Altered Gene Expression in Excitatory Neurons is Associated with Alzheimer's Disease and Its Higher Incidence in Women

A. Xavier Garcia
National Cancer Institute, National Institutes of Health
Jielin Xu
Cleveland Clinic
Feixiong Cheng
Cleveland Clinic
Eytan Ruppin
National Cancer Institute, National Institutes of Health
Alejandro A. Schäffer (alejandro.schaffer@nih.gov)
National Cancer Institute, National Institutes of Health

Research Article

Keywords:

Posted Date: August 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1953485/v1

Abstract

Background: Alzheimer's disease (AD) is a neurodegenerative disorder involving interactions between different cell types in the brain. Previous single-cell and bulk expression Alzheimer's studies have reported conflicting findings about the key cell types and cellular pathways whose expression is primarily altered in this disease. We reanalyze these data in a uniform, coherent manner aiming to resolve and extend past findings. We further examine if this reanalysis may shed light on the observation that females have substantially higher AD incidence than males.

Methods: We reanalyzed three published single-cell transcriptomics datasets including redoing some of their preprocessing. We used the software method MAST within the Seurat package to look for differentially expressed genes comparing AD cases to matched controls for both sexes together and for each sex separately. We used the GOrilla software to search for enriched pathways and other gene sets among the differentially expressed genes. Motivated by the male/female difference in incidence, we studied genes on the X-chromosome, focusing on genes in the pseudoautosomal region and on genes that are heterogeneous across either individuals or tissues for X-inactivation. We additionally obtained bulk AD datasets from the cortex from the Gene Expression Omnibus and used voom-limma for differential expression analysis.

Results: Our results resolve the apparent contradiction in the literature, showing that comparing AD patients to unaffected controls, excitatory neurons have more differentially expressed genes than do astrocytes and other cell types. Altered pathways between males and females in excitatory neurons involve synaptic transmission and related pathways. Further analysis suggests that numerous pseudoautosomal region (PAR) genes and X-chromosome heterogeneous genes, including such as *BEX1 and ELK1,* may contribute to the difference in sex incidence of Alzheimer's disease. Among autosomal genes, *GRIN1*, stood out as an overexpressed gene in cases vs. controls in all three single-cell data sets and as a functional candidate gene contributing to several important pathways upregulated in cases.

Conclusions: Taken together, these results point to a potential linkage between two longstanding questions concerning AD pathogenesis, involving which cell type is the most important and why females have higher incidence than males.

Background

Alzheimer's disease (AD) is a progressive disease where neurons in parts of the brain involved in thinking, learning and memory become damaged. In 2021, there were an estimated 6.2 million Americans aged 65 and older living with AD [1]. Alzheimer's is a complex disease to treat due to the involvement of many cell types, such as neurons, glial cells, and immune cells. Here, we aimed to analyze published datasets of single-cell and bulk gene expression in AD to learn more about two key basic questions regarding its pathogenesis: (1) what are the cell type specific transcriptional alterations that are associated with it pathogenesis and progression and which cell-type is most transcriptionally altered? And (2), given the considerable sex bias observed in the incidence of this disease [2], what are the underlying most notable sex-specific cell-type-specific transcriptional alterations?

There are at least two current theories about the pathogenesis of AD that focus on specific cell types. One theory proposed by Frere and Slutsky suggests that a core homeostatic machinery maintains the stability of central neural circuits, safeguarding from neurodegeneration, and when this machinery collapses, neurons develop a firing instability and impaired synaptic plasticity that leads to the neurodegeneration seen in AD and other diseases [3]. This theory thus suggests that the most important cells to study are neurons of various types. The second theory proposed by Michal Schwartz and colleagues suggests that AD has an immune-related etiology. Her lab has shown that the brain-immune crosstalk is impaired in neurodegenerative diseases, which could explain the loss of neuronal maintenance and support, and could potentially be rejuvenated by the immune system [4]. According to the latter theory, the most important cells to study are microglia.

There have been a few previous early attempts to utilize laser capture and next-generation sequencing technology to investigate the role of different brain cell types in AD. For example, previous work showed that expression data from neurons were significantly more accurate in predicting AD severity than comparable expression data from astrocytes and whole tissue

[5, 6]. Naturally, recent advances in single-cell technologies now enable a much better understanding of the roles and relative importance of different cell types in the pathogenesis of AD.

To better understand the transcriptional changes in cell types and the role they play in AD, Mathys et al. published the first single-cell transcriptomics dataset obtained from Alzheimer's patients and controls. They found that all cell types in the prefrontal cortex had transcriptional changes associated with AD [7]. Interestingly, they found neurons to have more differentially expressed genes (DEGs) downregulated while other cell types such as astrocytes and microglia, had more genes upregulated, in accordance with the central role of neurons proposed by Frere and Slutstky. While they found excitatory neurons to have the most DEGs, they did not pursue this finding further and appropriately indicated that their finding may be due to having sampled more excitatory neurons than cells of other types [7]. However, in a second single-cell Alzheimer's dataset (GSE157827, which we refer to as the Lau dataset), the authors found, apparently contrary to Mathys et al. findings, that astrocytes have more DEGs compared to the other cell types [8]. Lau et al. reported that their findings indicate that both endothelial and glial cells are key cell types that could be therapeutically targeted, which is more in accordance with the ideas of Schwartz et al. that non-neuronal brain cells play a major role in Alzheimer's pathogenesis.

Aiming to resolve this apparent quandary, we set out to reanalyze the Mathys and Lau datasets using consistent methods, partly different from the original publications. Furthermore, we analyzed a third single-cell transcriptomics Alzheimer's dataset that obtained samples collected from Alzheimer's patients with or without a *TREM2* variant [9]. Since *TREM2* is an immune gene functioning in microglia, this dataset was collected primarily to better understand the potential importance of microglia, as proposed by Schwartz et al., although neurons and other cell types were also studied.

In addition to the relative importance of different brain cell types, another longstanding puzzle about Alzheimer's pathogenesis is the observation of its substantially higher incidence in females than in males [2]. The biology and mechanisms of the sex differences in AD incidence are generally poorly understood even after a recent major mouse study targeting this question [10]. Indeed, Mathys et al. raised the possibility that their single-cell data could be used to study sex differences. Their sex-specific analysis concluded that there are subtle differences in the transcriptomics between sexes in each cell type, with neurons and oligodendrocytes having the most extreme differences [7]. However, they did not look at specific genes or enriched pathways in a sex-stratified way.

To study the observed sex differences at the single-cell resolution, Belonwu and colleagues reused the Mathys dataset to perform a sex-stratified analysis to identify sex-stratified cell type-specific perturbations in Alzheimer's patients. They found that neurons were more similar between males and females compared to glial cells, having more shared genes and pathways [11]. Belonwu et al. focused their analysis on three cell types (neurons, astrocytes, and microglia) but did not investigate more closely excitatory neurons beyond reporting that excitatory neurons have more DEGs than other cell types. Due to a technical flaw in their analysis that is explained below in the Results and Materials and Methods sections, their analysis was inadvertently limited to fewer than 200 genes. We instead performed a sex-stratified analysis using the differential expression method that was used to resolve the cell type contradiction that was previously discussed.

To validate the sex specific single-cell findings emerging from our analysis, we additionally analyzed three bulk expression datasets from the Gene Expression Omnibus (GEO; Materials and Methods). Notably, a recent paper studied the role of sex differences at the transcriptome level and how it influences complex traits analyzed in the most recent (v8) version of the Genotype-Tissue Expression (GTEx) v8 project [12]. Oliva and colleagues identified sets of sex-biased genes (genes with expression levels that differ significantly between males and females) for dozens of different tissues, which we further considered in our analysis.

To make our analysis of sex differences self-contained, we introduce some standard terminology about human X-chromosome genes. Near the Xp telomere is the pseudoautosomal region (PAR) containing a handful of genes such that females have and express two copies on X, while males have and express one copy on X and one or more copies on Y. The gene copies on X and Y do not recombine and hence can diverge in evolution [13]. One X-chromosome PAR gene whose importance in Alzheimer's has been recognized recently is *IL3RA* [14]. Among the non-PAR X-chromosome genes, most have

their expression between males and females somewhat balanced by the mechanism of X-chromosome inactivation in females [15], which is primarily regulated by the RNA gene *XIST* [16]. We find it useful to partition the non-PAR genes into three categories according to their X-inactivation status in females: always inactivated, always escaping X-inactivation, and heterogeneous with respect to X-inactivation. Heterogeneity of X-inactivation may be across individuals and/or across tissues. For this partition, we used a published gene classification [17].

In summary, here we consistently analyze three single-cell datasets to investigate the association of brain cell-types expression with the progression of AD, and further learn which cell-type specific pathways are enriched in AD. We also performed a sex-stratified enrichment analysis of DEGs in all three single-cell datasets to identify differences in pathways that may contribute to the observed sex bias in AD. We complement this sex-stratified analysis of single-cell data with additional in-depth analysis of three existing bulk RNA datasets from the cortex to reproduce findings from single-cell and bulk data. Finally, we investigate the roles X-chromosome genes may play in the sex bias of AD.

Materials And Methods

Single-cell expression analysis

Datasets

Syn18485175 (the Mathys dataset) contains 80,660 droplet-based single-nucleus RNA-seq transcriptomes profiled from the prefrontal cortex of 48 Religious Orders Study and Memory and Aging Project (ROSMAP) participants. Patients were classified as having high levels of B-amyloid and other pathological hallmarks of Alzheimer's or very low B-amyloid burden or other pathological hallmarks. Overall, there were 24 Alzheimer's patients and 24 controls that were sex- and age-matched.

GSE157827 (the Lau dataset) contains 169,496 single-nucleus RNA-seq transcriptomes profiled from the prefrontal cortex of 21 South West Dementia Brain Bank participants. Samples were classified as Alzheimer's patients or controls based on the Braak stage. Overall, there were 12 Alzheimer's patients and 9 controls that were sex- and age-matched.

Syn21670836 (the TREM2 dataset) contains 73,419 single-nucleus RNA-seq transcriptomes profiles from the prefrontal cortex of 42 Rush participants. Samples were classified as Alzheimer's patients with "TREM2-CV", Alzheimer's patients with "TREM2-R62H", and controls. Overall, there were 22 Alzheimer's patients and 11 controls that were sex- and age-matched with the "TREM2-CV" group.

Data preprocessing

To prepare the single-cell datasets for differential expression analysis, code was used that was provided by Belonwu et al. to preprocess Mathys and TREM2 datasets. We followed the "Methods" section for the publication of GSE157827 [8] to preprocess the Lau dataset.

For each of Mathys and TREM2 dataset, a Seurat object was created with the raw count data and included genes that were expressed in 3 or more cells, and cells with at least 200 detected genes. Normalization was performed with the default arguments in Seurat NormalizeData function. Each cell was identified by cell type using the meta-data provided by the authors.

For the Lau dataset, a Seurat object was created with the raw count data and included genes that were expressed in 5 or more cells, and cells with at least 200 detected genes. PercentageFeatureSet function was used with pattern "^MT-" to obtain the percentage of transcripts that map to mitochondrial genes. Samples were kept if they contained greater than 200 genes, less than 20,000 unique molecular identifiers, or more than 20 percent of mitochondrial genes. Normalization was performed with the default arguments in the NormalizeData function. Highly variable features was identified using FindVariableFeatures with arguments selection.method set to "vst" and nfeatures set to "1000". To integrate the 21 samples, FindIntegrationAnchors was used with argument dims set to "1:20" followed by IntegrateData with argument dims set to "1:20". Since cell-type annotation

was not provided, we performed FindClusters function with argument resolution set to “1” and random.seed set to “1”. To identify each cluster cell type, FindAllMarkers was used with argument logfc.threshold set to “0.25” and test.use set to “wilcox”. Using the cell-type markers provided by the authors, each cluster was identified as either astrocyte, endothelial cell, excitatory neuron, inhibitory neuron, microglia, and oligodendrocytes based on which cell type had more markers identified in the upregulated genes in each cluster.

Choice of MAST, as implemented in Seurat, as the analysis method for single-cell analysis

Since the data processing done by Belonwu et al. was through a Seurat object and the Seurat package [18] is widely used for single-cell transcriptomics data analysis, we decided to use the Belonwu et al. code (for steps before the overaggressive filtering) and to continue with other Seurat options for the revised differential expression analysis. Seurat’s *FindMarkers* function provides a way to identify DEGs between clusters while allowing for pre-filtering. In addition, Seurat also supports a variety of differential expression tests such as bimod, poisson, wilcox, MAST, and DESeq2 [18-21]. A large-scale comparison study of DE analysis methods found MAST to be the best-performing single-cell DE test [22]. We found MAST to be a widely used tool in the community with Google Scholar reporting 1,089 citations as of December 9th, 2021. In addition, MAST is one of four integrated packages that allows for the input of covariates within Seurat’s *FindMarkers* function. For these reasons, we decided to use MAST for our single-cell differential expression analysis.

Seurat’s *FindMarkers* function was used to perform the differential expression analysis with the default argument min.pct to filter out genes that were not detected at a minimum fraction of 0.1 in either Alzheimer’s or Controls. To control for sex and age, these covariates were passed into the argument “latent.vars”.

Down sampling of cell subsets for analysis of DEGs in single-cell data

The function subset was used to partition the Seurat object by cell type. The cell type with the least number of cells was identified and half of this amount was used to produce equal cell count across the cell types. The function replicate was used to perform the function sample 100 times to produce random numbers between one and the total number of cells per cell type with argument replace set to “F”. The random numbers were then used to further divide each Seurat object cell type to produce 100 replicates with same cell count across cell type. Seurat’s *FindMarkers* function was used to perform the differential expression analysis with the default argument min.pct to filter out genes that were not detected at a minimum fraction of 0.1 in either Alzheimer’s or Controls. To control for sex and age, these covariates were passed into the argument “latent.vars”. The mean total number of DEGs with an adjusted p-value less than 0.05 across the 100 replicates was obtained using function rowMeans and the standard deviation was obtained using function sd.

LogFC threshold variation test

After performing the differential expression analysis, the MAST output was filtered to only include genes with an absolute logFC greater than: 0, 0.05, 0.10, 0.15, 0.20, and 0.25.

Gene set enrichment analysis

We performed all gene set enrichment analysis using the web-based application GOrilla [23] with running mode set to “Single ranked list of genes”. Genes were ranked by logFC value and pasted into the text box. P-value threshold was set to “10^-3” and “Run GOrilla in fast mode” was unchecked.

TREM2 Ex cluster analysis

In Mathys and TREM2 datasets, Seurat’s *FindMarkers* function was used to perform the differential expression analysis between Ex and Ast clusters with the default argument min.pct to filter out genes that were not detected at a minimum fraction of 0.1. To control for condition, sex, and age, these covariates were passed into the argument “latent.vars”. The MAST output was filtered to only include genes with a logFC greater than 2.

In TREM2 dataset, Seurat's *FindMarkers* function was used to perform the differential expression analysis between Ex0 and Ex1 clusters with the default argument min.pct to filter out genes that were not detected at a minimum fraction of 0.1. To control for condition, sex, and age, these covariates were passed into the argument "latent.vars".

Bulk expression analysis

Datasets

Three datasets were obtained from Gene Expression Omnibus using GEO2R for the analysis on Alzheimer's samples [24]. These datasets are: GSE15222 [25]), GSE33000 [26], and GSE44770 [27].

GSE15222 contains samples obtained from 20 National Alzheimer's Coordinating Center (NACC) brain banks and from the Miami Brain Bank. Samples were classified as either no neuropathology present or neuropathologically confirmed late onstage AD. The diagnosis was defined by board-certified neuropathologist per standard NACC protocols. After removing samples that did not meet the study's criteria, the total samples are 363: 187 controls and 176 affected individuals. These samples cRNA were hybridized to Illumina Human Refseq-8 Expression BeadChip via standard protocols.

GSE33000 contains samples obtained from Harvard Brain Tissue Resource Center. Samples were classified as either no neuropathology present or AD patients. AD patients were phenotyped based on Braak stage, specific regional atrophy on a gross and microscopic scale, and ventricular enlargement. There are 467 samples: 157 controls and 310 affected individuals. These samples were profiled on a custom-made Agilent 44K array of 40,638 DNA probes.

GSE44770 contains samples obtained from the Harvard Brain Tissue Resource Center. Samples were classified as either normal non-demented subjects or subjects with late onstage AD. Braak stage, general and regional atrophy, gray and white matter atrophy and ventricular enlargement were used to confirm the Alzheimer's diagnosis. There are 230 samples: 101 controls and 129 affected individuals. These samples were hybridized to a custom microarray manufactured by Agilent Technologies consisting of 4,720 control probes and 39,579 probes targeting transcripts.

GTEx v8 data was downloaded via AnVIL (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2) after approval via dbGaP. The protected data was requested to preprocess the GTEx dataset using Oliva code (https://zenodo.org/record/3939042#.X05qnVVKguU). The GTEx v8 data contains a total of 54 tissues from 948 donors for a total of 17,382 samples [28]. The GTEx dataset is provided in two files: gene expression and the metadata.

GTEx data preprocessing

To prepare the GTEx data to run the existing preprocessing code, generate_input_data.R, from (Oliva et al. 2020), the protected data had to be substituted for the placeholder anonymized files provided in the paper supplementary data. Our code partitions the GTEx metadata by tissue, and only includes metadata on rna integrity number (RIN), total ischemic time, age, and sex. The code provided by Oliva and colleagues, generate_input_data.R, was then used on the GTEx gene expression dataset to produce samples and expression for each tissue. With the protected files in the correct folder, Oliva code, generate_surrogate_vars.R, was used to produce the surrogate variables for each tissue, based on RIN, total ischemic time, age, and sex.

GTEx differential expression analysis

To perform the differential expression analysis, the Oliva et al. code, perform_DE.R, was adapted. For each tissue, the data were normalized and rescaled using edgeR function calcNormFactors with method "TMM" [29]. The covariates were merged with the surrogate factors to produce the design matrix. The linear model used for this analysis was ~SMTSISCH+SMRIN+SEX+AGE+SV(n). Limma was used for the differential expression analysis, using empirical Bayes statistics for differential expression with the "trend" argument set to FALSE [30]. A modification made to Oliva code, perform_DE.R, was to save volcano plots using the package EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano). EnhancedVolcano arguments were adjusted to show genes, with an adjusted p-value less than 0.05 and absolute fold change value greater than zero, to be statistically significant.

Microarray data preprocessing

To prepare the microarray datasets obtained from GEO for differential expression analysis, code was used that was provided by Noori et al. to preprocess microarray datasets. Each dataset was normalized by Robust Multichip Average (RMA) using the package oligo [31]. For quality control, the package arrayQualityMetrics was used to identify any outliers and remove them if needed [32]. The probes were renamed to their gene symbol that was provided by getGEO with the argument “AnnotGPL” set to TRUE [24]. If the probes did not contain a gene symbol, these probes were removed from the analysis. The genes were then filtered to include only genes that were found to be sex-biased in an analysis of GTEx v8 [12] .

X-chromosome heterogeneous genes differential expression analysis

Before running the differential expression analysis, we extracted the rows of the expression matrix to include only those X-chromosome genes that were identified as heterogeneous in X inactivation. The covariates for the differential expression included sex, age, and condition. The linear model used for this analysis was ~SEX+AGE+CONDITION. Limma was used for the differential expression analysis, using empirical Bayes statistics for differential expression with the “trend" argument set to FALSE [30].

X-chromosome heterogeneous genes meta-analysis

After obtaining the differential expression results on each dataset, meta-analysis was performed to identify the significant genes across all three datasets. The results obtained per dataset were merged into one data frame and genes that were not found across the three datasets nor found to be significant across all three datasets were not included in subsequent steps. Using the package metap, the unadjusted p-values for each gene from each dataset were combined using the Fisher’s method (https://cran.r-project.org/web/packages/metap/index.html). To adjust for instances where the signs (+/-) of the log fold change (logFC) did not all agree across datasets, the combined p-value for these genes were replaced with the highest combined p-value where the signs agreed. The combined p-values were then adjusted for multiple comparisons using the false discovery rate (FDR) method.

Hypergeometric enrichment test for both single-cell and bulk analysis

For various analyses, we used hypergeometric tests to decide if one set genes (e.g., X-chromosome-heterogeneous genes) was over-represented in another set of genes (e.g., DEGs). To perform the hypergeometric enrichment test, we used the function *phyper*. We followed the following general guideline to perform each enrichment test: argument q is how many genes from gene set #1 were found in gene set #2, argument m is the number of genes in gene set #2, argument n is the number of genes in gene set #2 that we are not interested in, argument k is the number of genes in gene set #1, and lower.tail argument set to false.

For example, to compute the X-chromosome heterogeneous genes enrichment in Oliva cortex sex-bias gene set: argument q was how many X-chromosome heterogeneous genes were found in the 112 Oliva cortex sex-bias gene set, argument m was how many genes are in the Oliva cortex sex-bias gene set, argument n was how many genes were not sex-biased in Oliva cortex analysis, and k was how many X-chromosome heterogeneous genes were found overall in Oliva cortex analysis [12].

Results

Overview of the analysis

Three single-cell/ single-nucleus prefrontal cortex Alzheimer’s datasets were analyzed, containaing expression data for five cell types: astrocytes, excitatory neurons, inhibitory neurons, microglia, and oligodendrocytes (**Figure 1**, upper left). Later in the analysis, we also analyzed three bulk expression datasets from the cortex (**Figure 1**, lower left). To investigate which cell type may contribute more to the progression of AD, we performed a cell-type differential expression analysis on each dataset to identify the cell type that have the most differentially expressed genes (**Figure 1**, second panel from left, upper part). Once we

identified the cell type that had the most DEGs across the three datasets, we used the software program GOrilla to identify enriched pathways that may contribute to the progression of AD (**Figure 1**, upper right).

In addition to the cell-type differential expression analysis, we were also interested in a sex-stratified cell-type differential expression analysis to better understand what may be contributing to the observed sex bias in AD (**Figure 1**, second panel from left, lower part). For this analysis, we were interested in particularly two gene sets that can contribute to the sex difference observed: (a) GTEx cortex genes that are considered sex biased and were published by Oliva et al., and (b) X-chromosome genes (**Figure 1,** third panel from left). We performed a sex-stratified cell-type differential expression analysis on each dataset, then performed a sex-stratified enrichment analysis using the software GOrilla. In addition, we performed hypergeometric enrichment analyses (**Figure 1**, lower right) to determine if a certain sets of differentially expressed genes are enriched with the gene sets that can contribute to the observed sex bias.

To validate the role that X-chromosome genes have in the observed sex bias in AD, we analyzed three bulk cortex datasets (**Figure 1**, lower left) and performed an X-chromosome differential expression analysis (**Figure 1**, second panel from left, lowest part). After performing the differential expression analysis, we performed a meta-analysis to identify X-chromosomes that are statistically significant and are upregulated/downregulated in the same direction across the three datasets.

Excitatory neurons have a much greater number of differentially expressed genes associated with Alzheimer's disease than other cell types

To determine if excitatory neurons play a larger role in AD than other cell types, we started by analyzing a single-cell dataset from Synapse (syn18485175) [7] for DEGs; subsequently, we refer to these data as the "Mathys dataset". The metadata provided by Mathys et al. and the Religious Orders Study and Memory and Aging Project (ROSMAP) classified 48 patients by disease, classified cell by cell type, and provided sex and age for each patient.

Previously, Belonwu et al. performed a sex-stratified differential expression analysis on the Mathys dataset but reported very low numbers (having only two digits) of DEGs for a dataset with a large sample size [11]. Using voom-limma [30], which was developed specifically for bulk RNA-seq analysis, they filtered out more than 99% of the genes, which is likely to be too aggressive. Hence, we repeated this analysis and instead, used the MAST method ([19], Materials and Methods) in the version available via the Seurat package [18], that is widely used for single-cell transcriptomics data analysis, to identify the cell-type specific DEGs for six cell-types for which there were sufficient data; as evident, we find that excitatory neurons have far more significant DEGs than the other cell-types (Materials and Methods, **Table 1**, first row).

To investigate whether the far more significant DEGs found in excitatory neurons is because they have a larger cell count, we downsampled the cells for each cell type and repeated the analysis and took the mean across 100 replicates (Materials and Methods). We still found excitatory neurons to have more significant DEGs than all other cell types (**Table 1,** second row). 300 genes are differentially expressed in more than 50 out of 100 down sampling replicates, with *RASGEF1B*, *LINGO1*, and *SLC26A3* appearing in all 100 (**Table 2**), overall testifying that the notable transcriptional alterations observed in the excitatory neurons are likely to reflect the biology of the disease. Interestingly, *LINGO1* has been implicated in numerous neurodegenerative disorders and has been proposed as a potential for AD due to its critical role in the pathophysiology of AD by favoring the β-cleavage of APP and the generation of Aβ fragments. [33, 34]. In addition, *LINGO1* was one of the few genes that passed Belonwu et al. filtering and hence, they also detected *LINGO1* as significantly differentially expressed in the Mathys dataset [11].

Table 1. Total number of DEGs in the Mathys dataset with an FDR-adjusted p-value less than 0.05 before and after down sampling

	Astrocytes	Excitatory Neurons	Inhibitory Neurons	Microglia	Oligodendrocytes	Oligodendrocytes Progenitor Cells
Pre-downsampling DEGs	317	8272	2950	10	260	18
Post-downsampling DEGs	9	767	38	1	1	3

Table 2. Top 10 genes ranked by the number of times each gene appeared to be differentially expressed across the 100 subset replicates for the down sampling analysis of the Mathys dataset.

Gene	Replicate Count
RASGEF1B	100
LINGO1	100
SLC26A3	100
NGFRAP1	99
DHFR	97
GRIN1	93
PDE10A	89
BEX1	88
SPARCL1	88
IDS	87

After observing that excitatory neurons had a far greater number of DEGs compared to the other cell types, we analyzed the Lau dataset, to explore their findings that astrocytes have the most DEGs. The early pre-processing steps were done to match their paper, but we again used MAST within Seurat for the most important analysis steps (Materials and Methods). Reassuringly, the percentages of cells of each type that we found correspond well overall with those reported originally by Lau et al (**Supplementary Table S1**). Using our cell classifications, **Table 3** shows the number of DEGs found per cell type for each of the five pertaining cell types; we excluded endothelial cells because they were removed in Mathys analysis (Materials and Methods). These results thus reinforce our previous findings in the Mathys dataset that excitatory neurons had more significant DEGs compared to the other cell types. We performed the same down sampling test that we performed on Mathys dataset (Materials and Methods), finding that excitatory neurons contained more significant DEGs than all other cell types even after controlling for cell count, validating the results that we saw in (**Table 3**).

Table 3. Total number of DEGs with an FDR-adjusted p-value less than 0.05 before and after down sampling in the Lau dataset. The second row is the average of 100 replicates rounded to the nearest integer.

	Astrocytes	Excitatory Neurons	Inhibitory Neurons	Microglia	Oligodendrocytes
Pre-downsampling DEGs	4021	7,784	6515	350	3655
Post-downsampling DEGs	908	2106	984	118	975

Next, we studied the cell type question in a third dataset, syn21670836 [9], which we sometimes refer to as the "TREM2 dataset" (Materials and Methods). The TREM2 dataset contains samples collected from the prefrontal cortex of Alzheimer's patients and controls, like the two previous datasets of [7, 8]. The metadata provided by Zhou et al. classified patients by *TREM2* variant, classified cell by cell type, and provided sex and age for each patient. One unusual aspect of this dataset is that there were two clusters of excitatory neurons, which the authors denoted by Ex0 and Ex1.

We focused our analysis on the Alzheimer's patients with TREM2-CV, a common *TREM2* variant, because they were sex matched with the controls. As shown in **Table 4**, we again see that excitatory neurons have far more significant DEGs compared to the other cell types, if we combine Ex0 and Ex1. However, this time the down sampling test (Materials and Methods) revealed that astrocytes contained the most significant DEGs (**Table 4,** second row).

We reasoned that combining Ex0 and Ex1 for DEG analysis may be ill-advised and hence, re-performed the down sampling analysis in a cluster specific manner, focusing on the Ex0, which has a much larger cell count (Materials and Methods). This analysis showed that DE alterations in excitatory Ex0 genes are indeed the most widespread among the different cell-types, notably, reaffirming the findings of the analyses of the two previous datasets (**Table 4,** second row).

Table 4. Total number of DEGs with an FDR-adjusted p-value less than 0.05 before and after down-sampling in the TREM2 dataset. The second row represents the average of 100 replicates rounded to the nearest integer.

	Astrocytes	Excitatory Neurons (All)	Excitatory Neurons (Ex0)	Inhibitory Neurons	Microglia	Oligodendrocytes	Oligodendrocytes Progenitor Cells
Pre-downsampling DEGs	1,237	8,379	NA	2447	302	1621	177
Post-downsampling DEGs	198	68	547	77	70	47	55

We note that Zhou et al. did not provide any biological characterization of the difference between Ex0 and Ex1, just accepting these as distinct clusters produced automatically by Seurat [9]. To try to find a biological difference, we performed a differential expression analysis that we will present later in Results.

Our primary threshold for determining that a gene is differentially expressed is that the FDR-adjusted p-value is < 0.05, but another potentially important consideration involves the expression fold-change, quantified as the logFC. In the re-analyses reported above, we cautiously adhered to the original logFC cutoffs used to determine DEGs in each of the three datasets. However, for completeness we additionally tested how varying the logFC threshold may affect which cell type has the largest number of DEGs. We chose 6 different logFC cutoffs: 0, 0.05, 0.10, 0.15, 0.20, and 0.25. Interestingly, we see that indeed, at low logFC cutoffs, excitatory neurons have the most DEGs across the three datasets. However, we note that if one increases the

logFC cutoff, some cell types start to have more DEGs than do excitatory neurons in the Lau and TREM2 datasets (**Supplementary Tables S2-S4**).

Pathway enrichment analysis of DEGs in excitatory neurons

After discovering that excitatory neurons have the most DEGs, we performed an enrichment analysis to determine which of their pathways are upregulated or downregulated in Alzheimer's patients (Materials and Methods, **Supplementary Figures S1, S2**). Overall, we find different key enriched pathways in the different datasets we have studied (see Discussion section). In the Mathys dataset, pathways involved in synaptic signaling are significantly upregulated in Alzheimer's patients, which bears relevance to the core homeostatic machinery theory proposed by Frere and Slutsky (Frere & Slutsky, 2018). The top pathways that are significantly downregulated are involved in the electron transport chain, which also contributes to the stability of the core homeostatic machinery (Frere & Slutsky, 2018).

In the Lau dataset, the most significantly upregulated pathways were modulation of chemical synaptic transmission and negative regulation of cellular metabolic process (**Supplementary Figure S3**). However, opposite to what we saw in Mathys dataset, synaptic signaling was the most significant downregulated pathway in Alzheimer's patients. In addition, trans-synaptic signaling and cell-cell adhesion were downregulated in Alzheimer's patients (**Supplementary Figure S4**).

For the TREM2 dataset, we performed the enrichment analysis on the Ex0 cluster. Pathways involved in protein targeting are upregulated in Alzheimer's patients (Supplementary Figure S5), while synaptic signaling pathways are downregulated (**Supplementary Figure S6**). Interestingly, pathways involved in ion transport were some of the most significantly downregulated pathways, and ion transport is another homeostatic machinery pathway that contributes to the stability of the core homeostatic machinery (Frere & Slutsky, 2018).

To search for common potential AD gene targets, we looked for genes that were among the top 500 DEGs in each dataset ranked by adjusted p-value and are members of the dysregulated pathways identified by GOrilla. We focused on overexpressed genes since they may be easier to target. One gene that stood out is *GRIN1*, which is one of the most consistent DEGs in our downsampling of the excitatory neurons in the Mathys dataset (**Table 2**). This gene, which also known as GluRN1, encodes a glutamate receptor and is a strong functional candidate because of the key roles of glutamatergic synapses in the pathogenesis of AD [35-37]. Overexpression of GRIN1 protein in AD has been found in at least two studies [38, 39], but this has been hard to study in bulk samples because the glutamatergic neurons tend to die early in the disease [35, 40]. *GRIN1* contributes to the following upregulated pathways (GOrilla rankings in parentheses) in the Mathys dataset excitatory neurons GO:0050804 modulation of chemical synaptic transmission (1st), GO:0099177 regulation of trans-synaptic signaling (2nd), GO:0043269 regulation of ion transport (11th), GO:0032879 regulation of localization (16th), and GO:0099537 trans-synaptic signaling (32nd).

Finally, to try to find a biological difference between the two excitatory neuron clusters in the TREM2 dataset, we performed a MAST differential expression analysis between Ex0 and Ex1 clusters, controlling for sex, age, and condition. **Table 5** shows the top 10 DEGs between Ex0 and Ex1. Interestingly, the fifth gene that was expressed higher in Ex0 is *XIST*, which is known to control X-inactivation [41]. Based on this finding, we were interested in seeing if this means that Ex1 would have more upregulated X-chromosome inactivation heterogeneous genes upregulated than Ex0. To test this hypothesis, we analyzed only female samples. Indeed, we found that Ex1 had more X-chromosome inactivation heterogeneous genes upregulated compared to Ex0 (100 and 63, respectively). Since heterogeneity in X-chromosome inactivation occurs only in females, this finding led us to wonder whether these X chromosome genes are relevant to the difference in AD incidence between males and females. To pursue potential contributions of excitatory neurons to sex differences, we decided to do sex-stratified differential expression analysis.

Table 5. Top 10 DEGs between Ex0 and Ex1 ranked by descending logFC. A positive logFC value indicate higher expression in Ex0 cluster. Zero adj.p.values indicate a value less than 1.00e-310.

Gene	logFC	adj.p.value
MEG3	4.387	0
MALAT1	3.905	0
MIAT	3.532	0
MIR124-1HG	2.765	0
XIST	2.740	0
PNISR	2.521	0
LINC00632	2.520	0
KCNIP4-IT1	2.485	0
RNPC3	2.471	0
AH1	2.463	0

Sex-stratified enrichment analysis of differentially expressed pathways in excitatory neurons

We performed a sex-stratified enrichment analysis to determine if there are sex differences in enriched pathways (Materials and Methods). In each dataset we compared male cases to male controls and female cases to female controls (we did not compare male cases to female cases because this would not be a properly controlled comparison). In the Mathys dataset, the top pathways upregulated in Alzheimer's males and females were involved in cell adhesion and synaptic transmission, with trans-synaptic signaling being more prominent in females. Females also had upregulation of cellular component organization, which was not upregulated in males. Top pathways downregulated in both Alzheimer's males and females were involved with cellular metabolism and the electron transport chain. In addition, males had several downregulated pathways including ferric iron transport, transferrin transport, immune effector process and immune system process, which were not found in females.

In the Lau dataset, we found that both males and females had regulation of trans-synaptic signaling and modulation of chemical synaptic transmission pathways significantly upregulated in Alzheimer's patients. Interestingly, we again saw that only females had the pathways regulations of cellular component organization and negative regulation of cellular process significantly upregulated. In addition, we found metabolic process pathways to be significantly upregulated in female Alzheimer's patients but not in males.

In the TREM2 Ex0 cell cluster, we found that males had regulation of trans-synaptic signaling, negative regulation of metabolic process, regulation of ion transport, and cell-cell adhesion significantly upregulated in Alzheimer's patients. Interestingly, we found no upregulated pathways to be significantly enriched in the female Alzheimer's patients. The opposite was true for pathways among genes found to be downregulated in Alzheimer's patients. We found no significantly enriched pathways in male Alzheimer's patients. In females, we found ion transport, synaptic signaling, and ATP metabolic process to be significantly downregulated in Alzheimer's patients.

To search for common potential gene targets of interest in the sex-specific comparisons, we looked for genes that were among the top 500 DEGs for females in each data set ranked by adjusted p-value and contributed to the dysregulated pathways identified by GOrilla. One gene that stood out was *RAB3A*, which is involved in recycling of synaptic vesicles and was consistently downregulated in female patients compared to controls (ranking as the 224th, 36th, and 39th most significantly differentially expressed gene in the Mathys, Lau, TREM2 datasets, respectively). In contrast, *RAB3A* was not among the top 500 genes for any of the male patient vs. male controls comparisons and was slightly overexpressed in male patients vs. controls in the Lau and TREM2 datasets. A proteomic study showed that decreasing levels of RAB3A protein predict cognitive decline in AD [42]. An earlier in vitro RNAi screen showed that RAB3A has an important role in processing of APP into Aβ peptides [43]. The top gene implicated in that same screen was *RAB11A*, which was also significantly downregulated in all

three comparisons of female patients vs. controls (ranking 60th, 73rd, 763rd respectively); *RAB11A* was not nearly as significant and was not consistently differentially expressed in the male patient vs. male control comparisons. *RAB3A* and *RAB11A* contribute to various differentially regulated pathways including GO:0050804 modulation of chemical synaptic transmission, GO:0099177 regulation of trans-synaptic signaling, GO:0046903 secretion, etc.

In conclusion, key pathways that appear most often in the sex-specific comparisons are GO:0099177 regulation of trans-synaptic signaling and GO:0050804 modulation of chemical synaptic transmission. Genes in these pathways may be both upregulated and downregulated, consistent with the Frere and Slutsky hypothesis that loss of homeostasis is the key to Alzheimer's pathogenesis. When comparing upregulated and downregulated pathways, the most striking observation was that the TREM2 Ex0 cluster has only upregulated pathways in males and only downregulated pathways in females. Furthermore, among the 178 significant downregulated pathways in females and the 90 significantly upregulated pathways in females, 56 are in common, but changing in opposite directions. This reinforces our previous conclusion that the Seurat split between the Ex0 and Ex1 clusters in the TREM2 dataset unmasks some clues about sex differences in AD. The Mathys dataset has mostly upregulated pathways in both males and females, presenting an opportunity to compare datasets is the pathways upregulated in Mathys dataset males and TREM2 dataset males, but not in Mathys dataset females. There are three such pathways: GO:0009653 anatomical structure morphogenesis, GO:0045597 positive regulation of cell differentiation, GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules.

Revisiting the single-cell sex-specific findings in a bulk expression analysis

After discovering sex differences in enriched pathways in excitatory neurons, we decided to perform a differential expression analysis on bulk data to see if bulk data analysis would provide additional information on the observed sex differences. For this analysis, we were especially interested in X-chromosome PAR genes and X-chromosome heterogeneous for escaping X-inactivation and their potential role contributing to AD. We analyzed three datasets for AD that had gene expression from the cortex: GSE15222, GSE33000, and GSE44770 [25-27]. We performed linear modeling of the association of their DE with AD, controlling for sex and age (Materials and Methods).

After obtaining the differential expression analysis on each Alzheimer's dataset, meta-analysis was performed to identify the significant pseudoautosomal and inactivation-heterogeneous X-chromosome genes across all three datasets and the p-values were corrected for multiple testing (Materials and Methods). Among significant consistently differentially expressed genes genome-wide were five pseudoautosomal genes, *CD99* (genome-wide FDR adjusted p = 2.7e-11), *ZBED1* (p = 5.2e-11), *IL3RA* (p = 2.5e-09), *ASMTL* (p = 6.0e-09), and *GTPB6* (p = 8.6e-05). **Table 6** shows the top 10 statistically significant differentially expressed inactivation-heterogeneous X-chromosome genes adjusting for multiple testing (Materials and Methods). The top three genes were *BEX1*, *PRKX*, and *TSR2*. Interestingly, *BEX1* was one of the most consistent DEGs in our downsampling of the excitatory neurons in the Mathys dataset (**Table 2**), is significantly downregulated in cases versus controls in all three single-cell datasets and functionally has been shown to be involved in the regeneration of axons [44] and is downregulated in Alzheimer's patients compared to controls. *PRKX* encodes a serine threonine protein kinase and has been shown to play a crucial role in neural development [45]. *TSR2* has been shown to inhibit the transcriptional activity of NF-κB, which is one of the key transcription factors for the homeostatic model proposed by Frere and Slutsky, and *TSR2* was found to be downregulated in Alzheimer's patients [46]. *ELK1*, *ATP6AP2*, and *MCTS1* were also measured in all three single-cell datasets and significantly differentially expressed in excitatory neurons of all three single-cell datasets in the same direction as in the three bulk datasets (**Table 56**). It is worth noting that *XIST* did not pass the test that the logFC should consistently be in the same direction (Materials and Methods); hence, it does not appear in our meta-analysis.

Table 6: **Top 10 significant differentially expressed X-chromosome heterogeneous genes** between Alzheimer's and controls meta-analysis. Meta-analysis was done using Fisher's method with FDR correction for those X-chromosome inactivation heterogeneous genes that had logFC with the same sign in each of the three datasets.

Gene	GSE15222 logFC	GSE33000 logFC	GSE44770 logFC	Combined adj.p.value
BEX1	-0.580	-0.065	-0.087	4.45E-24
PRKX	0.545	0.081	0.080	2.50E-23
TSR2	-0.254	-0.047	-0.046	2.13E-22
FOXO4	0.598	0.080	0.046	2.61E-21
ELK1	0.333	0.092	0.107	5.57E-21
USP11	-0.537	-0.040	-0.054	3.15E-20
TBL1X	0.429	0.061	0.030	8.83E-20
ATP6AP2	-0.375	-0.037	-0.039	1.82E-19
IDS	-0.524	-0.026	-0.049	8.71E-19
MCTS1	-0.474	-0.022	-0.039	9.10E-18

Since we saw that X-chromosome heterogeneous genes were significantly differentially expressed between Alzheimer's and control in bulk data, we went back to the single-cell transcriptomics dataset and tested whether X-chromosome heterogeneous genes are over-represented among male or female excitatory neurons DEGs. In the Mathys dataset, we did not find significant enrichment of the X-chromosome heterogeneous genes in the DEGs for males (p-value = 0.21). However, quite strikingly, a hypergeometric test for females resulted in a p-value of 0, indicating that enrichment is statistically significant with a p-value less than 1.00e-310. This difference makes sense since some X-chromosome heterogenous genes differ in the expression between female individuals.

In the Lau dataset, we again did not find significant enrichment of the X-chromosome heterogeneous genes in the DEGs for males (p-value = 0.10). In the females, the hypergeometric test again resulted in a p-value of 0, indicating that enrichment is statistically significant with a p-value less than 1.00e-310.

In the TREM2 Ex0 cluster, we did not find significant enrichment of the X-chromosome heterogeneous genes in the DEGs for either males or females (p-value 0.51 and 0.21, respectively). This was expected based on the above analysis that puts the X-chromosome heterogeneous DEGs preferentially in the Ex1 cluster.

Another source of candidate genes to be involved in diseases, such as Alzheimer's, with a difference in prevalence by sex, comes from the analysis of Oliva et al of the newly published Genotype-Tissue Expression (GTEx) v8 project data. Out of the 112 sex-bias genes reported in that study for the Cortex, 31 are either X-chromosome inactivation heterogeneous or pseudoautosomal genes (19 and 12, respectively), with an enrichment p-value of 1.21e-40. Interestingly, 8 of the 19 X-chromosome inactivation heterogeneous genes are also identified in our X-chromosome meta-analysis in the bulk data. Out of the 31 genes mentioned above, 21 were found to be differentially expressed in the Mathys dataset, 18 were found to be differentially expressed in Lau dataset, and 24 were found to be differentially expressed in the TREM2 dataset (**Table 7**). Among these genes, *NAP1L3* and *CHM* are heterogeneous genes consistently downregulated in cases vs. controls.

Table 7. Thirty-two X-chromosome genes found in the Oliva published cortex sex-bias list and their differential expression status in Mathys, Lau, and TREM2 datasets. The second column indicates whether the gene is in the pseudoautosomal region or has heterogeneous X-inactivation status. LogFC value indicates if the gene was found to be statistically significant with an adjusted p-value less than 0.05. NS = Not Significant.

Gene	Heterogeneous or Pseudo	Mathys	Lau	TREM2
XIST	Heterogeneous	NS	-0.705	-0.686
ZFX	Heterogeneous	0.036	-0.020	-0.021
KDM5C	Heterogeneous	0.085	0.016	-0.074
DDX3X	Heterogeneous	0.005	0.048	-0.156
KDM6A	Heterogeneous	-0.025	-0.039	-0.024
EIF2S3	Heterogeneous	0.027	-0.035	-0.098
USP9X	Heterogeneous	-0.007	0.011	-0.163
UBA1	Heterogeneous	0.027	-0.003	-0.017
SYAP1	Heterogeneous	-0.027	NS	-0.021
RPS4X	Heterogeneous	-0.234	-0.030	0.056
EIF1AX	Heterogeneous	-0.109	0.126	-0.098
SMC1A	Heterogeneous	0.023	NS	0.004
STS	Heterogeneous	0.000	-0.057	-0.086
PNPLA4	Heterogeneous	NS	NS	NS
TXLNG	Heterogeneous	-0.067	0.008	-0.006
NAP1L3	Heterogeneous	-0.082	-0.086	-0.132
IRAK1	Heterogeneous	NS	NS	NS
PIN4	Heterogeneous	-0.108	-0.039	0.040
CHM	Heterogeneous	-0.093	-0.040	-0.096
CD99	Pseudo	NS	NS	NS
ZBED1	Pseudo	NS	NS	0.018
GTPBP6	Pseudo	0.002	0.039	0.184
ASMTL	Pseudo	-0.103	NS	-0.022
DHRSX	Pseudo	0.027	0.002	0.098
PPP2R3B	Pseudo	NS	NS	NS
PLCXD1	Pseudo	NS	NS	0.198
CD99P1	Pseudo	NS	NS	NS
LINC00685	Pseudo	NS	NS	NS
IL3RA	Pseudo	NS	NS	NS
AKAP17A	Pseudo	-0.046	NS	-0.065
SLC25A6	Pseudo	-0.199	0.029	-0.111

Taken together, our results suggest that among the commonly studied cell types, excitatory neurons have the most differentially expressed genes, resolving a previous contradiction in the literature. Analysis of the excitatory neuron data in three single-cell datasets and cortex data in three bulk datasets suggest that the differentially expressed genes in this cell type

and this tissue disproportionately include X-chromosome heterogeneous genes in females and pseudoautosomal genes when analyzing all cases versus all controls. Furthermore, different pathways are upregulated and downregulated in the two sexes in excitatory neurons. Thus, the longstanding questions of which cell type is most important in Alzheimer's pathogenesis and why females have higher incidence than males can be connected logically and biologically by analyzing differential gene expression in single-cell data.

Discussion

In this study, we focused on the roles of different neuronal cell types and their expression in the pathogenesis of AD. This is different from complementary studies that have investigated the possible pathogenetic role of synaptic alterations [35, 47] or genetic factors [48]. Our study was made feasible due to recent publicly available Alzheimer's single-cell transcriptomics datasets. Following previous work [7, 8, 11], we considered the number of differentially expressed genes (DEGs) between Alzheimer's and controls in specific cell types as a likely indicator of cell type functional importance. Analyzing these datasets in a comprehensive and uniform manner, we asked two fundamental questions; (a) what are the key transcriptomics differences that we can between the brains (cortical regions) of Alzheimer patients and controls and further (b) what the sex-dependent differences in the cell types and genes are that are differentially altered in Alzheimer's disease?

Regarding the first question, surprisingly, previous studies [7] and [8] that collected and analyzed Alzheimer's single-cell transcriptomics datasets reached contrasting conclusions, finding that excitatory neurons [7] or, in difference, astrocytes [8], have the most DEGs between Alzheimer's and controls. Our first key result resolves this discrepancy by homogeneously using the MAST [19] method on both data sets, showing that it is the excitatory neurons that have more DEGs. We further validated this result on a third single-cell data [9].

Identifying the cellular pathways that are over-represented among the DEGs, we found that synaptic signaling (upregulated in the Mathys dataset but downregulated in the Lau and TREM2 datasets) and mitochondrial functions related to energy production and ion transport were downregulated (in the Mathys and TREM2 data sets) stood out. In addition, synaptic signaling was downregulated in female patients by themselves but not in male patients in the TREM2 data set. These findings are consistent with the homeostatic model of [3], in which 'firing homeostasis' is central and depends partly on 'proteostasis' and 'energy homeostasis' for which the mitochondria play key roles.

We next turned to our second question and investigated the pathogenic role of sex differences by analyzing three single-cell transcriptomics datasets, three bulk gene expression datasets from cortex, and incorporating the recent genome-wide characterization of tissue-specific sex-biased genes published as part of the release of GTEx v8 [12]. Based on the analysis of the TREM2 data set, we hypothesized and validated that two gene sets on the X-chromosome, pseudoautosomal region (PAR) genes and heterogeneously X-inactivated genes, are significantly over-represented among the differentially expressed genes between Alzheimer's male and female patients, both in single-cell and bulk data.

Although they are few in number, PAR genes are natural candidates because the lack of recombination between X and Y chromosomes allows the X copy and the Y copy(ies) of PAR genes to diverge in evolution. A recent functional study established the importance on one PAR gene, namely *IL3A*, in Alzheimer's pathogenesis [14]. IL3RA is the cytokine-specific subunit for the heterodimeric receptor for the cytokine interleukin 3 (IL-3). McAlpine and colleagues showed in the most widely used mouse model of Alzheimer's that IL-3 is produced by astrocytes and the signal is received by the IL3RA/CSFR2 heterodimer in microglia, which has the functional effect of reducing Aβ aggregates. In addition, McAlpine et al found higher expression of IL3RA in the brains of individuals with Alzheimer's compared to controls.

Our hypotheses about the possible role of X-chromosome genes can be pursued further by reconsidering published GWAS data sets in females only via meta-analysis, as was done recently by Chung and colleagues, leading to the identification of variants in the autosomal gene *MGMT* as a female-specific risk factor [49]. Oliva and colleagues did attempt to connect the sex-biased genes they identified in GTEx to diseases by searching the NHGRI-EBI GWAS catalog [50] for published associations between the sex-biased genes and traits, but they found no such associations for AD. The approach of catalog

lookup is quick but is statistically suboptimal for at least two reasons. First, early GWAS studies ignored the X chromosome due to lack of statistical methods; this has been overcome with newer methods such as XWAS [51], but the NHGRI-EBI catalog does not include findings from XWAS studies or other X-chromosome-specific analyses. Second, meta-analysis of the original data, as Chung et al did, can identify markers that are significantly associated in the combined data, but not in individual studies curated in the GWAS catalog.

Our study has several major limitations. First, our analysis is focused on differential expression and statistical associations and not aimed at revealing underlying causal mechanisms. Second, our analyses of upregulated and downregulated gene pathways uncovered quite different results for each of the three single-cell data sets, possibly due to heterogeneity in their study designs and transcriptomics data collection methods. Third, except for the uncharacterized distinction between Ex0 and Ex1 in the TREM2 data set, all excitatory neurons were clustered together in the three single-cell data sets. To synthesize these datasets better with past findings of the potential importance of different types of excitatory neurons [52], it may be useful to distinguish which excitatory neurons are glutamatergic [36, 37] and express the transcription factor *RORB* [40]; however, distinguishing neuron subtypes is likely to reduce the sample size and the resulting statistical power. Fourth, the wet lab protocols and sequencing methods and quality control filters were different for the three datasets, and this could lead to substantial inter-dataset heterogeneity. Fifth, to have larger numbers of samples, we ignore important covariates, such as *APOE* genotype. Sixth, all our analysis is within cell types or tissues; it would be interesting to analyze predicted ligand-receptor interactions in addition. Finally, and most importantly, our analysis is focused on transcriptomics alterations and hence has very limited ability to shed light on classical and prevailing theories of Alzheimer's pathogenesis, involving many post-transcriptional mechanisms concerning the formation of amyloid plaques and neurofibrillary tangles and beyond [53–55].

Conclusions

In sum, our results resolve the apparent contradiction between [7] and [8], showing that comparing AD patients to unaffected controls, excitatory neurons have more differentially expressed genes than do astrocytes and other cell types. Analysis of enriched pathways in excitatory neurons points to differences in synaptic transmission and related pathways between males and females. Further analysis of differentially expressed genes between males and females in numerous single-cell and bulk transcriptomics datasets suggests that numerous PAR genes and X-chromosome heterogeneous genes may contribute to the difference in sex incidence of Alzheimer's disease. More generally, identifying cell types and genes whose expression is different between female and male patients opens new possibilities to understand the cellular and molecular etiology of this vexing illness.

Abbreviations

AD Alzheimer's disease

DEG(s) differentially expressed gene(s)

Ex and Ex1 two clusters of excitatory neurons identified as part of the TREM2 dataset

FDR false discovery rate

GEO Gene Expression Omnibus

GTEx Genotype-Tissue Expression project

logFC base 2 logarith of the fold change

PAR pseudoautosomal region of the X chromosome

Declarations

Availability of data and materials

We reanalyzed six data sets: three single-cell data sets and three bulk data sets. The Mathys data set and the TREM2 data set are available with an account and permission via Synapse (https://www.synapse.org/). The Lau single-cell data set and the three bulk data sets are available from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/gds) without any account or extra permission. Instructions on how to reproduce our analysis are provided in the directory https://ftp.ncbi.nlm.nih.gov/pub/catSMA/Alzheimer that contains the single file README.txt and the archive Alz_final_manuscript.tar.gz.

Acknowledgments

This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).

The results in this manuscript are in whole or in part based on data (Mathys dataset and TREM2 dataset) obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Samples for this study (for the Mathys dataset) were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

Figure 1 was created with Biorender.com (https://biorender.com/)

Funding

This research was supported in part by the NIH Intramural Research Program, National Cancer Institute. The research of A.X.G. was supported in part by the National Academy Enrichment Program at the NIH. This work was primarily supported by the National Institute of Aging (NIA) under Award Number R56AG074001, U01AG073323, R01AG066707, and R01AG076448 to F.C.

Contributions

Conceptualization ER, AAS. Data analysis: AXG with assistance from JX and AAS. Methods Design: All authors. Methods Implementation: AXG with assistance from AAS.

Writing – original draft: AXG, ER, AAS; Writing – review and editing: All authors.

Supervision and resources: FC, ER, AAS. All authors read and approved the submitted manuscript

Ethics declarations

Not applicable since our work is based entirely on reanalysis of publicly available data.

Consent for publication

Not applicable since our work is based entirely on reanalysis of publicly available data.

Competing interests

E.R. is a co-founder of MedAware, Metabomed and Pangea Biomed (divested), and an unpaid member of Pangea Biomed's scientific advisory board. The other authors have no competing interests.

References

1. **2021 Alzheimer's disease facts and figures.** *Alzheimers Dement* 2021, **17:**327-406.

2. Ullah MF, Ahmad A, Bhat SH, Abu-Duhier FM, Barreto GE, Ashraf GM: **Impact of sex differences and gender specificity on behavioral characteristics and pathophysiology of neurodegenerative disorders.** *Neurosci Biobehav Rev* 2019, **102:**95-105.
3. Frere S, Slutsky I: **Alzheimer's Disease: From Firing Instability to Homeostasis Network Collapse.** *Neuron* 2018, **97:**32-58.
4. Schwartz M, Peralta Ramos JM, Ben-Yehuda H: **A 20-year journey from axonal injury to neurodegenerative diseases and the prospect of immunotherapy for combating Alzheimer's disease.** *J Immunol* 2020, **204:**243-250.
5. Stempler S, Ruppin E: **Analyzing gene expression from whole tissue vs. different cell types reveals the central role of neurons in predicting severity of Alzheimer's disease.** *PLoS One* 2012, **7:**e45879.
6. Stempler S, Waldman YY, Wolf L, Ruppin E: **Hippocampus neuronal metabolic gene expression outperforms whole tissue data in accurately predicting Alzheimer's disease progression.** *Neurobiol Aging* 2012, **33:**2230 e2213-2230 e2221.
7. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, et al: **Single-cell transcriptomic analysis of Alzheimer's disease.** *Nature* 2019, **570:**332-337.
8. Lau SF, Cao H, Fu AKY, Ip NY: **Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease.** *Proc Natl Acad Sci U S A* 2020, **117:**25800-25809.
9. Zhou Y, Song WM, Andhey PS, Swain A, Levy T, Miller KR, Poliani PL, Cominelli M, Grover S, Gilfillan S, et al: **Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease.** *Nat Med* 2020, **26:**131-142.
10. Zhao N, Ren Y, Yamazaki Y, Qiao W, Li F, Felton LM, Mahmoudiandehkordi S, Kueider-Paisley A, Sonoustoun B, Arnold M, et al: **Alzheimer's risk factors age, APOE genotype, and sex drive distinct molecular pathways.** *Neuron* 2020, **106:**727-742.e726.
11. Belonwu SA, Li Y, Bunis D, Rao AA, Solsberg CW, Tang A, Fragiadakis GK, Dubal DB, Oskotsky T, Sirota M: **Sex-stratified single-cell RNA-Seq analysis identifies sex-specific and cell type-specific transcriptional responses in Alzheimer's disease across two brain regions.** *Mol Neurobiol* 2022, **59:**276-293.
12. Oliva M, Munoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, Parsana P, Kasela S, Balliu B, Vinuela A, et al: **The impact of sex on gene expression across human tissues.** *Science* 2020, **369:**eaba3066.
13. Olney KC, Brotman SM, Andrews JP, Valverde-Vesling VA, Wilson MA: **Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data.** *Biol Sex Differ* 2020, **11:**42.
14. McAlpine CS, Park J, Griciuc A, Kim E, Choi SH, Iwamoto Y, Kiss MG, Christie KA, Vinegoni C, Poller WC, et al: **Astrocytic interleukin-3 programs microglia and limits Alzheimer's disease.** *Nature* 2021, **595:**701-706.
15. Lyon MF: **Gene action in the X-chromosome of the mouse (Mus musculus L.).** *Nature* 1961, **190:**372-373.
16. Clerc P, Avner P: **Role of the region 3' to Xist exon 6 in the counting process of X-chromosome inactivation.** *Nat Genet* 1998, **19:**249-253.
17. Zhang Y, Castillo-Morales A, Jiang M, Zhu Y, Hu L, Urrutia AO, Kong X, Hurst LD: **Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving.** *Mol Biol Evol* 2013, **30:**2588-2601.
18. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al: **Integrated analysis of multimodal single-cell data.** *Cell* 2021, **184:**3573-3587 e3529.
19. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.** *Genome Biol* 2015, **16:**278.
20. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
21. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, Roederer M, Gottardo R: **Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments.** *Bioinformatics* 2013, **29:**461-467.

22. Soneson C, Robinson MD: **Bias, robustness and scalability in single-cell differential expression analysis.** *Nat Methods* 2018, **15:**255-261.
23. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.** *BMC Bioinformatics* 2009, **10:**48.
24. Davis S, Meltzer PS: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.** *Bioinformatics* 2007, **23:**1846-1847.
25. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, et al: **Genetic control of human brain transcript expression in Alzheimer disease.** *Am J Hum Genet* 2009, **84:**445-458.
26. Narayanan M, Huynh JL, Wang K, Yang X, Yoo S, McElwee J, Zhang B, Zhang C, Lamb JR, Xie T, et al: **Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases.** *Mol Syst Biol* 2014, **10:**743.
27. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, et al: **Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.** *Cell* 2013, **153:**707-720.
28. GTEx Consortium: **The GTEx Consortium atlas of genetic regulatory effects across human tissues.** *Science* 2020, **369:**1318-1330.
29. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.
31. Carvalho BS, Irizarry RA: **A framework for oligonucleotide microarray preprocessing.** *Bioinformatics* 2010, **26:**2363-2367.
32. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics–a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25:**415-416.
33. Fernandez-Enright F, Andrews JL: **Lingo-1: a novel target in therapy for Alzheimer's disease?** *Neural Regen Res* 2016, **11:**88-89.
34. de Laat R, Meabon JS, Wiley JC, Hudson MP, Montine TJ, Bothwell M: **LINGO-1 promotes lysosomal degradation of amyloid-β protein precursor.** *Pathobiol Aging Age Relat Dis* 2015, **5:**25796.
35. Griffiths J, Grant SGN: **Synapse pathology in Alzheimer's disease.** *Semin Cell Dev Biol* 2022.
36. Hardy J, Cowburn R, Barton A, Reynolds G, Lofdahl E, O'Carroll AM, Wester P, Winblad B: **Region-specific loss of glutamate innervation in Alzheimer's disease.** *Neurosci Lett* 1987, **73:**77-80.
37. Sokolow S, Luu SH, Nandy K, Miller CA, Vinters HV, Poon WW, Gylys KH: **Preferential accumulation of amyloid-beta in presynaptic glutamatergic terminals (VGluT1 and VGluT2) in Alzheimer's disease cortex.** *Neurobiol Dis* 2012, **45:**381-387.
38. Leuba G, Vernay A, Kraftsik R, Tardif E, Riederer BM, Savioz A: **Pathological reorganization of NMDA receptors subunits and postsynaptic protein PSD-95 distribution in Alzheimer's disease.** *Curr Alzheimer Res* 2014, **11:**86-96.
39. Zeidán-Chuliá F, de Oliveira BH, Salmina AB, Casanova MF, Gelain DP, Noda M, Verkhratsky A, Moreira JC: **Altered expression of Alzheimer's disease-related genes in the cerebellum of autistic patients: a model for disrupted brain connectome and therapy.** *Cell Death Dis* 2014, **5:**e1250.
40. Leng K, Li E, Eser R, Piergies A, Sit R, Tan M, Neff N, Li SH, Rodriguez RD, Suemoto CK, et al: **Molecular characterization of selectively vulnerable neurons in Alzheimer's disease.** *Nat Neurosci* 2021, **24:**276-287.
41. Panning B: **X-chromosome inactivation: the molecular basis of silencing.** *J Biol* 2008, **7:**30.
42. Bereczki E, Francis PT, Howlett D, Pereira JB, Höglund K, Bogstedt A, Cedazo-Minguez A, Baek JH, Hortobágyi T, Attems J, et al: **Synaptic proteins predict cognitive decline in Alzheimer's disease and Lewy body dementia.** *Alzheimers Dement* 2016, **12:**1149-1158.
43. Udayar V, Buggia-Prévot V, Guerreiro RL, Siegel G, Rambabu N, Soohoo AL, Ponnusamy M, Siegenthaler B, Bali J, Simons M, et al: **A paired RNAi and RabGAP overexpression screen identifies Rab11 as a regulator of β-amyloid production.** *Cell*

Rep 2013, **5:**1536-1551.

44. Khazaei MR, Halfter H, Karimzadeh F, Koo JH, Margolis FL, Young P: **Bex1 is involved in the regeneration of axons after injury.** *J Neurochem* 2010, **115:**910-920.
45. Blaschke RJ, Monaghan AP, Bock D, Rappold GA: **A novel murine PKA-related protein kinase involved in neuronal differentiation.** *Genomics* 2000, **64:**187-194.
46. He H, Zhu D, Sun J, Pei R, Jia S: **The novel protein TSR2 inhibits the transcriptional activity of nuclear factor-kappaB and induces apoptosis.** *Mol Biol (Mosk)* 2011, **45:**496-502.
47. Styr B, Slutsky I: **Imbalance between firing homeostasis and synaptic plasticity drives early-phase Alzheimer's disease.** *Nat Neurosci* 2018, **21:**463-473.
48. Fenoglio C, Scarpini E, Serpente M, Galimberti D: **Role of genetics and epigenetics in the pathogenesis of Alzheimer's disease and frontotemporal dementia.** *J Alzheimers Dis* 2018, **62:**913-932.
49. Chung J, Das A, Sun X, Sobreira DR, Leung YY, Igartua C, Mozaffari S, Chou YF, Thiagalingam S, Mez J, et al: **Genome-wide association and multi-omics studies identify MGMT as a novel risk gene for Alzheimer's disease among women.** *Alzheimers Dement* 2022.
50. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al: **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.** *Nucleic Acids Res* 2019, **47:**D1005-D1012.
51. Gao F, Chang D, Biddanda A, Ma L, Guo Y, Zhou Z, Keinan A: **XWAS: A software toolset for genetic data analysis and association studies of the X chromosome.** *J Hered* 2015, **106:**666-671.
52. Fu H, Hardy J, Duff KE: **Selective vulnerability in neurodegenerative diseases.** *Nat Neurosci* 2018, **21:**1350-1358.
53. Binder LI, Guillozet-Bongaarts AL, Garcia-Sierra F, Berry RW: **Tau, tangles, and Alzheimer's disease.** *Biochim Biophys Acta* 2005, **1739:**216-223.
54. Bloom GS: **Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis.** *JAMA Neurol* 2014, **71:**505-508.
55. Zhang YW, Thompson R, Zhang H, Xu H: **APP processing in Alzheimer's disease.** *Mol Brain* 2011, **4:**3.

Figures

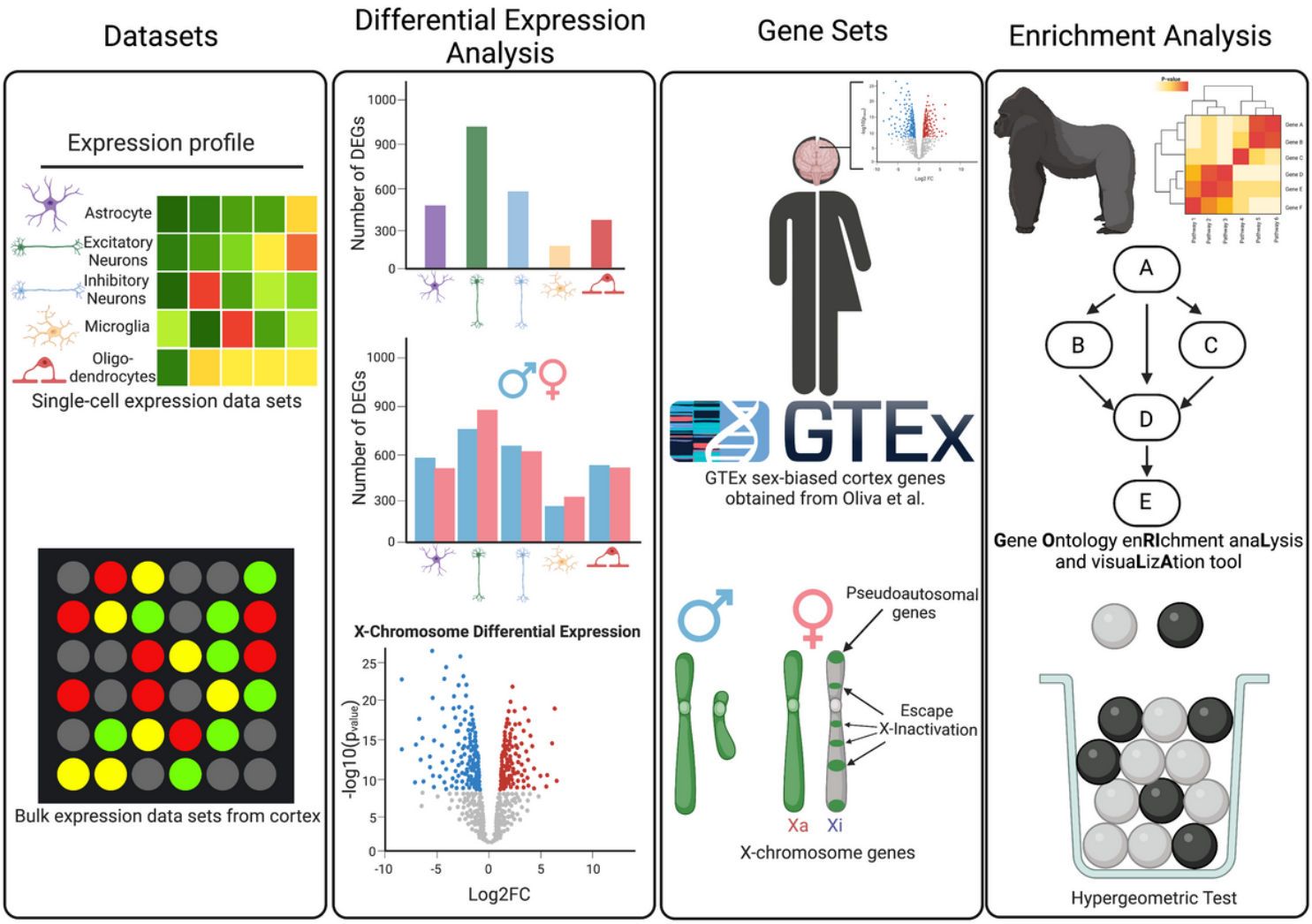

Figure 1

Overview of the analysis. *Single-cell/-nucleus transcriptomics and bulk gene expression Alzheimer's datasets were obtained for this project (left panel). We performed three different type of differential expression analysis: cell-type specific, sex-stratified cell-type specific, and X-chromosome genes (second panel from the left). Some of our analyses on sex bias incorporated published classifications of sex-biased genes from the GTEx project and a published classification of X-chromosome genes according to X-inactivation status (third panel from the left). After performing the differential expression analysis, we used the software GOrilla to perform the enrichment analysis (upper right). In addition, we performed hypergeometric tests (lower right) to identify enrichment of two gene sets in Alzheimer's patients: GTEx sex-biased cortex genes obtained from (Olivia et al. 2020) and the classification of X-chromosome genes.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiles.zip