







# Improving reproducibility of differentially expressed genes in single-cell transcriptomic studies of neurodegenerative diseases through meta-analysis

Received: 12 June 2025

Accepted: 24 July 2025

Published online: 12 August 2025

 Check for updates

Nathan Nakatsuka <sup>1,2,3</sup> ✉, Drew Adler<sup>4,5</sup>, Longda Jiang <sup>1</sup>, Austin Hartman <sup>1</sup>, Evan Cheng <sup>4,5</sup>, Eric Klann <sup>4,5</sup> & Rahul Satija <sup>1,2</sup>

False positive claims of differentially expressed genes (DEGs) in scRNA-seq studies are of substantial concern. We found that DEGs from individual Parkinson's (PD), Huntington's (HD), and COVID-19 datasets had moderate predictive power for case-control status of other datasets, but DEGs from Alzheimer's (AD) and Schizophrenia (SCZ) datasets had poor predictive power. We developed a non-parametric meta-analysis method, SumRank, based on reproducibility of relative differential expression ranks across datasets, and found DEGs with improved predictive power. Specificity and sensitivity of these genes were substantially higher than those discovered by dataset merging and inverse variance weighted p-value aggregation methods. Up-regulated DEGs implicated chaperone-mediated protein processing in PD glia and lipid transport in AD and PD microglia, while down-regulated DEGs were in glutamatergic processes in AD astrocytes and excitatory neurons and synaptic functioning in HD FOXP2 neurons. Lastly, we evaluate factors influencing reproducibility of individual studies as a prospective guide for experimental design.

As single-cell RNA-sequencing (scRNA-seq) technologies mature to process clinical samples, an increasing number of studies are profiling tissue from a multitude of disease states to identify cell type-specific transcriptional alterations associated with pathophysiology and general development. scRNA-seq case-control studies have generated data on a multitude of neuropsychiatric diseases, such as multiple sclerosis<sup>1-3</sup>, schizophrenia (SCZ)<sup>4-6</sup>, major depressive disorder<sup>7</sup>, autism<sup>8,9</sup>, Parkinson's disease (PD)<sup>10-15</sup>, alcohol use disorder<sup>16,17</sup>, Rett Syndrome<sup>18</sup>, vascular dementia<sup>19</sup>, and Huntington's disease (HD)<sup>20-23</sup>, though all with relatively few individuals per study and often not in the same brain region. For Alzheimer's Disease (AD) and COVID-19, however, scRNA-seq studies now have sample sizes in the hundreds<sup>24-27</sup>.

These studies have uncovered biological pathways perturbed in these conditions that represent potential therapeutic targets.

Nevertheless, there has been concern for possible false positive results in these studies<sup>28</sup>, and thus the statistical methodology required to perform case-control studies across multiple cell types remains an area of active interest<sup>29</sup>. Initial studies implemented case-control analyses by performing differential-expression testing on individual cells. This approach treats each cell as an independent replicate, which fails to account for correlations across cells from the same individual and can lead to a large false-positive bias. Subsequent studies have dealt with these issues by using mixed models with individuals as a fixed or random effect<sup>26</sup> or alternative regression models

<sup>1</sup>New York Genome Center, New York, NY, USA. <sup>2</sup>Center for Genomics and Systems Biology, New York University, New York, NY, USA. <sup>3</sup>Department of Psychiatry, New York University Grossman School of Medicine, New York, NY, USA. <sup>4</sup>Center for Neural Science, New York University, New York, NY, USA. <sup>5</sup>NYU Neuroscience Institute, New York University, New York, NY, USA. ✉ e-mail: [nnakatsuka@nygenome.org](mailto:nnakatsuka@nygenome.org)

previously developed for bulk RNA-seq<sup>30</sup> that can be used after pseudo-bulking clusters of single cells. Many of these methods can adequately control the false positive rate and yet are sufficiently powered in analyses of simulated differentially expressed genes (DEGs). Nevertheless, there has still been substantial worry about potential false positives in DEG results due to technical artifacts or simply biological variation present in only small numbers of individuals (particularly for studies with smaller sample sizes). This issue is likely of particular relevance for many neuropsychiatric diseases due to the high transcriptomic heterogeneity of the brain at baseline<sup>31</sup> and GWAS evidence for etiological diversity in many of these diseases<sup>32</sup>.

The field of human genetics, particularly genome-wide association studies (GWAS), can provide a model for the single-cell field in its high reproducibility<sup>33</sup> and well-established meta-analysis methods for combining information across multiple datasets<sup>34,35</sup>. The typical GWAS meta-analysis usually applies an inverse variance weighting to aggregate the effect sizes and standard errors derived from each study to obtain final effect sizes and *p* values for each genetic locus<sup>36</sup>. It is standard for new studies to have a separate test dataset to assess the reproducibility of significant genes found in the general analysis, testing for effect size, and at least ensuring the same direction of effect in the test dataset. Now that many large-scale case-control scRNA-seq studies have been undertaken for several diseases, the field is in a strong position to develop standardized meta-analysis methods that combine information across multiple datasets with the goal of finding genes with transcriptional expression (and later other epigenetic loci) robustly associated with disease.

In this study, we provide a systematic approach in this direction by first examining the reproducibility of 17 AD, 6 PD, 4 HD, 3 SCZ single-nucleus RNA-seq (snRNA-seq) studies and, as a positive control comparison due to its known strong transcriptional response, 16 scRNA-seq COVID-19 studies. We find by several measures that a large fraction of the genes found to be differentially expressed in single AD and SCZ datasets do not reproduce in other AD and SCZ datasets, while genes found in PD, HD, and COVID-19 datasets have moderate reproducibility. To address this challenge, we introduce a procedure for large-scale meta-analysis of scRNA-seq called SumRank that prioritizes the identification of DEGs that exhibit reproducible signals across multiple datasets and demonstrates that this approach substantially outperforms existing meta-analysis techniques in sensitivity and specificity of discovered DEGs. We demonstrate that SumRank identifies DEGs with high predictive power and reveals known and not previously known biology. We use a mouse model of AD to validate a gene of particular interest and demonstrate for the first time that *BCAT1* is downregulated specifically in oligodendrocytes, pointing to

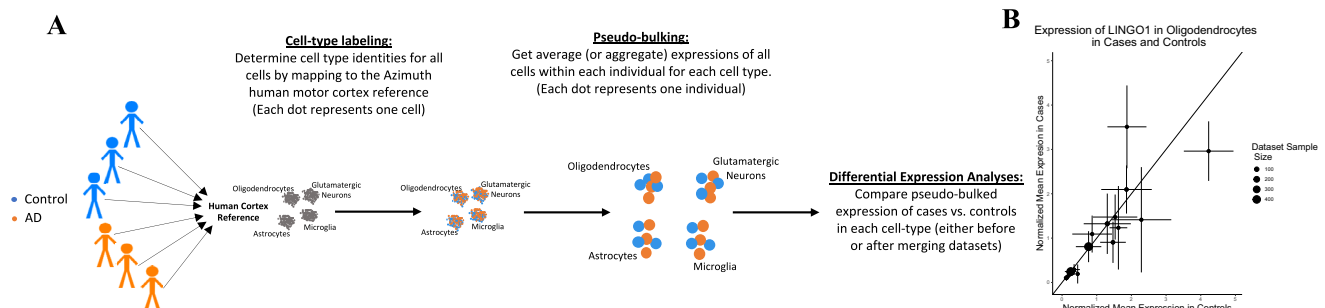
diminished branched-chain amino-acid metabolism in this cell type. We also show that SumRank DEGs are significantly enriched in genes associated with differentially accessible snATAC-seq peaks from a previous AD study, as well as human genetic associations. We find 56 DEGs shared amongst AD, PD, and HD. Moreover, we adapt SumRank to identify sex-specific DEGs. Finally, we assess factors that influence the reproducibility of an individual study's results as a prospective guide for experimental design. Our work demonstrates the importance and potential for large-scale meta-analyses to draw robust biological conclusions, especially for neuropsychiatric disorders.

## Results

### Reproducibility of DEGs in individual datasets is poor in AD and SCZ and moderate in PD, HD, and COVID-19

We first compiled data from 17 snRNA-seq studies of AD prefrontal cortex (Supplementary Data File 1). We performed standard quality control measures on each dataset (Methods) and then determined cell types by mapping them to an established snRNA-seq reference of human cortical tissue (motor cortex) from the Allen Brain Atlas<sup>37</sup> using the Azimuth toolkit<sup>38</sup>, which returns consistent cell type annotations for all datasets at multiple levels of resolution (Fig. 1). We then performed pseudobulk analyses for broad cell types, obtaining transcriptome-wide gene expression means or aggregate sums for each gene within each of the 7 cell types within each individual (aggregate sums were used for DESeq2<sup>30</sup> analyses while means were used for all other analyses). We used these values to identify cell-type-specific DEGs for AD vs. control samples in downstream analyses. Leveraging pseudobulk values removes the inherent lack of independence that characterizes multiple cells from the same individual, which would otherwise lead to substantial false positives for standard single-cell differential expression workflows. We also performed the same pipeline for six snRNA-seq studies of PD midbrain and four studies of HD caudate, determining cell types by mapping to the highest quality dataset (because there are no midbrain or caudate Azimuth atlases), and three snRNA-seq studies of SCZ prefrontal cortex. As a control experiment for a disease phenotype with a well-described and strong transcriptional response, we repeated this process for 16 scRNA-seq studies from PBMC samples from COVID-19 patients and healthy controls (Supplementary Data File 1 contains information about all datasets).

We evaluated the reproducibility of DEGs between disease and control samples by calculating DEGs based on pseudobulked values for each cell type and utilized the DESeq2<sup>30</sup> package for DEG detection using a *q* value-based FDR cutoff of 0.05, because DESeq2 with pseudobulking has been shown to have good performance in terms of



**Fig. 1 | Schematic of the procedure for obtaining differentially expressed genes.** **A** Schematic of mapping cells to determine cell types, pseudo-bulking, and obtaining cell type-specific differential expression (some cell types are removed for clarity). Orange represents AD individuals or cells, and blue represents controls. The first two sets of dots represent cells, while the third set of dots represents individuals (the sum or mean expression across all cells in a particular cell type for

that individual). **B** Example of a gene, *LINGO1*, previously highlighted as upregulated in oligodendrocytes, was shown not to be upregulated in most AD datasets. Values above the line (intercept = 0, slope = 1) are upregulated, while values below the line are downregulated. Points are means, and error bars are standard deviations. Violin plots of the expression of *LINGO1* in each individual across all datasets are shown in Supplementary Fig. 1. Source data are provided as a Source Data file.

specificity and sensitivity relative to other methods<sup>39</sup>. Strikingly, when using this criterion, over 85% of the AD DEGs we detected in one individual dataset failed to reproduce in any of the 16 others (Supplementary Table 1). Few genes (<0.1%) were consistently identified as DEGs in more than three of the 17 AD studies, and none were reproduced in over six studies. While we observed improved reproducibility in PD, HD, and COVID-19 datasets, we still failed to observe a single gene that was independently detected as exhibiting consistent cell type-specific differential expression in more than 4 of the 6 PD, 10 of 16 COVID-19, or 1 of the 3 SCZ studies (Supplementary Tables 2–5; note: the SCZ low overlap here was driven by having very few DEGs with this criteria, see Supplementary Note).

We frequently observed that genes that were identified as DEGs in multiple studies tended to rank highly, even in studies where they failed to pass the required threshold. For example, when we instead looked at the reproducibility of the top 200 genes for each cell type (ranked by *p* values), some genes were found in up to 9 of 17 AD, 6 of the 6 PD, 11 of 16 COVID-19, and 3 of the 3 SCZ datasets (Supplementary Tables 6–10). This suggests that at least some of the variability in DEG identification is driven by a lack of statistical power for any individual study. This further highlights the limitation of depending solely on one study to reliably identify DEGs that will reproduce in other studies, especially in intricate diseases such as AD. Illustrating this, we examined the gene *LINGO1*, a negative regulator of myelination previously spotlighted as a crucial oligodendrocyte DEG in a recent AD review<sup>40</sup>. While we reproduced this finding in a few individual datasets, our broader analysis suggests that *LINGO1* was not consistently upregulated in oligodendrocytes in the majority of datasets and was even downregulated in several studies (Fig. 1 and Supplementary Fig. 1),

highlighting challenges associated with identifying bona-fide and reproducible DEGs.

We also tested reproducibility by assessing the ability of the DEG sets from individual studies to differentiate between cases and controls in other studies. To standardize cross-dataset comparisons, we identified the same number of top-ranked DEGs (ranked by *p* value without requiring an explicit FDR cutoff) and derived a transcriptional disease score for each cell type in each individual. We obtained these by leveraging the UCell score<sup>41</sup>—a method that determines the relative rank of genes compared to others in a dataset. Our findings revealed that the DEGs identified by any individual AD dataset were not highly effective in predicting case-control status in other AD datasets (mean AUC of 0.68) or SCZ datasets (mean AUC of 0.55), though we observed improved power for PD, HD, and COVID-19 studies (mean AUCs of 0.77, 0.85, and 0.75, respectively) (Table 1, Supplementary Tables 11–15). Using a fixed FDR cutoff as an alternative for deriving transcriptional disease scores generally led to even poorer results (Supplementary Tables 16–18). However, we observed that DEGs identified by the 3 AD studies with a large number of individuals (>150 cases and controls each) exhibited superior predictive performance in alternative datasets (AUCs of 0.75 to 0.80) (Supplementary Table 11).

We wanted to evaluate reproducibility on a per-gene level rather than at only a combined gene set level, so we also tested the ability of individual DEGs to classify disease status for all samples across all studies. While the expected classification power for a single gene is expected to be low, we reasoned that the relative ranking of the genes could serve as an informative metric for evaluating different DEG sets. We therefore developed a single-gene metric of classification power ('Relative Classification Accuracy'), which was the normalized AUC of

**Table 1 | Comparisons of individual datasets and different meta-analysis methods in their predictive performances**

Disease	Gene set type	Mean AUC when using DEGs as a group to predict diagnoses of left-out datasets	Specificity: Percentage of DEGs in top 10% of RCA gene list	Mean relative classification accuracy of individual DEGs	Mean absolute log <sub>2</sub> fc of individual genes between cases and controls in each dataset
AD	Mean of individual datasets	0.67	34	43.4	0.15
AD	SumRank	0.78	73	64.4	0.33
AD	Merge	0.78	41	55.6	0.32
AD	Inverse variance	0.74	21	43.6	0.20
COVID-19	Mean of individual datasets	0.75	58	40.4	0.37
COVID-19	SumRank	0.91	78	58.6	0.79
COVID-19	Merge	0.90	72	57.0	0.97
COVID-19	Inverse variance	0.88	42	46.5	0.72
PD	Mean of individual datasets	0.77	57	53.0	0.31
PD	SumRank	0.88	87	71.0	0.52
PD	Merge	0.84	68	63.2	0.63
PD	Inverse variance	0.85	57	57.6	0.41
SCZ	Mean of individual datasets	0.55	37*	44.3*	0.24
SCZ	SumRank	0.62	51*	53.4*	0.35
SCZ	Merge	0.52	23*	43.8*	0.26
SCZ	Inverse variance	0.56	21*	38.4*	0.29
HD	Mean of individual datasets	0.85	48	50.2	0.58
HD	SumRank	0.84	68	62.3	0.96
HD	Merge	0.85	48	56.3	1.18
HD	Inverse variance	0.83	18	37.9	0.57

For all analyses here, the DEG lists included the same number of top genes (based on the number of SumRank genes with  $-\log_{10}(p)$  value at a cutoff identified in the main text). RCA Gene List is the list of genes ranked by their individual ability to distinguish cases from controls in all datasets (see text and Methods for more details). Relative Classification Accuracy is the mean AUC of individual genes in their ability to distinguish diagnosis status in each dataset, normalized within each disease. Mean absolute log<sub>2</sub>fc values were from comparisons of cases and controls in each dataset. \* indicates that the RCA Gene List is likely less reliable in SCZ due to the low number of datasets.

an individual gene for predicting case-control status (see Methods for more details); we ranked the genes by this metric and named the ranked list ‘RCA Gene List’. We identified the top 10% of genes in the RCA Gene List (1520, 1780, 1107, 1843, and 1742 for AD, PD, COVID-19, HD, and SCZ, respectively), reasoning that bona-fide DEGs should generally fall within this set. However, when returning to the sets of DEGs identified by individual datasets, we observed poor overlap within this list (mean of 34%, 57%, 58%, 48%, and 37% for AD, PD, COVID-19, HD, and SCZ). Even when examining the three largest AD datasets, we still observed poor performance for individual genes (37–51% in the top 10% of the RCA Gene List). Taken together, we conclude that analysis of individual datasets often fails to identify DEGs between cases and controls that reproduce in additional studies, and that this problem is exacerbated for diseases with more subtle or more heterogeneous transcriptional phenotypes such as AD. We therefore sought to explore approaches for meta-analysis that would leverage datasets from multiple studies to identify robust DEGs.

### A non-parametric meta-analysis uncovers DEGs with strong reproducibility across datasets

We tested two standard meta-analysis strategies. As one approach, we merged pseudobulk profiles together from all datasets and then conducted a differential expression analysis using DESeq2 while including the dataset ID as a batch covariate. As an alternative approach, we incorporated an inverse variance meta-analysis, a conventional approach for amalgamating GWAS summary statistics. For this, we fused the effect sizes and standard errors from each dataset’s DESeq2 results using metagen<sup>42</sup>. We used both approaches to calculate consensus DEG sets.

We found that the DEG sets identified by the merge and inverse variance strategies outperformed the DEG sets identified from individual dataset analyses. As an example, both methods correctly failed to identify significant differential expression for *LINGO1*. More broadly, the DEG gene sets had improved predictions of case-control status in omitted datasets with mean AUCs of 0.78 and 0.74, respectively, for AD and similar improvements for PD and COVID-19. Yet, even with enhanced AUCs, numerous genes identified by the meta-analyses showcased limited specificity, with <42% ranking within the top 10% of the RCA Gene list for AD (Table 1; Fig. 2). When examining the reason for this low specificity, we found an inherent weakness with these approaches: if a gene was highly significant in a small minority of datasets it would often pass significance thresholds after meta-analysis, even if no signal was observed in the remainder of the studies. We conclude that meta-analysis can improve the robustness of DEG identification, but existing methods remain prone to false positive identification.

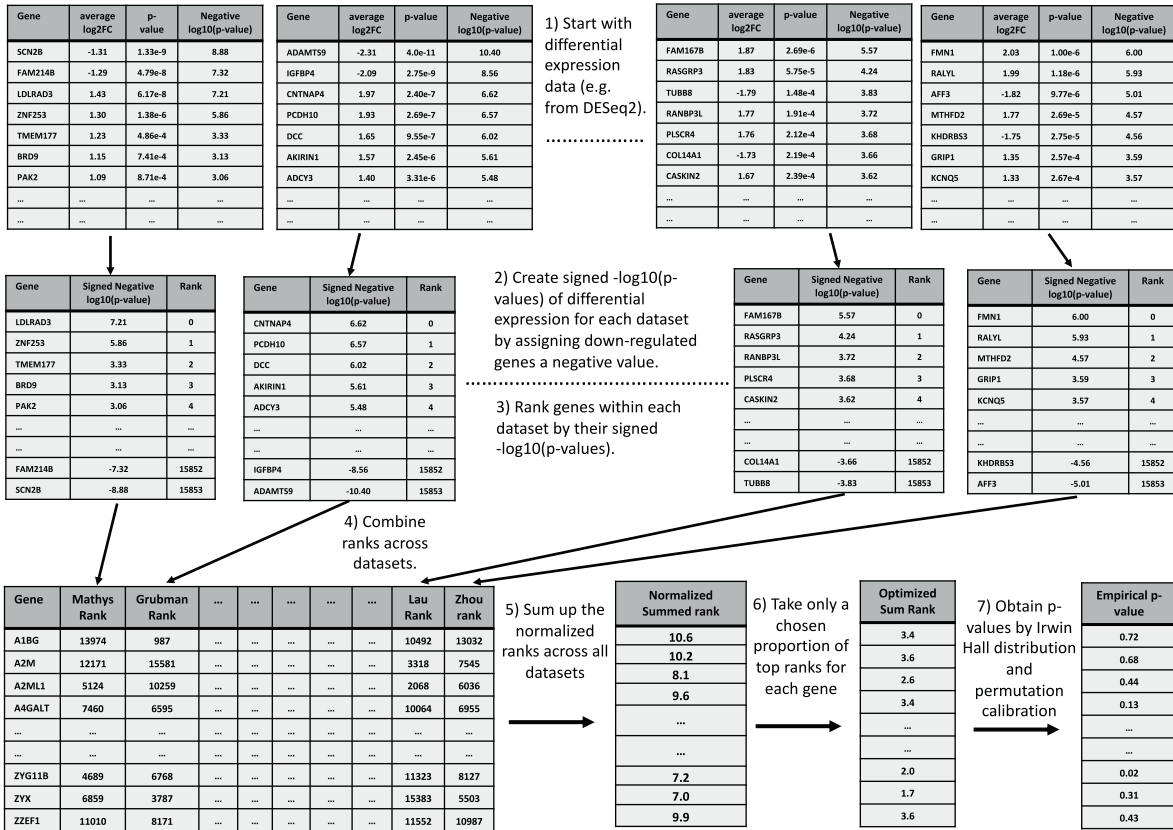
To address the issue of genes found with low reproducibility across datasets we developed a non-parametric meta-analysis method, which we call SumRank, that explicitly prioritizes reproducibility across multiple studies yet does not impose strict statistical cutoffs for any individual study (Fig. 2). This method takes the results of dataset-specific DE analysis, calculates ranks ( $p$  value based) for each gene in each dataset, and sums these ranks together across datasets. The resulting sum reflects a statistic that prioritizes genes that consistently exhibit evidence of differential expression across datasets. Given that requiring strong signals across all datasets can be overly strict—especially with large dataset numbers—we adjusted the SumRank statistic to consider only the ranks from a percentage of datasets. We set this percentage to 100% for meta-analyses based on fewer numbers of studies (PD, HD, and SCZ). For larger meta-analyses, we set this percentage based on cross-validation (65% and 55%, for AD and COVID-19, respectively), but found that our results remained consistent regardless of the exact threshold selected (Supplementary Data File 2). While the theoretical distribution of the SumRank statistic follows the Irwin-Hall distribution (see Methods), using only a subset of datasets causes

deviations from this distribution. To address this, we empirically modeled the distribution by performing 10,000 random permutations of case-control status. This allowed us to apply the identical differential expression and meta-analysis process to create a null distribution of SumRank statistics, which we used to compute empirical  $p$  values.

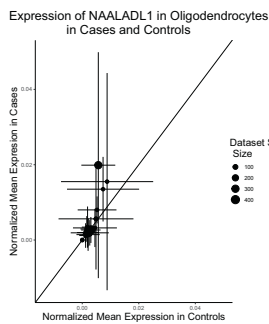
When we applied a Benjamini–Hochberg FDR cutoff of 0.05, we obtained 521 genes (394 up- and 127 downregulated across seven cell types) as significant in AD, 1597 genes in PD (1540 up- and 57 downregulated across eight cell types), 1026 genes in HD (628 up- and 398 downregulated across 15 cell types), and 1638 genes (1432 up- and 206 downregulated across 8 cell types) in COVID-19, but 0 genes in SCZ (Supplementary Data Files 3–7; Supplementary Figs. 3–10). With this cutoff, some cell types had no DEGs, so we looked for uniform  $-\log_{10}(p)$  value cutoffs that led to gene sets that maximized the ability to predict case-control status in left-out datasets. We found that for AD a  $-\log_{10}(p)$  value cutoff of 3.65 produced 814 genes (502 up- and 312 downregulated) with an AUC of 0.78, for PD a cutoff of 3.35 produced 1527 genes (1232 up- and 295 downregulated) with an AUC of 0.88, for HD a cutoff of 3.30 produced 1555 genes (740 up- and 815 downregulated) with an AUC of 0.84, for COVID-19 a cutoff of 3.90 produced 937 genes (730 up- and 207 downregulated) with an AUC of 0.91, and for SCZ a cutoff of 3.40 produced 98 genes (50 up- and 48 downregulated) with an AUC of 0.62, all higher AUCs than those from individual datasets or either of the previously tested meta-analysis procedures. Most encouragingly, we found that more than 73% of the AD DEGs fell within the top 10% of the RCA gene list, suggesting high specificity for individually identified genes. For standardization, we used the same number of genes from the SumRank meta-analyses (814, 1527, 1723, 937, and 98) for all other analyses reported in this paper. When thresholds based on corrected  $p$  values of the meta-analysis outputs were used (either through Bonferroni or  $q$  value-based FDR), it was not possible to find uniform  $p$  value cutoffs that allowed reasonable comparisons between the meta-analysis methods (in Supplementary Fig. 11, we show plots with the  $q$  value-based FDR thresholds for AD).

To assess whether clinical covariates affected reproducibility, we performed both DESeq2 and a logistic regression while regressing out all relevant covariates available for each dataset (sex, age, PMI, RIN, education level, ethnicity, language, age at death, batch, fixation interval, nCount\_RNA, and nFeature\_RNA). We did not observe any improvement in reproducibility with these analyses (Supplementary Table 16), suggesting that the datasets were generally well-controlled experiments with no systematic biases between cases and controls. We also used a newly developed single-cell differential expression method (Memento<sup>43</sup>) to assess whether this might improve reproducibility relative to using DESeq2 with pseudo-bulk data. We found that this also had slightly lower reproducibility relative to DESeq2, though SumRank combination of the Memento results improved reproducibility relative to individual Memento results (Supplementary Table 17). We also performed analyses at an increased cell resolution, looking at finer-grained subsets of the cortical neurons. We found 1611 significant (FDR < 0.05) DEGs (155 upregulated and 1456 downregulated) across the 14 neural cell types and 1,408 at a  $-\log_{10}(p)$  value cutoff of 3.65 (330 upregulated and 1078 downregulated; Supplementary Data File 3). The genes found at the broader neuron types were found repeatedly across the more specific types (e.g., *ADAMTS2*, *SCGN*, *HES4*, *CIRBP*, *PDE10A*, *VGF*), but the genes only found in the higher resolution types could represent true cell type-specific DEGs. However, when we used the more specific DEGs together with the glial genes, we obtained slightly decreased reproducibility (AUC = 0.77 for AD and 0.59 for SCZ). We believe this is potentially due to the predictive signal now being diluted across more cell types (increased model parameters), less accurate cell type mapping, or increasing missingness in the datasets at the higher cell

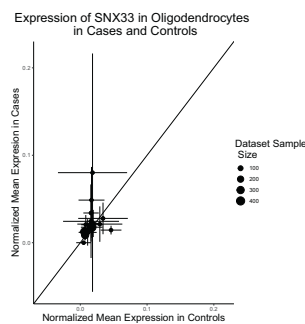
# A Sum Rank Method



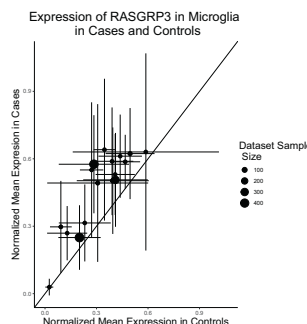
## B Merge Specific DEG



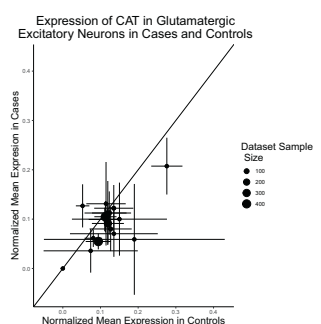
## C Inverse Variance Specific DEG



## D DEG Found by All Methods



## E SumRank Specific DEG



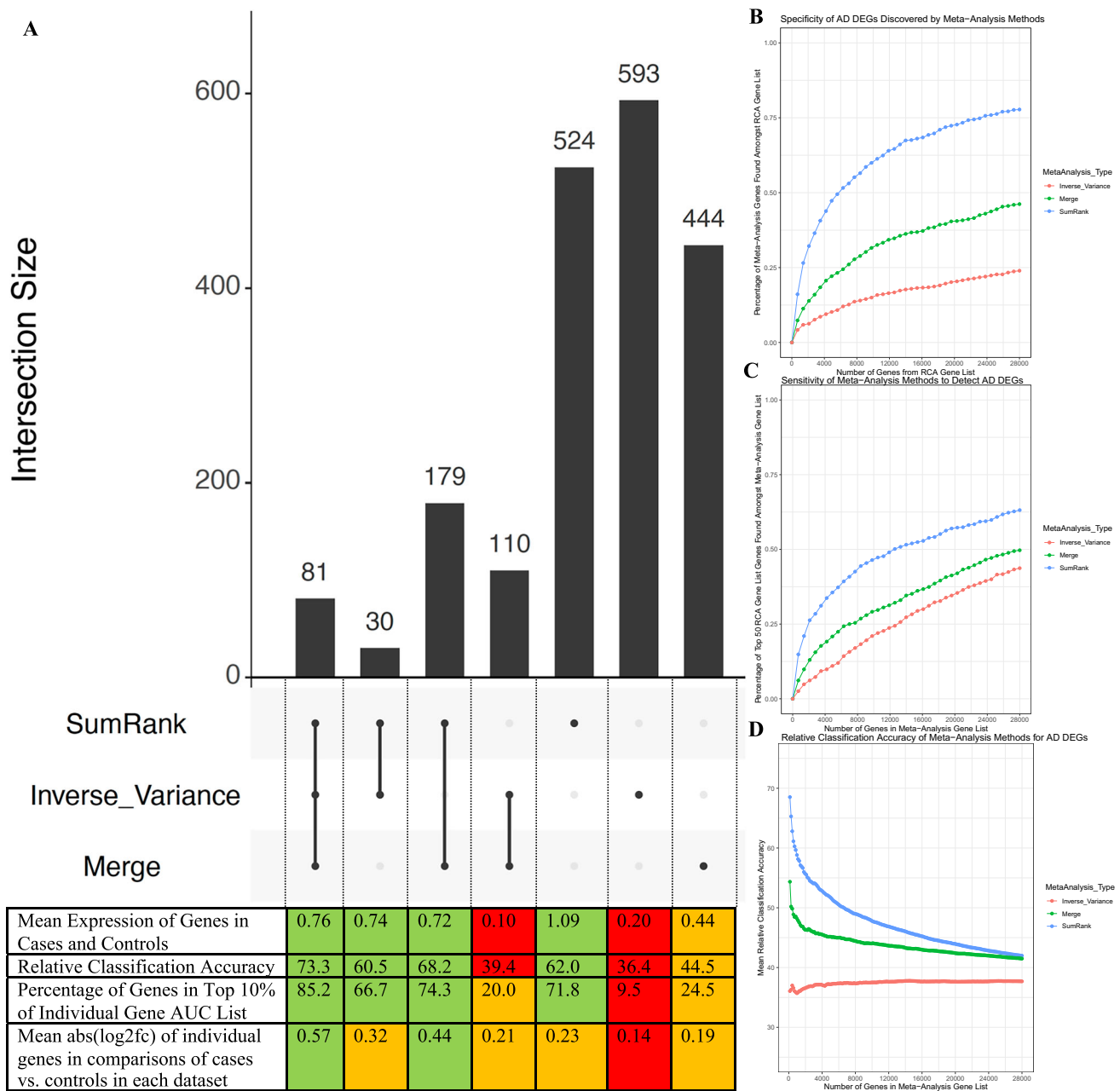
**Fig. 2 | Schematic and results of the SumRank method.** **A** Cartoon of the SumRank method: scoring each gene based on the sum of its ranks across all datasets (see text and Methods for more details). **B** Example of a gene (*NAALADL1*) putatively upregulated in AD oligodendrocytes based on the Merge method that is likely a false positive (very low expression and high variance). **C** Example of a gene (*SNX33*) putatively upregulated in AD oligodendrocytes based on the Inverse Variance method that is likely a false positive. **D** Example of a gene (*RASGRP3*) upregulated in

AD microglia based on all methods. **E** Example of a gene (*CAT*) downregulated in AD glutamatergic neurons based on the SumRank method that was not discovered by the Merge or Inverse Variance methods. Values above the line (intercept = 0, slope = 1) are upregulated, while values below the line are downregulated. Points are means, and error bars are standard deviations in all plots. Violin plots of the expression of *RASGRP3* in each individual across all datasets are shown in Supplementary Fig. 2. Source data are provided as a Source Data file.

resolution. We thus continued our subsequent analyses at the broader cell resolution.

To more carefully benchmark SumRank against alternative methods for meta-analysis, we compared the AD DEG gene sets for each method. We first focused on the 81 genes found across all three methods (SumRank, merge, Inverse Variance), reasoning that this represented a gold-standard DEG set (example in Fig. 2D and Supplementary Fig. 2). Consistent with this, we found that these genes tended to exhibit high Relative Classification Accuracy (Fig. 3). They also exhibited medium-high levels of expression (suggesting that they could be accurately quantified in individual datasets), and high mean

absolute  $\log_2(\text{fold-change})$  in comparisons of case vs control status in each dataset. We next examined genes that were identified by only a subset of methods. For example, we examined the genes that were identified by either the merge or inverse variance methods (or both), but not by the SumRank method. In contrast to our gold-standard gene set, these genes exhibited low RCA and reduced  $\log_2(\text{fold-change})$  (Fig. 3). They also tended to be lowly expressed. Taken together, these results suggest that many of these genes likely represent false positives and that the SumRank method correctly failed to identify them as DEGs. In contrast, the genes identified by SumRank (either exclusively or with one of the other meta-analysis methods) closely resembled the



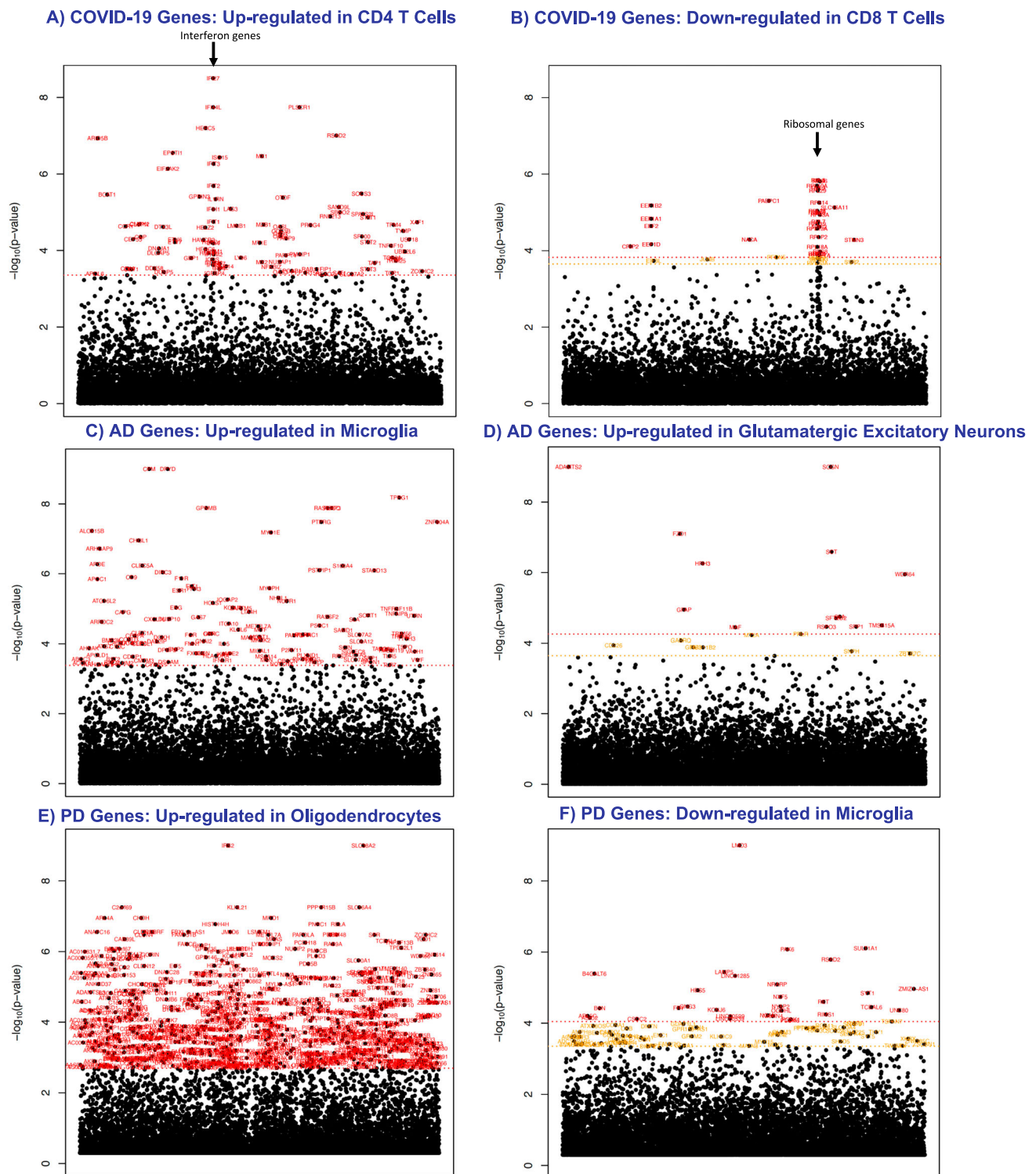
**Fig. 3 | Sensitivity and specificity of SumRank meta-analysis are better than the merge and inverse variance methods.** **A** UpSet R plot<sup>44</sup> showing intersection of AD genes discovered between the meta-analysis methods, the mean expression of the genes, relative classification accuracy (the normalized mean AUC of the individual genes in ability to predict diagnoses in all datasets), percentage of genes in top 10% of RCA Gene List, and mean abs(log2fc) from comparisons of cases vs. controls in each dataset. Color coding is based on the relative quality of the value, with green indicating the best values, orange indicating moderate, and red indicating poor. Bar graphs represent the number of genes (single values). Comparisons of meta-analysis methods in their **B** specificity, as measured by the

percentage of genes that intersect with the RCA Gene List (at different thresholds) with the same number of genes used in all meta-analyses (based on the 814 SumRank genes with  $-\log_{10}(p \text{ value}) > 3.65$ ). **C** sensitivity, as measured by the percentage of the top 50 RCA Gene List genes found amongst the meta-analysis DEGs at different thresholds, and **D** Relative Classification Accuracy, the mean AUC of individual genes in their ability to distinguish diagnosis status in each dataset (in this case averaged over all genes in the gene set). On the x axes of **B–D**, the number of genes is spread evenly across up- and downregulated genes and all the different cell types. A similar plot for COVID-19 is shown in Supplementary Fig. 11. Source data are provided as a Source Data file.

gold-standard gene set. We conclude that the SumRank method exhibits superior performance by avoiding both false-positives and false-negatives, excluding genes that do not reproduce across multiple datasets but also sensitively identifying genes whose aggregate signal across multiple datasets is reliably supportive of differential expression between cases and controls.

Examining the AD SumRank gene sets, we found that microglia, oligodendrocytes, GABA-ergic neurons, and astrocytes exhibited a

greater number of upregulated genes compared to downregulated ones. In contrast, glutamatergic neurons demonstrated more downregulated genes than upregulated, consistent with earlier findings<sup>44,45</sup> (Fig. 4, Supplementary Figs. 3, 4, 12). For AD, we detected the highest number of upregulated genes in astrocytes. In contrast, for PD, the highest number of upregulated genes was in oligodendrocytes, while for HD, FOXP2 neurons had a substantially higher number of downregulated genes than other cell types. For all diseases, over 75% of the



**Fig. 4 | Manhattan plots of differentially expressed genes. A, B AD COVID-19, and E, F PD. P values were obtained by permutation testing (see Methods for details). Significance threshold is in red with 0.05 FDR cutoff (Benjamini–Hochberg) to account for all genes being tested. In orange is a  $-\log_{10}(p)$  value cutoff that maximizes AUC (3.65 for AD, 3.90 for COVID-19, 3.35 for PD; not shown if it is higher**

than the FDR cutoff red line). The x axis is genes arranged in alphabetical order. Additional similar plots (including with SCZ) are found in Supplementary Figs. 3–10. Supplementary Data Files 3–7 show all genes with their  $p$  values. Source data are provided as a Source Data file.

DEGs were restricted to a single cell type (Supplementary Fig. 13). When examining the correlations of  $-\log(p)$  value for each cell type, we observed that cell types with greater similarities showed higher correlation (Supplementary Fig. 14). Furthermore, using the SumRank genes, we identified some predictive capacity for disease specificity

(Braak score) within AD patients ( $r=0.32$ ) when compared to separate datasets (mean  $r=0.12$ ) (Supplementary Data File 3). However, we found no predictive ability related to COVID-19 severity ( $r=0.03$ ) (Supplementary Data File 5). This was anticipated, as the severity of COVID-19 has minimal relation to transcriptional response<sup>46</sup>.

## Determining factors affecting reproducibility across diseases and datasets

The SumRank approach outperformed other methods in the context of PD, HD, and COVID-19, as shown in Table 1 and Supplementary Fig. 15. However, the margin of superiority was not as pronounced, likely due to the baseline increased reproducibility of PD, HD, and COVID-19 relative to AD. We thus sought to identify the factors underlying the differences in reproducibility between diseases. We restricted all AD datasets such that cases were only those with Braak scores of 5 or 6, and controls were only those with Braak scores of 0–2, to determine if patient selection was a major factor in reproducibility. The AUC with these selection criteria was 0.82, which, though higher than without these criteria, was still much lower than those of PD, HD, and COVID-19 (with Braak scores of 3–4, the AUC was 0.69). Given Braak scores are an imperfect measure of disease severity (since some individuals without dementia can have high Braak scores), it is possible that other metrics could decrease patient heterogeneity and increase DEG reproducibility, but alternatively, this might point to a general principle that AD might have more biological heterogeneity than PD, HD, and COVID-19, with potentially more factors contributing to the final phenotype clinically diagnosed as AD. Most strikingly, SCZ had a substantially lower reproducibility than all other diseases (Supplementary Note), which could represent substantial heterogeneity in the brains of patients with SCZ<sup>4</sup> due to inherent biology or different life experiences (e.g., more heterogeneous drug/medication use).

We next examined transcriptional effect size to assess its role in reproducibility (Supplementary Fig. 16). We found a significant ( $p = 0.0001$ ) positive correlation (Pearson's  $r = 0.72$ ) between effect size ( $\text{abs}(\log_2(\text{fold-change}))$ ) and reproducibility (average AUC for ability to predict case-control status in all datasets) for upregulated genes, meaning that genes with more differentiation between cases and controls are discovered more regularly across datasets (though for unclear reasons we find no significant relation ( $r = 0.04$ ,  $p = 0.86$ ) for downregulated genes). Consistent with this, PD and COVID-19, the most reproducible diseases, elicited the strongest transcriptional response, with mean  $\text{abs}(\log_2(\text{fold-change}))$ s of 0.93 (0.97 for upregulated genes and 0.77 for downregulated genes) and 0.86 (0.92 for upregulated genes and 0.39 for downregulated genes), respectively. In contrast, AD genes had a mean  $\text{abs}(\log_2(\text{fold-change}))$  of 0.49 (0.55 for upregulated genes and 0.40 for downregulated ones), and SCZ genes had a mean  $\text{abs}(\log_2(\text{fold-change}))$  of 0.25 (0.16 for upregulated genes and 0.35 for downregulated ones). We examined the relationship of variance (normalized to effect size by dividing by  $\log_2fc$ ) to reproducibility and found a small inverse correlation ( $r = -0.40$ ;  $p = 0.07$ ) between variance/ $\log_2fc$  and average AUC for upregulated genes (with downregulated genes  $r = -0.03$ ,  $p = 0.89$ ), providing suggestive evidence that reproducibility potentially increases with decreased variance.

We then attempted to identify experimental design factors that increased the performance and reproducibility of DEGs within the same disease. We down-sampled the individuals in the Fujita, MathysCell, and Hoffman datasets to see how varying sample numbers influenced reproducibility measures. We did not discover any clear saturation point, suggesting that reproducibility might continue to increase with even more individuals (Supplementary Fig. 17). This is consistent with our observation that for AD datasets, there is a positive correlation of Relative Classification Accuracy with sample size ( $r = 0.65$ ,  $p = 0.005$ ; Supplementary Table 16). In contrast, when we down-sampled the Stephenson COVID-19 dataset, reproducibility began to saturate at 70 individuals, and for the other COVID-19 datasets, sample sizes of only 7 cases and controls each had similar reproducibility as those with larger sample sizes (Supplementary Table 18). During this analysis we performed multiple random

iterations of the same number of samples and observed that even at 160 samples (80 cases and 80 controls), there was substantial variability in reproducibility, showing the large impact of biological variability on reproducibility (Supplementary Fig. 17). We also subsampled all AD datasets with sufficient sample size to 6 cases and 6 controls each and show that reproducibility is highly variable even at the same sample number (Supplementary Table 21). We then down-sampled the cell numbers of the AD datasets to assess their effect on reproducibility and found that reproducibility began to saturate  $-0.05$ – $0.1$  (Supplementary Fig. 18). This suggests that, particularly when doing analyses involving pseudo-bulking of broader cell types, single-cell experiments should generally prioritize sequencing more individuals rather than more cells per individual.

In addition to sample size, we noted that different studies used different phenotyping criteria to categorize disease and control individuals. For example, the Hoffman study<sup>26</sup> carefully selected AD individuals as those fulfilling a combination of neuropathological and clinical criteria. In contrast, the Fujita and MathysCell studies<sup>47,48</sup> intentionally encompassed a broader range of intermediate phenotypes amongst their cases, likely reducing DEG detection power (when using simple case-control DE designs) even with increased sample number. As a result, we found that the Hoffman dataset displayed the highest AUC of all individual AD datasets, driven not only by a large number of individuals, but also likely by the pronounced phenotypic contrasts that separate cases and controls.

We down-sampled AD datasets starting from either the most or least reproducible and found that adding datasets with even low reproducibility continues to increase or maintain the same overall reproducibility of the meta-analysis DEGs, and even down to three datasets, the reproducibility of the meta-analysis DEGs are higher than those of the individual datasets (Supplementary Tables 19, 20) and higher than the reproducibility of the 3 SCZ datasets. Consistent with this, when we only analyzed the 11 AD datasets with at least 10 cases each, the meta-analysis DEGs were not more reproducible than when all 17 datasets were analyzed (Supplementary Table 19). We performed a linear regression analysis of Braak Score on gene expression (while regressing out relevant covariates) to determine if reproducibility would improve with consideration of disease severity. Unfortunately, this did not improve reproducibility (Supplementary Table 19), potentially due to Braak scores being an imperfect correlate of disease severity. We used Gene Set Enrichment Analysis (GSEA)<sup>49</sup>, a threshold-free method to look at pathway enrichment that is generally more robust to power differences, and looked at the overlap of pathways. We found that there was a substantial decline in pathway overlap when down-sampling datasets (Supplementary Tables 22, 23) or individuals (Supplementary Fig. 17), demonstrating that power differences only account for some of the reproducibility differences, while biological variability (captured by the increased number of datasets and individuals) also is important. We also looked at gene set consistency at different thresholds by analyzing the top 25 genes from SumRank on all datasets and assessed the ranks of these genes after down-sampling datasets or individuals (Supplementary Fig. 19). We found that these genes were usually in the top 500 genes, though their average rank decreased as datasets or individuals decreased, demonstrating how both power and biological variability impact reproducibility. We also performed SumRank after mapping AD datasets to an AD-only reference (prefrontal cortex of AD patients from Mathys et al., Nature 2024<sup>50</sup>) to assess for possible AD-specific cell types/states. We found a slight decrease in reproducibility when only using the broader cell types (AUC = 0.76) as well as when using the more specific cell types (AUC = 0.65), which could be due to loss of cell type identity in AD<sup>51</sup>, leading to worse cell mapping. However, the DEGs were mostly similar and demonstrate some potential subtype-specific AD DEGs (Supplementary Data File 8).

## DEGs found in meta-analyses reveal disease pathophysiology

We explored the biological pathways associated with the genes identified in our meta-analyses, initially utilizing gene ontology (GO) via ClusterProfiler<sup>52</sup>. In the context of COVID-19, there was an upregulation of many interferon genes in CD4 and CD8 T cells, dendritic cells, monocytes, and natural killer cells (Fig. 4 and Supplementary Fig. 7). This was mirrored in the GO pathways which highlighted processes like “response to virus”, interferon response, and other related biological pathways (Supplementary Data File 9). We used gene sets generated from a new stimulation-based Perturb-seq experiment that provided more specific pathways than those generated by GOs<sup>53</sup> and found that the interferon-beta pathway in particular was upregulated in COVID-19 cell types more than the interferon-gamma, TNF-alpha, or TGF-beta 1 pathways (Supplementary Data File 10). Natural killer cells displayed upregulated pathways linked to nuclear division and chromosome segregation, stemming from the activation of cell cycle genes during cell proliferation (Supplementary Fig. 7; Supplementary Data File 9). B cells showcased elevated endoplasmic reticulum, protein folding, and protein modification pathways, which can be tied to the antibody production process. Across other cell types, there was a noticeable downregulation of many ribosomal genes, captured under the “cytoplasmic translation” pathway, potentially as a measure to thwart viral RNA translation (Supplementary Fig. 8).

For PD, the biological pathways upregulated were protein localization to the nucleus or mitochondria in oligodendrocytes and oligodendrocyte precursor cells (OPCs) and chaperone-mediated protein folding in oligodendrocytes, OPCs, endothelial cells, and astrocytes (Supplementary Data File 9; Supplementary Figs. 5 and 6), particularly due to upregulation of chaperonin genes, including *CCT3* and *MAPT*, which also harbor GWAS variants in PD<sup>54</sup>. This is consistent with a response to misfolded alpha-synuclein and Lewy bodies, which is the process leading to the destruction of dopaminergic neurons<sup>55</sup>. Interestingly, a top-downregulated microglia gene is *PAK6* (Fig. 4), a PD therapeutics target due to its role in phosphorylating *LRRK2*, a gene mutated in sporadic and inherited PD that activates substantia nigra microglia, triggering dopaminergic neuron death<sup>56</sup>.

For HD, we find upregulation of protein hydroxylation (*PLOD1*, *PLOD2*, *PLOD3*, *P4HA1*, and *P4HA2* genes) in ciliary ependymal cells, which is necessary for CSF transport. Interestingly, we find dramatic downregulation of synapse organization and neuron projection development genes in *FOXP2* neurons (Supplementary Fig. 12E; Supplementary Data File 9; Supplementary Note), pointing to their dysfunction. This is consistent with reports that *FOXP2* is degraded in HD due to its interaction with the mutant huntingtin protein, diminishing *FOXP2* neuron function<sup>57</sup>. Indeed, *FOXP2* overexpression rescues HD behaviors in mice<sup>58</sup>, while downregulation of *FOXP2* worsens HD symptoms<sup>59</sup>, and deficits in *FOXP2* lead to speech pathology that is found in late HD.

For AD microglia, cytokine production and immune response pathways were upregulated, likely promoting neurodegeneration and AD progression<sup>60</sup>. In endothelial cells, negative regulation of growth was upregulated, and in astrocytes, amino-acid catabolism was downregulated (Supplementary Data File 9). The pathways, however, were not consistent and were mixed with many other pathways of unclear relevance. Given this, we used STRING<sup>61</sup> to construct protein-protein interaction (PPI) networks using AD DEGs, searching for dense sub-networks that could improve our power to detect relevant biological pathways (Supplementary Data Files 3, 4, and 6). This yielded clearer signals. For example, the most densely connected network of upregulated microglia genes was enriched for pathways in regulation of lipid transport, driven by genes such as *APOE*, *TREM2*, and *SORT1*, which are all AD GWAS genes<sup>62–64</sup> (Supplementary Fig. 20). In GABAergic neurons, chemokine driven cellular response, cell fate specification, and wide pore channel activity (driven by the connexins *GJA3*, *GJA4*, *GJB2*, and *GJB4*) were enriched in the most connected

upregulated network. For the most connected downregulated networks, signaling by neurotrophic tyrosine kinase receptors, such as *BDNF*, *DUSP4*, *DUSP6*, and *VGF*, was enriched in glutamatergic neurons, while in astrocytes, glutamate metabolism was enriched, together showing how AD inhibits glutamatergic excitatory neuron growth and function. When we applied this approach to PD, oligodendrocytes, astrocytes, and OPCs show upregulation of networks enriched for chaperone-mediated protein folding and protein localization to mitochondria (Supplementary Fig. 21), while PD microglia show upregulation of lipid transport, revealing a common mechanism with AD microglia. Similarly, HD mural cells showed clear upregulation of chaperone folding genes as the most densely connected network, showing a possible common mechanism with PD glia.

In addition to these pathways, SumRank pointed to many genes with very clear reproducibility across a large majority of datasets that had not previously been highlighted by other AD papers in a cell-type-specific manner. For example, *PDE10A* was downregulated in excitatory and inhibitory neurons (Supplementary Data File 3). PDE inhibitors have long been proposed for AD<sup>65</sup>, and PDE10A inhibitors have shown some improvement in AD symptoms<sup>66</sup>. We also observed downregulation of *HES4* in inhibitory and excitatory neurons, *HES5* in OPCs, *VGF* in inhibitory and excitatory neurons, and microglia, and *VEGFA* in OPCs, all of which are involved in neuron<sup>67–69</sup> and endothelial growth<sup>70</sup>. Similarly, *SPPI1*, a gene associated with synapse loss<sup>71</sup>, was upregulated in endothelial cells and glutamatergic neurons, while *ADAMTS2*, a gene that breaks down extracellular matrix in the brain<sup>72</sup>, was upregulated in glutamatergic neurons. Together, this suggests that AD pathophysiology might involve inhibition of growth pathways, which could decrease synaptic plasticity and contribute to the cognitive dysfunction in AD; thus, therapeutics aimed at increasing these factors might be useful<sup>55</sup>. The importance of G protein-mediated signaling and amino acid and nucleotide metabolism dysregulation in AD was demonstrated by the fact that *RASGRP3* and *DPYD* were upregulated in microglia and *SLC38A2* was upregulated in oligodendrocytes, while *ARRDC3* was downregulated in astrocytes and *BCAT1* was downregulated in oligodendrocytes. Lastly, we observed that the *CAT* gene was downregulated specifically in glutamatergic neurons (in the SumRank analyses but not in the merge or inverse variance analyses; Fig. 2E). Catalase activity had previously been shown to be decreased in AD due to amyloid-beta<sup>73</sup>, and a catalase derivative has been proposed as a possible therapeutic for AD to decrease oxidative stress from free radicals<sup>74</sup>. These analyses suggest that *CAT* is specifically downregulated in glutamatergic neurons and not GABAergic inhibitory neurons or other cell types, consistent with the observation that excitatory neurons have increased oxidative stress and die at higher rates in AD.

Our approach of focusing on reproducible genes and predicting phenotypes in leave-one-out analyses provides some internal validation for our genes, but we wanted to compare with an independent system of AD. We thus performed experimental validation of one of the SumRank DEGs using the 5xFAD mouse line, which is a well-known model of late-onset AD<sup>75</sup> that overexpresses a mutant human amyloid-beta precursor protein, harbors multiple AD-associated mutations in human presenilin 1, and has been shown to have many phenotypic similarities to humans with AD, including amyloidosis and behavioral impairment. We looked to test a gene that was significant in the SumRank but not in the merge or inverse variance methods, and that had potential therapeutic relevance, but with no prior known cell type-specific data. We thus chose the *BCAT1* gene, which we found only by SumRank (not merge or inverse variance), to be downregulated in AD oligodendrocytes and is a cytosolic amino acid transaminase in both humans and mice. We performed multiplexed immunohistochemistry (IHC) staining on slices of the medial prefrontal cortex for *BCAT1* and measured the degree of staining in CC1 SOX10 double-positive, mature oligodendrocytes. We found that the 5xFAD mice had significantly

lower *BCAT1* expression in oligodendrocytes (Fig. 5), demonstrating for the first time in both humans and mice that *BCAT1* has oligodendrocyte-specific decreased expression in AD. *BCAT1* facilitates the transamination of branched-chain amino acids (BCAAs) to produce glutamate and GABA, which are essential for cognition, while elevated BCAAs can lead to neurotoxicity<sup>76</sup>; thus, decreased *BCAT1* expression could lead to AD progression through elevated BCAAs or decreased glutamate. Our findings also point to oligodendrocyte-specific manipulation of BCAA metabolism as a potential therapeutic for AD<sup>77</sup>. Notably, we also find *BCAT1* specifically downregulated in oligodendrocytes in HD.

We compared our meta-analysis DEGs to those exhibiting differential snATAC-seq peaks, given that differential expression should be correlated with differential chromatin accessibility. We used differentially accessible peak data from a study performing snATAC-seq on the prefrontal cortex of 44 AD cases and 48 controls<sup>51</sup> and found the genes whose promoters were nearest to the peaks. We found that the SumRank DEGs had significant ( $p < 2.2e-16$ ) enrichment for the top genes from differential accessibility analysis and even more enrichment than the DEGs from RNA of the same individuals from whom the snATAC-seq data were derived (Supplementary Fig. 22), providing an additional mode of validation. We then assessed the intersection of the 708 unique AD DEGs at the 3.65  $-\log_{10}(p)$  value cutoff with genes found in the largest AD GWAS<sup>62–64</sup> and found nine unique genes out of the 105 genes in GWAS to be shared (Supplementary Table 24;  $p = 1.3e-4$ , Fisher's exact test). When we looked at the intersection with AD whole-exome studies<sup>78–80</sup>, 4 of the 28 genes were shared ( $p = 1.1e-4$ , Fisher's exact test). Of the 1187 unique PD DEGs at the 3.35  $-\log_{10}(p)$  value cutoff, there were six unique genes out of the 72 genes in PD GWAS<sup>54</sup> shared ( $p = 2.0e-05$ ). Despite this indicating a statistically significant enrichment, it still represents a relatively minor overlap, suggesting that the genetic variants underlying predisposition to AD are often not the same as the genes whose expression are altered downstream of individuals with multiple decades of AD. We also evaluated the overlap of AD, HD, and PD genes and found 56 shared up-regulated genes amongst all 3 diseases and many more up- and downregulated genes shared between pairs of the diseases (Supplementary Data File 9). It is likely that these shared genes represent a common neurodegenerative biological pathway, but no significant GO enrichment was found.

### Adaptation of SumRank to uncover sex-specific DEGs

The female sex-bias in AD<sup>81</sup> motivated us to search for genes with sex-specific expression. We performed two types of analyses (Fig. 6; Methods). First, we used DESeq2's interaction term (SEX:Diagnosis) to look for genes with significant interaction between Sex and Diagnosis within each dataset. We then fed these values into SumRank, adding up the  $p$  value ranks of the genes across each dataset, considering only the top 65% of datasets (to be consistent with the general analyses), and using permutations (permuting sex) to calibrate the  $p$  values. This analysis will find all genes with significant differences in case vs. control gene expression between the sexes, but it could also find genes with decreased expression in one sex and unchanged expression in the other sex.

In order to focus on genes that have upregulated expression in one sex but are unchanged in the other, we devised another method that works by summing up four different scores to create a composite score. We performed differential expression and SumRank meta-analyses in DESeq2 to obtain  $p$  values for scores between males and females in only cases and in only controls, as well as cases vs controls in only males and in only females. Female-specific scores were calculated as the sum of the  $-\log_{10}(p)$  values of the cases vs. controls in females, with the  $-\log_{10}(p)$  values of the females vs. males in cases subtracted by the  $-\log_{10}(p)$  values of the cases vs. controls in males and the  $-\log_{10}(p)$  values of the females vs. males in controls. Male-specific

scores were calculated analogously, and we calibrated all  $p$  values empirically with permutations.

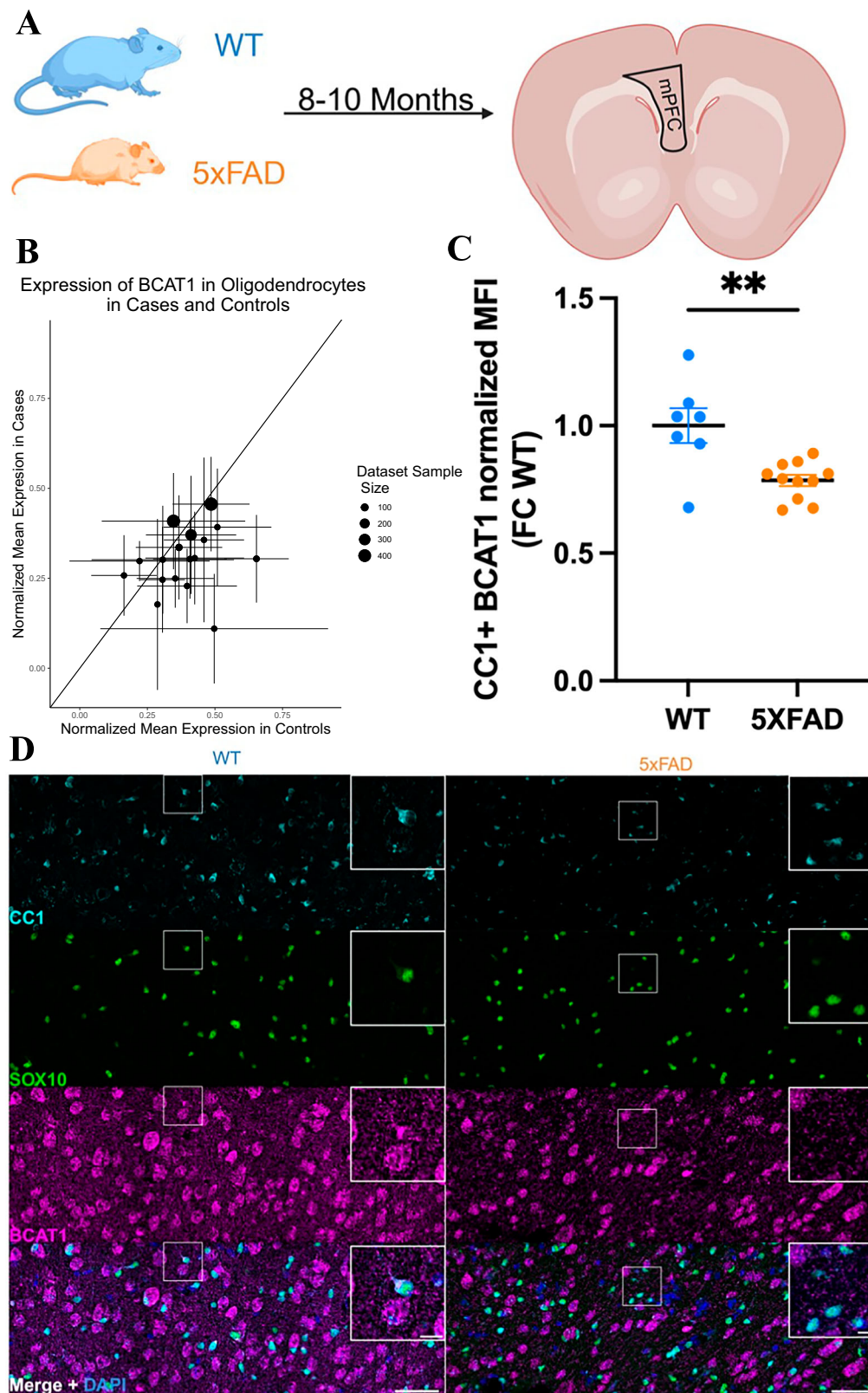
At  $q$  value or Benjamini–Hochberg-based FDR cutoffs of 0.05, no genes were significant with both methods, so we loosened our thresholds. We looked for genes that had  $-\log_{10}(p)$  values above 3.65 (the threshold chosen for the general analyses) in the Composite Score approach and were in the top 15 genes (0.1%) in the Sex Interaction approach. This led to the discovery of several female-specific genes, *SLITRK5* in OPCs, *ZFP36L1* and *DUSP1* in astrocytes, *DAPK2*, *APOE*, and *OR4N2* in GABA inhibitory neurons, and two male-specific genes, *MYC* and *IL16* in glutamatergic neurons (Fig. 6, Supplementary Fig. 23, Supplementary Data File 11). Of these, only *ZFP36L1* and *SLITRK5* were significant in the composite method at an FDR 0.05 cutoff. *ZFP36L1* is a 3'UTR binding protein that influences transcriptional regulation and has been found to be a DEG that is a candidate biomarker for AD<sup>82–84</sup>. Interestingly, the APOE risk factor is known to have a stronger association with females relative to males<sup>85</sup>. We also applied this method to COVID-19 and found *CLU* in dendritic cells and monocytes, *MTIE* in other T cells and *GOS2* in CD4 T cells as male-specific expressed and *CAMK1* in dendritic cells as female-specific expressed (Supplementary Data File 11).

The lack of clearly significant genes in any of the SumRank sex-specific analyses is likely due to insufficient power, because these analyses require at least twice as many individuals as the case-control analyses, given the extra consideration of sex. In addition, it is also probable that the sex-specific effect sizes are much smaller than the effect sizes differentiating cases vs. controls more generally, so overall, these results underscore the need for more data to better delineate these effects. We note that when we used the merge method with DESeq2 sex interaction, we found several genes that were significant at Bonferroni corrected  $p$  value thresholds of 0.05, but these genes were not significant and ranked very low in the SumRank methods due to only being significant in one or a few datasets (Supplementary Fig. 24), showing again the importance of reproducibility in these analyses.

## Discussion

Here, we assessed the reproducibility of DEGs across many AD, PD, HD, SCZ and COVID-19 datasets. We find that DEGs from single AD and SCZ datasets generally have poor reproducibility and thus cannot predict case-control status in other AD or SCZ datasets, though predictive power is improved with increased numbers of individuals in the study. In contrast, even small individual PD, HD, and COVID-19 studies have moderate predictive power for case-control status in other datasets. This study provides strong evidence that for diseases of high heterogeneity like AD and SCZ, the DEGs of case-control datasets of relatively small sample sizes (fewer than 100 total individuals), even when derived in a statistically rigorous manner, have a low likelihood of being reproduced in many other datasets and thus are more likely to be dataset specific artifacts rather than reliable indicators of disease pathology. In contrast, acute diseases or those with more uniform responses, such as PD, HD, and COVID-19, produce DEGs with moderate reproducibility across studies.

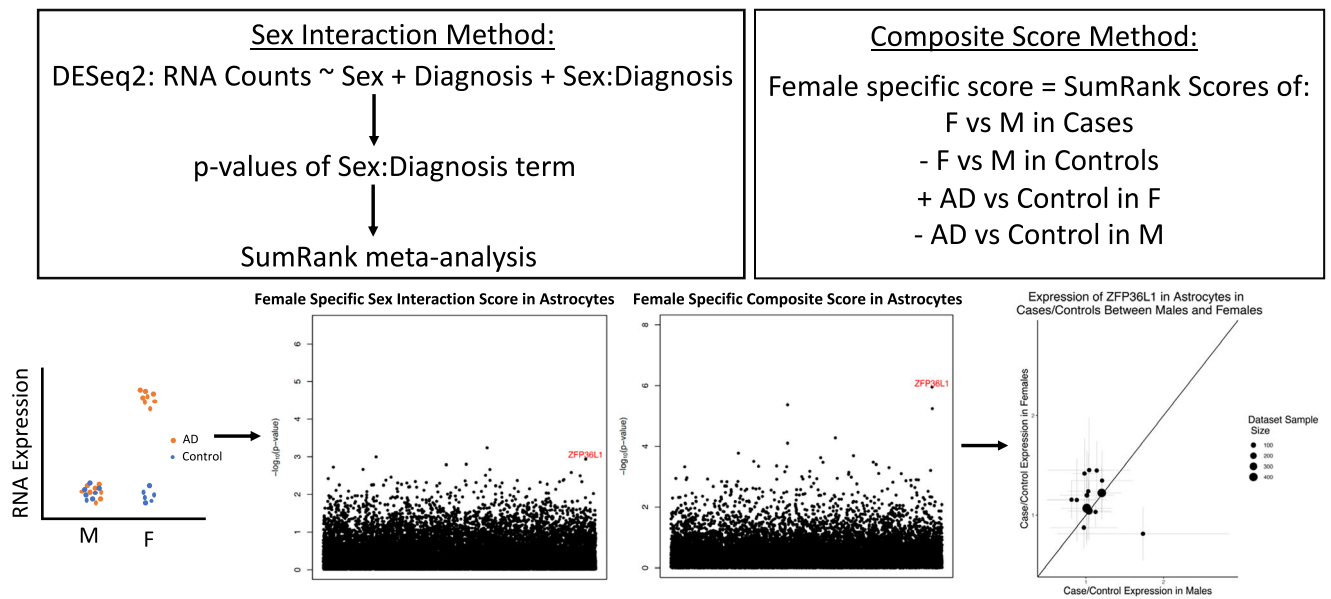
This presents a paradox in that for diseases with heterogeneous gene expression and low reproducibility, likely including many neuropsychiatric diseases, it is more important to ensure that genes are found reproducibly across multiple studies to avoid false positives. Motivated by this, we provide here a path towards GWAS-level reproducibility through the development of a meta-analysis method (SumRank) that prioritizes reproducibility across datasets. We show that SumRank outperforms merging of datasets with batch correction (the standard scRNA-seq method) and combining effect sizes with inverse variance weighting (the standard GWAS method). The DEGs found by SumRank have improved specificity as measured by the ability to predict case-control status in left-out datasets and demonstrate that many previously highlighted genes thought to be differentially



**Fig. 5 | Experimental validation of a meta-analysis AD DEG in a mouse model.**

**A** Schematic of an experiment measuring cell type-specific expression in the medial prefrontal cortex of 5xFAD mice from 8–10 months old. Created in BioRender. Adler, D. (2025) <https://BioRender.com/f15uw8o>. **B** Expression of *BCAT1* in oligodendrocytes in human postmortem snRNA-seq datasets. Values above the line (intercept = 0, slope = 1) are upregulated, while values below the line are downregulated. Points are means, and error bars are standard deviations. **C** Protein expression of *BCAT1* in oligodendrocytes of 5xFAD and WT mice

obtained by quantifying the mean fluorescent intensity (MFI) expressed as fold change (FC) over WT animals ( $n = 7$  WT, 11 5xFAD mice). Data represented as mean  $\pm$  s.e.m. Results are significant at  $p = 0.0026$  (students' two-tailed unpaired  $t$  test). **D** Representative multiplexed immunohistochemistry (IHC) staining of cortical slices from a 5xFAD mouse of *BCAT1* and 2 oligodendrocyte-specific markers (*SOX10* and *CC1*), along with the merged image (Scale bar = 50  $\mu$ M large images, 10  $\mu$ M insets). Source data are provided as a Source Data file.



**Fig. 6 | Schematic of the two methods used for assessing sex-specific expressed genes.** The sex interaction method uses the SumRank meta-analysis on the  $p$  values of the sex:diagnosis term from DESeq2, while the Composite Score method takes the composite of 4 different SumRank scores (shown here for female-specific scores; the male-specific score is defined analogously). On the bottom left is a schematic of an example female-specific expressed gene. The Manhattan plots highlight the *ZFP36L1* gene.  $P$  values were obtained by permutation testing (see

Methods for details). The ratios of mean expression of cases over mean expression of controls of *ZFP36L1* in females ( $y$  axis) and males ( $x$  axis) are plotted in the bottom right. Values above the line (intercept = 0, slope = 1) are up-regulated in females more than males, while values below the line are up-regulated in males more than females. Points are means, and error bars are standard deviations. Plots of the expression of *ZFP36L1* in individuals within each dataset are in Supplementary Fig. 23. Source data are provided as a Source Data file.

expressed in AD do not show differential expression across many datasets. The inverse variance method, though successfully utilized in GWAS, performs poorly for meta-analysis of scRNA-seq data due to dataset-specific artifacts that are carried through, such that some genes with very low  $p$  values in a small number of datasets are considered significant even though they are not differentially expressed in most datasets. This effect is much more pronounced in single-cell studies relative to GWAS due to the lower stability of RNA expression relative to DNA, leading to a greater propensity for very poorly calibrated  $p$  values. The merge method generally works much better than the inverse variance method (likely due to DESeq2's ability to have a dataset covariate correction), but still performs more poorly than the SumRank method for the same carried-over artifact issue. Moreover, the merge method is much slower than the other methods, as the merge process can take several hours, particularly for large datasets.

With the SumRank method, we were able to uncover COVID-19 biology, such as the division of NK cells and the downregulation of ribosomal genes. For PD, we consistently found upregulation of chaperone-mediated protein folding and protein localization to the nucleus and mitochondria across multiple cell types, with GWAS genes driving the signal. For HD, we find a dramatic downregulation of synaptic function in FOXP2 neurons. We also use PPI networks to increase biological pathway clarity, demonstrating upregulation of lipid transport in AD and PD microglia and downregulation of glutamate functioning via astrocytes and glutamatergic neurons, driven by key GWAS genes and consistent with known dysfunction of glutamatergic neurons in AD. Users can employ similar tools to improve biological insight from SumRank DEGs. We validate the *BCAT1* gene as downregulated in oligodendrocytes of humans and mice with AD.

A particularly interesting finding was that SCZ has the lowest reproducibility of the studied diseases, which could be due to its relatively increased biological complexity. SumRank only slightly increases reproducibility for SCZ (mean AUC = 0.55 for individual datasets and 0.62 for SumRank). We believe highly complex diseases

like SCZ will likely require either sub-phenotyping to find more biologically homogeneous cohorts or much larger sample sizes and/or more datasets to achieve clear reproducible DEGs, similar to GWAS for SCZ and autism, despite the high heritability of these diseases<sup>86</sup>. AD has lower reproducibility than PD, HD, and COVID-19, but still has hundreds of DEGs. Nevertheless, the biological inference was less clear due to the large number of pathways discovered with less support for each one. This could indicate true biological heterogeneity (i.e., relatively more distinct pathways all contributing to this disease), which would decrease power for pathway inference. Analyzing reproducibility of networks using approaches such as WGCNA<sup>87</sup> or with alternative data sources such as ATAC-seq, ChIP-Seq, metabolomic, or proteomic<sup>88</sup> data could improve inferences. We also note that for our analyses, we pseudobulk at both broad cluster levels as well as more narrow neuron levels. Our choice of cluster resolution will inevitably obscure heterogeneity within the clusters, while there will be cell type-specific DEGs at other cluster resolutions, some of which might be relevant for the disease. SumRank is a generic method that can be used at any cell type resolution, so additional studies can be performed to assess biological pathways that may be found in potentially more rare cell types or subpopulations.

Single-cell transcriptomic case-control studies have, to date, involved limited numbers of individuals for studies outside of AD and COVID-19, and for many neuropsychiatric disorders, it will likely take many years to reach the same cohort sizes and number of studies as in AD and COVID-19. It is thus critical to apply the lessons learned here to diseases with increasing numbers of individuals sequenced. Our results suggest that when designing scRNA-seq case-control studies, it is more important to sequence a larger number of individuals rather than more cells once there are over ~40 cells per cell type of interest (when pseudo-bulking). Investigators could also consider looking at extremes of phenotypes to increase power. Most importantly, it is critical for all studies, particularly small ones (fewer than 50 cases and controls each, based on observations from this study), to demonstrate clear reproducibility in the DEGs discovered and show that (ideally for

each individual gene) this reproducibility exceeds the reproducibility expected by chance.

We lastly highlight limitations of the SumRank method and single-cell meta-analysis methods in general, which will be important to overcome in the future to produce GWAS-quality meta-analyses. For the SumRank method in particular, a substantial limitation is the lack of weighting, which can cause power limitations. We were not able to come up with a reliable method for weighting the studies, because, for example, although there was a general correlation of predictability of DEGs (AUC) with number of individuals, the relationship was not uniform as some larger studies had poorer predictive power for reasons such as more heterogeneous phenotyping or poorer sequencing quality (e.g. multi-ome data in the Su COVID-19 dataset), so weighting by number of individuals, number of cells, or sequencing depth could lead to substantial biases. SumRank also requires multiple datasets of the same tissue region at the same context, or the method will usually lose power and fail to find context-specific signals. One possible way to find these signals is to first assess for more homogeneous clusters of individuals or datasets, apply SumRank to these subsets of the data, and then determine the characteristics of the individuals/datasets leading to the unique signals. The primary computational burden of SumRank is performing multiple rounds of differential expression after permutations. For 4 COVID-19 datasets on one cell type (Monocytes), the SumRank command itself takes only 36.8 minutes to run 1000 times on a standard laptop, while additional datasets add less than linear time (42.2 minutes for 8 COVID-19 datasets), while the merge process takes 44.4 and 119.3 minutes for 4 and 8 datasets, respectively, showing non-linear scaling, and the inverse variance process takes 111.7 and 149.7 minutes. However, running DESeq2 on 1000 differential expression permutations of the same datasets takes 9.3 hours (if not parallelized), which will be required for all methods if  $p$  values are obtained empirically. It is currently unclear how to speed up the differential expression process, but this study used clusters with parallelization to ensure a reasonable analysis time. SumRank was used in this study with a discrete cluster-based analysis of cases and controls with separate analysis of cell types, though in future work, it can be adapted to analyze cell and phenotype heterogeneity on a continuous spectrum and with cellular communities rather than discrete cell types. Lastly, a SumRank-specific limitation is the optimization of the choice of the number of datasets to use, which can increase computational time. Other limitations are generic to all single-cell meta-analysis methods. For example, there is currently no method to account for possible relatedness amongst the individuals either within or across datasets, unlike GWAS meta-analyses, which are now able to condition out relatedness without fully removing related individuals<sup>89</sup>. Accounting for relatedness is likely more difficult for RNA and other modalities relative to DNA, but future meta-analyses could potentially account for this by either having genotyping of all patients or looking for increased correlation in expression above the background. Similarly, population structure (e.g., individuals of a certain ethnic background being enriched in cases) could lead to spurious associations and must be accounted for in future analyses.

Refinement of GWAS methodologies, including addressing many of these issues, took over a decade<sup>90</sup>. Meta-analyses of single cell data face many challenges beyond those of genetic data, such as a greater propensity for dataset specific artifacts (due to the relative instability of RNA and potential for gene expression changes during technical processes), expression differences across tissues and tissue regions (increasing the noise when combining datasets), differences in life environments between cases and controls (e.g., medication use), and less clear principles for how genetic relatedness affects gene expression between individuals. On the other hand, the average effect sizes of RNA are usually much higher than genetic effect sizes, which are brought down due to natural selection, as evidenced by the mean effect size of individual DEGs for AD in this study being 1.40 relative to

1.05 for AD GWAS<sup>62</sup>. This means it is likely that lower sample sizes will be required for single-cell case-control analyses relative to GWAS. Overall, this study is intended to take a strong step in bringing single-cell case-control studies to GWAS levels of reproducibility, which we hope will clarify the cell-type-specific biological changes involved in different conditions, ultimately leading to more reliable drug targets to reverse disease pathophysiology<sup>91</sup>.

## Methods

### Ethics

This study complied with ethical regulations as indicated by the studies of each of the previously published scRNA-seq and snRNA-seq datasets. For mouse experiments, all procedures were approved by the NYU School of Medicine Institutional Animal Care and Use Committee and all ethical regulations provided by NYU were followed under protocol IA16-01351.

### Datasets

Count matrices were downloaded from GEO for GSE129308 (Otero-Garcia et al.<sup>92</sup>), GSE147528 (Leng et al.<sup>93</sup>), GSE140511 (Zhou et al.<sup>94</sup>), GSE138852 (Grubman et al.<sup>95</sup>), GSE174367 (Morabito et al.<sup>96</sup>), GSE157827 (Lau et al.<sup>97</sup>), GSE163577 (Yang et al.<sup>98</sup>), GSE183068 (Sayed et al.<sup>99</sup>), GSE148822 (Gerrits et al.<sup>100</sup>), GSE160936 (Smith et al.<sup>101</sup>), GSE167494 (Sadick et al.<sup>102</sup>), GSE157783 (Smajic et al.<sup>10</sup>), GSE184950 (Wang et al.<sup>15</sup>), GSE193688 (Adams et al.<sup>14</sup>), GSE243639 (Martirosyan et al.<sup>12</sup>), GSE148434 (Lee et al.<sup>13</sup>), GSE173731 (Garcia et al.<sup>103</sup>), GSE180928 (Lim et al.<sup>23</sup>), and GSE152058 (Matsushima et al.<sup>21</sup>). Other matrices were downloaded from Synapse (Mathys et al.<sup>44</sup>, Mathys et al.<sup>48</sup>, Hoffman et al.<sup>26</sup>, Fujita et al.<sup>24</sup>, Ruzicka et al.<sup>4</sup>), CellxGene (Gabbito et al.<sup>45</sup>), Zenodo (Batiuk et al.<sup>5</sup>: <https://zenodo.org/records/6921620>), NEMO (Ling et al.<sup>6</sup> and Handsaker et al.<sup>104</sup>), the Broad Institute Single Cell Portal (SCP1768: Kamath et al.<sup>11</sup>), or from the authors directly (Barker et al.<sup>105</sup>). Relevant meta-data were also retrieved from the corresponding publications. COVID-19 datasets were obtained from Tian et al.<sup>106</sup>.

### Quality control and data processing

Count matrices were first converted to Seurat objects using the Seurat V4 pipeline. Mitochondrial percentage, nCount\_RNA, and nFeature\_RNA were assessed for each dataset, and cells with outlier values were removed from the dataset (Supplementary Data File 1). Subsequently, SCTransform v2 was performed for normalization and variance stabilization of the data, then PCA was run with 30 PCs maintained, and UMAP was run on the PCA-reduced dataset with dims 1:30 selected. Cell types were then determined by mapping to the class and subclass groupings of the Azimuth motor cortex for AD and SCZ datasets and the Azimuth PBMC reference for COVID-19 datasets using FindTransferAnchors in Seurat with 1:30 PCA dimensions, and refDR reduction, with all other settings left at default. Mapping to the Azimuth reference ensures that even if the mapping is not perfect, there likely will be no bias since the mapping quality should be similar for the cases and controls within each dataset. For PD and HD datasets the cells were mapped to the Kamath et al.<sup>11</sup>, PD and Matsushima et al.<sup>21</sup>. HD datasets, respectively, due to lack of other reliable midbrain and caudate references. For the Lim HD dataset, we observed that the FOXP2 cells did not have reliable mapping, so we did not include them in our analyses (see Supplementary Note).

### Differential expression

Each dataset was pseudobulked by obtaining either the aggregate sum of all counts (for DESeq2 analyses) or the mean value (for all other analyses) for each cell type at the Azimuth class or subclass level for each individual in each dataset. Differential expression was done by comparing cases to controls within each cell type and using multiple different methods. For our general analyses, DESeq2<sup>30</sup> was used to

compare cases to controls with `logfc.threshold` and `min.pct` set to 0 to ensure that all genes were included (`pseudocount.use` was set at 1 due to the need for round count numbers for DESeq2). No normalization is needed prior to DESeq2 analyses, because DESeq2 performs internal normalization through its median of ratios method to account for sequencing depth and RNA composition. Mitochondrial genes were removed from all results, and the final gene set was chosen as the intersection of all of the datasets for the particular disease, leading to 15,201 genes for AD, 11,067 genes for COVID-19, 17,823 genes for PD, 14,833 genes for HD, and 17,420 genes for SCZ. To test down-regulation, differential expression was done between controls relative to cases with the same downstream process repeated as for the upregulated genes. Violin plots were made in Seurat using the `VlnPlot` command after subsetting to the cell type and gene of interest. DESeq2 was also used in separate differential expression analyses while regressing out relevant clinical covariates (any of the following if they were present in the dataset's metadata: sex, age, PMI, RIN, education level, ethnicity, language, age at death, batch, fixation interval, `nCount_RNA`, and `nFeature_RNA`) using `design = -Diagnosis+Clinical-Covariate`. Differential expression was also done using logistic regression with the "FindMarkers" function in Seurat V4 with `test.use = "LR"` and `latent.vars` set to the clinical covariates with all other settings set to default. Linear regression was performed in R, fitting a model of Braak score on gene expression and clinical covariates using the "lm" function in base R with all other settings set to default.

To test the ability of each gene to predict case-control status in each dataset (as a separate analysis from the general differential expression analyses above), we used logistic regression models of case-control status with and without each gene as implemented in the "FindMarkers" function in Seurat V4 with `test.use = "LR"`, `pseudocount.use=0.01`, `logfc.threshold=0`, `min.pct=0` (with all other settings at default) and obtained the `log2fc` and  $p$  values for each gene separately for each cell type and each dataset. We then took the mean of each gene's `abs(log2fc)` and signed  $-\log_{10}(p)$  values (negative for genes with negative `log2fc` values) in all datasets to obtain each gene's average ability to predict case-control status across all datasets (separately for each cell type).

To test the Ruzicka et al. differential expression pipeline, we converted the provided ACTIONet RDS object into a single-cell experiment object and separated the dataset into the McLean and MtSinai cohorts. We then created pseudobulk profiles with the mean of log-transformed counts within each individual and cell type. We filtered out the SZ3, SZ15, SZ24, SZ29, and SZ33 individuals and cells with a capture rate less than 0.05, as done by Ruzicka et al. We then removed the effect of batch and HTO variables using the `removeBatchEffect` function in `limma`<sup>107</sup> version 3.46.0, while incorporating age (split in half into older age and younger age), sex, postmortem interval, and the log transform of average number of UMIs per cell. We then used `muscat` version 1.18.0 to perform differential expression with the `limma-trend` model using `muscat` default filtering for genes and `min_cells=10` (see Supplementary Note for more details and explanation).

### SumRank meta-analysis

The genes of all datasets were ranked by their signed  $-\log_{10}(p)$  values, with genes having negative `log2(fold-change)`s being set to negative so that downregulated genes would be at the bottom and up-regulated genes at the top. The ranks of each gene for each dataset were then normalized by first subtracting one from them and then dividing by one less than the total number of genes (so that the highest-ranked gene was 0 and the lowest-ranked gene was 1). To improve power, by removing the influence of datasets that might have poor scores for artifactual reasons, only the ranks of the top datasets were considered for each gene. The number of datasets chosen for consideration was based on the ability of its resulting gene set to most accurately predict

case-control status in left-out datasets (measured by AUC; see below), with the additional specification that at least half of the datasets be used. AUC (area under the receiver operating characteristic (ROC) curve) is the area under the ROC curve, which plots sensitivity against specificity. We took the sum of the normalized ranks of the top datasets for each gene. If the sum was greater than the number of datasets divided by two, we set the value to the number of datasets divided by two (to ensure that genes that were consistently not differentially expressed would not be considered significant).

The AUC represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative, with 1.0 being a perfect score and 0.5 being the lowest score. This metric allows us to compare the performance of different models, assuming the datasets are roughly balanced between cases and controls. The Irwin-Hall distribution is the theoretical null distribution for the SumRank statistic, because it assumes that the genes in each study are uniformly distributed and each study is independent of the other, and the Irwin-Hall distribution is the sum of independent, uniformly distributed random variables. We thus initially obtain  $p$  values for each gene using an Irwin-Hall distribution (two-sided) as implemented in the unified version 1.1.6<sup>108</sup> package, `dirwin-hall` function, with the number of datasets as the number of uniform distributions specified. However, it is possible that genes are not uniformly distributed, given the complexities of gene expression, and we also choose only a subset of datasets for each gene, so for both of these reasons, the distribution will deviate from Irwin-Hall. We thus calibrated the  $p$  values by permutations (see below).

### Merge meta-analysis

After quality control, the Seurat objects for each dataset were first subsetted to the relevant cell type and then merged using the Seurat merge function with all settings at default. The count matrices for the merged objects had 1 added to them (for a pseudocount) and were then converted to DESeq dataset types with the `DESeqDataSetFromMatrix` command with `design = -Diagnosis+Dataset`, to provide some accounting for dataset-specific batch effects (this design regresses out dataset-specific artifacts). DESeq2 differential expression was then performed, and results were extracted for the Diagnosis variable ( $p$  values and `log2` fold-changes for each gene).

### Inverse variance meta-analysis

Differential expression effect sizes (`log2` fold-change) and standard errors for each gene and each dataset were obtained from the DESeq2 output as described in the Differential Expression section above (with `design = -Diagnosis` and no other variables regressed out). These summary statistics were then put into the `metagen` function from the meta version 6.5.0 R package<sup>42</sup> to obtain combined effect sizes across the datasets with `sm = "OR"` (to specify odds ratio was used), `fixed=FALSE`, `random=TRUE` (to specify using a random effects model, given the expected heterogeneity in the datasets), `method.tau = "REML"` (restricted maximum likelihood method to obtain the estimator from inverse variance weighting), `hahn=TRUE` (Hartung and Knapp statistic adjustment), `control=list(stepadj=0.1,maxiter=10000)`. These parameters are recommended to minimize the risk of false positives<sup>42</sup>. The effect sizes were obtained from `TE.random`, and the  $p$  values obtained from `pval.random` (two-sided). When we attempted to improve the inverse variance method by only taking a certain percentage of top datasets, we found that this did not increase the AUC, so we retained all datasets for this analysis.

### Permutations for obtaining empirical $p$ values

To calibrate  $p$  values for case-control differential expression, permutations of case and control status were performed either 1000 or 10,000 times by sampling without replacement from the diagnosis labels of each individual (1000 times for the sex analyses and 10,000

times for the general case-control analyses). We chose 10,000 permutations for the case-control analyses, since this allows us to obtain  $p$  values  $< 1e-8$ , which is  $1/(10,000 \times 15,000)$ , where 15,000 is the approximate number of genes tested (1000 permutations allows us to obtain  $p$  values  $< 1e-7$ ; since no gene reached near that  $p$  value for the sex-specific analyses, we believed that 1000 permutations would be sufficient). The relevant analysis procedures were then done in the standard way (as specified above) to obtain negative  $\log p$  values for each gene. The null distribution for the real data was then taken to be the full list of all  $-\log_{10}(p)$  values across all permutations and all genes (i.e., the length of the list was the number of permutations times the number of genes).  $P$  values for the real data were then calculated as the proportion of times the  $-\log_{10}(p)$  values of the null distribution list were higher than the value of the gene for the real data.

For the analyses of sex differences, the permutations were done the same way, except permuting the sexes within the controls and cases separately (and no permutations of diagnosis status). The sex-specific analyses (see below) were then conducted in the same manner, and empirical  $p$  values for the real data were obtained with the same method as for the case-control differential expression.

### Leave-one-out analyses

The accuracy of genes obtained from each analysis was assessed by the ability of the genes to predict case-control or disease severity in left-out datasets. For each analysis where this approach was conducted, the analysis was conducted with all datasets except one that was left out (alternating so that analyses were done with each dataset left out). The resulting gene sets were then used to create a “transcriptional score” for each individual specific to each cell type using the `AddModuleScore_UCell` from the `UCell` package (v1.3)<sup>41</sup> with `maxRank` set to 16,000 to ensure that all genes were used for the analyses. Scores of 0 were set to NA. `UCell` scores were normalized such that for each cell type, the minimum of the scores was subtracted from each score, and the results were then divided by the range of the scores for that cell type (maximum score minus minimum). Missing scores were then set to the mean of the scores of that cell type. When the gene set included multiple genes, a composite transcriptional score was created for each individual as the sum of the `UCell` scores across each cell type for up-regulated genes minus the sum of the `UCell` scores across each cell type for downregulated genes (note: endothelial cells in AD datasets were not used due to incomplete coverage on all datasets for this cell type and the observation that including it decreased AUC).

A logistic regression model was created from the `UCell` scores of each individual and their diagnosis statuses using all datasets except the left-out dataset. This model was then tested on the `UCell` scores and diagnosis statuses of the left-out dataset, with AUC determined from “`auc`” function of the `pROC` R package version 1.18.4<sup>109</sup>. To determine the ability of the genes to predict disease severity, a linear regression model was created from the `UCell` scores of each individual and their disease severities (Braak scores for AD, on a scale of 0 to 6, and a scale from 0 to 3 for COVID-19, with 1 indicating mild, 2 indicating moderate, 3 indicating severe based on clinical status of the patients). For the disease severity calculations, only disease cases were used to prevent confounding from the ability to predict general case vs. control status. For COVID-19 analyses, only datasets that had all cell types were used. For AD analyses, the Barker dataset was not used for disease severity calculations, because this dataset specifically focused on individuals with high Braak scores (some of whom had normal cognition and some of whom had impaired cognition).

We used the matrix of `UCell` scores for each individual across all datasets and all cell types and performed a heatmap using R with the settings `symm=T` and all other settings set to default. RCA Gene Lists were obtained specifically for each cell type by using each individual gene to create a `UCell` score for each dataset and then following the same process as above. We separated the genes into up- and

downregulated sets based on whether the mean expression of the gene was higher in cases relative to controls or vice versa in all datasets. We then ranked each list by its mean AUC in predicting the case-control status of the individuals in each dataset. These lists were called “RCA Gene List” throughout the paper. Relative Classification Accuracy was defined as the AUCs from the RCA Gene List, normalized by subtracting the minimum value for the particular disease and dividing by the range of AUCs for that disease.

Hoffman, Fujita, MathysCell, and Stephenson dataset individual down-samplings were performed by taking a random sample (with replacement) of cases and controls 20 times for each number of cases and controls and repeating the standard individual dataset analyses as described above. Cell number down-sampling was performed by randomly taking different proportions (0.001, 0.005, 0.001, 0.05, 0.1, 0.5) of cells from each dataset and then performing differential expression and SumRank meta-analyses as described above. AD datasets were also down-sampled one at a time, either from the most reproducible (as measured by gene set AUC) or the least reproducible. SumRank meta-analysis was then performed with these down-sampled sets of datasets, with 65% of datasets chosen unless this number was  $< 7$ , in which case either 7 datasets were chosen or all datasets were chosen (if this was  $< 7$ ).

### Sex-specific analyses

Two methods were used to determine sex-specific differential expression. The first method we call the Sex Interaction Method. In this method, differential expression was performed for each dataset with DESeq2 using the counts matrix of the data subsetted to cell type and using design = `-Sex+Diagnosis+Sex:Diagnosis`. This design tests the effect of each term (Sex, Diagnosis, and Sex:Diagnosis) on gene expression. We took the effect sizes of the interaction term (Sex:F.DiagnosisAD) and obtained their  $p$  values for each dataset. The signed  $-\log_{10}(p)$  values for each dataset were then combined using the SumRank meta-analysis method described above, with  $p$  values calibrated empirically using permutations as described above.

The second method is called the Composite Score Method. In this method, four different scores were combined to create a composite score. Standard differential expression was performed in DESeq2 between males and females in only cases (Score 1) and in only controls (Score 2), as well as cases vs controls in only males (Score 3) and in only females (Score 4). SumRank meta-analyses were then performed for each of these individual analysis types to obtain  $-\log_{10}(p)$  values. For female specificity, the composite score was calculated as the sum of the  $-\log_{10}(p)$  values of the females vs. males in cases, with the  $-\log_{10}(p)$  values of the cases vs. controls in females subtracted by the  $-\log_{10}(p)$  values of the females vs. males in controls and the  $-\log_{10}(p)$  values of the cases vs. controls in males (Score 1 + Score4 - Score3 - Score 2). For male specificity, the composite score was calculated as the sum of the  $-\log_{10}(p)$  values of the males vs. females in cases, with the  $-\log_{10}(p)$  values of the cases vs. controls in males subtracted by the  $-\log_{10}(p)$  values of the males vs. females in controls and the  $-\log_{10}(p)$  values of the cases vs. controls in females. These  $p$  values were then calibrated empirically with permutations as described above. We looked for genes that had  $-\log_{10}(p)$  values  $> 3.65$  in one of the analyses and were in the top 15 (0.1%) of genes in the other analysis.

For several of the COVID-19 datasets, some of the sex statuses of the individuals were not listed, so these were obtained by creating a composite RNA score of Y chromosome genes (*NLGN4Y*, *LINCO0278*, *TTY14*, *TMSB4Y*, *EIF1AY*, *USP9Y*, *KDMSD*, *ZFY*, *UTY*, *DDX3Y*, and *RPS4Y1*), which were able to differentiate sexes in the dataset well (total expression of these genes  $> 10$  was defined as genetic male).

The ratio of mean expression of cases over mean expression of controls for females and males was calculated for plotting. The standard deviations for these were calculated by the error propagation

formula as  $Ratio^* \sqrt{\left(\frac{sd(A)}{Mean(A)}\right)^2 + \left(\frac{sd(B)}{Mean(B)}\right)^2}$ , where Ratio is  $\frac{mean(A)}{mean(B)}$ , and  $A$  is the expression in cases, while  $B$  is the expression in controls. Standard deviations were calculated separately for males and females, and both were plotted.

### Human genetic comparisons

Significant genes from GWAS of AD<sup>62–64</sup> and PD<sup>54</sup> were inferred as the genes most proximal to the genome-wide significant genetic variants from the studies or those prioritized through various metrics by the study authors. Significant genes from AD whole-exome association studies<sup>78–80</sup> were inferred as the genes with exons harboring the significant genetic variant. We assessed the statistical significance of overlap of the meta-analysis genes with human genetic genes by Fisher's exact test (two-sided) as implemented in R (fisher.test function).

### GO analyses

Cluster Profiler 4.0<sup>52</sup> was used to find biological pathways with statistically significant enrichment from the meta-analysis gene sets. The organism was set to human (org.Hs.eg.db), ont (subontology) was set to BP (biological process), and pvaluecutoff was set to 0.05. The up- and downregulated gene sets were analyzed with these settings, with the rest of the settings at default.

COVID-19 pathways were also analyzed by comparing the overlap of the up-regulated genes in each cell type to the gene sets derived from a database generated by Perturb-Seq experiments in which six cell lines were stimulated with different perturbations (interferon-beta, interferon-gamma, transforming growth factor beta 1, and tumor necrosis factor-alpha) and then had expression of individual genes knocked down with CRISPR guides to assess the effect of each gene on the perturbation response. This provided more specific gene sets for these pathways than could be obtained by standard GO<sup>53</sup>. The specific pathways were coded as IFNG\_REMOVE\_IFNB; IFNB\_REMOVE\_IFNG; IFNG\_REMOVE\_TNFA; TNFA\_REMOVE\_IFNG; IFNB\_REMOVE\_TNFA; TNFA\_REMOVE\_IFNB; TNFA\_REMOVE\_TGFB1; TGFB1\_REMOVE\_TNFA, where each gene set was the genes involved in the specific perturbation pathway that were not involved in other pathways. The overlap of the meta-analysis up-regulated genes with the top 100 genes from each pathway was examined to determine more specifically the pathways involved in COVID-19 in each cell type, where the dominant pathway was determined as the pathway with the highest overlap after removing the genes from other pathways with high overlap.

### Overlap calculations

GSEA was used as implemented in the Cluster Profiler 4.0<sup>52</sup> and MSigDB v7.5.1<sup>49</sup> packages. GO gene sets were obtained with the msigdb command with species "Homo sapiens", category "C5", subcategory "GO:BP". GSEA was then performed on the ranked, signed  $-\log_{10}(p)$  values of differential expression obtained from pseudo-bulked DESeq2 analyses (as in the Differential Expression section above) using the GSEA command with pvalueCutoff = 1.1 and all other settings at default. The pvalueCutoff was set to 1.1 to ensure all pathways were present in the final output. Overlap of GSEA pathways was also performed after down-sampling datasets or individuals. In the first analysis, the top 20 upregulated GSEA pathways (ranked by  $p$  value) were obtained from the upregulated genes after SumRank was performed with all AD datasets performing GSEA as above. This was then compared with the top 100 upregulated pathways after performing SumRank with different numbers of datasets removed. These analyses were done with all broad cluster types (oligodendrocytes, astrocytes, OPCs, glutamatergic neurons, endothelial cells, GABA-ergic inhibitory neurons, and microglia) and for both up- and downregulated genes and the average proportion of

overlap was calculated. In the second analysis, the top 20 upregulated GSEA pathways were obtained from the upregulated genes of a random subset with 340 individuals (170 cases and 170 controls) of the MathysCell dataset. This was then compared with the top 100 upregulated pathways based on differential expression with different down-samplings of individuals from the MathysCell dataset, and proportion overlap was calculated in the same way. The same analyses were then done with downregulated genes.

Gene set consistency was analyzed by obtaining the top 25-ranked genes after SumRank was performed with all AD datasets. The ranks of these genes were then obtained in the signed  $-\log_{10}(p)$  ordered SumRank genes after down-sampling datasets. These analyses were done with all broad cluster types and for both up- and downregulated genes, and each individual rank was plotted. The top 25 genes were obtained from the signed  $-\log_{10}(p)$  ordered genes of a random subset with 340 individuals (170 cases and 170 controls) of the MathysCell dataset. The ranks of these genes were then assessed in the signed  $-\log_{10}(p)$  ordered SumRank genes after down-sampling individuals. For each down-sampling, the mean rank across all broad cluster types and up- and downregulated genes was plotted.

### Mapping to Mathys et al. Nature, 2024 AD atlas

AD datasets were mapped to an atlas composed of the AD individuals from Mathys Nature 2024<sup>50</sup> using the same FindTransfersAnchors approach as described above for major\_cell\_type and cell\_type\_high\_resolution of the Mathys dataset. The data were then pseudo-bulked at the individual level within each cell type, and reproducibility was assessed as described above.

### snATAC-Seq analyses

Differentially accessible peaks from pseudo-bulked broad cell types were obtained from the Supplementary file of Xiong et al.<sup>51</sup>, who performed snATAC-seq on AD and control individuals and assessed for differential chromatin accessibility after pseudo-bulking to broad cell types (assigning cell types after integrating ATAC and RNA data). The genes corresponding to these peaks were found by overlapping the peak locations with the nearest gene promoter using the distancetoNearest function in the GenomicRanges R package version 3.2.0<sup>110</sup> using hg38 annotations obtained through the AnnotationHub version 3.2.0 R package<sup>111</sup>. The genes corresponding to differentially accessible peaks were then ranked by  $p$  value after subsetting to up- or downregulated, and then overlapped for each cell type with meta-analysis DEGs after taking different percentages of the top ATAC-seq genes. Overlap enrichment was calculated by Fisher's exact test. To find the DEGs from the Xiong RNA, we performed DESeq2 differential expression on the snRNA-seq of only the individuals that also had snATAC-seq performed on them. snATAC-seq peaks were visualized in UCSC Genome Browser hg38<sup>112</sup>.

### Memento analyses

Memento version 0.12<sup>43</sup> was used for differential expression analyses of PD data to compare with DESeq2 and logistic regression of pseudo-bulk data. Non-pseudobulked single-cell PD data were analyzed using the "fixed effects binary testing with multiple samples" method according to the Memento documentation tutorial with capture\_rate of 0.07 and min\_perc\_group of 0.9, with all other settings left as default. Genes that were filtered out were set to effect size of 0 with  $p$  value of 1. The de\_coef value and de\_pval were used for effect sizes and  $p$  values, respectively, and SumRank was performed on the resulting Memento differential expression results as described above.

### PPI network mapping

DEGs (BH corrected  $p$  value < 0.05) were input into STRING version 12.0<sup>61</sup>, which created PPI networks of the DEGs, connecting each DEG to another DEG if a PPI was known to exist between them. These

networks were input into Cytoscape version 3.10.3<sup>113</sup> and the “Analyze Network” function was used to find network characteristics for each gene (e.g., betweenness centrality). The ClueGO app of Cytoscape was used to analyze enriched GO pathways of the genes in the most densely connected networks.

### Mice

Mice were bred in-house or obtained from the Jackson Laboratory (JAX). Animals were housed at New York University (NYU) Medical Center Animal Facility under specific pathogen-free conditions. Mice were housed in a 12-h light–dark cycle in a temperature-controlled and humidity-controlled environment with water and food provided ad libitum. Both males and females were used in this study. The following mouse strain was used: B6.Cg-Tg(APPs<sup>w</sup>FILon,PSEN1<sup>M146L</sup>\*L286V)6799Vas/Mmjax (5x*FAD*; JAX 034848). For analysis of *BCAT1* staining in oligodendrocytes, 8–10-month-old mice were used. For euthanasia, mice were deeply anaesthetized with an intraperitoneal injection of ketamine (100 µg g<sup>-1</sup>) and xylazine (10 µg g<sup>-1</sup>) and perfused with 20 ml ice-cold 1× PBS followed by 20 ml 4%PFA.

### Tissue collection and processing

Mice were perfused with cold 1× PBS followed by 4%PFA. Brains were removed, post-fixed overnight, cryopreserved in 30% sucrose, and cryo-embedded in OCT. 40 µm coronal cryosections were generated between bregma 1.335–.745. For staining at least two sections containing mPFC were used for multiplexed IHC.

### IHC, imaging, and quantification

Coronal brain slices were rinsed 3× in PBS for 10 min each prior to antigen retrieval. For antigen retrieval, slides were immersed in 1M citrate buffer, microwaved until boiling, and incubated for 15 minutes at 99 °C in a water bath. Afterwards, slides were returned to room temperature, rinsed 2 × 10 min in PBS and blocked in 10% normal donkey serum (Jackson ImmunoResearch AB\_2337258), 1% BSA, 25% tritonX100, with Mouse on Mouse IG blocking reagent (Vector Labs BMK-2202) in 1× PBS for 2hrs at room temperature. Sections were then stained with the following primary antibodies; Mouse anti CCI (1:200, Sigma OP80), Goat anti SOX10 (1:200, R&D Systems AF2864-SP), and Rabbit anti *BCAT1* (1:200, Proteintech 13640-1-AP) overnight in blocking solution with Mouse on Mouse protein concentrate instead of IG blocking reagent (Vector Labs BMK-2202) at 4 °C. The next day sections were then washed 3× with PBST and incubated for 2hrs at RT with the following secondary antibodies all at 1:500; Alexa488 Donkey anti goat (Jackson ImmunoResearch 705-545-003), Alexa568 Donkey anti Mouse (Invitrogen A-31571), Alexa647 Donkey anti Rabbit (Jackson ImmunoResearch 711-605-152) in blocking solution with Mouse on Mouse protein concentrate (Vector Labs BMK-2202). Sections were then washed 3× with PBST and mounted with Fluoromount-G Mounting Medium, with DAPI (Invitrogen 00-4959-52). Z-stack tiled images of the mPFC were acquired using an LSM 800 Confocal microscope (Zeiss) using a ×40 oil immersion objective (Na 1.3). Quantitative analysis was conducted on at least 2 slices per animal using the Fiji package for ImageJ software by a researcher blind to the experimental groups. After applying a median filter (two pixel radius) to the *BCAT1* channel, SOX10+ CCI+ double-positive oligodendrocyte cytoplasm were drawn by hand with the polygon tool. *BCAT1* mean fluorescent intensity was quantified per cell, normalized over *BCAT1* background staining, and averaged per animal. Data was expressed as FC over WT samples, normalized for each batch of staining.

### Statistics and reproducibility

No statistical method was used to predetermine sample size. Sample sizes were chosen based on the available number of mice and human data. No data were excluded from the analyses (all individual cases and controls were used from each study) except for the Ruzicka et al.

dataset, where exclusion of individuals was done in the same way as in the Ruzicka et al. study. The experiments were not randomized. Quantitative analysis of the microscopy images was done by a researcher blind to the experimental groups.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All data are publicly available online (see Supplementary Data File 1 for details). Count matrices are available from GEO at [GSE129308](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129308) (Otero-Garcia et al.<sup>92</sup>), [GSE147528](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147528) (Leng et al.<sup>93</sup>), [GSE140511](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140511) (Zhou et al.<sup>94</sup>), [GSE138852](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138852) (Grubman et al.<sup>95</sup>), [GSE174367](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174367) (Morabito et al.<sup>96</sup>), [GSE157827](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157827) (Lau et al.<sup>97</sup>), [GSE163577](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163577) (Yang et al.<sup>98</sup>), [GSE183068](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183068) (Sayed et al.<sup>99</sup>), [GSE148822](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148822) (Gerrits et al.<sup>100</sup>), [GSE160936](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160936) (Smith et al.<sup>101</sup>), [GSE167494](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167494) (Sadick et al.<sup>102</sup>), [GSE157783](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157783) (Smajic et al.<sup>10</sup>), [GSE184950](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184950) (Wang et al.<sup>15</sup>), [GSE193688](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193688) (Adams et al.<sup>14</sup>), [GSE243639](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243639) (Martirosyan et al.<sup>12</sup>), [GSE148434](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148434) (Lee et al.<sup>13</sup>), [GSE173731](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE173731) (Garcia et al.<sup>103</sup>), [GSE180928](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180928) (Lim et al.<sup>23</sup>), and [GSE152058](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152058) (Matsushima et al.<sup>21</sup>). Other matrices are available on Synapse (Mathys et al.<sup>44</sup>, Mathys et al.<sup>48</sup>, Hoffman et al.<sup>26</sup>, Fujita et al.<sup>24</sup>, Ruzicka et al.<sup>4</sup>), CellxGene (Gabbitto et al.<sup>45</sup>), Zenodo (Batiuk et al.<sup>5</sup>: <https://zenodo.org/records/6921620>), NEMO (Ling et al.<sup>6</sup> and Handsaker et al.<sup>104</sup>), the Broad Institute Single Cell Portal (SCP1768: Kamath et al.<sup>11</sup>), or from the authors directly (Barker et al.<sup>105</sup>). COVID-19 datasets are available at: <https://atlas.fredhutch.org/fredhutch/covid/>. Source data are provided with this paper.

### Code availability

Code for all new analyses in this paper, including runnable software for the SumRank method, are available in a Github repository: [https://github.com/nathan-nakatsuka/scRNA\\_Reproducibility](https://github.com/nathan-nakatsuka/scRNA_Reproducibility). A permanent Zenodo repository for this study is <https://doi.org/10.5281/zenodo.15809261>.

### References

- Schirmer, L. et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
- Jäkel, S. et al. Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543–547 (2019).
- Kihara, Y. et al. Single-nucleus RNA-seq of normal-appearing brain regions in relapsing-remitting vs. secondary progressive multiple sclerosis: implications for the efficacy of fingolimod. *Front. Cell. Neurosci.* **16**, 918041 (2022).
- Ruzicka, W. B. et al. Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science* **384**, eadg5136 (2024).
- Batiuk, M. Y. et al. Upper cortical layer–driven network impairment in schizophrenia. *Sci. Adv.* **8**, eabn8367 (2022).
- Ling, E. et al. A concerted neuron–astrocyte program declines in ageing and schizophrenia. *Nature* **627**, 604–611 (2024).
- Nagy, C. et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).
- Velmeshev, D. et al. Single-cell genomics identifies cell type–specific molecular changes in autism. *Science* **364**, 685–689 (2019).
- Gandal, M. J. et al. Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. *Nature* **611**, 532–539 (2022).
- Smajic, S. et al. Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. *Brain* **145**, 964–978 (2022).
- Kamath, T. et al. Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson’s disease. *Nat. Neurosci.* **25**, 588–595 (2022).

12. Martirosyan, A. et al. Unravelling cell type-specific responses to Parkinson's disease at single cell resolution. *Mol. Neurodegener.* **19**, 1–24 (2024).
13. Lee, A. J. et al. Characterization of altered molecular mechanisms in Parkinson's disease through cell type-resolved multiomics analyses. *Sci. Adv.* **9**, eabo2467 (2023).
14. Adams, L., Song, M. K., Yuen, S., Tanaka, Y. & Kim, Y.-S. A single-nuclei paired multiomic analysis of the human midbrain reveals age-and Parkinson's disease-associated glial changes. *Nat. Aging* **4**, 364–378 (2024).
15. Wang, Q. et al. Molecular profiling of human substantia nigra identifies diverse neuron types associated with vulnerability in Parkinson's disease. *Sci. Adv.* **10**, eadi8287 (2024).
16. van den Oord, E. J., Xie, L. Y., Zhao, M., Aberg, K. A. & Clark, S. L. A single-nucleus transcriptomics study of alcohol use disorder in the nucleus accumbens. *Addict. Biol.* **28**, e13250 (2023).
17. Brenner, E. et al. Single cell transcriptome profiling of the human alcohol-dependent brain. *Hum. Mol. Genet.* **29**, 1144–1153 (2020).
18. Renthal, W. et al. Characterization of human mosaic Rett syndrome brain tissue by single-nucleus RNA sequencing. *Nat. Neurosci.* **21**, 1670–1679 (2018).
19. Mitroi, D. N., Tian, M., Kawaguchi, R., Lowry, W. E. & Carmichael, S. T. Single-nucleus transcriptome analysis reveals disease-and regeneration-associated endothelial cells in white matter vascular dementia. *J. Cell. Mol. Med.* **26**, 3183–3195 (2022).
20. Lee, H. et al. Cell type-specific transcriptomics reveals that mutant huntingtin leads to mitochondrial RNA release and neuronal innate immune activation. *Neuron* **107**, 891–908. e898 (2020).
21. Matsushima, A. et al. Transcriptional vulnerabilities of striatal neurons in human and rodent models of Huntington's disease. *Nat. Commun.* **14**, 282 (2023).
22. Al-Dalahmah, O. et al. Single-nucleus RNA-seq identifies Huntington disease astrocyte states. *Acta Neuropathol. Commun.* **8**, 1–21 (2020).
23. Lim, R. G. et al. Huntington disease oligodendrocyte maturation deficits revealed by single-nucleus RNAseq are rescued by thiamine-biotin supplementation. *Nat. Commun.* **13**, 7791 (2022).
24. Fujita, M. et al. Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat. Genet.* **56**, 605–614 (2024).
25. Su, Y. et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479–1495. e1420 (2020).
26. Hoffman, G. E. et al. Efficient differential expression analysis of large-scale single cell transcriptomics data using dreamlet. *bioRxiv*, <https://www.biorxiv.org/content/10.1101/2023.03.17.533005v2.full> (2023).
27. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
28. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 1–15 (2021).
29. Murphy, A. E., Fancy, N. & Skene, N. Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset. *Elife* **12**, RP90214 (2023).
30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
31. Cembrowski, M. S. Single-cell transcriptomics as a framework and roadmap for understanding the brain. *J. Neurosci. methods* **326**, 108353 (2019).
32. Wendt, F. R., Pathak, G. A., Tylee, D. S., Goswami, A. & Polimanti, R. Heterogeneity and polygenicity in psychiatric disorders: a genome-wide perspective. *Chronic Stress* **4**, 2470547020924844 (2020).
33. Marigorta, U. M., Rodríguez, J. A., Gibson, G. & Navarro, A. Replicability and prediction: lessons and challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
34. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
35. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 1–6 (2010).
36. Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
37. Bakken, T. E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
38. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587. e3529 (2021).
39. Junttila, S., Smolander, J. & Elo, L. L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Brief. Bioinformatics* **23**, bbac286 (2022).
40. Murdock, M. H. & Tsai, L.-H. Insights into Alzheimer's disease from single-cell genomic approaches. *Nat. Neurosci.* **26**, 181–195 (2023).
41. Andreatta, M. & Carmona, S. J. UCell: robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
42. Schwarzer, G., Carpenter, J. R. & Rücker, G. *Meta-analysis with R*. Vol. 4784 (Springer, 2015).
43. Kim, M. C. et al. Method of moments framework for differential expression analysis of single-cell RNA sequencing data. *Cell* **187**, 6393–6410. e6316 (2024).
44. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
45. Gabitto, M. I. et al. Integrated multimodal cell atlas of Alzheimer's disease. *bioRxiv*, <https://www.biorxiv.org/content/10.1101/2023.05.08.539485v3> (2023).
46. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2023).
47. Fujita, M. et al. Cell-subtype specific effects of genetic variation in the aging and Alzheimer cortex. *bioRxiv*, <https://www.biorxiv.org/content/10.1101/2022.11.07.515446v1.full> (2022).
48. Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385. e4327 (2023).
49. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
50. Mathys, H. et al. Single-cell multiregion dissection of Alzheimer's disease. *Nature* **632**, 858–868 (2024).
51. Xiong, X. et al. Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422–4437. e4421 (2023).
52. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
53. Jiang, L. et al. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nat. Cell Biol.* **27**, 505–517 (2024).
54. Kim, J. J. et al. Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. *Nat. Genet.* **56**, 27–36 (2024).
55. Cattaneo, A. et al. The expression of VGF is reduced in leukocytes of depressed patients and it is restored by effective anti-depressant treatment. *Neuropsychopharmacology* **35**, 1423–1428 (2010).
56. Giusto, E. et al. Prospective role of PAK6 and 14-3-3γ as biomarkers for Parkinson's disease. *J. Parkinson's Dis.* 1–12 (2024).
57. Pogoda, A., Chmielewska, N., Maciejak, P. & Szyndler, J. Transcriptional dysregulation in Huntington's disease: the role in pathogenesis and potency for pharmacological targeting. *Curr. Med. Chem.* **28**, 2783–2806 (2021).

58. Hachigian, L. J. et al. Control of Huntington's disease-associated phenotypes by the striatum-enriched transcription factor Foxp2. *Cell Rep.* **21**, 2688–2695 (2017).
59. Rodríguez-Urgellés, E. et al. Postnatal Foxp2 regulates early psychiatric-like phenotypes and associated molecular alterations in the R6/1 transgenic mouse model of Huntington's disease. *Neurobiol. Dis.* **173**, 105854 (2022).
60. Miao, J. et al. Microglia in Alzheimer's disease: pathogenesis, mechanisms, and therapeutic potentials. *Front. Aging Neurosci.* **15**, 1201982 (2023).
61. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
62. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
63. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
64. De Rojas, I. et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat. Commun.* **12**, 3417 (2021).
65. Xi, M. et al. Therapeutic potential of phosphodiesterase inhibitors for cognitive amelioration in Alzheimer's disease. *Eur. J. Med. Chem.* **232**, 114170 (2022).
66. Sikora, J. et al. Quetiapine and novel PDE10A inhibitors potentiate the anti-BuChE activity of donepezil. *J. Enzym. Inhib. Med. Chem.* **35**, 1743–1750 (2020).
67. Kageyama, R., Ohtsuka, T. & Kobayashi, T. Roles of Hes genes in neural development. *Dev. Growth Differ.* **50**, S97–S103 (2008).
68. Bai, G. et al. Epigenetic dysregulation of hairy and enhancer of split 4 (HES4) is associated with striatal degeneration in postmortem Huntington brains. *Hum. Mol. Genet.* **24**, 1441–1456 (2015).
69. Bozdagi, O. et al. The neurotrophin-inducible gene Vgf regulates hippocampal function and behavior through a brain-derived neurotrophic factor-dependent mechanism. *J. Neurosci.* **28**, 9857–9869 (2008).
70. Ali, M. & Bracko, O. VEGF paradoxically reduces cerebral blood flow in Alzheimer's disease mice. *Neurosci. Insights* **17**, 26331055221109254 (2022).
71. De Schepper, S. et al. Perivascular cells induce microglial phagocytic states and synaptic engulfment via SPP1 in mouse models of Alzheimer's disease. *Nat. Neurosci.* **26**, 406–415 (2023).
72. Gurses, M. S., Ural, M. N., Gulec, M. A., Akyol, O. & Akyol, S. Pathophysiological function of ADAMTS enzymes on molecular mechanism of Alzheimer's disease. *Aging Dis.* **7**, 479 (2016).
73. Nandi, A., Yan, L.-J., Jana, C. K. & Das, N. Role of catalase in oxidative stress-and age-associated degenerative diseases. *Oxid. Med. Cell. Longev.* **2019**, 9613090 (2019).
74. Nell, H. J. et al. Targeted antioxidant, catalase–SKL, reduces beta-amyloid toxicity in the rat brain. *Brain Pathol.* **27**, 86–94 (2017).
75. Forner, S. et al. Systematic phenotyping and characterization of the 5xFAD mouse model of Alzheimer's disease. *Sci. Data* **8**, 270 (2021).
76. Siddik, M. A. B. et al. Branched-chain amino acids are linked with Alzheimer's disease-related pathology and cognitive deficits. *Cells* **11**, 3523 (2022).
77. Nong, X. et al. The mechanism of branched-chain amino acid transferases in different diseases: Research progress and future prospects. *Front. Oncol.* **12**, 988290 (2022).
78. Bis, J. C. et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* **25**, 1859–1875 (2020).
79. Holstege, H. et al. Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease. *Nat. Genet.* **54**, 1786–1794 (2022).
80. Prokopenko, D. et al. Whole-genome sequencing reveals new Alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development. *Alzheimer's Dement.* **17**, 1509–1527 (2021).
81. Guo, L., Zhong, M. B., Zhang, L., Zhang, B. & Cai, D. Sex differences in Alzheimer's disease: Insights from the multiomics landscape. *Biol. Psychiatry* **91**, 61–71 (2022).
82. Zhao, S., Ye, B., Chi, H., Cheng, C. & Liu, J. Identification of peripheral blood immune infiltration signatures and construction of monocyte-associated signatures in ovarian cancer and Alzheimer's disease using single-cell sequencing. *Heliyon* **9**, e17454 (2023).
83. Patel, H., Dobson, R. J. & Newhouse, S. J. A meta-analysis of Alzheimer's disease brain transcriptomic data. *J. Alzheimer's Dis.* **68**, 1635–1656 (2019).
84. Tian, Y. et al. Identification of diagnostic signatures associated with immune infiltration in Alzheimer's disease by integrating bioinformatic analysis and machine-learning strategies. *Front. Aging Neurosci.* **14**, 919614 (2022).
85. Walters, S. et al. Associations of sex, race, and apolipoprotein E alleles with multiple domains of cognition among older adults. *JAMA Neurol.* **80**, 929–939 (2023).
86. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
87. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 1–13 (2008).
88. Johnson, E. C. et al. Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nat. Neurosci.* **25**, 213–225 (2022).
89. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
90. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59 (2021).
91. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
92. Otero-Garcia, M. et al. Molecular signatures underlying neurofibrillary tangle susceptibility in Alzheimer's disease. *Neuron* **110**, 2929–2948. e2928 (2022).
93. Leng, K. et al. Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. *Nat. Neurosci.* **24**, 276–287 (2021).
94. Zhou, Y. et al. Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).
95. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
96. Morabito, S. et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).
97. Lau, S.-F., Cao, H., Fu, A. K. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl. Acad. Sci.* **117**, 25800–25809 (2020).
98. Yang, A. C. et al. A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature* **603**, 885–892 (2022).
99. Sayed, F. A. et al. AD-linked R47H-TREM2 mutation induces disease-enhancing microglial states via AKT hyperactivation. *Sci. Transl. Med.* **13**, eabe3947 (2021).

100. Gerrits, E. et al. Distinct amyloid- $\beta$  and tau-associated microglia profiles in Alzheimer's disease. *Acta neuropathol.* **141**, 681–696 (2021).
101. Smith, A. M. et al. Diverse human astrocyte and microglial transcriptional responses to Alzheimer's pathology. *Acta Neuropathol.* **143**, 75–91 (2022).
102. Sadick, J. S. et al. Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. *Neuron* **110**, 1788–1805. e1710 (2022).
103. Garcia, F. J. et al. Single-cell dissection of the human brain vasculature. *Nature* **603**, 893–899 (2022).
104. Handsaker, R. E. et al. Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease. *Cell* **188**, 623–639. e619 (2025).
105. Barker, S. J. et al. MEF2 is a key regulator of cognitive potential and confers resilience to neurodegeneration. *Sci. Transl. Med.* **13**, eabd7695 (2021).
106. Tian, Y. et al. Single-cell immunology of SARS-CoV-2 infection. *Nat. Biotechnol.* **40**, 30–41 (2022).
107. Smith, G. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* Springer, New York, 397–420 (2005).
108. Quijano Xacur, O. A. The unified distribution. *J. Stat. Distrib. Appl.* **6**, 1–12 (2019).
109. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 1–8 (2011).
110. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
111. Morgan, M. & Shepherd, L. AnnotationHub: client to access annotationHub resources. *R. package version 2*, 2017 (2017).
112. Perez, G. et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2025).
113. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
114. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

## Acknowledgements

We thank all members of the Satija lab and members of the CEGS Center for Integrated Cellular Analysis in New York City for helpful discussions and constructive criticism. We thank Li-Huei Tsai and Ravikiran Raju for providing data from Barker et al., 2021. We thank Brad Ruzicka and Shahin Mohammadi for providing their code for the analyses of Ruzicka et al. 2024 and answering questions about their publication. This work was supported by the Chan Zuckerberg Initiative (EOSS-0000000082 and HCA-A-1704-01895 to R.S.) and the NIH (RM1HG011014-02, 1OT2OD026673-01, DP2HG009623-01, R01HD096770 and R35NS097404 to R.S., NIH-NINDS R01NS122316 and R21NS121786 to E.K., and NIH-NIMH T32MH019524 to D.A.).

## Author contributions

N.N. performed computational analyses. D.A. and E.C. performed mouse modeling and antibody staining experiments. N.N., D.A., L.J., A.H., E.C., E.K., and R.S. interpreted the data. N.N. and R.S. conceived the study and wrote the paper with the assistance of other co-authors. E.K. directed mouse modeling aspects of the study, and RS directed all other aspects of the study.

## Competing interests

In the past three years, R.S. has received compensation from Bristol-Myers Squibb, ImmunAI, Resolve Biosciences, Nanostring, 10x Genomics, Neptune Bio, and the NYC Pandemic Response Lab. R.S. is a co-founder and equity holder of Neptune Bio. The other authors declare that they have no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-62579-z>.

**Correspondence** and requests for materials should be addressed to Nathan Nakatsuka.

**Peer review information** *Nature Communications* thanks Feng Bao, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025