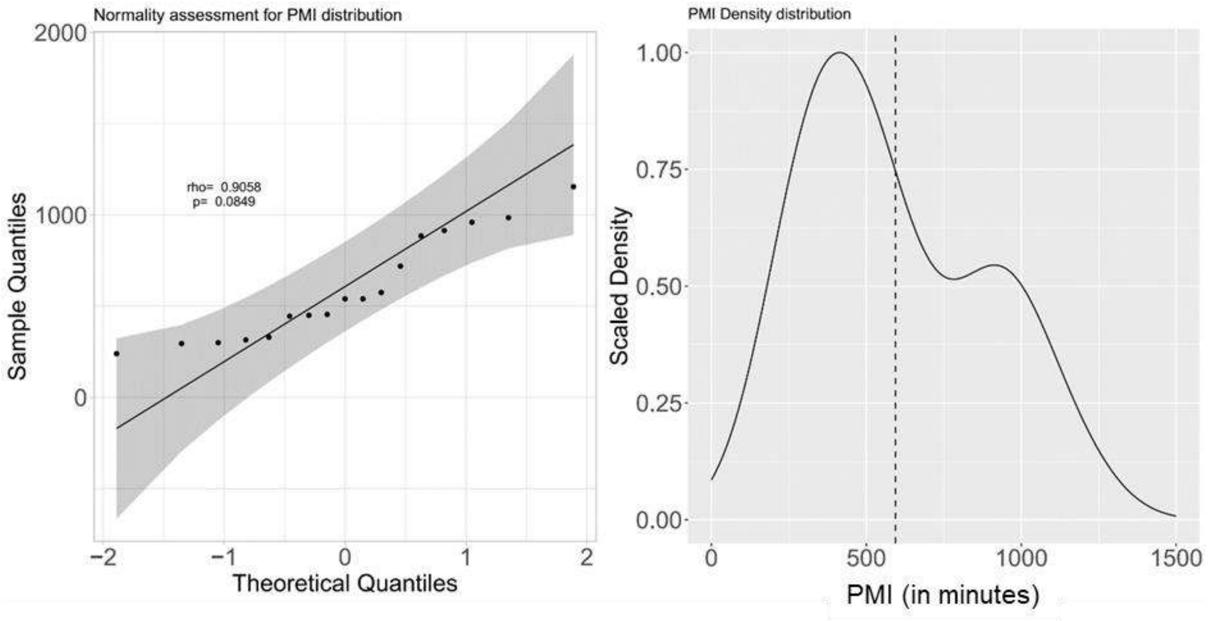


Figure S11 | Normality assessment for PMI of human samples

Q-Q plot (left) and PMI probability distribution (right) plot to assess normality for human PMI. The dashed line in the distribution plot represents the mean.

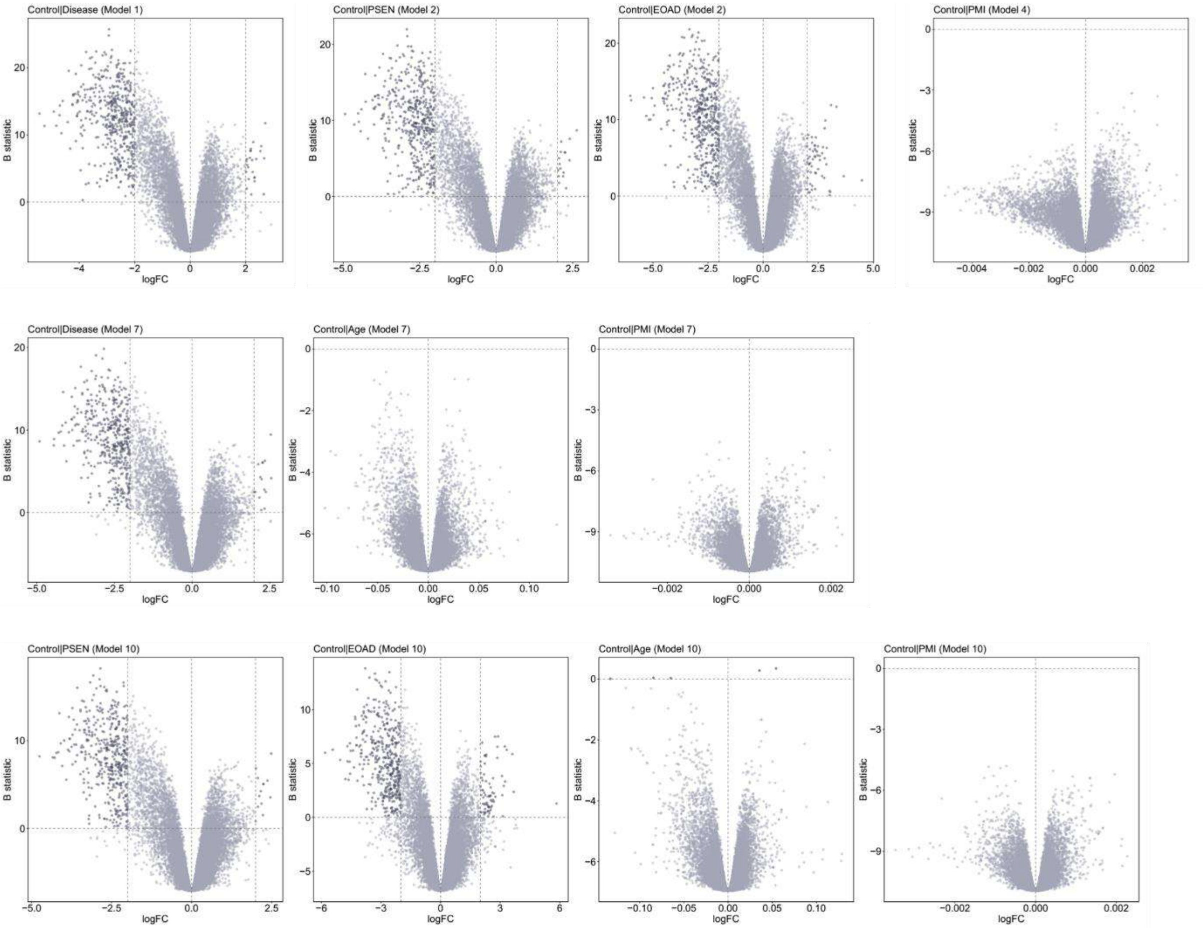


Figure S12 | Volcano plots that supported PMI exclusion from linear models

Volcano plots of differentially expressed, i.e. up- (positive logFC) and down-regulated (negative logFC), genes in human AD patients compared with non-diseased individuals. Thresholds of magnitude (vertical dashed lines) for categorical variables and significance (horizontal dashed line) for both categorical and continuous variables, for the DEG (represented in darker colour) were considered according to section 8.7.3 (logFC > 2; B > 0).

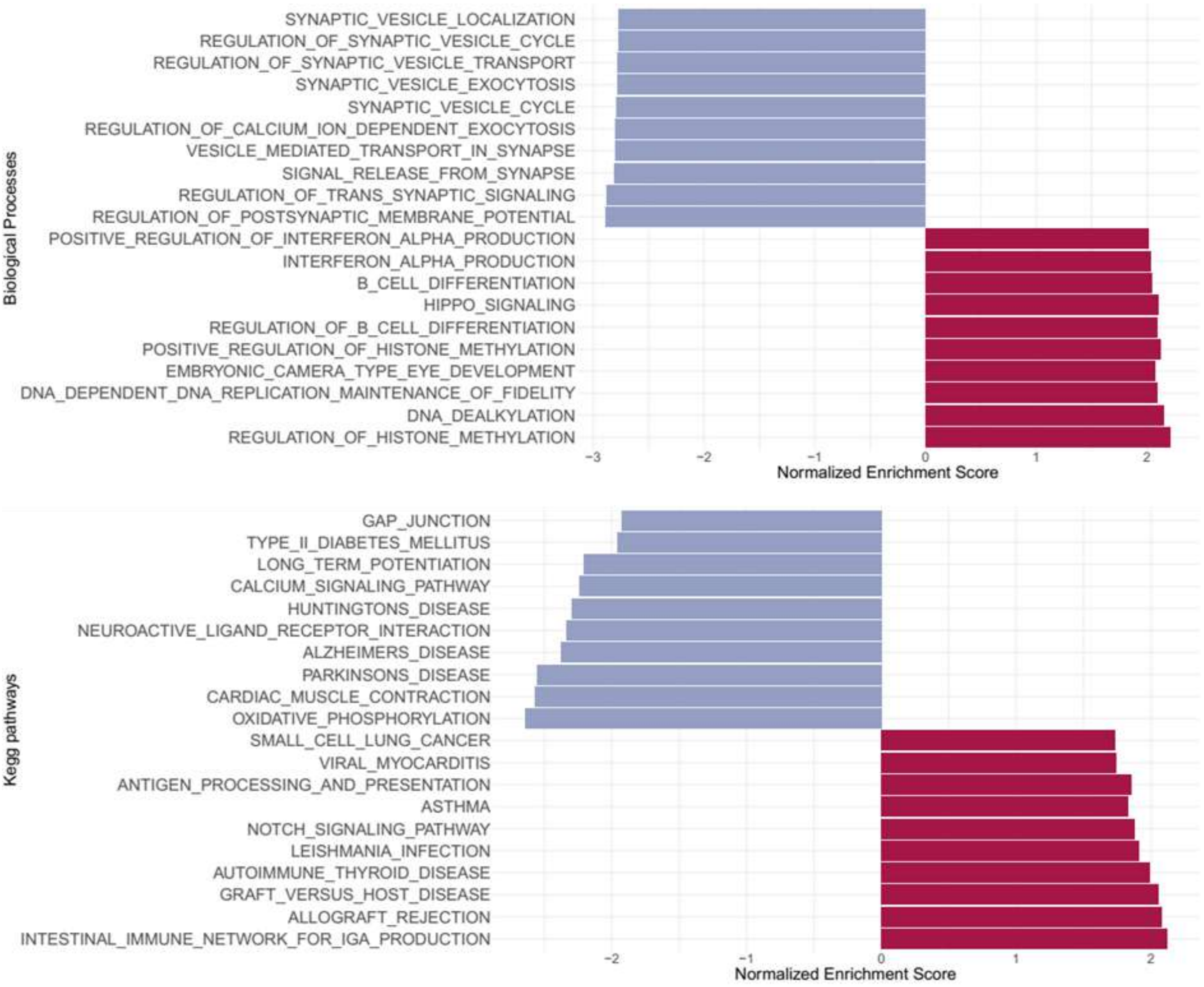


Figure S13 | *PSEN* disrupted Biological Processes and Kegg pathways (Human)

Representation of the 10 most significant down- (blue) and up-regulated (magenta) Biological Processes (top) and Kegg pathways (bottom), based on DEG regarding the *PSEN* coefficient, for the human dataset. Pathways were considered significantly enriched if FDR < 0.05.

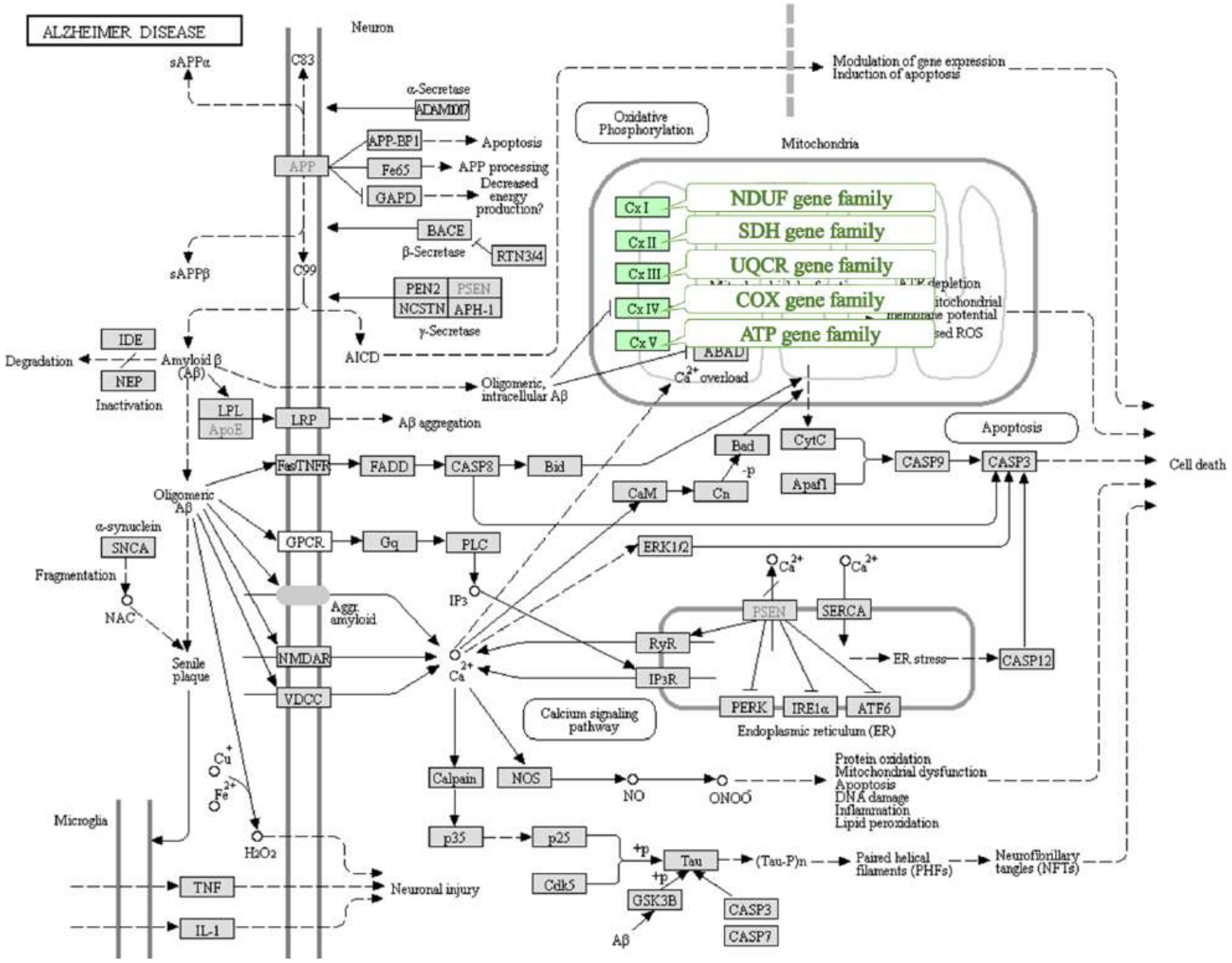


Figure S14 | Alzheimer's disease KEGG pathway

Schematic image representing the AD pathway from KEGG database, with highlight on the NDUF, SDH, UQCR, COX and ATP gene families (in green) that are negatively affected by oligomeric intracellular Aβ – based on the online resource [251].

b. Mouse dataset

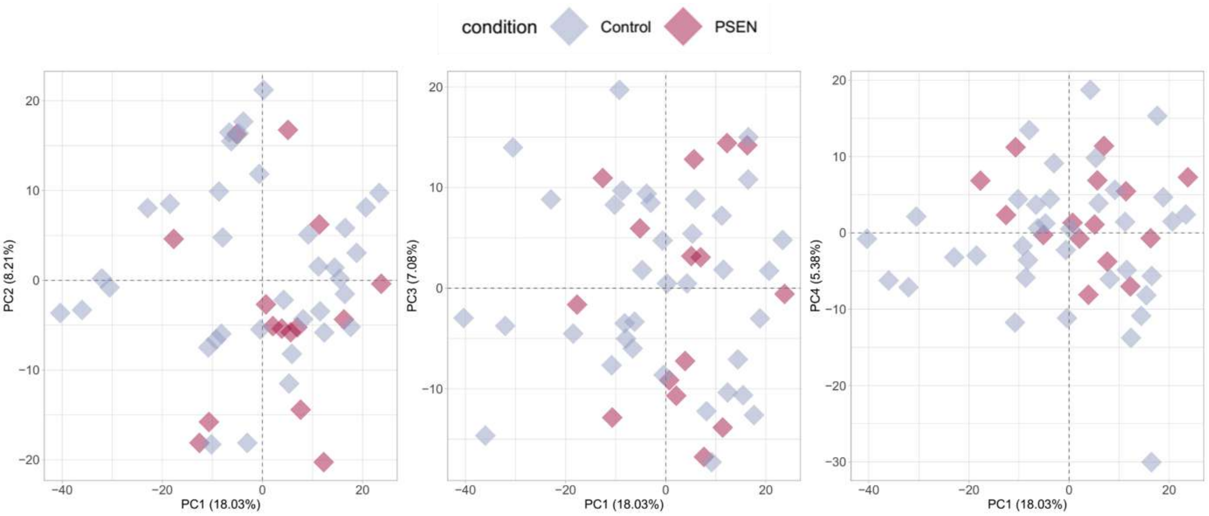


Figure S15 | PCA after outlier exclusion coloured by condition (single-mutant Mouse)

Principal components of normalized single-mutant mouse GE data, with points coloured by AD condition.

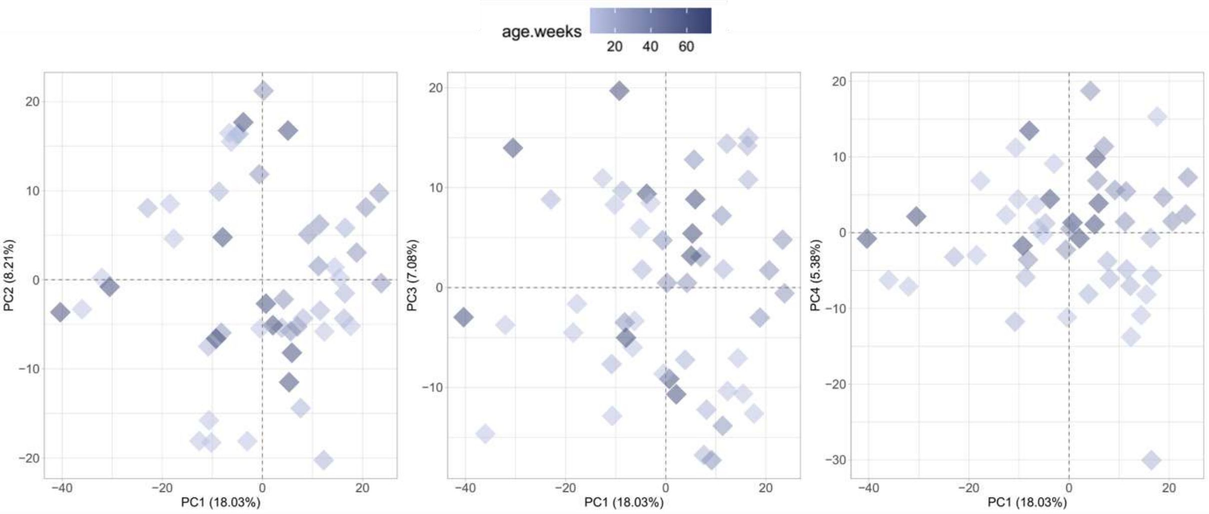


Figure S16 | PCA after outlier exclusion coloured by age (single-mutant Mouse)

Principal components of normalized single-mutant mouse GE data, with points coloured by age.

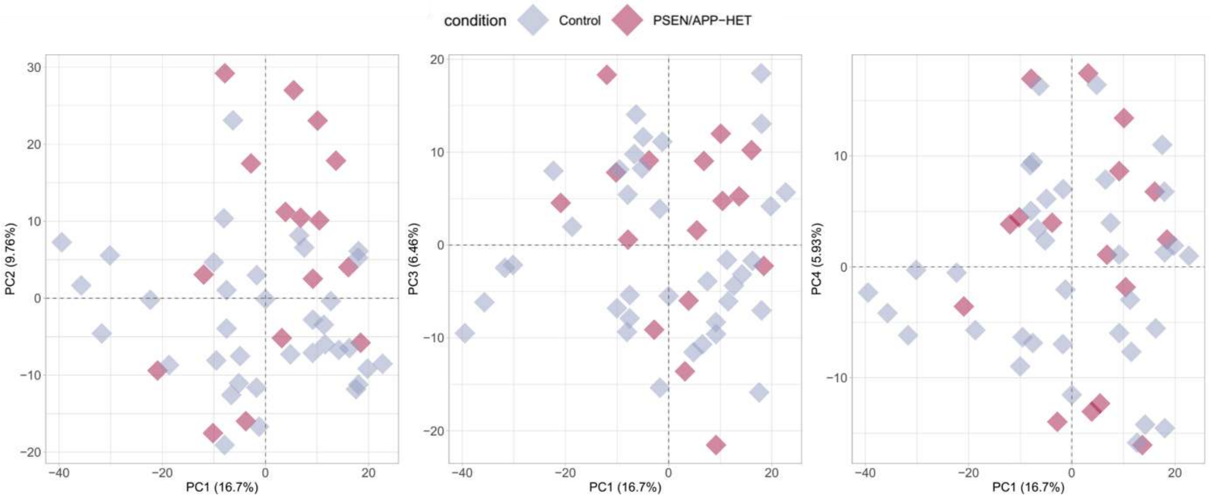


Figure S17 | PCA after outlier exclusion coloured by condition (HET Mouse)

Principal components of normalized HET mouse GE data, with points coloured by AD condition.

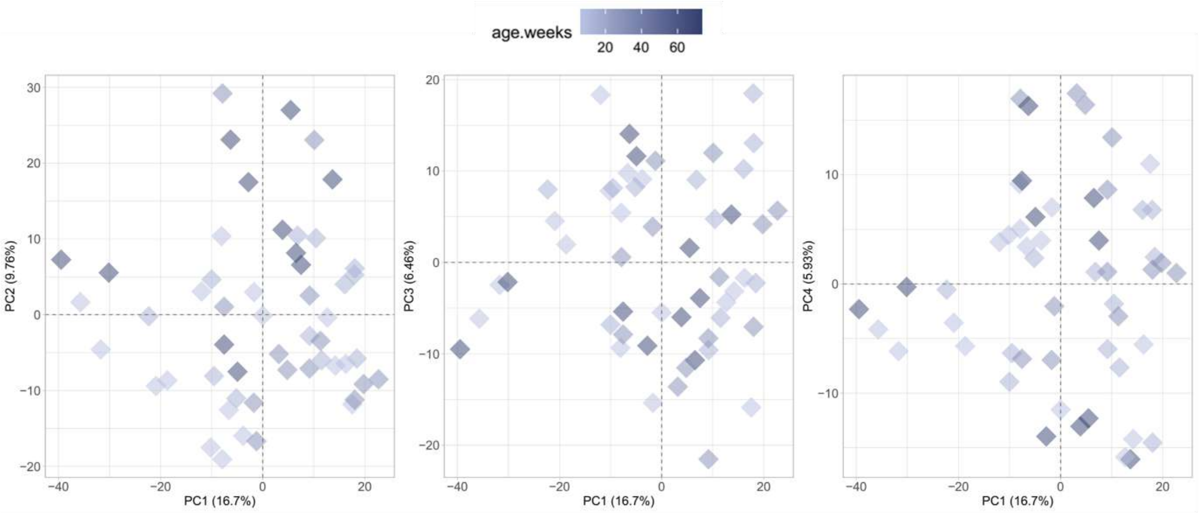


Figure S18 | PCA after outlier exclusion coloured by age (HET Mouse)

Principal components of normalized HET mouse GE data, with points coloured by age.

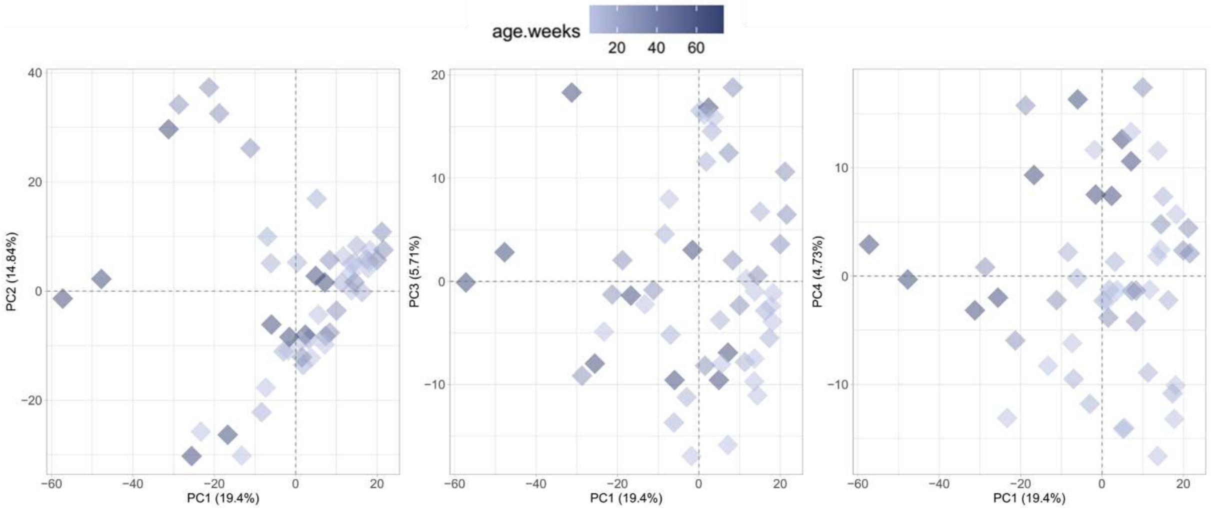


Figure S19 | PCA after outlier exclusion coloured by age (HO Mouse)

Principal components of normalized HO mouse GE data, with points coloured by age.

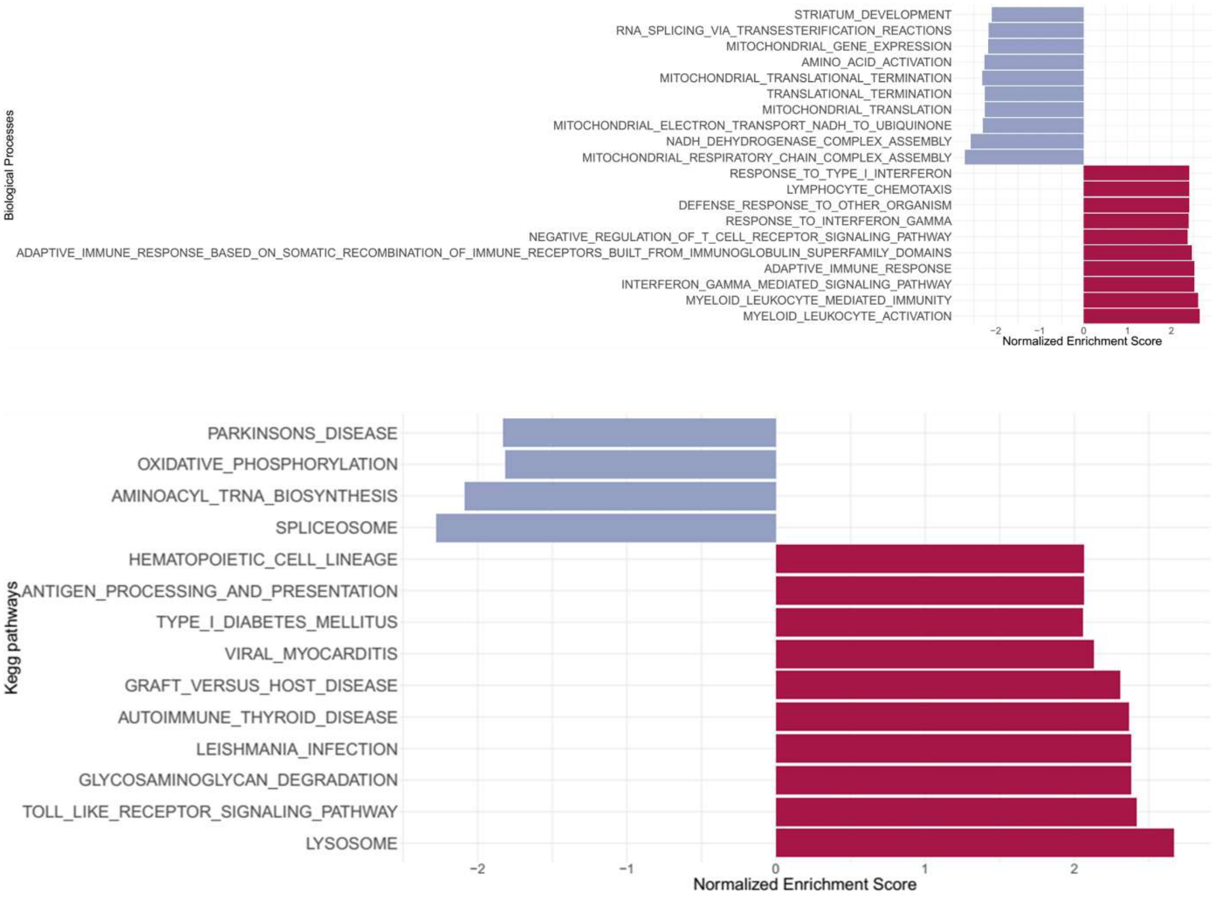


Figure S20 | *PSEN* disrupted Biological Processes and Kegg pathways (Mouse)

Representation of the 10 most significant down- (blue) and up-regulated (magenta) Biological Processes (top) and Kegg pathways (bottom) in mouse *PSEN* brains (double-mutant homozygous dataset). Pathways were considered significantly enriched if FDR < 0.05.

c. Joint dataset

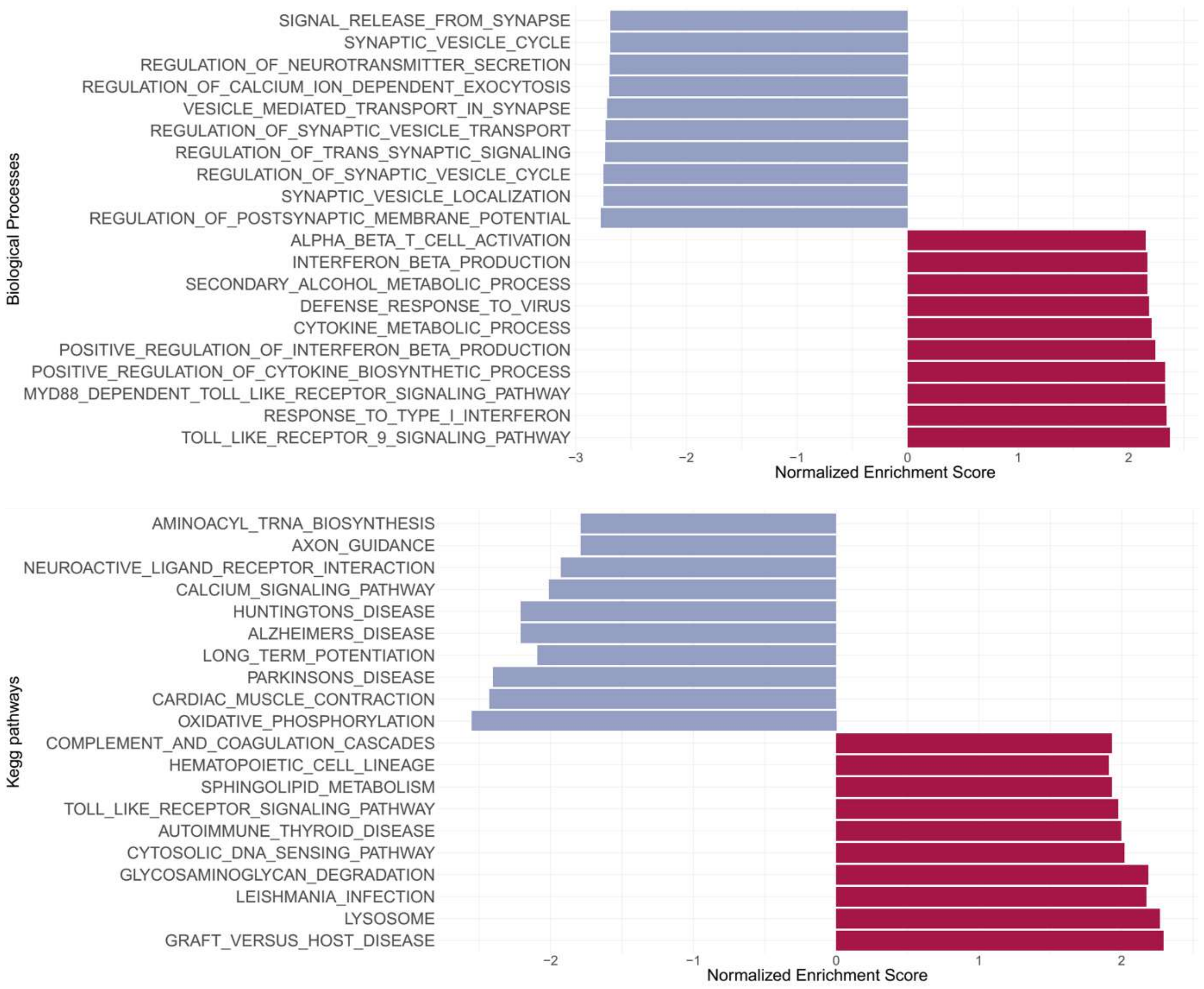


Figure S21 | *PSEN* disrupted Biological Processes and Kegg pathways (Joint dataset)

Representation of the 10 most significant down- (blue) and up-regulated (magenta) Biological Processes (top) and Kegg pathways (bottom) based on DEG regarding the *PSEN* coefficient, for the joint dataset. Pathways were considered significantly enriched if FDR < 0.05.

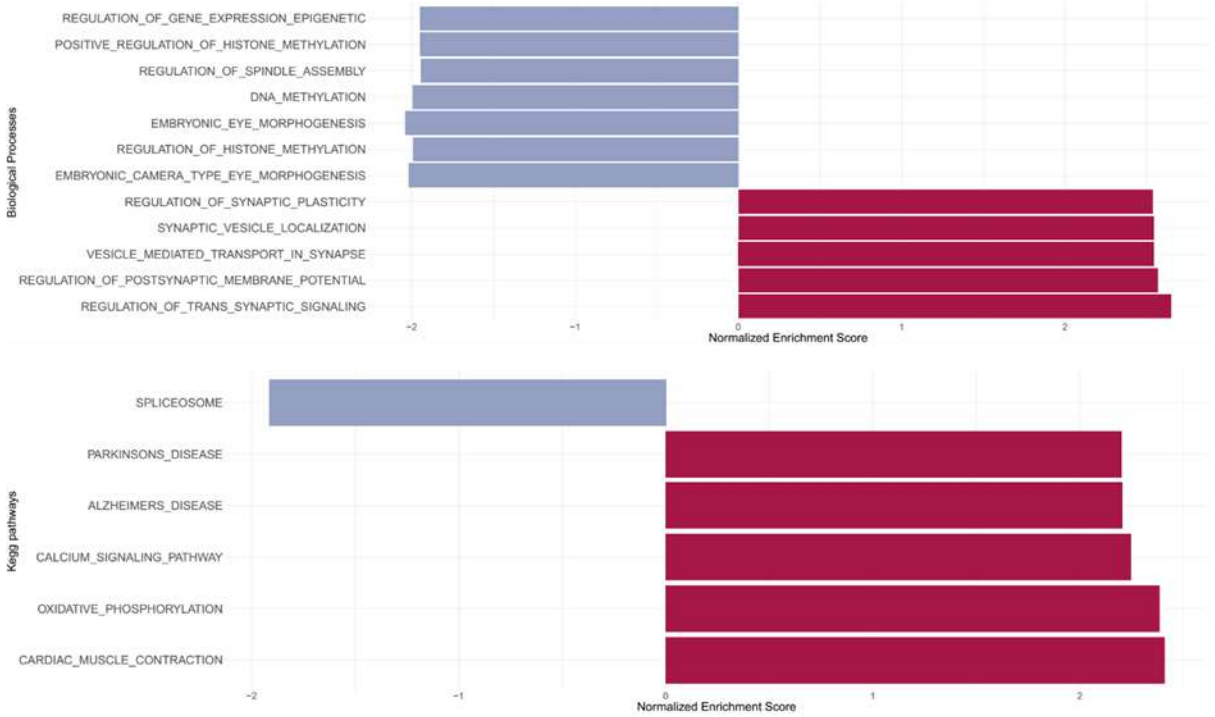


Figure S22 | PSEN/Species *Interaction* disrupted Biological Processes and Kegg pathways (Joint dataset)

Representation of the 10 most significant down- (blue) and up-regulated (magenta) Biological Processes (top) and Kegg pathways (bottom) based on DEG regarding the PSEN/Species *Interaction* coefficient, for the joint dataset. Pathways were considered significantly enriched if FDR < 0.05.