nature communications

Article

https://doi.org/10.1038/s41467-025-63753-z

Disentangling associations between complex traits and cell types with *seismic*

Received: 12 November 2024

Accepted: 27 August 2025

Published online: 01 October 2025

Check for updates

Qiliang Lai[1], Ruth Dannenfelser[1], Jean-Pierre Roussarie[2] ✉ & Vicky Yao[1,3,4] ✉

Integrating single-cell RNA sequencing with Genome-Wide Association Studies (GWAS) can uncover cell types involved in complex traits and disease. However, current methods often lack scalability, interpretability, and robustness. We present *seismic*, a framework that computes a novel specificity score capturing both expression magnitude and consistency across cell types and introduces influential gene analysis, an approach to identify genes driving each cell type-trait association. Across over 1000 cell-type characterizations at different granularities and 28 polygenic traits, *seismic* corroborates known associations and uncovers trait-relevant cell groups not apparent through other methodologies. In Parkinson's and Alzheimer's, *seismic* unveils both cell- and brain-region-specific differences in pathology. Analyzing a pathology-based Alzheimer's GWAS with *seismic* enables the identification of vulnerable neuron populations and molecular pathways implicated in their neurodegeneration. In general, *seismic* is a computationally efficient, powerful, and interpretable approach for mapping the relationships between polygenic traits and cell-type-specific expression, offering new insights into disease mechanisms.

Genome Wide Association Studies (GWAS) have shown exceptional promise for identifying genetic variants across populations that are key contributors to human diseases and a range of phenotypic traits. Simultaneously, the rise of large-scale single-cell RNA sequencing (scRNA-seq) datasets has revolutionized our ability to analyze gene expression profiles at the level of individual cell types and states. These advancements present a unique opportunity to integrate the population-level genetic associations revealed by GWAS with the molecular precision offered by scRNA-seq to pinpoint specific trait-associated cell types. Intuitively, a natural way to unify these datasets is to aggregate trait-associated variant statistics from GWAS at the gene level, which can then be subsequently analyzed for cell-type specificity using scRNA-seq.

Several computational methods have been developed to identify trait-associated cell types[1–6] by integrating reference scRNA-seq atlases with single-nucleotide polymorphism (SNP)-level association statistics derived from GWAS. We focus on the major class of methods that use MAGMA[7] to convert SNP-level GWAS associations into gene-level statistics while accounting for the complexities of linkage disequilibrium. The MAGMA software itself can be adapted to analyze cell type-trait relationships[3,6] via gene set enrichment analysis, which we term S-MAGMA (see Methods) to avoid confusion with the upstream linkage disequilibrium correction process. Broadly, these MAGMA-based methods have been successfully used to help elucidate disease-associated cell types[3,6,8–10], but several technical limitations remain (Table 1). One such limitation is the requirement of arbitrary thresholds, either for selecting the number of genes associated with a GWAS trait (e.g., top 1000 trait-associated genes as in scDRS[1]), genes to characterize a cell type (e.g., top 10% most specific genes for S-MAGMA[3] as used in[10]), or a score threshold for cell-type association enrichment (e.g., 5% score quantile as in scDRS[1]). Biologically, these numbers would naturally vary case-by-case, and most methods

[1]Department of Computer Science, Rice University, Houston, US. [2]Department of Anatomy & Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, US. [3]Ken Kennedy Institute, Rice University, Houston, US. [4]Rice Synthetic Biology Institute, Rice University, Houston, US.
✉e-mail: jproussa@bu.edu; vy@rice.edu

Table 1 | Comparison of *seismic* with other MAGMA-based cell-type association methods

	seismic	scDRS	FUMA	S-MAGMA
No gene thresholding based on trait association	✓		✓	✓
No gene thresholding based on cell-type expression	✓	✓	✓	
Accounts for gene expression variability	✓	✓		
Scalable runtime	✓		✓	✓
Tool includes results visualizations	✓	✓	✓	
Identifies cell-type level influential genes	✓			

recommend trying several. Methods also often fail to account for gene expression variability, focusing instead more simply on mean expression within a cell type, rendering them more susceptible to noise. Furthermore, as the size of single-cell datasets continues to grow, scalability becomes a major concern, necessitating a method that can handle large numbers of cells and cell types. Finally, a critical gap in previous methods is that though they may identify several significant cell type-trait associations, they only output statistical significance without further gene-level interpretation. Although some methods attempt to resolve the issue using global correlation[1] or modularity analysis[4], they fail to quantify each gene's contribution to the observed significance association of an interested cell type, limiting the actionable insights to be derived from these analyses.

Here, we present *seismic*, a framework that enables robust and efficient discovery of cell type-trait associations and provides the first method to simultaneously identify the specific genes and biological processes driving each association. Notably, *seismic* eliminates the need to select arbitrary thresholds to characterize trait or cell-type association through the use of a cell-type-specificity scoring method that accounts for background gene expression variability. We apply *seismic* and existing MAGMA-based cell-type association methods on both simulated and real data to demonstrate that *seismic* is well-calibrated, efficient, and powerful.

Through a deep exploration of neurological disease-associated brain cell types, we find that cell type definition from input scRNA-seq data is an important, yet currently underappreciated, factor that influences downstream findings. Previous studies[6,8,11,12] have typically used broad characterizations of cell types, such as "telencephalon projecting excitatory neurons" and "frontal cortex neurons," without accounting for finer regional or tissue-specific distinctions. This coarse characterization can obscure valuable biological insights, especially when cell diversity is high. While broad cell type characterizations may be adequate when studying relatively homogeneous cell populations (e.g., microglial cells in the brain), this approach falls short for highly diverse cell populations like neurons. For instance, neurodegenerative diseases preferentially target neurons with very distinct regional and cell type identities. In Alzheimer's disease, neurons in the entorhinal cortex are especially vulnerable, whereas neurons in even neighboring brain regions, such as the dentate gyrus and CA2/CA3 in the hippocampus, are not[13]. We show that using finer granularities for cell type characterization reveals more specific cell-type trait links, which better reflects true biological mechanisms. Notably, *seismic* consistently outperforms other methods in identifying disease-associated cell types across these different cell type characterizations. Furthermore, we demonstrate the importance of considering different GWAS endpoints to reveal disease mechanisms, reporting, to our knowledge, the first computational identification of a neuronal association with an Alzheimer's disease biomarker (tau level in cerebrospinal fluid). Together, our results expand current notions of best practices for cell type-trait association analyses and provide a methodological toolkit to take fuller advantage of both scRNA-seq and GWAS data to unravel the intricate interplay between tissue/cell type and complex traits.

Results

Identifying cell type-trait associations using *seismic*

Many cell-type-trait association methods consider the same inputs—variant-trait information from GWAS resolved to gene-trait relevance using MAGMA[7] and single-cell expression data—to find statistically significant associations between cell types and traits (Fig. 1A). However, these methods may rely on arbitrary gene thresholds or cell-type mean expression profiles to identify trait-implicated cell types (Table 1), without accounting for the global relationship of cell type specificity and disease risk. Here, we introduce a novel integration framework, Single-cell Expression Integration System for Mapping genetically Implicated Cell types (*seismic*), that overcomes the limitations of previous methods to provide a threshold-free, fast, and interpretable method for combining single cell expression data with gene-trait relationships (Fig. 1B).

At the core of *seismic* is a cell type-specificity score ("Methods"), which calculates the specificity and consistency of expression for each gene in a cell type relative to all other cell types. The *seismic* specificity score is designed to compare the relative probability of a gene in a cell type with consistently higher expression than background cells among all cell types ("Methods"), thus providing a global view of gene specificity in a cell type. We empirically assess these scores in various pancreatic cell types in the Tabula Muris FACS datasets[14]. Even though these cell types exhibit highly correlated expression patterns, we find that established marker genes[15] are all ranked among the highest in their corresponding cell types (Supplementary Fig. 1A, B). Moreover, marker genes consistently show significantly higher specificity scores in their target cell types compared to those same marker genes in non-pancreatic cell types or housekeeping genes across all cell types (Supplementary Fig. 1C), highlighting the score's ability to capture biologically meaningful cell-type-specific patterns.

The *seismic* specificity score is also robust to different characterizations of cell types, whether it is broader groupings or more specific subclusters, as it is robust to arbitrary re-labeling of homogeneous populations, while exploiting genuine substructure to achieve high resolution in identifying trait-associated cell types (Supplementary Note 1, Supplementary Fig. 2). Furthermore, the *seismic* specificity score demonstrates robustness to noise in cell cluster characterizations. In real-world datasets, cell type annotations derived from unsupervised clustering and subsequent manual curation can inevitably contain some mislabeled cells or mixtures of closely related cells[16]. Through a label permutation simulation, we find that the *seismic* specificity score shows superior resilience to cell type label noise compared to other common specificity metrics (Supplementary Fig. 3), demonstrating that the cell type specificity profiles are relatively less sensitive to slight inaccuracies in upstream cell type definitions.

After calculating the *seismic* specificity score for a collection of cell types in a scRNA-seq dataset, the *seismic* framework then applies a regression model to test for significant associations between the specificity scores and MAGMA gene z-scores, under the assumption that the genetically implicated cell types specifically expresses more of these genes with higher trait relevance ("Methods"). For significantly associated cell types, the *seismic* framework also introduces influential observation analysis to the corresponding regression model, enabling what we term 'influential gene analysis.' To our knowledge, influential gene analysis is the first method to systematically rank and identify genes driving purported cell type-trait associations.

Systematic benchmarking and runtime analysis

To assess how well *seismic* and three of the most commonly used cell type-trait identification methods (scDRS[1], FUMA[2], and S-MAGMA[3]) are

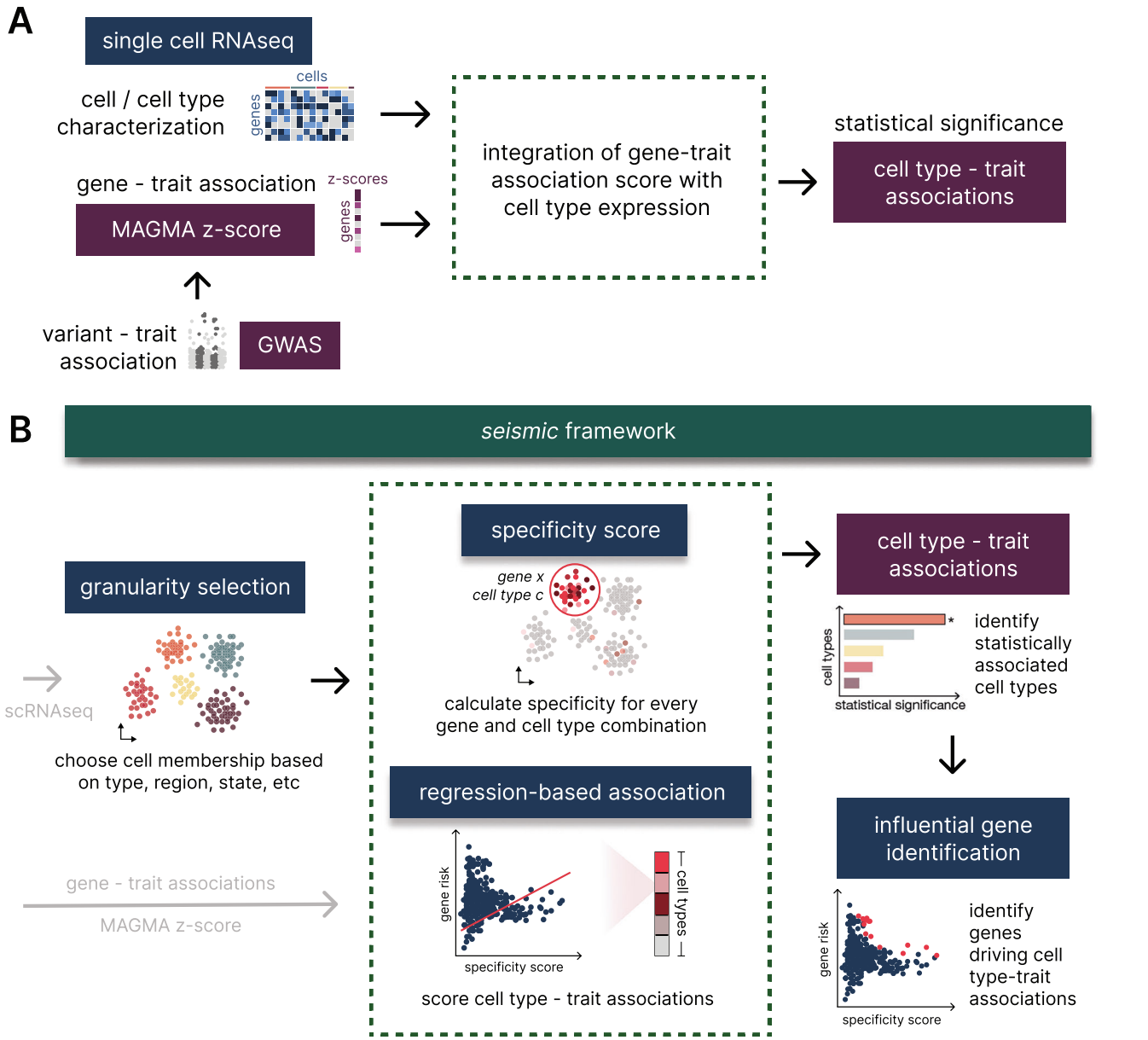

Fig. 1 | General framework for identifying cell type-trait associations and overview of *seismic*. **A** Current approaches for computationally linking cell types to trait typically integrate single-cell expression data with gene-trait associations from MAGMA to prioritize genetically susceptible cell types by statistical significance. Each method has a different procedure for how these inputs are combined to produce the final set of identified associations. **B** The *seismic* framework expansion of this general workflow. *seismic* allows for a flexible set of cell type labels or granularities, ranging from broader cell classes to specifically defined cell types, before calculating a novel gene specificity score for each of the corresponding cell characterizations, capturing both the magnitude and consistency of gene expression. These cell-type-level gene specificity scores are then integrated with MAGMA z-scores using a regression model to assess the statistical significance of cell type-trait associations, under the assumption that disease-critical genes are more specific to the implicated cell type (see "Methods"). Unlike existing methods, individual gene contributions to the cell type-trait association can be quantified via influential gene analysis and can pinpoint the genes and underlying biological processes that drive significant associations.

calibrated to false positives, we perform a systematic simulation to detect the frequency of type I errors. We first randomly select 10 sets of MAGMA trait z-scores from GWAS (Supplementary Data 1) and subsample 10 expression datasets, each containing 10,000 cells from the Tabula Muris (TM) FACS scRNA-seq data[17] (Supplementary Data 2). For each subsampled expression dataset, we randomly select 100 cells as a cell type of interest ("Methods"). Next, across 10,000 runs, we randomize the gene labels in the expression data and compare the *p*-values reported by each method for the association between the randomly assigned target cell type and trait. We find that all methods generally control type I error, with FUMA being markedly conservative, potentially limiting its detection power. *seismic* is, on average, conservative and has stable performance. In contrast, using the analytically transformed *p*-values from scDRS, we see slightly inflated *p*-values at extreme quantiles, and S-MAGMA can also, at times, report inflated *p*-values (Fig. 2A). The *seismic* and scDRS implementations enable examination of the effect of randomization of MAGMA trait z-scores, and we observe the same trends, where *seismic* still has generally well-calibrated *p*-values, and scDRS has slight inflation at tail quantiles (Supplementary Fig. 4).

Complex, polygenic traits frequently involve multiple disease-associated cell types and subtle expression perturbations across a large number of genes[18–20]. In order to evaluate the extent to which *seismic* can correctly identify trait-associated cell types reflective of

https://doi.org/10.1038/s41467-025-63753-z

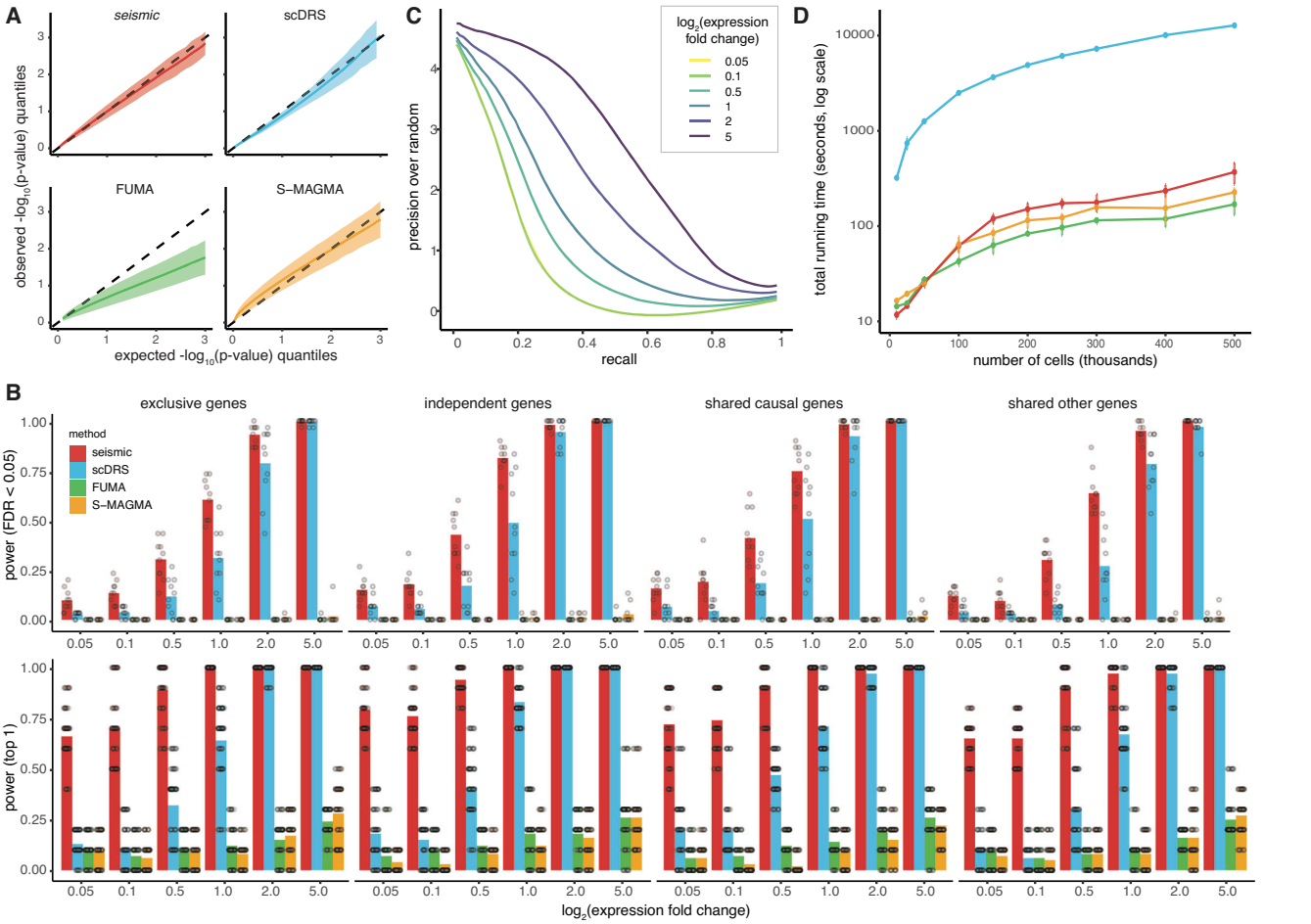

Fig. 2 | Systematic benchmarking results for *seismic* and three other commonly used MAGMA-based methods for cell type-trait association detection. **A** Assessment of type I error calibration for *seismic*, scDRS[1], FUMA[2], and S-MAGMA[7]. Results are based on random cell type assignments and expression datasets across 10,000 simulation runs with 10 randomly selected traits and 10 subsets of 10,000 cells (Methods). *P*-values are calculated using each method's default analysis settings as described in Methods. Quantile-quantile plots show the comparison of expected $-\log_{10}(p-\text{value})$ quantiles (the dashed line) in comparison with observed $-\log_{10}(p-\text{value})$ quantiles. The solid line represents the mean of observed $-\log_{10}(p-\text{value})$ quantiles across the 10 random traits, while the shaded ribbon represents the mean ± one standard deviation (SD) of the value. The dashed line represents theoretically perfectly calibrated *p*-values, and observations falling below this line indicate that *p*-value estimates are more conservative. **B** Comparative power analysis of cell type-trait association methods under four different gene sampling strategies. To simulate a complex disease, we designate 3 random cell types as causal. Four gene sampling strategies were deployed to simulate the cell types with different correlation patterns (see Methods for details). We examine power in two ways: (top) the mean frequency across 10 random traits where the 3 perturbed cell types with trait-specific expression are identified as significantly associated across 10 independent simulations (FDR < 0.05), and (bottom) the ratio of simulations where the top associated cell type is one of the 3 causal cell types. Each dot summarizes the frequency of 10 expression and target cell type randomizations for a random trait, while bars and error bars represent the mean and the range of mean ± standard deviation across $n = 10$ random traits. Results are shown over varying effect sizes, or expression fold changes (in log2 scale), with 50% of perturbed genes designated as causal. **C** Relative precision-recall curves showing $\log_2(\text{precisionoverrandom})$, illustrating the power of *seismic*'s influential gene analysis to identify simulated causal disease genes at different effect sizes. The metric is calculated as the mean precision divided by the random prior across simulations at each level of recall, when 50% of perturbed genes are causal disease genes. **D** Total runtime in seconds (log scale) for each method to identify associated cell types for a trait in scRNA-seq datasets of various sizes (number of cells). Each data point represents the mean runtime from 5 independent runs, with error bars denoting the range of mean ± one standard deviation. Each of the runs was performed on a different dataset generated by subsampling (or upsampling) cells from the Tabula Sapiens[28] dataset to the specified numbers.

this complexity, we simulate several different scenarios: (1) single or multiple associated cell types; (2) distinct or overlapping genes driving the trait association across cell types; (3) strong or subtle expression perturbations linked with the trait. We use scDesign3[21] to generate realistic synthetic count-level data that mimics trait-specific expression perturbations at different effect sizes ("Methods"). Applying the four methods to this simulated data, we compare their power to correctly identify the perturbed cell types as trait-associated by calculating the proportion of simulations where the perturbed cell types are significantly associated with the trait (FDR < 0.05), as well as whether the highest-ranked cell type is a correct association. *seismic* consistently has the highest power of all methods across scenarios and effect sizes (Fig. 2B and Supplementary Figs. 5, 6). Interestingly, scDRS is also relatively powerful at higher effect sizes but deteriorates in the challenging scenario where the expression perturbation is milder. In contrast, FUMA and S-MAGMA both show limited power across most scenarios (Fig. 2B); while the correct disease-associated cell type rarely reaches statistical significance, they are at times ranked highly. Focusing on the situation where there is only a single causal cell type, we also simulate different types of polygenic signals by varying the ratio of causal genes to confounding genes. We observe that FUMA and S-MAGMA still have lower power in this simpler scenario. scDRS is somewhat underpowered when effect sizes are small, even if there are a large number of causal genes, but it is powerful in the rarer scenario

where there are fewer causal genes of very high effect size. *seismic* is consistently powerful, especially in detecting associations when many genes are involved, even if the expression perturbations are subtle (Supplementary Fig. 5).

To further complement the power simulation and examine robustness to false positives, we also quantified the frequency that the non-perturbed cell types (cell types without simulated trait-specific signals) reach a nominally significant statistical significance. *seismic* still maintained well-controlled false positive ratios consistent with the expected theoretical threshold (Supplementary Fig. 7) across multiple traits whose GWAS exhibit diverse sample sizes and genetic architectures. This analysis further confirms that *seismic*'s enhanced power does not come at the cost of inflated type I error.

Beyond accurate, sensitive identification of causal cell types, *seismic*'s novel influential gene analysis prioritizes the causal genes that drive the model results. Even for the same trait or disease, the causal genetic variants and biological processes may vary across associated cell types. For instance, for blood pressure regulation, genes related to the nitric oxide pathway primarily affect endothelial cells and their role in vascular function[19,22], while others predominantly influence the calcium response and contractile function in smooth muscle cells[19,23]. To evaluate *seismic*'s ability to distinguish the true causal gene set from all other genes, we analyze the influential genes that *seismic* identifies for the target cell type in the simulated data (Methods). We find that *seismic* successfully prioritizes causal genes whether in the single or multiple target cell type paradigm, achieving more than 20-fold higher precision for the top-ranked genes compared to random chance (Fig. 2C, Supplementary Fig. 8).

As single-cell technologies continue to improve and we move closer towards atlas-scale datasets, computational methods need to be able to scale well with the number of cells. To this end, we also benchmark the four methods for their runtime as the number of cells in the input single-cell expression dataset increases (Fig. 2D, Supplementary Data 3). *seismic*, FUMA, and S-MAGMA scale comparatively well, handling hundreds of thousands of cells with runtimes in the scale of minutes, whereas scDRS is over 30-fold slower, taking hours to run. We speculate that the dramatic difference in speed between scDRS and all other methods is likely due to its reliance on Monte Carlo-based subsampling for empirical statistics, whereas all other methods directly assess associations at the cell type level and do not rely on simulations for significance assessment.

Together, these comprehensive benchmarking results highlight *seismic*'s capabilities as a powerful, versatile, and efficient tool for analyzing cell type-trait associations. Specifically, *seismic* controls appropriately for type I errors, while having enhanced sensitivity in detecting true causal associations, even when there are only subtle expression changes across cell types and genes. Overall, *seismic* is the only method to exhibit high detection power with computational efficiency that scales to handle large datasets.

Methodological comparisons across traits and cell types

To examine *seismic*'s ability to capture known cell type-trait associations across a broad range of GWAS traits, we assemble 27 studies spanning neurological diseases and disorders, immune-related conditions, and a variety of other traits, including demographic, cardiovascular, and metabolic endpoints (Supplementary Data 1). We test for cell type associations using the expression values and annotations from the Tabula Muris FACS dataset[17], which includes nearly 45,000 cells, covering 130 cell type characterizations across 17 tissues (Supplementary Data 2). In total, we find 653 pairs of cell type-trait associations that pass the significance threshold of FDR ≤0.05 (Supplementary Data 4). Notably, *seismic* identifies associations linking leukocytes with immune diseases, neurons with neuropsychiatric diseases, smooth muscle cells with cardiovascular diseases, pancreatic cells with type 2 diabetes, and hepatocytes with metabolic traits (Fig. 3A), recapitulating known biological cell type-trait associations. Moreover, *seismic* is robust against variations in gene window size (average Pearson's correlation between all pairs of windows> 0.98 across all 27 traits, Supplementary Fig. 9).

We then apply S-MAGMA, FUMA, and scDRS to the same Tabula Muris FACS dataset and GWAS traits, checking for consistency with *seismic*'s results and any differences in associations (Supplementary Data 4, Supplementary Figs. 10–16). Notably, 89% of associations identified by *seismic* are also detected by at least one of these frameworks (Fig. 3B, Supplementary Figs. 10–16), where *seismic* captures most of FUMA's reported associations (95%), followed by scDRS (88%), then S-MAGMA (81%). This high degree of overlap highlights *seismic*'s robustness and alignment with established methods. The high correlation is further illustrated in a detailed between-method comparison, which reveals that *seismic* consistently achieves the highest trait-wise concordance among the methods, as measured by Spearman's correlation. Specifically, for 26 of the 27 traits examined, *seismic* and one other framework achieve the highest concordance (Fig. 3C, Supplementary Fig. 17). Notably, *seismic* shows the highest average between-method Spearman's correlation across all traits (0.69), compared with 0.60 for scDRS, 0.66 for FUMA, and 0.60 for S-MAGMA. For the 330 common association pairs found by all frameworks, *seismic* exhibits the most significant false discovery rates (FDR) in 80% of these pairs (263 out of the 330 pairs), demonstrating its power (Supplementary Data 4).

Besides these common findings, *seismic* also identifies additional association patterns that may better capture the underlying biology. For erythrocyte count, while scDRS ranks several cells from the intestine as most relevant, *seismic* and S-MAGMA identify several hematopoietic lineage cell types in marrow to be most associated, more accurately reflecting the developmental process of red blood cells. *seismic* also observes broad associations between neuropsychiatric diseases and various pancreatic islet cell types, which is especially noticeable in depression. These somewhat outlandish associations are also recapitulated by the other methods, and interestingly, previous studies have found potential associations between pancreatic and neuropsychiatric diseases[24–27], which has led to increased interest in a potential pancreas-brain axis. In total, *seismic* only misses 2 cell type-association pairs identified by all other methods (Fig. 3B, Supplementary Data 4), the fewest compared to other methods (56, 29, and 53 pairs for scDRS, FUMA, and S-MAGMA, respectively). The two undetected associations—between microglia and autoimmune disease, as well as pre-activation T cell subtype and ulcerative colitis—are close to the multiple hypothesis threshold (with FDR = 0.066 and 0.096, respectively, Supplementary Data 4).

To assess the degree to which technical factors may affect *seismic*'s specificity scores, we split the Tabula Muris dataset by individual donor ID and recalculated the scores for each cell type. Specificity scores remained highly consistent across donors for the majority of cell types (Supplementary Fig. 18A), with cell types with larger numbers of cells exhibiting higher correlation. This concordance also extended to the downstream trait-implicated cell type analysis, where a high correlation of statistical significance was observed (mean Pearson's correlation = 0.88 for cell-type associations ($-\log_{10}(p-\text{value})$) between pairs of donors across all traits, Supplementary Fig. 18B), despite some traits having target cell types not captured in several donors (e.g., liver tissue absent in 4 of 6 donors).

To explore generalizability to other large scRNA-seq datasets, we also apply *seismic* to the Tabula Muris (TM) droplet dataset[17] (a dataset obtained by droplet-based single-cell sequencing rather than profiling individually sorted cells as in TM FACS, Supplementary Fig. 19) and the Tabula Sapiens (TS) human scRNA-seq dataset[28] (Supplementary Fig. 20). Comparing cell types that overlap between TM FACS and these 2 additional scRNA-seq datasets, one using a different

https://doi.org/10.1038/s41467-025-63753-z

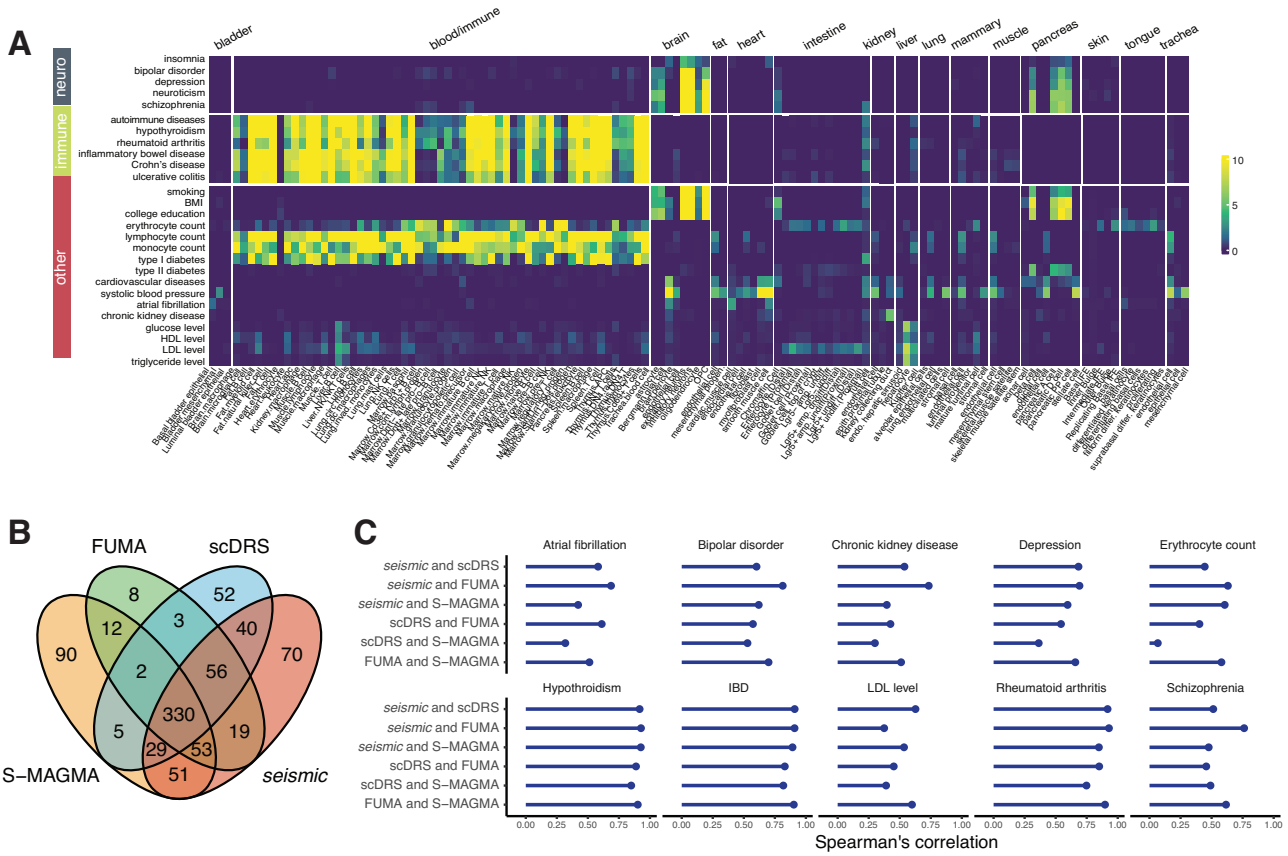

Fig. 3 | Characterization of cell type-trait associations across 27 diverse GWAS. **A** *seismic* cell type-trait associations covering 15 tissue types and 130 cell types from the Tabula Muris FACS dataset[17] and 27 GWAS studies covering neurological, immune, and various other disease or demographic endpoints. **B** Venn diagram comparing the significantly associated cell type-trait relationships between methods (based on a 0.05 FDR threshold) assessed by the four MAGMA-based computational methods, S-MAGMA, FUMA, scDRS and *seismic*. **C** Pairwise Spearman's correlations of statistical significance of all (130) cell types ($-\log_{10}(p-\text{value})$) for each pair of any two methods across a selected set of 12 diverse traits (other trait correlations are in Supplementary Fig. 16).

technology, and the other using human cells, we find consistent cell type-trait associations (Supplementary Fig. 21). The mean Spearman's correlation is 0.78 between TM FACS and TM droplet, and 0.67 between TM FACS and TS across all traits, in terms of statistical significance (in $-log_{10}(p\text{-value})$). We note that neither the TM droplet nor the TS contains brain tissue, and the mean Spearman's correlation is 0.84 and 0.75, respectively, if neuropsychiatric traits are excluded from the comparison. The high concordance of *seismic* with other methods is also consistent across datasets (Supplementary Figs. 17, 22, 23). Such consistency underscores *seismic*'s robustness in identifying trait-associated cell types across datasets of larger size, varying coverage, as well as *seismic*'s adaptability to different species.

seismic associations are consistent across tissue-cell type granularities

Having examined *seismic*'s consistency in identifying a wide variety of trait-associated cell types, we turn our attention to evaluate the accuracy of *seismic's* ability to distinguish known vulnerable neuron types for a well-characterized neurological disease, Parkinson's disease (PD). PD pathophysiology is well-established, with dopaminergic neurons residing in the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) characterized as being particularly vulnerable to degeneration[29]. Using a large mouse brain dataset[30] encompassing up to 231 distinct cell types from 9 regions of the adult mouse brain, in conjunction with a recent PD GWAS study[31] with over 480,000 participants, we test whether *seismic*, scDRS, FUMA, and S-MAGMA can recover known PD associations (Fig. 4, Supplementary Data 5).

The rich brain region and cell type annotations in[30] provide a unique opportunity to test how changes in cell type granularity affect the reported cell type-trait annotations. We examine 5 different granularities of cell types, ranging from 14 broad subclass labels to 231 highly-specific cell annotations (brain region + fine cluster). *seismic* is the only method to significantly prioritize PD-relevant dopaminergic neurons across all cell type granularities (Fig. 4). FUMA and scDRS also rank relevant cell types in some granularities highly, but mostly fail to reach statistical significance after multiple hypothesis test correction. Notably, S-MAGMA completely misses these vulnerable cell types. We note also that most previous cell type-trait association analyses that use datasets such as[30] typically perform their analyses at a broader cell type level (usually what we have termed the 'brain region + class' granularity). Though using finer resolution annotations increases the number of multiple hypotheses compared, we demonstrate that it may be a worthwhile trade-off when using a more powerful association detection method, as it can lead to more precise biological insights.

The *seismic* framework offers a novel perspective on AD pathogenesis

The choice of an endpoint for cell-type trait association can allow for the dissection of cell contribution to various endophenotypes of disease. This is particularly true for diseases with multicellular pathogenesis like Alzheimer's disease (AD), where genetic studies have mapped clinical and alternative traits, making it also an ideal test case for demonstrating the power of *seismic*. Furthermore, while selective neuronal vulnerability and pathological lesion formation

https://doi.org/10.1038/s41467-025-63753-z

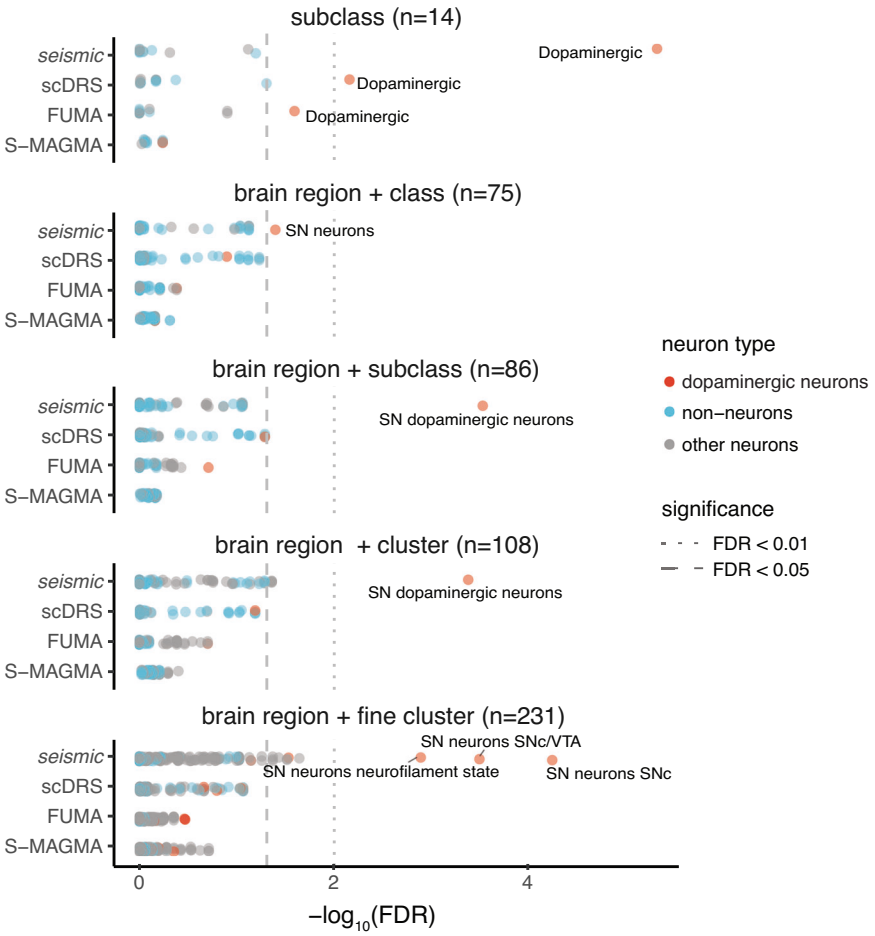

Fig. 4 | Identification of genetically vulnerable brain cell types for Parkinson's disease across multiple analysis granularities. Cell type-trait associations with a Parkinson's disease GWAS study[31] using cell type characterizations at different granularities from Saunders et al.[30]. Specifically, cells are partitioned into five different resolutions with increasingly more specific cell type characterizations, the broadest of which is subclass labels (top), to the most specific, which is brain region + fine clusters (bottom). Points represent individual cell types, colored by the neuron type (the most vulnerable dopaminergic neurons, other neurons, or non-neurons). Vertical dashed lines indicate FDR thresholds (0.01 and 0.05). Any cell types with FDR < 0.01 are labeled, except for the subclass and brain region + class granularities, where cell types with FDR < 0.05 are labeled.

have been thoroughly described in AD[32–35], the precise molecular mechanisms driving neurodegeneration leading to cognitive decline remain poorly understood. Formally, AD is characterized by two pathological hallmarks, extracellular amyloid plaques composed of Aβ peptide and intracellular neurofibrillary tangles (NFTs) formed by aggregated tau protein. NFTs appear according to a stereotypical spatial pattern, first emerging in layer II of the entorhinal cortex (EC), later appearing in deeper layers of EC and CA1 in the hippocampus, before subsequently spreading to other neocortical and subcortical regions. Progression of NFTs is accompanied by neurodegeneration in the affected area. In spite of the strong correlation between clinical symptoms of the disease and neuronal processes (NFT formation, neurodegeneration, synapse loss), many GWAS studies have primarily identified associations with immune cells such as microglia[36–38]. This leads to questions of whether microglia are the primary drivers of the disease or merely responsible for the clinical symptoms of the disease. If the latter, one would expect to find non-microglia associations for GWAS with non-clinical, pathology-based endpoints, which could open new research avenues for understanding pathogenic mechanisms.

We use *seismic* to test whether GWAS for different AD-related endpoints might yield divergent cell-type associations. Given that AD pathology typically exhibits selective regional vulnerability, the analysis was conducted at the most fine-grained resolution ('brain region + fine cluster' in Fig. 4). We first used an AD GWAS that includes a large cohort of >63,000 patients diagnosed via clinical observations (clinical GWAS)[39]. This large study is representative of the AD GWAS typically used in cell-type trait association studies. We also explore *seismic* results for a GWAS for an alternative AD endophenotype comprised of around 3100 patient samples of cerebrospinal fluid (CSF) tau levels[40], which serves as a biomarker of AD progression (tau GWAS)[41]. Though the tau GWAS has a much smaller patient cohort, we hypothesized that it may deliver clues for pathological mechanisms that have remained elusive with the clinical GWAS.

Applying *seismic* on the clinical AD study along with the expression data from Saunders et al.[30], we identify microglial cells from various brain regions as the most associated with clinical GWAS (Fig. 5A), demonstrating the pervasive neuroinflammation patterns underlying clinical symptoms in AD patients. This is consistent with previous studies, and indeed, we see that scDRS also identifies significant associations between clinical AD with microglial cells; neither FUMA nor S-MAGMA find statistically significant associations, though FUMA does rank microglial associations as highest among cell types (Supplementary Fig. 24). It is noteworthy that microglial cells from both vulnerable (hippocampus) and resistant (striatum) regions of the brain are similarly associated with the trait, suggesting that they are not the primary driver of regional differences in pathology. Using the tau GWAS with the same

https://doi.org/10.1038/s41467-025-63753-z

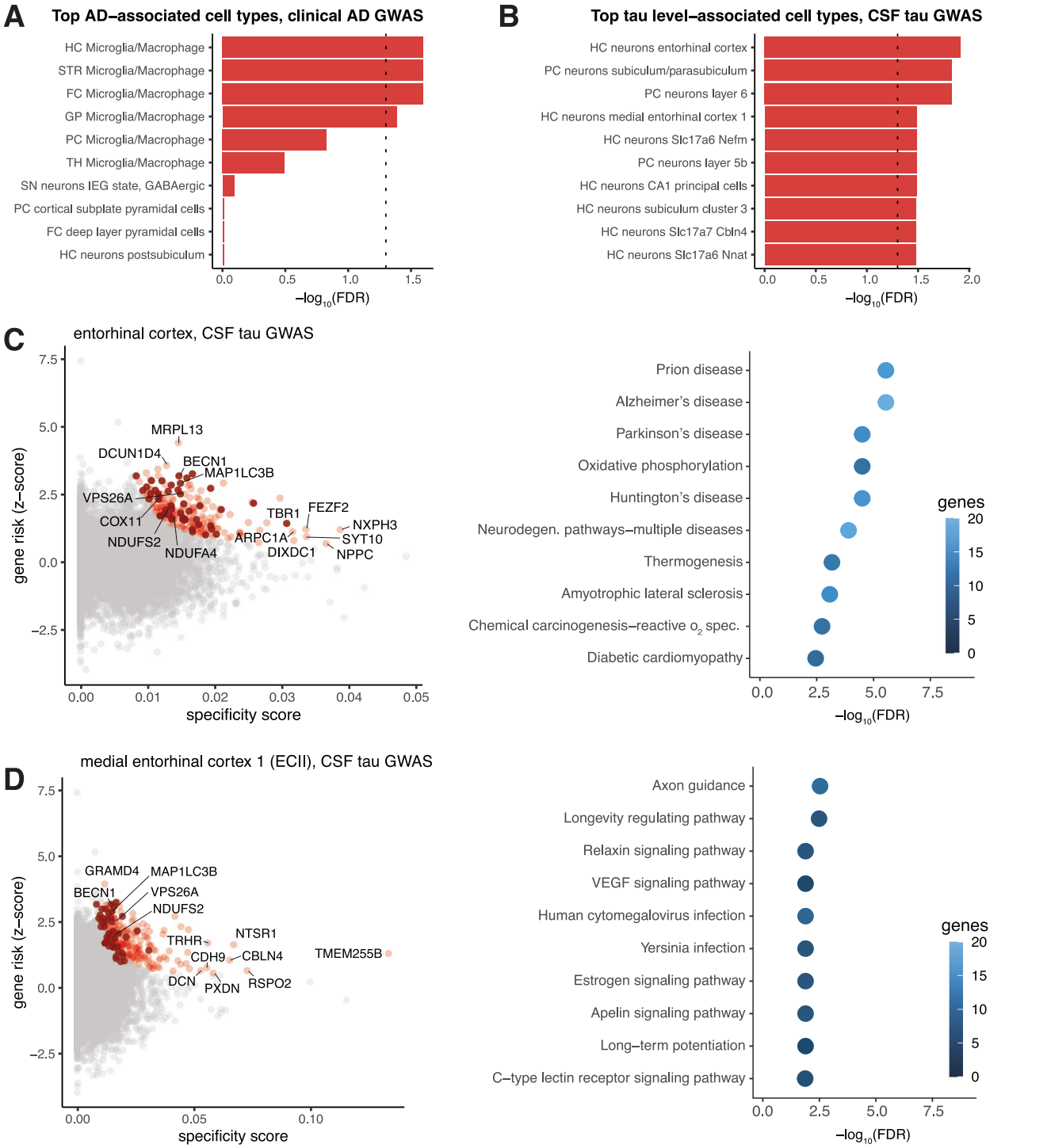

Fig. 5 | *seismic* reveals distinct divergent cellular mechanisms underlying Alzheimer's disease (AD) progression and its hallmark tau accumulation. A Top 10 associated cell types (at the brain region and fine cluster granularity) when applying *seismic* to an AD GWAS study with a clinical endpoint are shown, ordered by *p*-values. Brain region abbreviations: FC, frontal cortex; GP, globus pallidus; HC, hippocampus; PC, posterior cortex; SN, substantial nigra; STR, striatum; TH, thalamus. **B** Top 10 associated cell types (at the brain region and fine cluster granularity) when using a GWAS that measures an AD biomarker (CSF tau levels). Brain region abbreviations: HC, hippocampus; PC, posterior cortex. Additional term specification: "HC neurons entorhinal cortex" are entorhinal cortex (EC) excitatory neurons positive for *Nxph3*; "HC neurons medial entorhinal cortex 1" are EC layer II (ECII) excitatory neurons positive for *Cbln1*. **C** Influential gene analysis for the genes driving the association of neurons from deep layers of the entorhinal cortex (HC neurons entorhinal cortex) and the tau GWAS. Genes colored in red are influential, and dark red indicates influential genes driving the associations for both these and ECII neurons; several top influential genes are labeled. KEGG terms enriched for these genes are shown in the enrichment plot to the right. **D** Influential gene analysis for the genes associated with neurons in the medial entorhinal cortex (HC neurons medial entorhinal cortex 1, which we refer to as ECII) and the tau GWAS. As in (**C**), influential genes are colored red, and influential genes shared between deep layers of the EC and ECII are darkened. The enrichment plot shows KEGG pathways for the corresponding influential ECII genes.

Saunders et al. expression dataset, we see that scDRS seems to suggest generic astrocyte associations. Though FUMA and S-MAGMA still struggle with identifying statistically significant associations, they do prioritize diverse neuronal associations, including one of the neuronal populations known to be vulnerable to tau accumulation in AD ("HC neurons entorhinal cortex"). Meanwhile, using *seismic*, we find significant associations with several of the neuronal populations most vulnerable to tau accumulation in AD (Fig. 5B), including deep ("HC neurons entorhinal cortex") and superficial ("HC neurons medial entorhinal cortex 1") layers of the entorhinal cortex, as well as CA1 pyramidal cells. While these populations have been established as vulnerable based on pathology studies, such associations have not previously been found using trait-association studies based on AD-related GWAS. These results suggest that tau pathology in AD may be more intrinsically linked to neuronal susceptibility than inflammation alone.

With *seismic*'s influential gene analysis, we can more closely inspect the genes corresponding to increased risk driving the top cell type-trait association. For hippocampal microglia, we identify 88 genes as positively influential for clinical AD diagnosis, and find these to be enriched for immune-related GO processes (Supplementary Fig. 25, Supplementary Data 6). Some of these genes are expected, like *SPI1*, *MS4A6A*, and *TREM2*, which are microglia-specific and known to have significantly reported associated SNPs with clinical AD. There are also several other interesting genes identified by *seismic*, such as *LAPTM5*, an amyloid plaque responsive gene[42], and phagocytosis regulator *VAV1*[43]. Much less expected are the influential genes found for entorhinal cortex neurons and the tau GWAS (218 genes for neurons from deep layers of the entorhinal cortex, 199 genes for neurons from entorhinal cortex layer II) since, as mentioned, previous GWAS have not yielded clear neuronal genes associated with AD-related traits (Fig. 5C, D, Supplementary Data 6).

There is some overlap of influential genes shared by all AD vulnerable neurons (different layers of the entorhinal cortex and hippocampus CA1) associated with the tau GWAS (Supplementary Fig. 26). However, few pathways are consistently enriched across all three neuron types (Supplementary Data 6). Instead, both entorhinal cortex populations are enriched in genes involved in axon guidance, while CA1 and entorhinal cortex layer II (ECII) show enrichment in genes related to long-term potentiation (Fig. 5C, D, Supplementary Data 6), suggesting that distinct cellular processes contribute to the association of these cell types with CSF tau. Notably, the enrichment of both long-term potentiation and axon guidance genes in ECII aligns well with our previous study[44], which used orthogonal datasets and analysis strategies to demonstrate that regulators of structural and electrophysiological features of the axon underlie ECII vulnerability. This convergence of evidence strongly points to ECII axons as a defining Achilles' heel for these neurons. Beyond these axonal and synaptic pathways, we find that genes driving associations with CSF tau levels in deep layers of the EC show strong enrichment in several metabolic pathways (e.g., cellular respiration, electron transport chain), while those driving associations with tau levels in layer II neurons are more enriched in proteostasis pathways (e.g., protein destabilization) (Supplementary Fig. 27, Supplementary Data 6). Metabolic[45] and proteostatic[46,47] contributions to vulnerability of EC are among the very pathways previously suggested to underlie EC vulnerability, and several identified influential genes, such as *VPS26A*, have connections with EC vulnerability in both cell types. The fact that *seismic* does not find any association between microglia and CSF tau further suggests that microglia might not be the main drivers of tau pathology, but rather the drivers of the clinical manifestations of AD. Additionally, we have found that using the tau GWAS enables *seismic* to uncover neuronal associations as well as genes and pathways with important mechanistic and therapeutic potential. These results demonstrate the value of more targeted endophenotype GWAS—albeit smaller—for complex diseases.

Discussion

Atlas-scale single-cell RNAseq datasets have been generated with the promise of many exciting future applications. One way in which we can realize this potential is by combining them with quantitative genetic studies, helping to disentangle the tissue- and cell-specificity of complex traits and diseases. So far, several existing tools integrating scRNA-seq with GWAS studies to uncover cell type-trait associations have emerged[1–3], but these tools have several drawbacks (Table 1) that *seismic* addresses, including removing the need for thresholds, accounting for gene expression variability, and being scalable with the number of cells in an expression dataset. Importantly, *seismic* proposes influential gene analysis as a means to derive deeper biological insights for cell type-trait analyses.

Through *seismic*, we underscore the critical, yet often overlooked, influence of cell type granularity and GWAS endpoint selection on the outcome of cell type-trait association studies—two factors that, while intuitively important, have not been systematically explored in previous analyses. *seismic* not only brings this observation to light but also demonstrates precisely how these choices shape the biological relevance and depth of the insights gained. Here, we observe that *seismic* can find results that are largely robust to changes in cell type granularity (Fig. 4), while changes in granularity seem to more strongly affect the cell type-trait associations detected by other methods, both in the statistical power of detecting true associations and in the prioritization of specific cell types. We thus strongly recommend running any method at several cell type granularities to assess result robustness and using the finer cell type definitions to reveal more mechanistic interactions. Furthermore, as evidenced by the AD case study, it may be fruitful to include more targeted GWAS endpoints when studying complex disease. One of the interesting future directions we envision for *seismic* is an extension to automatically test for cell-type associations at different cell-type granularities simultaneously. Such a feature would further complement the expanding cell reference atlases and facilitate faster identification of important cell-trait signal.

The application of *seismic* on the two AD-related GWAS studies (Fig. 5) begins to address a long-standing conundrum with AD-linked GWAS: large clinical studies consistently identify only microglial associations, despite AD being characterized by neuronal pathology and selective neuronal vulnerability—patterns not reflected in the regional homogeneity of microglia. This has dampened enthusiasm for investigating the mechanisms of tau accumulation within neurons, an important facet of AD pathogenesis that offers significant therapeutic potential. Limitations of large clinical GWAS, such as the inclusion of individuals with significant silent pathology in control groups[48] or mixed pathologies in the diseased group[49], likely dilute signals related to AD pathology and accentuate neuroinflammation signals common across various neuropathologies. More critically, clinical GWAS may capture processes tied closely to symptomatic presentation rather than the broader, decades-long pathological cascade of AD.

The cognitive reserve theory and pathological evidence from longitudinal cohort studies suggest that while NFTs are necessary for AD symptom onset, the severity of cognitive symptoms does not always align with the extent of neuropathology. In fact, cognitive decline can be influenced by factors that have little to do with pathogenic mechanisms[50–52], suggesting that clinical endpoints may not be fully adequate for capturing the neuronal signals essential for disease progression. GWAS using pathological or biomarker endpoints can avoid these pitfalls, but they are smaller and have less power. Here, we demonstrate that *seismic*'s ability to leverage neuron-type-specific signals in lower-powered GWAS can overcome these challenges. In what we believe is the first reported neuronal association for an

AD-related GWAS with normal reference expression data, we show that a proxy for tau pathology, tau levels in the CSF, is genetically associated with the most vulnerable neurons in AD. While previous GWAS-based cell-type enrichment methods have predominantly identified microglial associations[1,6,38], our approach successfully captures the neuronal component that reflects established pathological patterns. Since CSF tau is released from neurons undergoing degeneration or accumulating tau, it likely reflects processes occurring within these vulnerable populations. Among all the neuron types profiled by Saunders et al., *seismic* identifies every type that exhibits NFT pathology during preclinical stages of AD, underscoring the power of this framework. The association of vulnerable neurons with CSF tau, but not microglia, suggests that solely focusing on microglia may overlook key aspects of AD pathology. By using GWAS for other AD-linked endpoints, we hope to eventually develop a more holistic view of microglia's role, particularly in their crosstalk with the vulnerable neurons responsible for AD neuropathology.

Our CSF tau GWAS analysis also contributes to the ongoing debate on tau spreading. The identification of CA1 pyramidal cells and entorhinal cortex layer V neurons, which accumulate tau pathology after ECII neurons, suggests that these cells may also have an intrinsic vulnerability to NFTs, rather than simply being passive recipients of tau spreading from ECII neurons. This again highlights the potential for hard-wired susceptibility in certain neuron types.

Even though in our study, *seismic* has demonstrated strong performance in capturing trait-associated cell types, several caveats deserve further attention. Firstly, as with any other cell type-trait association method, statistical significance suggests a strong association, but not necessarily causality between cell types and traits. Secondly, there can still be false negatives using *seismic*. In our analysis of the Tabula Muris FACS dataset across 27 traits, *seismic* misses 22 association pairs that are identified by at least two other methods (Fig. 3). We note that this is the lowest number among all methods (135, 139, 118 for scDRS, FUMA, S-MAGMA, respectively), and a closer inspection of these 22 associations reveal that several may be spurious (e.g., Crohn's disease associated with limb muscle skeletal muscle satellite cell, all results in Supplementary Data 4), suggesting that some of these 22 may be false positives by competing methods rather than false negatives from *seismic*. Nonetheless, a small number are better supported associations, such as the microglia-autoimmune disease link, which *seismic* may underdetect because it will better emphasize more sharply specific gene-expression signatures as opposed to more diffuse or broadly expressed programs. A third limitation is the usage of a scRNA-seq mouse brain atlas[30] in our PD and AD analyses, as opposed to a human scRNA-seq brain expression atlas. Unfortunately, current limitations in data quality and scale hinder us from using human brain-level data right now[53]. TS does not include brain tissue, but we do find that *seismic* identifies associations that are consistent between the mouse TM FACS dataset and the human TS dataset in overlapping cell types. We also note that, as found in previous studies[1], the mouse datasets typically yield cleaner association patterns. As atlas-scale human data continues to improve, we expect the signal we can detect with *seismic* to also improve. Finally, *seismic* currently leverages MAGMA-derived gene-level associations primarily based on proximal SNPs, potentially overlooking regulatory variants that act from greater distances. Future enhancements to *seismic*, detailed in Supplementary Note 3, could incorporate explicit modeling of cellular dependencies, batch effects, hierarchical cell-type relationships, and distal regulatory SNP-gene interactions. These advancements would further expand *seismic*'s utility for diverse complex trait-disease scenarios.

In conclusion, we have developed a new methodological toolkit to take better advantage of scRNA-seq and GWAS data to model the interplay between tissue, cell type and complex traits. We make all code for *seismic* and the accompanying analyses available through GitHub (https://github.com/ylaboratory/seismic-analysis) and installable as an R package, `seismicGWAS` (https://github.com/ylaboratory/seismic). The processed expression data and GWAS summary statistics are now publicly available on the Zenodo repository (https://zenodo.org/records/15582078).

Methods

The *seismic* framework

To identify cell type-trait associations, *seismic* takes 2 inputs: (1) MAGMA[7] z-scores processed from GWAS summary statistics for a given trait; and (2) a scRNA-seq dataset that covers cell types of interest (Fig. 1A). More details regarding how we processed GWAS and scRNA-seq datasets can be found below (see 'Data preprocessing'). Many scRNA-seq datasets have an inherent hierarchical labelling structure (e.g., cells can be grouped by tissue of origin and also further divided by cell subclass or cell state). In other words, cell types can be categorized at different levels of *granularity*, for example, adding resolution to a traditional cell type characterization by also considering its tissue subregion. In most applications, we recommend choosing finer granularities, which typically translates to higher resolution results.

One of the motivating assumptions of *seismic* is that the genes that best characterize a cell type are not necessarily the genes that have the highest expression in the cell type, but instead the genes that are *most specific* to that cell type. Optimally, these cell-type-specific genes would have consistently higher expression in all cells within the cell type and much lower or even no expression in other cells. The key insight of *seismic* lies in how to translate this optimal criterion into a continuous specificity score that can be calculated from a given scRNA-seq dataset and collection of cell types.

The *seismic* specificity score. The optimal criterion can be broken down into two sub-criteria: (1) consistently higher expression in a cell type of interest in comparison to other cells; and (2) expression in all cells within the cell type. These two sub-criteria are naturally related, but capture different aspects of cell-type-specificity. The first sub-criteria captures the variability and magnitude of expression, while the second focuses on the proportion of cells in a cell type where the gene is expressed (in a binary sense, ignoring the magnitude of expression). Here, we describe how continuous scores representing each of these sub-criteria are calculated.

Let E be an $N \times M$ matrix representing scRNA-seq expression data with N genes and M cells. With $j \in \{1, \ldots, M\}$ representing the index for each cell in E and $c \in \{1, \ldots, C\}$ representing the index for each of C cell types at the selected granularity, we represent cell type set membership as a labeled set of cells $L^{(c)} = \{j | j \text{ is labeled as cell type } c\}$. For a given cell type c, we define $E^{(c)}$ as the sub-matrix of E with column indices provided by $L^{(c)}$. $E^{(c')}$ is the sub-matrix of E for the set complement of $L^{(c)}$, which represents the expression of all cells not labeled as c.

We first seek to estimate the probability a gene has consistently higher expression in a cell type of interest. In other words, we want to estimate a score $p_i^{(c)} = P(X_i^{(c)} \geq X_i^{(c')})$, where $X_i^{(c)}$ is a random variable representing the expression level of gene i in cell type c. We see that:

$$P(X_i^{(c)} \geq X_i^{(c')}) = P(X_i^{(c')} - X_i^{(c)} \leq 0) = P\left(\frac{\left(X_i^{(c')} - X_i^{(c)}\right) - \left(\bar{X}_i^{(c')} - \bar{X}_i^{(c)}\right)}{\sqrt{\frac{\sigma_i^{(c')^2}}{|L^{(c')}|} + \frac{\sigma_i^{(c)^2}}{|L^{(c)}|}}} \leq \frac{\bar{X}_i^{(c)} - \bar{X}_i^{(c')}}{\sqrt{\frac{\sigma_i^{(c)^2}}{|L^{(c)}|} + \frac{\sigma_i^{(c')^2}}{|L^{(c')}|}}} \right), \quad (1)$$

where $\bar{X}_i^{(c)}$ and $\bar{X}_i^{(c')}$ are the sample mean expression for gene i in cell type c and all other cells, respectively, $\sigma_i^{2(c)}$ and $\sigma_i^{2(c')}$ are the corresponding sample variances, and $|L^{(c)}|$ and $|L^{(c')}|$ are the number of cells in cell type c and all other cells, respectively. Thus, if there is a sufficiently large number of cells in cell type c (and otherwise), we know

that based on the Central Limit Theorem, we can estimate $p_i^{(c)}$ as:

$$p_i^{(c)} = P\left(X_i^{(c)} \geq X_i^{(c')}\right) \approx \Phi\left(\frac{\bar{X}_i^{(c)} - \bar{X}_i^{(c')}}{\sqrt{\frac{\sigma_i^{(c)2}}{|L^{(c)}|} + \frac{\sigma_i^{(c')2}}{|L^{(c')}|}}}\right), \quad (2)$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. We note that this formulation carries similarities with the derivation for the test statistic for a two-sample z-test; however, we are primarily interested in calculating a good estimate for $p_i^{(c)}$ and not in calculating p-values that would correspond to a null hypothesis for no difference in expression. In this current formulation, $p_i^{(c)}$ is bounded from [0, 1], where higher scores reflect higher expression in cell type c for gene i, even after considering the variability of expression.

Next, we consider how to estimate the probability of expression across cells in a cell type of interest if we consider expression as a binary variable. We see quickly that this is akin to estimating the ratio or proportion of cells in a cell type where we see non-zero gene expression. Thus,

$$r_i^{(c)} = \frac{1}{|L^{(c)}|}\sum_{j=1}^{|L^{(c)}|} I\left(E_{ij}^{(c)} > 0\right), \quad (3)$$

where I is the indicator function. As with $p_i^{(c)}$, $r_i^{(c)}$ it is also bounded from [0, 1], but here, higher scores reflect that gene i is expressed in a higher proportion of cells in cell type c.

Finally, we define the *seismic* specificity score as:

$$s_i^{(c)} = \frac{p_i^{(c)} r_i^{(c)}}{\sum_{c=1}^{C} p_i^{(c)} r_i^{(c)}}. \quad (4)$$

Here, we see that $s_i^{(c)}$ is still bounded between [0, 1], but now, $p_i^{(c)}$ and $r_i^{(c)}$ are also rescaled across all cell types to further highlight the specificity. A score close to 1 can be achieved if gene i both has consistently higher expression in cell type c compared with other cells, even after considering expression variability and is also expressed in all cells within the cell type. As part of the *seismic* framework, we calculate the specificity score $s_i^{(c)}$ for all genes i and cell types c in the dataset.

Quantifying cell type-trait associations. We assume that if a cell type is highly associated with a trait, then the cell-specificity of its gene expression should have explanatory power for gene risk. In other words, if we observe that as the specificity of a gene to a particular cell type increases, there is correspondingly increased association with a disease, that would suggest a strong cell type-trait association. This can be formulated as a linear model:

$$Z^{(c)} = \beta_0 + \beta_1 S^{(c)} + \epsilon \quad (5)$$

Here, $Z^{(c)}$ denotes the vector of gene z-scores for cell type c given by MAGMA gene analysis, and $S^{(c)}$ is the vector of *seismic* specificity scores for cell type c. Only genes that both have a z-score from MAGMA and are captured in the processed scRNA-seq datasets are considered.

After fitting the linear model, we can test the null hypothesis of $\beta_1 = 0$ against the one-sided alternative hypothesis $\beta_1 > 0$, resulting in a p-value for each cell type c. To correct for multiple hypothesis testing, we calculate and report Benjamini-Hochberg false discovery rates (FDRs)[54]. The one-sided test here tests for a positive linear relationship between cell-type-specificity and gene risk; we note that it is also possible that the relationship can be non-linear, which could potentially be addressed in the future using kernel methods. However, the advantage of this formulation is that we can statistically quantify the association in a directly interpretable way and it enables influential gene analysis (below).

Alternatively, *seismic* also provides a nonparameteric statistical test option (Spearman's rank correlation coefficient test) if the user wishes to explore potential non-linear relationships between the *seismic* score with MAGMA gene-level z-score. For a more detailed discussion of choice for a linear model over a non-linear model for prioritizing trait-implicated cell types, please refer to Supplementary Note 2.1.

Influential gene analysis. For a significant cell type-trait association pair, some genes with may be particularly "influential" for the linear model. Here, we are using the statistical definition of an influential observation, specifically that removing the observation (i.e., gene) would have a strong effect on the model. Here, we calculate the difference in betas (DFBETAS) statistic[55], which is a scaled measure of how much the model parameters will change when removing a single observation. Specifically, for each gene and β_1:

$$DFBETAS_i = \frac{\hat{\beta}_1 - \hat{\beta}_{(-i)1}}{\sqrt{MSE_{(-i)}(S^{(c)^T} S^{(c)})_{ii}^{-1}}}, \quad (6)$$

where $\hat{\beta}_1$ is the regression coefficient estimated with *all* genes, $\hat{\beta}_{(-i)1}$ is the new regression coefficient calculated with gene i removed, $MSE_{(-i)}$ is the mean squared error of the updated linear model without gene i, and $(S^{(c)^T} S^{(c)})_{ii}^{-1}$ is the i^{th} diagonal element of the $(S^{(c)^T} S^{(c)})^{-1}$ matrix calculated using *all* genes. As recommended by Belsley et al., we use the size-adjusted threshold for selecting top influential genes as $|DFBETAS| \geq \frac{2}{\sqrt{N}}$, where N is the number of genes modeled[55]. For gene set enrichment analysis of the resultant influential genes, we use clusterProfiler[56], where a negative geometric test was used for testing the statistical significance.

Existing MAGMA-based cell type-trait association methods

For scDRS[1] runs, we use the default parameters (1000 genes) with 1000 Monte Carlo (MC) samples and the top 5% quantile of trait scores across cells within a cell type as the test statistic for a given cell type. Because scDRS requires disease gene sets and an expression dataset as input, our processed data ('Data preprocessing') required some additional processing. Disease gene sets were generated from MAGMA z-scores using the provided *munge-gs* command. For cross-species analyses, scDRS deals with cross-species gene matches by their gene names; as such, we annotated the gene name entries of the MAGMA z-score files according to biomaRt mappings between Entrez IDs and gene symbols[57]. If a gene is mapped to multiple z-scores, we collapse by calculating the mean gene z-score. Rare cell types (those with fewer than 20 cells) were removed from expression datasets to enable consistent comparisons across methods. scDRS recommends the use of empirical P-values for assessment of statistical significance; however, the resolution is limited by the number of MC simulation runs, especially in a multiple-hypothesis test setting. Thus, we use the alternate recommendation of transforming disease-cell type association MC z-scores to their analytical p-value, followed by FDR calculations.

FUMA[2] provides a website for GWAS preprocessing and cell type-trait association analysis, but it is challenging to systematically analyze new datasets, especially for the null simulations and runtime analyses. We thus processed scRNA-seq datasets into the input format required by MAGMA (with expression values represented as log(CPM+1) as in FUMA) and used FUMA's recommended parameters and commands to obtain p-values for each trait and cell type using the MAGMA software.

MAGMA-specific gene set analysis (S-MAGMA[3]) takes as input a list of top genes based on the non-log-transformed CPM expression as in

Bryois et al.[6]. Specifically, we used rescaled gene expression (where each gene's expression value is divided by the total expression of that gene across all cell types), and the top 10% of genes that are most specific as defined by Byrois et al.[6] were used for downstream testing.

scRNA-seq datasets. We processed and analyzed 4 large atlas-scale scRNA-seq datasets, specifically Tabula Muris FACS (TM FACS)[17], Tabula Muris droplet (TM droplet)[17], Tabula Sapiens (TS)[28], and Saunders et al.[30], capturing over 1.5 million cells across over 60 tissue regions (Supplementary Data 2). The TM FACS, TM droplet, and TS datasets each had different quality control filters before we downloaded them. Specifically, the two TM datasets had removed mitochondrial genes and filtered outlier cells. We additionally filter out cells with fewer than 2000 unique molecular identifier (UMI) counts. The TS dataset had filtered out cells with fewer than 2500 counts but did not consider mitochondrial gene expression, so we filter out cells where >10% counts are from mitochondrial genes. The Saunders et al. dataset provided cell-level annotations for doublets, outliers, singletons, and unannotated cells, all of which we excluded. We further filtered out cells if either >10% of counts were from mitochondrial genes or if there were fewer than 1000 reads in the cell. Raw gene expression counts were normalized using cell-specific size factors estimated by *scran*[58]. Subsequent analyses use log2(normalized counts) with a pseudocount of 1. For all datasets, we filtered out genes that were either expressed in fewer than 10 cells or had a lower than 0.01 mean log-expression across all cell types. We also only retained cell types that had at least 20 cells, to ensure sufficient data for stable statistical estimation of specificity scores and to minimize potential noise from poorly represented clusters. We selected these default parameters based on empirical evaluations, but it may be possible that dataset-specific adjustments can further emphasize the dataset's characteristics and capture more signal (see Supplementary Note 2.2).

For comparisons using mouse scRNA-seq data, we used mouse-to-human gene mappings from biomaRt[57]. Genes without mapping to human genes were discarded, and for genes with multiple mappings to a single human gene, the mean specificity score was used.

We also manually examined the cell type label annotations for all datasets to resolve or filter out cells with unclear annotations. For example, certain cells had clearly confusing labels, such as hepatocytes in the heart tissue in TS, and so were excluded from further analyses. For the TM dataset, we used the provided annotations, pooling cells into different cell types based on either the existing Cell Ontology terms or the more detailed annotation where available. For the Tabula Sapiens (TS) dataset, cell types were manually assessed against corresponding terms in the Cell Ontology[59,60], and any cell types that could not be confidently resolved to an ontology term were excluded. To further understand the potential impact of inter-individual variation on results derived from the Tabula Muris FACS dataset, we performed a validation analysis by splitting the data by donor ID and recalculating specificity scores and trait associations separately for each donor (Supplementary Fig. 18). For the primary cross-trait analyses involving the TM and TS datasets (Fig. 3), cell types were defined based on the combination of tissue of origin and the curated cell type annotations described above, allowing for investigation of both tissue- and cell-type-specific associations. For other analyses, cell type annotations without tissue information were directly used.

For the Saunders et al. dataset, cells were annotated to different regions that had been sequenced separately (frontal cortex, posterior cortex, substantia nigra, hippocampus, thalamus, cerebellum, globus pallidus, entopedeuncular nucleus), as well as cell classes, clusters and subclusters. Cell classes represent broad cell types of a region, which Saunders et al. further refined by iterative clustering into progressively more specific cell clusters and subclusters. For neurons, the subclusters had also been more precisely annotated to their respective structure in the region based not only on computational inference but also on immunohistochemical validation. To establish better consistency for cell types at each granularity, we manually cleaned and refined these annotations. Specifically, for the 'subclass' granularity, non-neuronal cells retained their corresponding annotations (e.g., microglia, oligodendrocytes), while neuronal cells were categorized by neurotransmitter type (i.e., excitatory, inhibitory, dopaminergic, cholinergic). For the 'brain region + class' granularity, cells were grouped by both brain region and broader cell class (i.e., neuronal cells were considered as one entity and not subdivided by neurotransmitter type). Similarly, for the 'brain region + subclass' granularity, cells were grouped by both brain region and subclass. For the 'brain region + cluster' granularity, cells were grouped by both brain region and cleaned cluster annotations provided by Saunders et al. (which had finer resolution than subclass and corresponds to specific cortical layers or neuronal connectivity). Saunders et al. also further divided clusters into subclusters, the finest resolution annotations that they provide. These subcluster annotations were often redundant, so after resolving typos, we combined subclusters that had closely related annotations when they were derived from the same cluster within a dissected brain region. 'Brain region + fine cluster' is the granularity that corresponds to combining these combined subcluster annotations with brain region information. For interneurons, which span multiple brain regions, subclusters were combined by the primary gene marker. The number of cell types thus varied from 14 at the 'subclass' granularity to 231 at the 'brain region + fine cluster' granularity (Supplementary Data 4).

GWAS datasets. We processed and analyzed GWAS summary statistics across a total of 30 complex traits (Supplementary Data 1). Because the majority of GWAS summary statistics were reported using the GRCh37 reference build, any studies with summary statistics based on GRCh38 were converted to GRCh37 using LiftOver[61] for consistency. SNP rsID annotations used dbSNP build 151. Duplicated SNPs were dropped and SNPs without annotations in dbSNP were retained and named by their respective chromosomes and positions.

We used MAGMA (v1.09b) to annotate SNPs to genes and compute gene-level z-scores for each trait, which reflect the overall gene-level association to the trait after factors such as linkage disequilibrium and population stratification are regressed out based on individuals of European ancestry from Phase 3 of the 1000 Genomes Project[62]. SNP assignment to genes used a 35 kb upstream and 10 kb downstream window around the gene body, as recommended in previous studies[6]. We also examined the effects of systematically varying the MAGMA gene window size on *seismic* results (Supplementary Fig. 9).

Label noise simulation

To evaluate the robustness of *seismic* specificity score, we simulated various levels of cell type label noise using the TM FACS dataset to mimic the scenario when cell cluster errors exist. First, we randomly selected 10 tissue-specific cell types from the TM FACS dataset as target cell types. Then, we systematically introduced label noise by reassigning cells from other non-target cell types originating from the same tissue as the target cell type. These cells can be represented as contributor of clustering errors, which exhibit similar but slightly different expression patterns. Increasing proportions of these cells (ranging from 5% to 100% of the target cell types' original cell count) were added to the target cell type label set, simulating different degrees of such noise. We then recalculated cell type specificity scores for the new perturbed target cell types using *seismic* and compared three other cell-type-specificity metrics: differential expression score (DE score[63]), gene specificity used in S-MAGMA (Bryois gene specificity[6]), and specificity index[64]. The robustness of each scores was then assessed by the normalized L1 similarity (defined as 1 minus the normalized L1 distance) between the specificity score vector computed before and after the introduction of

label noise. Higher similarity values indicate greater robustness to annotation noise.

Null simulation

Null simulations were performed using random subsamples of the Tabula Muris (TM) FACS dataset[17] and randomly selected GWAS endpoints. More specifically, we randomly selected 10 GWAS endpoints from the 27 GWAS traits examined in Fig. 3 (Supplementary Data 1), then randomly subsampled 10 subsets of 10,000 cells from TM FACS to use as input expression data for each method, where 100 cells were randomly selected as the target cell type. In our first null simulation setup, for each of 10,000 runs, the gene indices for the expression matrix were randomized. This random shuffle essentially breaks any biologically relevant signal between the single cell expression and the gene z-scores. Keeping the MAGMA z-score fixed, the p-value for the random cell type and trait association was calculated for each method and compared with the expected p-values.

We also performed the more direct null simulation using shuffled z-scores on *seismic* and scDRS. (FUMA and S-MAGMA were excluded from this comparison because they require complete MAGMA gene analysis files (as opposed to more easily manipulated z-score files as in *seismic* and scDRS).) In this setup, for each of the 10,000 runs, we randomized the MAGMA z-score vector used as input to *seismic* and scDRS together with the same randomly selected target cell types from before (without further randomization of the expression matrix) and examined the reported p-values for cell type-trait association.

Power simulation

To evaluate each framework's ability to distinguish truly associated cell types from irrelevant ones, we simulated disease-specific cell expression profiles by perturbing the corresponding disease gene sets used in the null simulation. Specifically, we utilized the subsampled TM FACS dataset and a subset of trait MAGMA z-score vectors as input, leveraging scDesign3's capability to generate realistic synthetic count-level scRNA-seq expression matrices for trait-specific expression patterns[21]. In our simulation procedure, we randomly selected 100 cells as the target cell type from subsets of the TM FACS dataset with shuffled gene indices. The scDesign3 model was fitted using default parameters and model distribution specifications. To enhance the fit of the μ matrix (the expected mean expression parameter for the cell-by-gene matrix), we incorporated cell library size (total UMI counts) and cell type labels (cell ontology class) as covariates. We selected 10 random target gene sets per expression dataset based on the MAGMA z-score vector of the corresponding GWAS summary statistics. These sets comprised 1000 genes expressed in at least 10 cells, with a portion being trait-specific causal genes with high MAGMA z-scores, indicating a high density of significant variants in proximal regions. Causal genes were sampled based on MAGMA z-score weighting, while the other confounding genes were sampled uniformly. To simulate various levels of polygenic signal, we adjusted the ratio of causal genes to the total gene set size, ranging from 0.2 to 0.8. New count-level data were generated to replace the submatrix of the target cell type and gene set in each randomization, yielding 100 new expression matrices. We also varied the strength of expression perturbation by specifying different effect size gradients by modifying the fitted μ matrix to exact folds of the original value and simulating the new submatrix using scDesign3.

After simulating datasets with synthetic cell type association signals, we ran each method with the previously described pipeline and parameters. We report the power of each method in identifying the perturbed cells as true trait associations (where the FDR is less than 0.05 after Benjamini-Hochberg correction), defined as the proportion of 100 randomizations in which the goal is achieved. We also evaluated the power of each method to rank the correct cell type based on reported p-values (top 1 as well as top 5). Influential gene analysis is also performed using these simulations to evaluate *seismic*'s ability to distinguish true causal genes from the others.

For a complex trait, multiple cell types may be involved and exhibit trait specificity or disease vulnerability. To simulate this more sophisticated scenario, we also generate test cases with more than one causal cell type that display trait-specific expression patterns. Four simulation schemes were used to capture the correlation patterns across multiple associated cell types: exclusive genes, independent genes, shared causal genes, and shared other genes. In the 'exclusive genes' scheme, target genes are sampled uniquely for each cell type, where no overlap exists among the target gene sets of different cell groups. The 'independent genes' scheme allows for potential overlap by sampling target genes independently. The 'shared causal genes' scheme keeps causal genes identical across all target cell types, while the 'shared other genes' scheme maintains identical non-causal genes across cell types. For each scheme, we generate synthetic datasets following the previously described scDesign3-based approach, adjusting the sampling strategy accordingly.

Runtime analysis

We sampled a range of 10,000 to 500,000 cells from the Tabula Sapiens[28] (TS) data set to compare the end-to-end runtime for identifying associated cell types from a scRNA-seq dataset for 5 random traits (Supplementary Data 1). To assess runtime scalability beyond the original (around 300,000 cells) available in the TS dataset, we generated larger datasets (e.g., 400,000 and 500,000 cells) by resampling extra cells that exceed the total number of cells in the dataset. For each resampled cell that is appended the dataset, we concatenated a unique numerical suffix to its cell type annotation, in order to preserve the relative cell type proportions to the number of cells observed in the original data. In the command line mode of scDRS, it also performs extra analyses and reads in files multiple times, which would further increase its runtime. To isolate the runtime of only the cell type-trait association process, we wrote a Python script that excluded these unrelated steps. All time for reading in the dataset for scDRS was also excluded from the comparison. To control the CPU usage, podman containers were used to limit the CPU usage to a single core.

Reporting summary

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

Data availability

Processed expression data and GWAS summary statistics are available through Zenodo [https://zenodo.org/records/15582078].

Code availability

All code for *seismic* and the accompanying analyses are available through GitHub (https://github.com/ylaboratory/seismic-analysis) and installable as an R package, `seismicGWAS` (https://github.com/ylaboratory/seismic).

References

1. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
2. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with fuma. *Nat. Commun.* **8**, 1826 (2017).
3. de Leeuw, C. A., Stringer, S., Dekkers, I. A., Heskes, T. & Posthuma, D. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nat. Commun.* **9**, 3768 (2018).

4. Jia, P., Hu, R., Yan, F., Dai, Y. & Zhao, Z. scgwas: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. *Genome Biol.* **23**, 220 (2022).
5. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
6. Bryois, J. et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).
7. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. Magma: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
8. Mullins, N. et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).
9. Abrantes, A. et al. Gene expression changes following chronic antipsychotic exposure in single cells from the mouse striatum. *Mol. psychiatry* **27**, 2803–2812 (2022).
10. 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).
11. Trpchevska, N. et al. Genome-wide association meta-analysis identifies 48 risk variants and highlights the role of the stria vascularis in hearing loss. *Am. J. Hum. Genet.* **109**, 1077–1091 (2022).
12. Bell, S., Tozer, D. J. & Markus, H. S. Genome-wide association study of the human brain functional connectome reveals strong vascular component underlying global network efficiency. *Sci. Rep.* **12**, 14938 (2022).
13. Mrdjen, D. et al. The basis of cellular and regional vulnerability in alzheimer's disease. *Acta Neuropathol.* **138**, 729–749 (2019).
14. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
15. Hrovatin, K. et al. Delineating mouse β-cell identity during lifetime and in diabetes with a single cell atlas. *Nat. Metab.* **5**, 1615–1637 (2023).
16. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
17. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: the tabula muris consortium. *Nature* **562**, 367 (2018).
18. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
19. Padmanabhan, S., Caulfield, M. & Dominiczak, A. F. Genetic and molecular aspects of hypertension. *Circ. Res.* **116**, 937–959 (2015).
20. Kathiresan, S. & Srivastava, D. Genetics of human cardiovascular disease. *Cell* **148**, 1242–1257 (2012).
21. Song, D. et al. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat. Biotechnol.* **42**, 247–252 (2024).
22. Salvi, E. et al. Target sequencing, cell experiments, and a population study establish endothelial nitric oxide synthase (enos) gene as hypertension susceptibility gene. *Hypertension* **62**, 844–852 (2013).
23. Ren, M. et al. The biological impact of blood pressure-associated genetic variants in the natriuretic peptide receptor c gene on human vascular smooth muscle. *Hum. Mol. Genet.* **27**, 199–210 (2018).
24. Mayr, M. & Schmid, R. M. Pancreatic cancer and depression: myth and truth. *BMC Cancer* **10**, 1–6 (2010).
25. Balliet, W. E. et al. Depressive symptoms, pain, and quality of life among patients with non-alcohol-related chronic pancreatitis. *Pain research and treatment* **2012** (2012).
26. Desai, G. S. et al. The pancreas-brain axis: insight into disrupted mechanisms associating type 2 diabetes and alzheimer's disease. *J. Alzheimer's. Dis.* **42**, 347–356 (2014).
27. Röder, P. V., Wu, B., Liu, Y. & Han, W. Pancreatic regulation of glucose homeostasis. *Exp. Mol. Med.* **48**, e219–e219 (2016).
28. Consortium*, T. S. et al. The tabula sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
29. Hirsch, E., Graybiel, A. M. & Agid, Y. A. Melanized dopaminergic neurons are differentially susceptible to degeneration in parkinson's disease. *Nature* **334**, 345–348 (1988).
30. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
31. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
32. Braak, H. & Braak, E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).
33. Arnold, S. E., Hyman, B. T., Flory, J., Damasio, A. R. & Van Hoesen, G. W. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with alzheimer's disease. *Cereb. Cortex* **1**, 103–116 (1991).
34. Fukutani, Y. et al. Neurons, intracellular and extracellular neurofibrillary tangles in subdivisions of the hippocampal cortex in normal ageing and alzheimer's disease. *Neurosci. Lett.* **200**, 57–60 (1995).
35. Gómez-Isla, T. et al. Neuronal loss correlates with but exceeds neurofibrillary tangles in alzheimer's disease. *Ann. Neurol.: Off. J. Am. Neurol. Assoc. Child Neurol. Soc.* **41**, 17–24 (1997).
36. Efthymiou, A. G. & Goate, A. M. Late-onset alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener.* **12**, 1–12 (2017).
37. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
38. Jagadeesh, K. A. et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA sequencing and human genetics. *Nat. Genet.* **54**, 1479–1492 (2022).
39. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
40. Deming, Y. et al. Genome-wide association study identifies four novel loci associated with alzheimer's endophenotypes and disease modifiers. *Acta Neuropathol.* **133**, 839–856 (2017).
41. Zou, K., Abdullah, M. & Michikawa, M. Current biomarkers for alzheimer's disease: from CSF to blood. *J. Pers. Med.* **10**, 85 (2020).
42. Salih, D. A. et al. Genetic variability in response to amyloid beta deposition influences alzheimer's disease risk. *Brain Commun.* **1**, fcz022 (2019).
43. Hall, A. B. et al. Requirements for vav guanine nucleotide exchange factors and rho gtpases in fcγr-and complement-mediated phagocytosis. *Immunity* **24**, 305–316 (2006).
44. Roussarie, J.-P. et al. Selective neuronal vulnerability in alzheimer's disease: a network-based analysis. *Neuron* **107**, 821–835 (2020).
45. Khan, U. A. et al. Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease. *Nat. Neurosci.* **17**, 304–311 (2014).
46. Fu, H. et al. A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. *Nat. Neurosci.* **22**, 47–56 (2019).
47. Rodriguez-Rodriguez, P. A cell autonomous regulator of neuronal excitability modulates tau in alzheimeras disease vulnerable neurons. *Brain* **147**, 2384–2399 (2024).
48. Balusu, S., Praschberger, R., Lauwers, E., De Strooper, B. & Verstreken, P. Neurodegeneration cell per cell. *Neuron* **111**, 767–786 (2023).

49. Escott-Price, V. & Hardy, J. Genome-wide association studies for alzheimer's disease: Bigger is not always better. *Brain Commun.* **4**, fcac125 (2022).
50. Wilson, R. S. et al. Loneliness and risk of alzheimer disease. *Arch. Gen. psychiatry* **64**, 234–240 (2007).
51. Boyle, P. A. et al. Effect of purpose in life on the relation between alzheimer disease pathologic changes on cognitive function in advanced age. *Arch. Gen. psychiatry* **69**, 499–504 (2012).
52. Buchman, A. S. et al. Physical activity, common brain pathologies, and cognition in community-dwelling older adults. *Neurology* **92**, e811–e822 (2019).
53. Armand, E. J., Li, J., Xie, F., Luo, C. & Mukamel, E. A. Single-cell sequencing of brain cell transcriptomes and epigenomes. *Neuron* **109**, 11–26 (2021).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
55. Belsley, D. A., Kuh, E. & Welsch, R. E.*Regression diagnostics: Identifying influential data and sources of collinearity* (John Wiley & Sons, 2005).
56. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a J. Integr. Biol.* **16**, 284–287 (2012).
57. Smedley, D. et al. Biomart–biological queries made easy. *BMC Genomics* **10**, 1–12 (2009).
58. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. (2016).
59. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, 1–20 (2012).
60. Haendel, M. A. et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in uberon. *J. Biomed. Semant.* **5**, 1–13 (2014).
61. Nassar, L. R. et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
62. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
63. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
64. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).

Acknowledgments

We would like to thank members of the laboratory for helpful discussions. We would also like to thank the research participants and employees of 23andMe, Inc., for making this work possible by providing access to the Parkinson's disease GWAS summary statistics. This work was supported by the National Institute of Aging [RF1AG054564 to JPR, with subaward to VY; R21AG085464 to JPR] and the Cancer Prevention & Research Institute of Texas [RR190065 to VY]. JPR is supported by the Cure Alzheimer's Fund and the Karen Toffler Charitable Trust. VY is a CPRIT Scholar in Cancer Research.

Author contributions

Q.L. implemented the method, built the package, and performed analyses. R.D. refactored the package code and prepared figures with Q.L. J.P.R. and V.Y. conceived of the study and supervised the work. All authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63753-z.

Correspondence and requests for materials should be addressed to Jean-Pierre Roussarie or Vicky Yao.

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