

---

1  
2 **Polygenic regression uncovers trait-relevant cellular contexts through pathway activation**  
3 **transformation of single-cell RNA sequencing data**

4 Yunlong Ma<sup>1,2,#</sup>, Chunyu Deng<sup>3,#</sup>, Yijun Zhou<sup>1,2</sup>, Yaru Zhang<sup>1,2</sup>, Fei Qiu<sup>1</sup>, Dingping Jiang<sup>1</sup>, Gongwei  
5 Zheng<sup>1</sup>, Jingjing Li<sup>1</sup>, Jianwei Shuai<sup>2</sup>, Yan Zhang<sup>3,\*</sup>, Jian Yang<sup>4,5,\*</sup>, Jianzhong Su<sup>1,2,\*</sup>

6  
7 <sup>1</sup> School of Biomedical Engineering, School of Ophthalmology & Optometry and Eye Hospital,  
8 Wenzhou Medical University, Wenzhou, 325027, China

9 <sup>2</sup> Oujian Laboratory, Zhejiang Lab for Regenerative Medicine, Vision and Brain Health, Wenzhou  
10 325101, Zhejiang, China

11 <sup>3</sup> School of Life Science and Technology, Harbin Institute of Technology, Harbin 150080, China

12 <sup>4</sup> School of Life Sciences, Westlake University, Hangzhou, 310012, Zhejiang, China

13 <sup>5</sup> Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China

14 # These authors contributed equally to this work.

15  
16 \* **Correspondence:** (J.S.) [sujz@wmu.edu.cn](mailto:sujz@wmu.edu.cn), (J.Y.) [jian.yang@westlake.edu.cn](mailto:jian.yang@westlake.edu.cn), or (Y. Z.)  
17 [zhangtyo@hit.edu.cn](mailto:zhangtyo@hit.edu.cn)

18  
19 **Keywords:** GWAS, Single-cell RNA sequencing, Genetic variants, Risk genes, Trait-relevant cell  
20 types

## 30 **Summary**

31 Advances in single-cell RNA sequencing (scRNA-seq) techniques have accelerated functional  
32 interpretation of disease-associated variants discovered from genome-wide association studies  
33 (GWASs). However, identification of trait-relevant cell populations is often impeded by inherent  
34 technical noise and high sparsity in scRNA-seq data. Here, we developed scPagwas, a computational  
35 approach that uncovers trait-relevant cellular context by integrating pathway activation transformation  
36 of scRNA-seq data and GWAS summary statistics. scPagwas effectively prioritizes trait-relevant genes,  
37 which facilitates identification of trait-relevant cell types/populations with high accuracy in extensive  
38 simulated and real datasets. Cellular-level association results identified a novel subpopulation of naïve  
39 CD8+ T cells related to COVID-19 severity, and oligodendrocyte progenitor cell and microglia subsets  
40 with critical pathways by which genetic variants influence Alzheimer's disease. Overall, our approach  
41 provides new insights for the discovery of trait-relevant cell types and improves the mechanistic  
42 understanding of disease variants from a pathway perspective.

43

## 44 **Introduction**

45 Genome-wide association study (GWAS) data on complex diseases and numerous genotype-phenotype  
46 associations have tremendously accumulated in the past decades<sup>1-4</sup>. However, functional interpretation  
47 of these variants identified by GWASs remains challenging. It is still unclear how these variants  
48 regulate key biological pathways in relevant tissues/cell types to mediate disease development. The  
49 advent of single-cell RNA sequencing (scRNA-seq) technology has provided an unprecedented  
50 opportunity to characterize cell populations and states from heterogeneous tissues<sup>5</sup>. Unveiling trait-  
51 relevant cell populations from scRNA-seq data is crucial for exploring the mechanistic etiology of  
52 complex traits (including diseases)<sup>6</sup>. Thus, linking scRNA-seq data with genotype-phenotype  
53 association information from GWASs has considerable potential to provide new insights into the  
54 polygenic architecture of complex traits at a high resolution<sup>7</sup>.

55

56 Several studies have revealed significant enrichment of complex traits in relevant tissue types by  
57 integrating tissue-specific gene expression profiles with GWAS summary statistics<sup>8-10</sup>. Inspired by  
58 these tissue-type enrichment methods, several methods<sup>11-15</sup>, including LDSC-SEG, RolyPoly, and

---

59 MAGMA-based approaches, have been employed to incorporate GWAS and scRNA-seq data to  
60 identify predefined cell types associated with complex traits. However, these approaches largely  
61 neglect the considerable intra-heterogeneity within each cell type and thus are not suitable for making  
62 inferences at single-cell resolution. Recently, scDRS<sup>16</sup> was developed to distinguish disease-associated  
63 cell populations at the single-cell level; however, its accuracy relies heavily on a set of disease-specific  
64 genes identified from GWAS data using gene-based association test methods<sup>17-20</sup>, such as MAGMA<sup>20</sup>.  
65 Although the gene-scoring methods focus on the top significant genotype-phenotype associations and  
66 have been applied to bulk tissue or aggregated data analysis, it is still challenging to use these methods  
67 to make accurate per-cell-based inferences in scRNA-seq data. The top genetic association signals at  
68 specific loci may be absent from most cells because of the extensive sparsity and technical noise in  
69 single cell data<sup>21,22</sup>. To date, no method exists to optimize effective and robust trait-relevant genes  
70 applicable to scRNA-seq data for accurate inference of disease-associated cells at a fine-grained  
71 resolution.

72

73 Dynamic cell activities and states are often caused by the combined actions of interacting genes in a  
74 given pathway or biological process<sup>23</sup>. Compared with leveraging the expression level of individual  
75 genes, pathway activity scoring methods that collapse the functional actions of different genes  
76 involved in the same biological pathways can prominently enhance statistical power and biological  
77 interpretation for determining particular cellular functions or states<sup>21,24-26</sup>. Recent studies have shown  
78 that such pathway-based scoring methods exhibit a greater reduction of technical and biological  
79 confounders of scRNA-seq data<sup>27-29</sup>. Moreover, multiple lines of evidence have suggested that  
80 clinically informative variants associated with complex diseases mainly occur in systems of closely  
81 interacting genes, and even variants with weak association signals clustered in the same biological  
82 pathway could provide critical information to understand the genetic basis of complex diseases<sup>30,31</sup>.  
83 Thus, integrating coordinated transcriptional features in biological pathways from scRNA-seq data  
84 and polygenic risk signals from GWAS summary statistics is a promising approach to prioritize trait-  
85 relevant genes and distinguish critical cells by which genetic variants influence diseases.

86

87 Here, we present a pathway-based polygenic regression method (scPagwas) that integrates scRNA-seq  
88 and GWAS data for the discovery of cellular context critical for complex diseases and traits. scPagwas  
89 performs a linear regression of GWAS signals on pathway activation transformed from scRNA-seq  
90 data to identify a set of trait-relevant genes, which are subsequently used to infer the most trait-relevant  
91 cell subpopulations. We show that scPagwas outperforms the state-of-the-art methods using extensive  
92 simulated and real scRNA-seq datasets. Through scPagwas-based analyses of different diseases, we  
93 provide new biological insights into how disease-associated naïve CD8<sup>+</sup> T cells are involved in  
94 COVID-19 severity and subsets of microglia and oligodendrocyte progenitor cells (OPCs) can  
95 contribute to Alzheimer’s disease (AD) risk.

## 96 97 **Results**

### 98 *Overview of scPagwas*

99 Given extensive evidence<sup>11,13,32,33</sup> indicating a positive correlation between genetic associations for a  
100 trait of interest and expression levels of genes in bulk tissue or specific cell type, we apply this principle  
101 to scRNA-seq data and take advantage of gene expression signatures shared in a biological pathway.  
102 scPagwas first links single-nucleotide polymorphisms (SNPs) in the GWAS summary data to each  
103 pathway by annotating SNPs to their proximal genes of the corresponding pathway (Figure 1A). In  
104 each cell, scPagwas calculates the correlation between the genetic effects of SNPs and gene expression  
105 levels within a given pathway to estimate regression coefficient  $\tau$  (Figure 1B), which reflects the  
106 strength of association between pathway-specific gene expression activity and the variance of SNP  
107 effects<sup>34</sup>. Meanwhile, scPagwas transforms the normalized gene-by-cell matrix to a pathway activity  
108 score (PAS)-by-cell matrix, which is constructed using the first principal component (PC1) of gene  
109 expression in each pathway via the singular value decomposition (SVD) method<sup>22,35</sup> (Figure 1C, see  
110 Methods).

111  
112 Following previous studies<sup>13,34</sup>, we compute the product of  $\hat{\tau}$  and PAS for each pathway, hereinafter  
113 referred to as genetically-associated PAS (gPAS), to capture the pathway-based genetic variances  
114 of traits of interest at the single-cell level (Figure 1D). Then, we use the central limit theorem method<sup>36</sup>  
115 to identify significant trait-relevant pathways based on the ranking of gPASs of pathways across

116 individual cells within each cell type (see Supplementary Methods). In the meanwhile, we compute  
117 the sum of gPASs over all pathways in each cell, correlate it with the expression level of each gene  
118 across cells, and prioritize the trait-relevant genes by ranking the correlations (Figure 1D). Finally, a  
119 trait-relevant score (TRS) of each cell is computed by averaging the expression level of the trait-  
120 relevant genes and subtracting the random control cell score via the cell-scoring method used in  
121 Seurat<sup>37</sup> (see Methods). By treating the set of cells in a predefined cell type as a pseudo-bulk  
122 transcriptomic profile, scPagwas can also be employed to infer significant trait-relevant cell-types  
123 using the block bootstrap method.

124

125 The input of scPagwas includes gene sets of pathways, a gene-by-cell matrix of scRNA-seq data, and  
126 summary statistics from a GWAS or meta-analysis for a quantitative trait or disease (case-control  
127 study). The typical output includes 1) per cell-based TRSs and the corresponding P values; 2) trait-  
128 associated cell types from the block bootstrap analysis; 3) trait-relevant pathways based on the ranked  
129 gPASs; 4) trait-relevant genes based on the ranked PCCs (Figure 1E).

130

### 131 *scPagwas effectively identifies trait-relevant genes*

132 Because trait-relevant genes are vital for inferring the TRS of each cell, we compared the biological  
133 functions of the top 1,000 trait-relevant genes identified from scPagwas with those identified with the  
134 widely-used gene-scoring method, MAGMA<sup>20</sup>, and three other well-established eQTL-based methods  
135 including TWAS<sup>17</sup>, S-PrediXcan<sup>19</sup>, and S-MultiXcan<sup>18</sup>. A panel of highly-heritable hematopoietic traits  
136 was used for benchmark analysis (Supplementary Tables S1-S2). We found that trait-relevant genes  
137 identified by scPagwas were more highly enriched in functional gene sets related to blood cell traits  
138 than those identified by the other four gene-scoring methods (Figure 2A and Supplementary Table S3).  
139 For example, the lymphocyte count-relevant genes prioritized by scPagwas showed highly significant  
140 enrichment in biological processes related to immune response, including T cell activation, adaptive  
141 immune response, leukocyte differentiation, and leukocyte cell-cell adhesion, whereas those  
142 prioritized by MAGMA lacked enrichment in any functional term (false discovery rate, FDR < 0.01,  
143 Figure 2B).

144

145 In scRNA-seq data, the sparsity and technical noise of individual genes can lead to high computational  
146 costs and inadequate association inference at single-cell level<sup>21,22</sup>. Using five distinct scRNA-seq  
147 datasets, we found that the use of pathway information with scPagwas could remarkably reduce the  
148 sparsity compared with that of individual gene-based evaluations (Supplementary Figure S1). The  
149 average expression magnitudes of individual genes showed a strong positive correlation with their  
150 variances (Figure 2C, blue line). In contrast, pathway activation scores transformed from transcriptome  
151 profiles significantly reduced the technical noise of variances of single-cell data, which facilitated the  
152 identification of biologically-relevant genes and improved downstream analyses<sup>38</sup> (Figure 2C and  
153 Supplementary Figure S2). These results demonstrate that scPagwas not only reduces sparsity and  
154 technical noise but also prioritizes more functional genes associated with trait of interest for per-cell-  
155 based inference to identify trait-relevant cells.

156

### 157 *Assessment of scPagwas in discerning trait-relevant cells*

158 We first assessed the power and precision of scPagwas in identifying trait-relevant cells using a real  
159 GWAS dataset and simulated scRNA-seq datasets. We adopted a highly heritable and relatively simple  
160 trait, monocyte count, for benchmark analysis, with the GWAS summary statistics from a large-scale  
161 study (N = 563,946, Supplementary Table S1). We synthesized an scRNA-seq dataset from  
162 fluorescence-activated cell-sorted bulk hematopoietic populations as the ground truth (see Methods),  
163 which contained a known relevant cell type (monocytes, n = 1,000 cells) and non-relevant cell types  
164 (T and B cells, dendritic cells (DCs), and natural killer (NK) cells, n = 1,000 cells in total, Figure 3A).  
165 We found that scPagwas (using the cell-scoring method of Seurat<sup>37</sup> by default) could accurately  
166 distinguish monocyte count-relevant cells from all simulated cells (precision = 95.9%, Figure 3B). We  
167 further examined whether the scPagwas-identified trait-relevant genes could improve the power of the  
168 latest cell-scoring method, scDRS<sup>16</sup>, by comparing the results with those using the default gene-based  
169 method, MAGMA. The precision of scDRS in identifying the trait-relevant cells increased from 0.940  
170 when using the the top 1,000 genes prioritized by MAGMA to 0.957 when using the top 1,000 genes  
171 prioritized by scPagwas (Figure 3C-D).

172

173 Moreover, compared with the gene-based methods that incorporate eQTL information (i.e., S-

174 MultiXcan, S-PrediXcan, and TWAS), the scPagwas-identified trait-relevant genes considerably  
175 enhanced the performance of the scDRS in distinguishing cells relevant to the monocyte count trait  
176 (scPagwas precision = 95.7% vs other three methods precision = 62.9%–90%, Supplementary Figure  
177 S3A). Using the same simulated single-cell dataset, we further examined the lymphocyte count trait  
178 also with a large-scale GWAS dataset (N = 171,643 samples) to benchmark the performance of  
179 scPagwas against the other four gene-based methods when using scDRS to score cells. We observed a  
180 consistent result that scPagwas yielded the best performance (scPagwas precision = 81.8% vs other  
181 four methods precision < 50%, Supplementary Figure S4A). In addition, we evaluated the performance  
182 of scPagwas in identifying predefined cell types related to a trait of interest in simulated data (see  
183 Supplementary Methods). We observed that scPagwas could effectively identify trait-relevant cell  
184 types under different genetic architectures when the number of included pathways was more than 100  
185 (Supplementary Figures S5–S6).

186

187 Next, we assessed whether scPagwas could distinguish monocyte count trait-related enrichment in a  
188 real ground truth scRNA-seq dataset (Figure 3E) that contained monocytes (CD14+ and CD16+  
189 monocytes, n = 5,000 cells) and non-monocyte cells (T cells (n = 3,000), B cells (n = 1,000), DCs (n  
190 = 200), and NK cells (n = 800)) from a bone marrow mononuclear cell (BMMC) scRNA-seq dataset  
191 (Supplementary Table S2)<sup>39</sup>. Consistent with the simulation results, scPagwas robustly identified the  
192 known trait-relevant cell populations with higher precision using Seurat as the cell-scoring method  
193 (precision = 98.2%, Figure 3F). Compared with the default setting of scDRS that uses the top 1,000  
194 MAGMA-identified genes, applying the top 1,000 scPagwas-identified genes to scDRS considerably  
195 enhanced the discovery of monocyte count-relevant cells with the improvement of the precision from  
196 0.787 to 0.984 (Figure 3G–H).

197

198 Analogous to the simulation results, we only found a moderate enrichment of monocyte count-relevant  
199 cells by applying the top genes prioritized by the three eQTL-based methods (S-MultiXcan, S-  
200 PrediXcan, and TWAS) to scDRS analysis (precision = 61.7–71.0%, Supplementary Figure S3B). This  
201 observation remained reproducible for lymphocyte count with the inclusion of the same real ground  
202 truth scRNA-seq dataset (Supplementary Figure S4B). When further evaluating whether the number

203 of included top trait-relevant genes influences the power of scoring trait-relevant cells, scPagwas using  
204 the scDRS method achieved a stable and robust performance after choosing the top 100 trait-relevant  
205 genes. In contrast, the use of scDRS with MAGMA resulted in a variable and moderate performance  
206 that was largely influenced by the number of included genes prioritized by MAGMA or the three other  
207 eQTL-based methods (Supplementary Figure S7A–B). Additionally, when applying genes prioritized  
208 by scPagwas to two other cell-scoring methods, e.g., Vision<sup>29</sup> and AUCCell<sup>27</sup>, we consistently found that  
209 these cell-scoring methods yielded a high performance in identifying cells relevant to monocyte count  
210 with either the simulated or real RNA-seq data (Supplementary Figure S3C-D). Taken together, our  
211 results reveal that scPagwas enables trait relevance to be accurately and robustly characterized at the  
212 single-cell resolution.

213

### 214 ***scPagwas accurately identifies blood cell trait-relevant cell populations at distinct stages of human*** 215 ***hematopoiesis***

216 scPagwas was used to identify hematological trait-relevant cell populations in a large BMMC scRNA-  
217 seq dataset (n = 35,582 cells, Figure 4A) that contained the full spectrum of human hematopoietic  
218 differentiation from stem cells to their progeny<sup>39</sup>. To explore the genetic associations for 10 highly-  
219 heritable blood cell traits in various cellular contexts, we aggregated the TRSs of individual cells within  
220 the same annotated cell type to assess the enrichments of hematopoietic traits at distinct stages of  
221 human hematopoiesis using the unsupervised clustering method (Supplementary Table S1). According  
222 to the aggregation results, different cell populations from the same lineage were predisposed to have  
223 consistent associations across relevant traits (Figure 4B and Supplementary Figure S8A). For example,  
224 red blood cell traits, including hemoglobin concentration, mean corpuscular hemoglobin, and mean  
225 corpus volume, tended to have similar associations within the same module based on the TRS of the  
226 cell type, consistent with previous findings<sup>22</sup>.

227

228 The TRSs of cells for three representative traits are shown in low-dimensional t-distributed stochastic  
229 neighbor embedding (t-SNE) space (Figure 4C-E). Remarkably, cell lineages relevant to corresponding  
230 blood cell traits yielded considerably high TRSs under different conditions (Figure 4C–E and  
231 Supplementary Figure S8B-D), indicating that the cell specificity of these genetic effects was well



232 captured by scPagwas. For monocyte count, scPagwas identified not only monocyte-related cell  
233 compartments with increased TRSs but also granulocyte-monocyte progenitor cells showing increased  
234 enrichment (Figure 4C). Furthermore, several cell compartments related to CD8+ T cells, CD4+ T  
235 cells, NK cells, and B cells yielded increased TRSs for lymphocyte count (Figure 4D), and early and  
236 late erythrocytes, CMP-LMMP, and hematopoietic stem cells exhibited increased TRSs for the mean  
237 corpus volume (Figure 4E).

238

239 When applying the top 1,000 scPagwas-prioritized genes to scDRS in the BMMC scRNA-seq dataset,  
240 cells relevant to three representative traits were enriched and had increased TRSs (Supplementary  
241 Figure S9A–C, left panel). However, the use of scDRS with the top 1,000 MAGMA-prioritized genes  
242 did not show such trait-relevant enrichment (Supplementary Figure S9A–C, right panel). Using an  
243 independent peripheral blood mononuclear cell (PBMC) scRNA-seq dataset with a larger number of  
244 cells ( $n = 97,039$  cells)<sup>40</sup>, consistently, scPagwas using either cell scoring method, Seurat or scDRS,  
245 accurately distinguished monocyte and lymphocyte count-relevant cell compartments, whereas there  
246 was no specific trait relevance using scDRS with the top MAGMA-prioritized genes (Supplementary  
247 Figures S10–S11). Collectively, these results suggest that scPagwas can recapitulate known  
248 associations between blood cell traits and the cellular context and identify novel trait-associated cell  
249 subpopulations and states.

250

### 251 ***scPagwas identifies novel immune subpopulations associated with severe COVID-19 risk***

252 Understanding the effects of host genetic factors on immune responses to severe infection can  
253 contribute to the development of effective vaccines and therapeutics to control the COVID-19  
254 pandemic. scPagwas was applied to discern COVID-19-associated immune cell types/subpopulations  
255 by integrating a large-scale GWAS summary dataset on severe COVID-19 ( $N = 7,885$  cases and  
256  $961,804$  controls) with a large PBMC scRNA-seq dataset ( $n = 469,453$  cells) containing healthy  
257 controls and COVID-19 patients with various clinical severities (Supplementary Tables S1–S2).  
258 scPagwas identified that three immune cell types, including naïve CD8+ T cells ( $P = 4.6 \times 10^{-17}$ ),  
259 megakaryocytes ( $P = 7.8 \times 10^{-6}$ ), and CD16+ monocytes ( $P = 1.14 \times 10^{-4}$ ), demonstrated significant  
260 associations with severe COVID-19 ( $FDR < 0.05$ , Figure 5A–D and Supplementary Table S4), whereas

261 these three cell types only showed suggestive associations inferred by the three cell type-level  
262 inference methods (LDSC-SEG<sup>11</sup>, MAGMA-based approach<sup>41</sup>, and RolyPoly<sup>13</sup>) (see Supplementary  
263 Methods). Both CD16<sup>+</sup> monocytes and megakaryocytes have been reported to be associated with  
264 aggressive cytokine storm among severe COVID-19 patients<sup>42,43</sup>.

265  
266 Of note, scPagwas identified a novel cell subpopulation of naïve CD8<sup>+</sup> T cells related to severe  
267 COVID-19 (Figure 5A–E). scPagwas identified that five biological pathways relevant to COVID-19  
268 severities showed high specificity for naïve CD8<sup>+</sup> T cells and included the prolactin signaling pathway,  
269 thyroid hormone signaling pathway, and type I diabetes mellitus (FDR < 0.05, Supplementary Figure  
270 S12), which have been reported to potentially play crucial roles in COVID-19<sup>44-46</sup>. Recent single-cell  
271 sequencing studies<sup>47-49</sup> have demonstrated that naïve CD8<sup>+</sup> T cells show prominent associations with  
272 COVID-19 severity. Moreover, naïve CD8<sup>+</sup>T cells are essential for recognizing newly-invaded viral  
273 antigens including SARS-CoV-2, leading to initiation of the adaptive immune response by  
274 differentiating naïve T cells into subpopulations of cytotoxic effector CD8<sup>+</sup> T cells or memory CD8<sup>+</sup>  
275 T cells<sup>48,50,51</sup>.

276  
277 As shown in Figure 5E, the naïve CD8<sup>+</sup>T cells were grouped into four clusters. We found that trait-  
278 relevant cells with high scPagwas TRSs were mainly in clusters 0 and 1 (Figure 5F and Supplementary  
279 Figure S13). Of note, cluster 0 showed high expression of memory effector marker genes (*GZMK*,  
280 *AQP3*, *GZMA*, *PRF1*, and *GNLY*), while cluster 1 demonstrated high expression of exhaustive effector  
281 marker genes (*LAG3*, *TIGIT*, *GZMA*, *GZMB*, *PRDMI*, and *IFNG*) (Supplementary Figure S14).  
282 Further analysis showed that the molecular signature scores of the effector marker genes across cells  
283 were significantly positively correlated with the TRSs (Figure 5G and Supplementary Table S5),  
284 indicating that severe COVID-19-associated T cells tend to activate effector signatures involved in the  
285 anti-viral immune response. These new cell subpopulations may play important roles in modulating  
286 the immune response in severe COVID-19 patients.

### 288 *scPagwas distinguishes heterogeneous cell populations associated with AD*

289 AD is a detrimental neurodegenerative disease that causes a gradual increase in neuronal death and

290 loss of cognitive function. scPagwas was applied to uncover cell subpopulations associated with AD  
291 by integrating a human brain entorhinal cortex single-nucleus RNA-seq (snRNA-seq) dataset  
292 containing five brain cell types ( $n = 11,786$  cells, Figure 6A and Supplementary Table S2) with an AD  
293 GWAS summary dataset ( $N = 21,982$  cases and  $41,944$  controls, Supplementary Table S1). We found  
294 that OPCs and microglia with higher TRSs showed stonger enrichments in AD (Figure 6B).  
295 Consistently, at the cell type level, both OPCs and microglia were significantly associated with AD  
296 ( $FDR < 0.05$ , Figure 6C and Supplementary Table S6). For independent validation, three large single-  
297 cell datasets (Supplementary Table S1), including two human brain snRNA-seq datasets ( $n = 101,906$   
298 and  $14,287$  cells) and one mouse brain scRNA-seq dataset ( $n = 160,796$  cells), were used for scPagwas  
299 analysis. These results also indicated that OPCs and microglia were significantly associated with AD  
300 ( $P < 0.05$ , Supplementary Table S7).

301

302 Remarkably heterogeneous associations between OPCs and AD were detected by scPagwas  
303 (heterogeneous  $FDR = 3.33 \times 10^{-4}$ , Supplementary Figure S15), which is consistent with the recent  
304 finding of functionally diverse states of OPCs<sup>52</sup>. Disruption of OPCs is related to accelerated myelin  
305 loss and cognitive decline and is considered an early pathological sign of AD<sup>53</sup>. Analogous to our  
306 results, a recent genetic study demonstrated that OPCs exhibit significant associations with  
307 schizophrenia<sup>54</sup>, which was repeated by scPagwas using the same schizophrenia GWAS and scRNA-  
308 seq data as in the Agarwal et al. study (Supplementary Figure S16). Consistently, multiple lines of  
309 genetic evidence have indicated a critical role of microglia in the pathogenesis of AD<sup>55-58</sup>.

310

311 Moreover, the top significant trait-relevant pathways of the microglial association were related to  
312 immune pathways, including Th17 cell differentiation and influenza A (Figure 6D). Genes in the  
313 immune pathway of Th17 cell differentiation in disease-associated microglia have been identified as  
314 involved in AD risk<sup>59</sup>. The top-ranked significant pathways for OPC associations were related to brain  
315 development and synaptic transmitters, including glutamatergic synapses, taste transduction, and the  
316 prolactin signaling pathway (Figure 6D). Alteration of glutamatergic synapses has the potential to  
317 inhibit OPC proliferation and may be related to disruption of myelination, which is a prominent feature  
318 of AD<sup>60</sup>. These results suggest that these trait-relevant cell types could contribute to AD risk via distinct

319 biological pathways.

320

321 We further identified the 1,000 top-ranked trait-relevant genes for AD by computing the correlation  
322 between the expression of a given gene and the summed gPASs of each cell across all 11,786 brain  
323 cells (see Methods and Figure 6E). To assess the association between these prioritized genes and AD,  
324 we adopted the RISmed method<sup>61</sup>, which searches for supporting evidence from reported studies in  
325 the PubMed database. A significant positive correlation was observed between the scPagwas results  
326 and PubMed search results ( $r = 0.23$ ,  $P = 2.17 \times 10^{-13}$ , Figure 6F), which was notably higher than that  
327 from matched random gene sets (permuted  $P < 0.01$ , Supplementary Figure S17). These risk genes  
328 were significantly enriched in several important functional cellular components related to  
329 neurodegenerative diseases, including postsynaptic specialization and neuron-to-neuron synapses  
330 (FDR  $< 0.05$ , Supplementary Figure S18 and Table S8). To further evaluate the phenotypic associations  
331 of these top 1,000 scPagwas-identified risk genes, we leveraged two large and independent bulk-based  
332 expression profiles on AD patients ( $N = 222$ ) and matched controls ( $N = 219$ ). We found that 42.1%  
333 (421/1,000) of these genes were significantly up-regulated in AD patients (two-sided T test  $P < 0.05$ ,  
334 Supplementary Table S9). Of note, this proportion was significantly higher than that of randomly  
335 selected length-matched genes (permuted  $P = 0.01$ , Supplementary Figure S19).

336

337 The highest-ranked genes, *CREB1* ( $P = 1.48 \times 10^{-18}$  and  $2.34 \times 10^{-5}$ ) and *GSK3B* ( $P = 5.3 \times 10^{-7}$  and 0.043),  
338 exhibited significantly higher expression in AD patients than in controls in both datasets (Figure 6G–  
339 J). Genetic variants in *CREB1* have been associated with brain-related phenotypes, including  
340 neuroticism<sup>62</sup>, major depressive disorder<sup>63</sup>, and cognitive performance<sup>64</sup>. Inhibition of *GSK3B*  
341 expression decreases microglial migration, inflammation, and inflammation-associated neurotoxicity<sup>65</sup>.  
342 In addition, activation of the kinase *GSK3B* promotes TAU phosphorylation, which corresponds to  
343 amyloid- $\beta$  (A $\beta$ ) accumulation and A $\beta$ -mediated neuronal death<sup>66</sup>. In summary, scPagwas not only  
344 identified subpopulations of microglia and OPC relevant to AD but also uncovered the key AD-  
345 associated pathways and risk genes.

346

347 **Discussion**

348 Here, we introduce scPagwas, a pathway-based polygenic regression method that incorporates GWAS  
349 summary statistics and scRNA-seq data to identify trait-relevant individual cells. scPagwas exhibits  
350 well-calibrated and powerful performance benchmarked with extensive simulated and real datasets.  
351 scPagwas can capture the essential trait-relevant features of single-cell data and provide previously  
352 unrecognized functional insights by linking trait-relevant genetic signals to the cellular context. It  
353 should be noted that scPagwas does not require parameter tuning for cell type annotations and  
354 significantly enhances the discovery of trait-relevant enrichment at the single-cell resolution compared  
355 to existing methods<sup>11-15</sup>. scPagwas is suitable for analyzing genetic enrichment of rare or previously  
356 unknown cell populations in large-scale single-cell datasets<sup>67</sup>.

357  
358 High sparsity and technical noise are the principal issues in analyzing single-cell sequencing data<sup>22,24</sup>.  
359 The activity of individual genes cannot represent cell functionality because it highly relies on the  
360 activity of other partner genes in a given pathway<sup>23</sup>. Additionally, the combination of biological  
361 functions of different genes in the same pathway has been reported to reduce inflated zero counts and  
362 technical noise<sup>21,24-26,28</sup>. Furthermore, disease-associated genetic variants that are mainly involved in  
363 systems of highly communicating genes and even variants with weak associations grouped in a given  
364 biological pathway could play important roles in uncovering the genetic mechanisms of complex  
365 diseases or traits<sup>30,31</sup>. Leveraging these pathway-based advantages, scPagwas could reduce the high  
366 sparsity and technical noise across millions of scRNA-seq profiles from different tissues and organs  
367 from mice and humans. Crucially, scPagwas not only recapitulated well-established cell type-disease  
368 associations, including the associations of immune cell types with hematological traits and microglia  
369 with AD, but also detected notable enrichments that have not been reported in previous studies and are  
370 biologically plausible, supporting the powerful potential of scPagwas for the discovery of novel  
371 mechanisms.

372  
373 Based on the correlation between genetically-influenced pathway activity and gene expression,  
374 scPagwas prioritized more trait-relevant genes than other widely used gene-scoring methods, including  
375 MAGMA<sup>20</sup>, S-PrediXcan<sup>19</sup>, S-MultiXcan<sup>18</sup>, and TWAS<sup>17</sup> (Figure 2). Although the use of scDRS with  
376 the MAGMA-identified genes was more powerful than with the genes identified by the other three

377 methods (i.e., S-PrediXcan, S-MultiXcan, and TWAS), scPagwas yielded the best performance in  
378 distinguishing trait-relevant cells. When the top 1,000 scPagwas-identified trait-relevant genes were  
379 applied to scDRS, the precision for distinguishing trait-relevant cells was significantly enhanced,  
380 indicating that it is important to prioritize a group of robust trait-relevant genes for scoring cells. This  
381 explains why previous cell-scoring methods<sup>16,27,29</sup> based on the top-ranked MAGMA genes only  
382 achieved moderate performance. Moreover, scPagwas could discern trait-relevant cell populations as  
383 well as early progenitor cells for blood cell traits, which is consistent with the fact that pathway-based  
384 scoring methods are useful in determining disease-associated but heterogeneous early developmental  
385 progenitor cells<sup>28,68,69</sup>.

386  
387 Several limitations of this study should be noted. First, identification of a statistical association  
388 between complex diseases/traits and individual cells does not imply causality but may reflect indirect  
389 identification of causal associations, parallel to previous methods<sup>11,13,14,16</sup>. Nevertheless, even under  
390 such circumstances, the scPagwas-identified trait-relevant cells are inclined to be biologically relevant  
391 to the causal cells because of their similar genetic co-expression patterns. Second, for the current study,  
392 we selected canonical pathways identified in the Kyoto Encyclopedia of Genes and Genomes (KEGG)  
393 database<sup>70</sup> because these pathways have been experimentally validated. Third, to be compatible with  
394 Seurat software<sup>37</sup>, the most extensively used tool for scRNA-seq data analysis, scPagwas by default  
395 employed the cell-scoring method of the *AddModuleScore* function of Seurat to directly compute the  
396 TRS, which makes it convenient to integrate scPagwas into existing scRNA-seq analysis pipelines.  
397 According to our current results, other state-of-the-art cell-scoring methods, including the scDRS<sup>16</sup>,  
398 AUCCell<sup>27</sup>, and Vision<sup>29</sup>, also showed a good and robust performance for distinguishing trait-relevant  
399 cells when applying scPagwas-identified trait-relevant genes. Finally, we annotated SNPs into genes  
400 and their corresponding pathways based on the proximal distance of a 20 kb window. It may be possible  
401 to establish the link between SNPs and genes using other methods, such as functionally-informed SNP-  
402 to-gene linking approaches<sup>71</sup>, in the future.

403  
404 In conclusion, scPagwas demonstrates promise for uncovering significant trait-relevant individual  
405 cells. Our pathway-based inference strategy will increase the identification of key cell subpopulations

406 with reasonable biological interpretation for traits of interest. From a discovery viewpoint, the  
407 identification of reproducible trait-relevant individual cells will help to achieve the first step toward  
408 an in-depth experimental investigation of novel cell types or states with potential physiological roles  
409 in health and disease.

410

## 411 **Figure titles and legends**

412 **Figure 1. Overview of scPagwas approach.** A. Linking single-nucleotide polymorphisms (SNPs)  
413 from GWAS summary statistics into corresponding pathway. The linkage disequilibrium (LD) matrix  
414 for SNPs is calculated based on the 1,000 Genomes Project Phase 3 Panel. B. Statistical model. The  
415 pathway-specific polygenic regression analysis between the SNP effect sizes and the adjusted gene  
416 expression within a given pathway is used to infer an estimated coefficient  $\tau$  for each pathway. C.  
417 Transforming gene-by-cell matrix to pathway activity score (PAS)-by-cell matrix via using the singular  
418 value decomposition (SVD) method. The first principal component (PC1) represents the PAS for each  
419 pathway. D. The Pearson correlation model. The genetically-associated PAS (gPAS) for each pathway  
420 is defined as the product between the estimated coefficient  $\tau$  and weighted PAS (see Methods). The  
421 bottom panel represents the Pearson correlation analysis of the summed gPASs of all pathways in a  
422 given cell with the expression of a given gene for all individual cells, and ranking the Pearson  
423 correlation coefficients (PCCs) to prioritize top trait-relevant genes. Then, scPagwas uses the cell-  
424 scoring method in Seurat to collapse the expression of top  $n$  trait-relevant genes (default top 1,000  
425 genes) for calculating the trait-relevant score (TRS) of each cell. E. scPagwas outputs. The typical  
426 outputs includes (i) trait-relevant cells, (ii) trait-relevant cell types, and (iii) trait-relevant  
427 pathways/genes.

428

429 **Figure 2. The reproducible and functional results of scPagwas.** A. GO-term enrichment analyses  
430 of top-ranked 1,000 genes from scPagwas and other four gene-based methods (i.e., MAGMA, TWAS,  
431 S-PrediXcan, and S-MultiXcan) for 10 highly-heritable blood cell traits. Different color dots represent  
432 number of significant GO-terms of biological processes (BP, FDR < 0.01) enriched by top-ranked  
433 genes from scPagwas and other four methods (see Supplementary Table S3). B. Example of the  
434 distribution of scPagwas-identified risk genes and MAGMA-identified risk genes ranked by their

435 average expression across all cells for the lymphocyte count trait. The percentages of risk genes for  
436 top-half (over-expression genes) and bottom-half (down-expression genes) cells are shown in the plot  
437 accordingly. GO enrichment results of the lymphocyte count trait classified by two groups of over-  
438 expression genes (FDR < 0.01, 35 significant GO-terms enriched by scPagwas vs. 0 GO-terms by  
439 MAGMA) and down-expression genes (FDR < 0.01, 13 significant GO-terms enriched by scPagwas  
440 vs. 2 GO-terms by MAGMA) are shown in the plot. C. Plot demonstrating the variance of gene-level  
441 expression magnitude and pathway-level expression magnitude in the BMMC scRNA-seq dataset (n  
442 = 35,582 cells). Fitted line with red color represents pathway-level expression magnitude, which shows  
443 a mean-variance fit that demonstrates the relationship between average expression of genes in a given  
444 pathway (x axis) and its corresponding variance (y axis). Fitted line with blue color represents gene-  
445 level expression magnitude, which shows a mean-variance fit that demonstrates the relationship  
446 between average gene expression (x axis) and gene variance (y axis). The black dots in the plot indicate  
447 outliers. See also Supplementary Figure S2.

448

449 **Figure 3. Assessment of the performance of scPagwas in both simulated and real scRNA-seq**  
450 **datasets.** A. UMAP embedding plot shows the cellular component of a synthesized ground truth  
451 scRNA-seq dataset (monocytes: n = 1,000 cells, and T, B, DC, and NK: n = 1,000 cells in total). E.  
452 UMAP plot shows the cellular component of a real ground truth scRNA-seq dataset. The real BMMC  
453 scRNA-seq dataset contains 10,000 cells with seven cell types: monocytes (11\_CD14.Mono.1,  
454 12\_CD14.Mono.2, and 13\_CD16.Mono, n = 5,000 cells), DC (09\_pDC and 10\_cDC, n = 200 cells),  
455 T cells (19\_CD8.N and 20\_CD4.N1, n = 3,000 cells), B cells (17\_B, n = 1,000 cells), and NK cells  
456 (25\_NK, n = 800 cells). B, C, F, G) Illustration of the performance of top 1,000 scPagwas-identified  
457 genes for identifying monocyte count trait-relevant cells based on two cell-scoring methods (i.e.,  
458 Seurat and scDRS) in the synthesized (B, C) and real (F, G) scRNA-seq datasets. D, H). Illustration of  
459 the performance of top-ranked 1,000 putative disease genes identified by the gene-scoring method of  
460 MAGMA for identifying monocyte count trait-relevant cells based on the scDRS method in the  
461 synthesized (D) and real (H) scRNA-seq datasets. The UMAP projections of every cell colored by its  
462 TRS. The vertical bar exhibits cells descendingly ranked according to their corresponding TRSs (top-  
463 ranked 1,000 genes), where red color indicates monocyte cells and blue color indicates non-monocyte



464 cells. The accuracy of each method represents the percentage of monocyte count trait-related cells (i.e.,  
465 monocytes) for top-half cells that are ranked by TRS for all cells in a descending manner. See also  
466 Supplementary Figures S3-S4.

467

468 **Figure 4. Application of scPagwas to multiple blood cell traits for identifying trait-relevant cells.**

469 The 10 hematological traits were analyzed using scPagwas (Seurat) on a large BMMC scRNA-seq  
470 dataset. A. The tSNE plot shows the cell type labels. B. The average TRSs for cells belonging to the  
471 same cell type are shown in the heatmap. Unsupervised clustering analysis was conducted and cell-  
472 type category were grouped into six main clusters, including DCs, B cells, Monocytes (Mono), NKs,  
473 T cells, and early/progenitor cells. cDC, classical dendritic cell; pDC, plasmacytoid dendritic cell; Unk,  
474 Unknown; CD4.M, CD4+ memory T cells; CD8.CM, CD8+central memory T cells; CD8.EM,  
475 CD8+effector memory T cells; CD8.N, CD8+ naïve T cells; CD4.N1/N2, CD4+naïve T cells; CLP,  
476 common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte-macrophage  
477 progenitor; LMPP, lymphoid-primed multipotent progenitor; HSC, Hematopoietic stem cell; Baso,  
478 basophil; Eryth, erythrocyte; Neut, neutrophil. C-E) Per-cell TRS calculated by scPagwas (Seurat) for  
479 three representative traits including monocyte count (C), lymphocyte count (D), and mean corpus  
480 volume (E) are shown in tSNE coordinates (left) and per cell type (right). Boxplots (left to right: n =  
481 1,425, 2,260, 903, 377, 2,097, 1,653, 446, 111, 1,050, 544, 325, 1,800, 4,222, 292, 420, 710, 1,711, 62,  
482 1521, 2,470, 2,364, 3,539, 796, 2,080, 2,143, and 161 cells) show the median with interquartile range  
483 (IQR) (25-75%); whiskers extend 1.5× the IQR. See also Supplementry Figure S9.

484

485 **Figure 5. scPagwas identifies trait-relevant immune cell types and subpopulations for severe**

486 **COVID-19.** A-D. Benchmarking analysis of uncovering trait-relevant cell types by using scPagwas,  
487 LDSC-SEG, MAGMA-based approach, and RolyPoly for COVID-19 patients with various clinical  
488 severities of severe (A), moderate (B), mild (C), and healthy controls (D), respectively. The horizontal  
489 dashed red lines represent the significant threshold (Bonferroni corrected  $P < 0.05$ ), and the horizontal  
490 dashed blue lines indicate the raw significant threshold (i.e., raw  $P < 0.05$ ). E. The tSNE visualization  
491 of 766 naïve CD8+ T cells with four cell clusters. F. scPagwas TRS for the phenotype of severe  
492 COVID-19 risk is displayed for all naïve CD8+T cells in the tSNE plot. G. The correlation of scPagwas

493 TRSs with molecular signature scores of effector marker genes across all naïve CD8+T cells. See also  
494 Supplementary Figures S12-S14.

495

496 **Figure 6. scPagwas discerns human brain cell types and subpopulations in association with AD.**

497 A. The UMAP plot of scRNA-seq profiles of 11,786 human brain cells containing five brain cell types.

498 Cells are colored by the cell type annotation. B. scPagwas TRS for the phenotype of AD risk is

499 displayed for all cells in the UMAP plot. OPC and microglial cells are highlighted with dashed lines.

500 C. Benchmarking analysis of uncovering significant AD-associated cell types by using scPagwas,

501 LDSC-SEG, MAGMA-based approach, and RolyPoly. The horizontal dashed red line represents the

502 significant threshold ( $P < 0.05$ ). D. Dot plot demonstrating the trait-relevant pathways across five brain

503 cell types identified by scPagwas. Dot size represents the log-ranked P value for each pathway, and

504 color intensity indicates the proportion of cells within each cell type genetically influenced by a given

505 pathway (pathway-level coefficient  $\beta > 0$ ). E. Trait-relevant genes ranked by the Pearson correlation

506 coefficients (PCCs) using scPagwas across all individual cells. F. Correlations of trait-relevant genes

507 for AD ranked from scPagwas results and PubMed search results. The Pearson correlation is calculated

508 between scPagwas results and PubMed search results ( $\log_2(n+1)$ ). The top-ranking trait-relevant genes

509 are labelled. G-H. Violin plots show the differential gene expression (DGE) analyses of *GSK3B* and

510 *CREB1* between AD patients and controls in two independent bulk-based expression profiles. The two

511 bulk transcriptomic datasets (GSE109887 with 78 samples, and GSE15222 with 363 samples) contain

512 222 AD patients and 219 controls. The two-sided Student's T test was used for assessing the statistical

513 significance. See also Supplementary Table S9.

514

## 515 **Resource Availability**

### 516 **Lead contact**

517 Further information and requests for resources and reagents should be directed to and will be fulfilled

518 by the Lead Contact, Jianzhong Su ([sujz@wmu.edu.cn](mailto:sujz@wmu.edu.cn)).

519

### 520 **Materials availability**

521 This study did not generate new unique reagents

522

## 523 **Data and Code Availability Statements**

524 All GWAS summary datasets were downloaded from three publicly accessible databases of the IEU  
525 open GWAS project (<https://gwas.mrcieu.ac.uk/>), the COVID-19 Host Genetics Initiative  
526 ([www.covid19hg.org/results](http://www.covid19hg.org/results)), the Psychiatric Genomics Consortium website (<https://pgc.unc.edu/>),  
527 and the NHGRI-EBI GWAS Catalog (<https://www.ebi.ac.uk/gwas/>). The healthy BMMC scRNA-seq  
528 dataset was downloaded from a website ([https://jeffgranja.s3.amazonaws.com/MPAL-  
529 10x/Supplementary\\_Data/Healthy-Data/scRNA-Healthy-Hematopoiesis-191120.rds](https://jeffgranja.s3.amazonaws.com/MPAL-10x/Supplementary_Data/Healthy-Data/scRNA-Healthy-Hematopoiesis-191120.rds)). The healthy  
530 PBMC scRNA-seq dataset used to validate scPagwas performance was downloaded from the  
531 ArrayExpress database (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10026/>). The  
532 mouse brain scRNA-seq dataset was downloaded from the Mouse Brain Atlas  
533 ([https://storage.googleapis.com/linnarsson-lab-loom/15\\_all.loom](https://storage.googleapis.com/linnarsson-lab-loom/15_all.loom)). Human brain snRNA-seq dataset  
534 #1 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138852>), Human brain snRNA-seq  
535 dataset #2 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160936>), Human brain snRNA-  
536 seq dataset #3 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140231>), and two bulk-  
537 based transcriptomic profiles on AD (1. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109887>;  
538 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15222>) were downloaded from the GEO  
539 database. The PBMC scRNA-seq dataset on COVID-19 severity was downloaded from the  
540 ArrayExpress database (<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9357>). The  
541 scRNA-seq dataset on the human cell landscape (HCL) was downloaded from the GEO database  
542 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134355>). scPagwas is implemented as an  
543 R package and is available on GitHub (<https://github.com/dengchunyu/scPagwas>). The code to  
544 reproduce the results is available in a dedicated GitHub repository  
545 ([https://github.com/dengchunyu/scPagwas\\_reproduce](https://github.com/dengchunyu/scPagwas_reproduce)).  
546

547

## 548 **Method details**

### 549 *scPagwas methodology*

550 The workflow of the scPagwas method is shown in Figure 1. In brief, scPagwas employs

551 an optimized polygenic regression model to identify the associations of a subset of cells  
552 with a complex disease or trait of interest. The framework of the method is described in  
553 detail in the following steps.

554

### 555 ***Linking SNPs to their corresponding pathways***

556 Based on previous evidence<sup>6</sup> indicating that most eQTLs consistently lie in a 20-kb window  
557 centered on the transcription start site of a gene, a window size of 20 kb is adopted as the  
558 default parameter of scPagwas to assign SNPs from GWAS summary statistics to associated  
559 genes. We use the notation  $g(k)$  to represent a gene  $g$  with an SNP  $k$ . With the  
560 assignment of SNPs to corresponding gene, there are a few SNPs with multiple associated  
561 genes. We duplicate these SNPs and consider them as independent SNP-gene pairs  
562 following an earlier study<sup>13</sup>. In our data applications, SNPs with minor allele frequencies  
563 smaller than 0.01 or on the sex chromosomes (ChrX-Y) were removed.

564

565 Based on pathways in the KEGG database<sup>70</sup>, we annotate these SNPs to associated gene  $g$   
566 in corresponding pathway, and use the notation  $S_i = \{k : g(k) \in P_i\}$  to indicate the set of  
567 SNPs within the pathway  $i$ . The notation  $P_i$  indicates the set of genes in the pathway  $i$ .  
568 scPagwas provides other functional gene sets, such as Reactome<sup>72</sup> and MSigDB<sup>73</sup>, as  
569 alternative options. In our data applications, the 1,000 Genomes Project Phase 3 Panel<sup>74</sup>  
570 was applied to calculate the linkage disequilibrium (LD) among SNPs available in the  
571 GWAS summary statistics, and the major histocompatibility complex region (Chr6: 25–35  
572 Mbp)<sup>75</sup> was removed because of the extensive LD in this region.

573

### 574 ***PAS matrix transformation***

575 scPagwas uses the variance-stabilizing transformation method<sup>76</sup> with a scale factor of  
576 10,000 to normalize a sparse gene-by-cell matrix from scRNA-seq data as follows:

577  $e_{g,j} = \log(a_{g,j} \cdot 1e4 / \sum_g a_{g,j})$ , where  $a_{g,j}$  is the raw expression for gene  $g$  in cell  $j$  and  $e_{g,j}$  is

578 the normalized expression of gene  $g$  in cell  $j$ . Pathways such as those from the KEGG  
579 database<sup>70</sup> can be used as a gene set to calculate PASs. The SVD method can greatly improve  
580 the computational efficiency<sup>22,77</sup> of analyzing a sparse matrix with high dimensionality and  
581 can be used to generate eigenvalues without calculating the covariance matrix. We apply  
582 the SVD method to transform a normalized gene-by-cell matrix into a pathway-by-cell  
583 matrix with reduced dimensional space.

584

585 For each pathway  $i$ , we extract a  $N \times M_i$  sub-matrix  $A_i$  from the normalized single-cell  
586 matrix  $A$ , with  $N$  being the number of cells and  $M_i$  being the number of genes in  
587 pathway  $i$ . Applying the SVD method,  $A_i$  can be decomposed as follows:

588 
$$A_i^T = U \Sigma V^T$$

589 where  $U$  is an  $N \times N$  orthogonal matrix,  $\Sigma$  is a diagonal matrix with all zeroes except  
590 for the elements on the main diagonal, and  $V^T$  is an  $M_i \times M_i$  orthogonal matrix. For the  
591 right orthogonal matrix  $V = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{M_i})$ , the  $t$ th column vector  $\mathbf{v}_t$  represents the  $t$ th  
592 principal component. In reference to previous studies<sup>28,35</sup>, we use the projection of the  
593 characteristics of genes in each pathway on the direction of the PC1 eigenvalue to define  
594 PAS  $s_{i,j}$  for the pathway  $i$  in cell  $j$ , which reflects the main coordinated expression  
595 variability of genes in a given pathway among single-cell data.

596

### 597 ***Polygenic regression model***

598 According to a previous method<sup>13,34</sup>, we assume a linear regression model,  $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$ ,  
599 where  $\mathbf{y}$  is an  $n$ -vector of phenotypes,  $\mathbf{X}$  denotes the  $n \times m$  matrix of genotypes  
600 (standardized to mean 0 and variance 1 for each SNP vector),  $\mathbf{b}$  indicates the per-  
601 normalized-genotype effect sizes vector of  $m$  SNPs when fitted jointly, and  $\boldsymbol{\varepsilon}$  is the  
602 stochastic environmental error term. The released GWAS summary dataset contains per-  
603 SNP effect estimates, denoted as  $\hat{\boldsymbol{\beta}}$ . These estimates indicate the marginal regression

604 coefficients from univariate models and can be calculated using the transformation equation

605  $\hat{\beta}_k = \mathbf{X}_k^T \mathbf{y}$ , where  $\mathbf{X}_k^T$  represents standardized genotypes for SNP  $k$  across  $n$  GWAS

606 samples. After substituting the polygenic model  $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$  into the estimation equation

607  $\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$ , the estimated marginal effect sizes of SNPs can be written as:

$$608 \quad \hat{\boldsymbol{\beta}} = \mathbf{R}\mathbf{b} + \mathbf{X}^T \boldsymbol{\varepsilon}$$

609 where  $\mathbf{R}$  denotes the LD matrix.

610

611 As previously mentioned,  $S_i$  denotes an SNP set that contains SNPs mapped to genes in a

612 pathway  $i$ . The polygenic model assumes that the effect sizes of SNPs in pathway  $i$  are

613 random effects, which follow the multi-variable normal distribution

614  $\mathbf{b}_{S_i} \sim \text{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{I}_{|S_i| \times |S_i|})$ , where  $\sigma_i^2$  is the variance of effect sizes for SNPs in the pathway

615 and  $\mathbf{I}$  is the  $|S_i| \times |S_i|$  identity matrix. Based on the prior assumption of above polygenic

616 model, the distribution for the vector of the estimated effects of SNPs ( $\hat{\boldsymbol{\beta}}_{S_i}$ ) associated with

617 a pathway follows:

$$618 \quad \hat{\boldsymbol{\beta}}_{S_i} \sim \text{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{R}_{S_i}^2 + \sigma_e^2 \mathbf{R}_{S_i})$$

619

620 In reference to the extension of stratified LD score regression to continuous annotations<sup>34</sup>,

621 the per-normalized SNP estimates  $\hat{\boldsymbol{\beta}}$  is a mean 0 vector whose variance  $\sigma_i^2$  depends on

622 continuous-valued annotations (in this case, expression levels of genes in a given pathway).

623 Based on the assumption that a positive correlation between genetic associations and gene

624 expression levels in each cell associated with a trait of interest, the variance  $\sigma_i^2$  is modeled

625 using the linear weighted sum method for each SNP  $k$ :

$$626 \quad \sigma_i^2 = \tau_0 + \sum_j \tau_{i,j} \tilde{e}_{g^{(k)},j}^i$$

627 where  $\tau_0$  is an intercept term,  $\tau_{i,j}$  is the coefficient for the pathway  $i$  in cell  $j$ , which  
 628 measures the strength of association between pathway-specific gene expression activity and  
 629 the variance of GWAS effect sizes at the single-cell level, and  $\tilde{e}_{g,j}^i$  is the adjusted gene  
 630 expression for each gene  $g$  in the given pathway  $i$  calculated as  $\tilde{e}_{g,j}^i = s_{i,j} \hat{e}_{g,j}$  with  $s_{i,j}$   
 631 being the PAS of pathway  $i$ . For each gene  $g$  in the pathway  $i$ , the gene expression  $e_{g,j}$   
 632 is rescaled using the min-max rescaling method:

$$633 \quad \hat{e}_{g,j} = \frac{e_{g,j} - \text{MIN}(e_{g,j})}{\text{MAX}(e_{g,j}) - \text{MIN}(e_{g,j})}$$

634 where  $\text{MAX}(e_{g,j})$  denotes the maximum gene expression in pathway  $i$  and  $\text{MIN}(e_{g,j})$   
 635 denotes the minimum gene expression in pathway  $i$ .

636  
 637 To optimize the coefficients for each pathway in cells under the polygenic regression model,  
 638 scPagwas adopts the method-of-moments approach, which can prominently improve the  
 639 computational efficiency and the estimated uniform convergence<sup>13</sup>. Then, the observed and  
 640 expected squared effects of SNPs relevant to each pathway are fitted, and the following  
 641 equation is used to estimate the expected value:

$$642 \quad E(\hat{\beta}_k^2) = \sigma_i^2 (\mathbf{R}_{S_i}^2)_{k,k} + \sigma_e^2$$

643 where  $(\mathbf{R}_{S_i}^2)_{k,k}$  represents the  $k$ th diagonal element of matrix  $\mathbf{R}_{S_i}^2$ . Then, the coefficient  $\tau_{i,j}$   
 644 can be estimated using the following linear regression:

$$645 \quad E(\hat{\beta}_k^2) = (\mathbf{R}_{S_i}^2)_{k,k} (\tau_0 + \sum_j \tau_{i,j} \tilde{e}_{g^{(k),j}}^i) + \sigma_e^2$$

646  
 647 Of note, the estimated coefficient  $\hat{\tau}_{i,j}$  represents the per-SNP contribution of one unit of the  
 648 pathway-specific activity to heritability. We define a gPAS for each pathway  $i$  that is  
 649 calculated by the product between the estimated coefficient  $\hat{\tau}_{i,j}$  and weighted PAS using

650 the following equation:  $\text{gPAS}_{i,j} = \hat{\tau}_{i,j} \sum_{g \in P_i} \frac{\hat{e}_{g,j} s_{i,j}}{M_i}$ , where  $M_i$  is the number of genes in the

651 pathway  $i$ . Essentially, gPAS is a pathway-activity-based prediction of the genetic variance of a  
652 normal distribution of *cis*-GWAS effect sizes for pathway  $i$  (Figure 1D). Note that the larger gPASs  
653 would have larger *cis*-GWAS effects for each cell, and gPASs can be used to rank trait-relevant  
654 pathways (see Supplementary Methods).

655

### 656 ***Identification of trait-relevant genes and individual cells***

657 To optimize genes relevant to complex diseases/traits at single-cell resolution, we determine which  
658 gene  $g$  exhibits expression that is highly correlated with the summed gPASs across individual cells  
659 using the Pearson correlation method. To maximize the power, the expression of each gene  $g$  is  
660 inversely weighted by its gene-specific technical noise level, which is estimated by modeling the mean-  
661 variance relationship across genes in the scRNA-seq data<sup>78</sup>. By arranging the PCCs for all genes in  
662 descending order, we select the top-ranked risk genes as trait-relevant genes (default top 1,000 genes)  
663 according to a previous method<sup>16</sup>.

664

665 Subsequently, we quantify the aggregate expression of predefined trait-relevant genes in each cell to  
666 generate raw TRSs. For a given cell  $j$  and a trait-relevant gene set  $B$ , the cell-level raw TRS,  
667  $TRS_j$ , is defined as the average relative expression of the genes in  $B$ . However, such raw TRSs may  
668 be confounded by cell complexity, as cells with higher complexity would have more genes identified  
669 and consequently tend to have higher TRSs for any given gene set. To properly control for the effect  
670 of cell complexity, we calculate a control cell score with a control gene set  $B^{Ctrl}$ , which is randomly  
671 selected in a manner that maintained a comparable distribution of expression levels to that of the  
672 predefined gene set. The process included two steps: 1) using the average expression levels to group  
673 all analyzed genes into 25 bins of equal size and 2) randomly selecting 100 genes from the same  
674 expression bin for each gene in the predefined gene set. The final TRS is defined as the initial raw TRS  
675 after subtracting its corresponding control cell score:  $TRS_j = \sum_{g \in B} e_{g,j} / |B| - \sum_{g \in B^{Ctrl}} e_{g,j} / |B^{Ctrl}|$ . The

676 *AddModuleScore* cell-scoring method in Seurat<sup>37</sup> is employed to calculate the TRS with default  
677 parameters.

678



679 To further assess whether a cell is significantly associated with the trait of interest, we employ an  
680 approach<sup>36</sup> to determine the statistical significance of individual cells by calculating the ranking  
681 distribution of trait-relevant genes. Initially, the percent ranks of these trait-relevant genes across the  
682 cells yielded  $\mathbf{r}_g = (\frac{r_{g,1}}{N}, \frac{r_{g,2}}{N}, \dots, \frac{r_{g,N}}{N})$ , where  $r_{g,j}$  is the expression rank of gene  $g$  in cell  $j$  and  
683  $N$  is the total number of cells. The percent ranks of genes follow a uniform distribution  $U(0,1)$ .  
684 Under the null hypothesis that there is no relationship between the percent rank of genes, a statistic  
685  $T_j$  for each cell is calculated using the formula

$$686 \quad T_j = \frac{\sum_g r_{g,j} / N - G / 2}{\sqrt{G / 12}}$$

687  
688 In view of a large number of cells is included in the single-cell data, the distribution of  $T_j$  can be  
689 deduced using the central limit theorem<sup>36</sup>:  $T_j \sim N(0,1)$ , where  $G$  is the number of selected trait-  
690 relevant genes. The hypothesis for the significance test is  $H_0: T_j = 0$  vs  $H_1: T_j > 0$ . The P value  
691 for each cell  $j$  can be written as  $p_j = \Pr(T_j \leq t)$ .

692

### 693 ***Inference analysis of trait-relevant cell types***

694 scPagwas can also identify trait-relevant cell types, where the set of cells is treated as a  
695 pseudo-bulk transcriptomic profile and the expression of a gene across cells is averaged  
696 within a given cell type. For the cell type association, the block bootstrap method<sup>79</sup> is used  
697 to estimate the standard error and compute a t-statistic with a corresponding P value for  
698 each cell type. Because the goal of the block bootstrap is to maintain data structures when  
699 sampling from the empirical distribution, we leverage all pathways in the KEGG database<sup>70</sup>  
700 to partition the genome into multiple biologically-meaningful blocks and sample these  
701 pathway-based blocks with replacement. Under default parameters, scPagwas performs 200  
702 block bootstrap iterations for each cell-type association analysis. The optional parameters  
703 are provided for the block bootstrap. Detailed information on scPagwas cell type-level

704 inference analysis can be found in the Supplementary Methods.

705

### 706 ***Simulations***

707 We used scDesign2 (version 1.0.0)<sup>80</sup> to simulate a ground truth scRNA-seq dataset containing five  
708 cell types including monocytes, DCs, and B, NK, and T cells to assess the performance of scPagwas  
709 in identifying monocyte count trait-relevant individual cells. DC, a type of cell differentiated from  
710 monocytes<sup>81</sup>, was chosen as a non-trait-relevant cell type, which could be a confounding factor for  
711 distinguishing monocytes from all simulated cells. In the model-fitting step, we first fitted a  
712 multivariate generative model to a real dataset via the fluorescence-activated cell-sorted bulk  
713 hematopoietic populations downloaded from the GEO database (Accession No. GSE107011)<sup>82</sup>.  
714 Because there were five sorted cell types, we divided the datasets into five subsets according to the  
715 cell types and fitted a cell type-specific model to each subset. In the data-generation step, we generated  
716 a synthetic scRNA-seq dataset from the fitted model to represent trait-relevant cell populations  
717 (monocytes) and non-trait-relevant cell populations (non-monocyte cells including DCs and B, NK,  
718 and T cells) for the monocyte count trait. Finally, we obtained 2,000 cells with synthetic scRNA-seq  
719 data with cell proportions of 0.5 (monocytes), 0.05 (DCs), 0.2 (B cells), 0.05 (NK cells), and 0.2 (T  
720 cells).

721

### 722 ***scRNA-seq datasets***

723 Eight independent scRNA-seq or snRNA-seq datasets spanning 1.4 million human (*Homo sapiens*)  
724 and mouse (*Mus musculus*) cells were used in this study (Supplementary Table S1). For blood cell  
725 traits, we collected two scRNA-seq datasets based on human BMBCs (n = 35,582 cells)<sup>39</sup> and human  
726 PBMCs (PBMC #1, n = 97,039 cells)<sup>40</sup> to identify trait-relevant cell subpopulations or types. For AD,  
727 we collected four single-cell datasets including a mouse brain scRNA-seq dataset (n = 160,796 cells)<sup>83</sup>,  
728 a human brain entorhinal cortex snRNA-seq dataset (Human brain #1, n = 11,786 cells)<sup>84</sup>, a human  
729 brain snRNA-seq dataset (Human brain #2, n = 101,906 cells)<sup>85</sup>, and another human brain snRNA-seq  
730 dataset (Human brain #3, n = 14,287 cells)<sup>54</sup>. To identify severe COVID-19-related immune cell  
731 populations, we collected a large-scale PBMC scRNA-seq dataset (PBMC #2, n = 469,453 cells)  
732 containing 254 peripheral blood samples from patients with various COVID-19 severities (mild N =

733 109 samples, moderate N = 102 samples, and severe N = 50 samples) and 16 healthy controls<sup>86</sup>. The  
734 scRNA-seq dataset from the human cell landscape (HCL, n = 513,707 cells in 35 adult tissues)<sup>87</sup>, as  
735 well as the previously mentioned four scRNA-seq datasets (i.e., BMMC, PBMC #1, Human brain #1,  
736 and Mouse brain), were used to assess the performance of scPagwas in reducing the sparsity and  
737 technical noise.

738

### 739 ***GWAS summary datasets for complex diseases and traits***

740 We obtained GWAS summary statistics for 10 blood cell traits (average N = 307,772) and AD (21,982  
741 cases and 41,944 controls) from the IEU OpenGWAS, for schizophrenia (67,390 cases and 94,015  
742 controls) from the Psychiatric Genomics Consortium, and for severe COVID-19 (7,885 cases and  
743 961,804 controls) from the COVID-19 Host Genetics Initiative (Supplementary Table S2). The 10  
744 blood cell traits included monocyte count, lymphocyte count, lymphocyte percent, mean corpus  
745 volume, neutrophil count, white blood count, eosinophil count, basophil count, mean corpuscular  
746 hemoglobin, and hemoglobin concentration.

747

### 748 **Acknowledgement**

749 We acknowledge funding support from the National Natural Science Foundation of China (32200535  
750 to Y.M.; 61871294 and 82172882 to J.S), the Scientific Research Foundation for Talents of Wenzhou  
751 Medical University (KYQD20201001 to Y.M.), the Science Foundation of Zhejiang Province  
752 (LR19C060001 to J.S), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang  
753 (2021R01013 to J.Y.), Research Program of Westlake Laboratory of Life Sciences and Biomedicine  
754 (202208013 to J.Y.), and Westlake Education Foundation (101566022001 to J.Y.). We thank Dr.  
755 Zhenhui Chen from Wenzhou Medical University for providing helpful suggestions and manuscript  
756 revisions. We also thank all of the authors who have deposited and shared GWAS summary data in  
757 public databases and the authors who publicly released the scRNA-seq and snRNA-seq datasets.

758

### 759 **Author contributions**

760 J.S., J.Y., and Y.M. conceived and designed the study. Y.M., C.D. and Y.Z. developed the method. Y.M.,  
761 C.D., Y.Z, Y.R.Z., F.Q., D.J., G.Z., J.Q., and J.L. managed data collection. Y.M., C.D., Y.Z, Y.R.Z., and

762 J.L. conducted the bioinformatics analysis and data interpretation. Y.M., J.S., C.D., Y.Z, J.Y., and Y.Z.  
763 wrote the manuscript. All authors reviewed and approved the final manuscript.

764

## 765 **Declaration of interests**

766 The authors declare no competing interests.

767

## 768 **References:**

- 769 1. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O.,  
770 and O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* *562*,  
771 203-209.
- 772 2. Ma, Y., Huang, Y., Zhao, S., Yao, Y., Zhang, Y., Qu, J., Wu, N., and Su, J. (2021). Integrative genomics analysis  
773 reveals a 21q22. 11 locus contributing risk to COVID-19. *Human molecular genetics* *30*, 1247-1258.
- 774 3. Graham, S.E., Clarke, S.L., Wu, K.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S.,  
775 Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids.  
776 *Nature* *600*, 675-679. [10.1038/s41586-021-04064-3](https://doi.org/10.1038/s41586-021-04064-3).
- 777 4. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.Y.,  
778 Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in  
779 schizophrenia. *Nature* *604*, 502-508. [10.1038/s41586-022-04434-5](https://doi.org/10.1038/s41586-022-04434-5).
- 780 5. Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will  
781 revolutionize whole-organism science. *Nat Rev Genet* *14*, 618-630. [10.1038/nrg3542](https://doi.org/10.1038/nrg3542).
- 782 6. Hekselman, I., and Yeger-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in heritable traits  
783 and diseases. *Nat Rev Genet* *21*, 137-150. [10.1038/s41576-019-0200-9](https://doi.org/10.1038/s41576-019-0200-9).
- 784 7. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurler, M.E., Kathiresan, S., Kenny, E.E.,  
785 Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. *Nature* *577*, 179-  
786 189. [10.1038/s41586-019-1879-7](https://doi.org/10.1038/s41586-019-1879-7).
- 787 8. Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci  
788 with gene-expression data identifies specific pathogenic immune cell subsets. *Am J Hum Genet* *89*, 496-  
789 506. [10.1016/j.ajhg.2011.09.002](https://doi.org/10.1016/j.ajhg.2011.09.002).
- 790 9. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson,  
791 S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene  
792 functions. *Nat Commun* *6*, 5890. [10.1038/ncomms6890](https://doi.org/10.1038/ncomms6890).
- 793 10. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich,  
794 M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*  
795 *518*, 197-206. [10.1038/nature14177](https://doi.org/10.1038/nature14177).
- 796 11. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C.,  
797 Shores, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant  
798 tissues and cell types. *Nat Genet* *50*, 621-629. [10.1038/s41588-018-0081-4](https://doi.org/10.1038/s41588-018-0081-4).
- 799 12. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation  
800 of genetic associations with FUMA. *Nat Commun* *8*, 1826. [10.1038/s41467-017-01261-5](https://doi.org/10.1038/s41467-017-01261-5).
- 801 13. Calderon, D., Bhaskar, A., Knowles, D.A., Golan, D., Raj, T., Fu, A.Q., and Pritchard, J.K. (2017). Inferring  
802 Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am J Hum Genet* *101*, 686-

- 803 699. [10.1016/j.ajhg.2017.09.009](https://doi.org/10.1016/j.ajhg.2017.09.009).
- 804 14. Watanabe, K., Umičević Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P., and Posthuma, D. (2019). Genetic  
805 mapping of cell type specificity for complex traits. *Nat Commun* *10*, 3222. [10.1038/s41467-019-11181-1](https://doi.org/10.1038/s41467-019-11181-1).
- 806 15. Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D.,  
807 Miller, J.A., Muñoz-Manchado, A.B., et al. (2018). Genetic identification of brain cell types underlying  
808 schizophrenia. *Nat Genet* *50*, 825-833. [10.1038/s41588-018-0129-5](https://doi.org/10.1038/s41588-018-0129-5).
- 809 16. Zhang, M.J., Hou, K., Dey, K.K., Sakaue, S., Jagadeesh, K.A., Weinand, K., Taychameekiatchai, A., Rao, P.,  
810 Pisco, A.O., Zou, J., et al. (2022). Polygenic enrichment distinguishes disease associations of individual cells  
811 in single-cell RNA-seq data. *Nat Genet*. [10.1038/s41588-022-01167-z](https://doi.org/10.1038/s41588-022-01167-z).
- 812 17. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., and  
813 Wright, F.A. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature*  
814 *genetics* *48*, 245-252.
- 815 18. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating predicted  
816 transcriptome from multiple tissues improves association detection. *PLoS genetics* *15*, e1007889.
- 817 19. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah,  
818 K.P., Garcia, T., and Edwards, T.L. (2018). Exploring the phenotypic consequences of tissue specific gene  
819 expression variation inferred from GWAS summary statistics. *Nature communications* *9*, 1-20.
- 820 20. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis  
821 of GWAS data. *PLoS Comput Biol* *11*, e1004219. [10.1371/journal.pcbi.1004219](https://doi.org/10.1371/journal.pcbi.1004219).
- 822 21. Frost, H.R. (2020). Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene  
823 set scoring. *Nucleic Acids Res* *48*, e94. [10.1093/nar/gkaa582](https://doi.org/10.1093/nar/gkaa582).
- 824 22. Yu, F., Cato, L.D., Weng, C., Liggett, L.A., Jeon, S., Xu, K., Chiang, C.W.K., Wiemels, J.L., Weissman, J.S., de  
825 Smith, A.J., and Sankaran, V.G. (2022). Variant to function mapping at single-cell resolution through  
826 network propagation. *Nature Biotechnology*. [10.1038/s41587-022-01341-y](https://doi.org/10.1038/s41587-022-01341-y).
- 827 23. Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* *461*,  
828 218-223. [10.1038/nature08454](https://doi.org/10.1038/nature08454).
- 829 24. Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O.,  
830 Lauffenburger, D.A., Heyn, H., et al. (2020). Robustness and applicability of transcription factor and pathway  
831 analysis tools on single-cell RNA-seq data. *Genome Biol* *21*, 36. [10.1186/s13059-020-1949-z](https://doi.org/10.1186/s13059-020-1949-z).
- 832 25. Zhang, Y., Ma, Y., Huang, Y., Zhang, Y., Jiang, Q., Zhou, M., and Su, J. (2020). Benchmarking algorithms for  
833 pathway activity transformation of single-cell RNA-seq data. *Comput Struct Biotechnol J* *18*, 2953-2961.  
834 [10.1016/j.csbj.2020.10.007](https://doi.org/10.1016/j.csbj.2020.10.007).
- 835 26. Zhang, Y., Zhang, Y., Hu, J., Zhang, J., Guo, F., Zhou, M., Zhang, G., Yu, F., and Su, J. (2020). scTPA: a web  
836 tool for single-cell transcriptome analysis of pathway activation signatures. *Bioinformatics* *36*, 4217-4219.  
837 [10.1093/bioinformatics/btaa532](https://doi.org/10.1093/bioinformatics/btaa532).
- 838 27. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F.,  
839 Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and  
840 clustering. *Nat Methods* *14*, 1083-1086. [10.1038/nmeth.4463](https://doi.org/10.1038/nmeth.4463).
- 841 28. Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.B., Zhang, K., Chun, J., and  
842 Kharchenko, P.V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set  
843 overdispersion analysis. *Nat Methods* *13*, 241-244. [10.1038/nmeth.3734](https://doi.org/10.1038/nmeth.3734).
- 844 29. DeTomaso, D., Jones, M.G., Subramaniam, M., Ashuach, T., Ye, C.J., and Yosef, N. (2019). Functional  
845 interpretation of single cell similarity maps. *Nat Commun* *10*, 4376. [10.1038/s41467-019-12235-0](https://doi.org/10.1038/s41467-019-12235-0).
- 846 30. Mooney, M.A., Nigg, J.T., McWeeney, S.K., and Wilmot, B. (2014). Functional and genomic context in

- 
- 847 pathway analysis of GWAS data. *Trends in Genetics* *30*, 390-400.
- 848 31. Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association  
849 studies. *Nature Reviews Genetics* *11*, 843-854.
- 850 32. Watanabe, K., Umičević Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P., and Posthuma, D. (2019). Genetic  
851 mapping of cell type specificity for complex traits. *Nature communications* *10*, 1-13.
- 852 33. Battle, A., Brown, C.D., Engelhardt, B.E., and Montgomery, S.B. (2017). Genetic effects on gene expression  
853 across human tissues. *Nature* *550*, 204-213. [10.1038/nature24277](https://doi.org/10.1038/nature24277).
- 854 34. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B.,  
855 Neale, B.M., and Gusev, A. (2017). Linkage disequilibrium-dependent architecture of human complex traits  
856 shows action of negative selection. *Nature genetics* *49*, 1421-1427.
- 857 35. Tomfohr, J., Lu, J., and Kepler, T.B. (2005). Pathway level analysis of gene expression using singular value  
858 decomposition. *BMC bioinformatics* *6*, 1-11.
- 859 36. Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the national  
860 Academy of Sciences* *42*, 43-47.
- 861 37. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic  
862 data across different conditions, technologies, and species. *Nat Biotechnol* *36*, 411-420. [10.1038/nbt.4096](https://doi.org/10.1038/nbt.4096).
- 863 38. Wu, Y., and Zhang, K. (2020). Tools for the analysis of high-dimensional single-cell RNA sequencing data.  
864 *Nature Reviews Nephrology* *16*, 408-421.
- 865 39. Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke,  
866 M., and Zheng, G.X. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-  
867 phenotype acute leukemia. *Nature biotechnology* *37*, 1458-1465.
- 868 40. Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak,  
869 W., Worlock, K.B., and Yoshida, M. (2021). Single-cell multi-omics analysis of the immune response in  
870 COVID-19. *Nature medicine* *27*, 904-916.
- 871 41. Bryois, J., Skene, N.G., Hansen, T.F., Kogelman, L.J.A., Watson, H.J., Liu, Z., Brueggeman, L., Breen, G., Bulik,  
872 C.M., Arenas, E., et al. (2020). Genetic identification of cell types underlying brain complex traits yields  
873 insights into the etiology of Parkinson's disease. *Nat Genet* *52*, 482-493. [10.1038/s41588-020-0610-9](https://doi.org/10.1038/s41588-020-0610-9).
- 874 42. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19  
875 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* *184*, 1895-1913.e1819.  
876 [10.1016/j.cell.2021.01.053](https://doi.org/10.1016/j.cell.2021.01.053).
- 877 43. Manne, B.K., Denorme, F., Middleton, E.A., Portier, I., Rowley, J.W., Stubben, C., Petrey, A.C., Tolley, N.D.,  
878 Guo, L., Cody, M., et al. (2020). Platelet gene expression and function in patients with COVID-19. *Blood*  
879 *136*, 1317-1329. [10.1182/blood.2020007214](https://doi.org/10.1182/blood.2020007214).
- 880 44. Kumari, K., Chainy, G.B., and Subudhi, U. (2020). Prospective role of thyroid disorders in monitoring COVID-  
881 19 pandemic. *Heliyon* *6*, e05712.
- 882 45. Croce, L., Gangemi, D., Ancona, G., Liboà, F., Bendotti, G., Minelli, L., and Chiovato, L. (2021). The cytokine  
883 storm and thyroid hormone changes in COVID-19. *Journal of Endocrinological Investigation* *44*, 891-904.
- 884 46. Sen, A. (2020). Repurposing prolactin as a promising immunomodulator for the treatment of COVID-19:  
885 Are common Antiemetics the wonder drug to fight coronavirus? *Medical hypotheses* *144*, 110208.
- 886 47. Rydzynski Moderbacher, C., Ramirez, S.I., Dan, J.M., Grifoni, A., Hastie, K.M., Weiskopf, D., Belanger, S.,  
887 Abbott, R.K., Kim, C., Choi, J., et al. (2020). Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute  
888 COVID-19 and Associations with Age and Disease Severity. *Cell* *183*, 996-1012.e1019.  
889 [10.1016/j.cell.2020.09.038](https://doi.org/10.1016/j.cell.2020.09.038).
- 890 48. Quiros-Fernandez, I., Poorebrahim, M., Fakhr, E., and Cid-Arregui, A. (2021). Immunogenic T cell epitopes

- 891 of SARS-CoV-2 are recognized by circulating memory and naïve CD8 T cells of unexposed individuals.  
892 *EBioMedicine* *72*, 103610. [10.1016/j.ebiom.2021.103610](https://doi.org/10.1016/j.ebiom.2021.103610).
- 893 49. Nguyen, T.H.O., Rowntree, L.C., Petersen, J., Chua, B.Y., Hensen, L., Kedzierski, L., van de Sandt, C.E.,  
894 Chaurasia, P., Tan, H.X., Habel, J.R., et al. (2021). CD8(+) T cells specific for an immunodominant SARS-  
895 CoV-2 nucleocapsid epitope display high naïve precursor frequency and TCR promiscuity. *Immunity* *54*,  
896 1066-1082.e1065. [10.1016/j.immuni.2021.04.009](https://doi.org/10.1016/j.immuni.2021.04.009).
- 897 50. Kaech, S.M., and Ahmed, R. (2001). Memory CD8+ T cell differentiation: initial antigen encounter triggers  
898 a developmental program in naïve cells. *Nat Immunol* *2*, 415-422. [10.1038/87720](https://doi.org/10.1038/87720).
- 899 51. Jergović, M., Coplen, C.P., Uhrlaub, J.L., Besselsen, D.G., Cheng, S., Smithey, M.J., and Nikolich-Žugich, J.  
900 (2021). Infection-induced type I interferons critically modulate the homeostasis and function of CD8(+)  
901 naïve T cells. *Nat Commun* *12*, 5303. [10.1038/s41467-021-25645-w](https://doi.org/10.1038/s41467-021-25645-w).
- 902 52. Spitzer, S.O., Sitnikov, S., Kamen, Y., Evans, K.A., Kronenberg-Versteeg, D., Dietmann, S., de Faria Jr, O.,  
903 Agathou, S., and Káradóttir, R.T. (2019). Oligodendrocyte progenitor cells become regionally diverse and  
904 heterogeneous with age. *Neuron* *101*, 459-471. e455.
- 905 53. Vanzulli, I., Papanikolaou, M., De-La-Rocha, I.C., Pieropan, F., Rivera, A.D., Gomez-Nicola, D., Verkhratsky,  
906 A., Rodríguez, J.J., and Butt, A.M. (2020). Disruption of oligodendrocyte progenitor cells is an early sign of  
907 pathology in the triple transgenic mouse model of Alzheimer's disease. *Neurobiology of Aging* *94*, 130-  
908 139.
- 909 54. Agarwal, D., Sandor, C., Volpato, V., Caffrey, T.M., Monzón-Sandoval, J., Bowden, R., Alegre-Abarrategui,  
910 J., Wade-Martins, R., and Webber, C. (2020). A single-cell atlas of the human substantia nigra reveals cell-  
911 specific pathways associated with neurological disorders. *Nature communications* *11*, 1-11.
- 912 55. Sims, R., van der Lee, S.J., Naj, A.C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B.W., Boland,  
913 A., Raybould, R., Bis, J.C., et al. (2017). Rare coding variants in *PLCG2*, *ABI3*, and *TREM2* implicate microglial-  
914 mediated innate immunity in Alzheimer's disease. *Nat Genet* *49*, 1373-1384. [10.1038/ng.3916](https://doi.org/10.1038/ng.3916).
- 915 56. Corces, M.R., Shcherbina, A., Kundu, S., Gloudemans, M.J., Frésard, L., Granja, J.M., Louie, B.H., Eulalio, T.,  
916 Shams, S., Bagdatli, S.T., et al. (2020). Single-cell epigenomic analyses implicate candidate causal variants  
917 at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet* *52*, 1158-1168. [10.1038/s41588-020-00721-x](https://doi.org/10.1038/s41588-020-00721-x).
- 919 57. Yang, A.C., Vest, R.T., Kern, F., Lee, D.P., Agam, M., Maat, C.A., Losada, P.M., Chen, M.B., Schaum, N., Khoury,  
920 N., et al. (2022). A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature*.  
921 [10.1038/s41586-021-04369-3](https://doi.org/10.1038/s41586-021-04369-3).
- 922 58. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and  
923 Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell  
924 RNA-sequencing and human genetics. *Nature Genetics*, 1-14.
- 925 59. Xu, J., Zhang, P., Huang, Y., Zhou, Y., Hou, Y., Bekris, L.M., Lathia, J., Chiang, C.-W., Li, L., and Pieper, A.A.  
926 (2021). Multimodal single-cell/nucleus RNA sequencing data analysis uncovers molecular networks  
927 between disease-associated microglia and astrocytes with implications for drug repurposing in Alzheimer's  
928 disease. *Genome research* *31*, 1900-1912.
- 929 60. Sahel, A., Ortiz, F.C., Kerninon, C., Maldonado, P.P., Angulo, M.C., and Nait-Oumesmar, B. (2015). Alteration  
930 of synaptic connectivity of oligodendrocyte precursor cells following demyelination. *Frontiers in cellular  
931 neuroscience* *9*, 77.
- 932 61. Shang, L., Smith, J.A., and Zhou, X. (2020). Leveraging gene co-expression patterns to infer trait-relevant  
933 tissues in genome-wide association studies. *PLoS Genet* *16*, e1008734. [10.1371/journal.pgen.1008734](https://doi.org/10.1371/journal.pgen.1008734).
- 934 62. Yao, X., Glessner, J.T., Li, J., Qi, X., Hou, X., Zhu, C., Li, X., March, M.E., Yang, L., Mentch, F.D., et al. (2021).

- 935 Integrative analysis of genome-wide association studies identifies novel loci associated with  
936 neuropsychiatric disorders. *Translational psychiatry* *11*, 69. 10.1038/s41398-020-01195-5.
- 937 63. Howard, D.M., Adams, M.J., Clarke, T.-K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R.I., Hagenaaars,  
938 S.P., Ward, J., Wigmore, E.M., et al. (2019). Genome-wide meta-analysis of depression identifies 102  
939 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience*  
940 *22*, 343-352. 10.1038/s41593-018-0326-7.
- 941 64. Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghziian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko,  
942 J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide  
943 association study of educational attainment in 1.1 million individuals. *Nat Genet* *50*, 1112-1121.  
944 10.1038/s41588-018-0147-3.
- 945 65. Huang, Y., and Mucke, L. (2012). Alzheimer mechanisms and therapeutic strategies. *Cell* *148*, 1204-1222.  
946 10.1016/j.cell.2012.02.040.
- 947 66. Ochalek, A., Mihalik, B., Avci, H.X., Chandrasekaran, A., Téglási, A., Bock, I., Giudice, M.L., Táncos, Z., Molnár,  
948 K., László, L., et al. (2017). Neurons derived from sporadic Alzheimer's disease iPSCs reveal elevated TAU  
949 hyperphosphorylation, increased amyloid levels, and GSK3B activation. *Alzheimers Res Ther* *9*, 90.  
950 10.1186/s13195-017-0317-z.
- 951 67. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. (2020). *Nature* *583*, 590-595.  
952 10.1038/s41586-020-2496-1.
- 953 68. Garofano, L., Migliozi, S., Oh, Y.T., D'Angelo, F., Najac, R.D., Ko, A., Frangaj, B., Caruso, F.P., Yu, K., Yuan, J.,  
954 et al. (2021). Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with  
955 therapeutic vulnerabilities. *Nat Cancer* *2*, 141-156. 10.1038/s43018-020-00159-4.
- 956 69. Weiss, T., and Weller, M. (2021). Pathway-based stratification of glioblastoma. *Nat Rev Neurol* *17*, 263-264.  
957 10.1038/s41582-021-00474-z.
- 958 70. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*  
959 *28*, 27-30.
- 960 71. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen,  
961 T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes.  
962 *Nature* *593*, 238-243. 10.1038/s41586-021-03446-x.
- 963 72. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M.,  
964 Haw, R., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res* *48*, D498-d503.  
965 10.1093/nar/gkz1031.
- 966 73. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular  
967 Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* *1*, 417-425.  
968 10.1016/j.cels.2015.12.004.
- 969 74. Consortium, G.P. (2015). A global reference for human genetic variation. *Nature* *526*, 68.
- 970 75. Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P.,  
971 and Culley, O.J. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*  
972 *546*, 370-375.
- 973 76. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell  
974 gene expression data. *Nat Biotechnol* *33*, 495-502. 10.1038/nbt.3192.
- 975 77. Yu, F., Sankaran, V.G., and Yuan, G.-C. (2022). CUT&RUNTools 2.0: a pipeline for single-cell and bulk-level  
976 CUT&RUN and CUT&Tag data analysis. *Bioinformatics* *38*, 252-254.
- 977 78. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M.,  
978 Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* *177*, 1888-



- 
- 979 1902.e1821. [10.1016/j.cell.2019.05.031](https://doi.org/10.1016/j.cell.2019.05.031).
- 980 79. Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other  
981 measures of statistical accuracy. *Statistical science*, 54-75.
- 982 80. Sun, T., Song, D., Li, W.V., and Li, J.J. (2021). scDesign2: a transparent simulator that generates high-fidelity  
983 single-cell gene expression count data with gene correlations captured. *Genome biology* 22, 163.
- 984 81. Goudot, C., Coillard, A., Villani, A.-C., Gueguen, P., Cros, A., Sarkizova, S., Tang-Huau, T.-L., Bohec, M.,  
985 Baulande, S., and Hacohen, N. (2017). Aryl hydrocarbon receptor controls monocyte differentiation into  
986 dendritic cells versus macrophages. *Immunity* 47, 582-596. e586.
- 987 82. Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carre, C., Burdin, N., Visan, L., Ceccarelli, M., and  
988 Poidinger, M. (2019). RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution  
989 of human immune cell types. *Cell reports* 26, 1627-1640. e1627.
- 990 83. Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E.,  
991 Borm, L.E., La Manno, G., et al. (2018). Molecular Architecture of the Mouse Nervous System. *Cell* 174, 999-  
992 1014.e1022. [10.1016/j.cell.2018.06.021](https://doi.org/10.1016/j.cell.2018.06.021).
- 993 84. Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-  
994 Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with  
995 Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci* 22, 2087-2097.  
996 [10.1038/s41593-019-0539-4](https://doi.org/10.1038/s41593-019-0539-4).
- 997 85. Smith, A.M., Davey, K., Tsartsalis, S., Khozoie, C., Fancy, N., Tang, S.S., Liaptsi, E., Weinert, M., McGarry, A.,  
998 Muirhead, R.C.J., et al. (2022). Diverse human astrocyte and microglial transcriptional responses to  
999 Alzheimer's pathology. *Acta Neuropathol* 143, 75-91. [10.1007/s00401-021-02372-6](https://doi.org/10.1007/s00401-021-02372-6).
- 1000 86. Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C.L., Voillet, V., Duvvuri, V.R., Scherler, K., Troisch, P., et  
1001 al. (2020). Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. *Cell*  
1002 183, 1479-1495.e1420. [10.1016/j.cell.2020.10.037](https://doi.org/10.1016/j.cell.2020.10.037).
- 1003 87. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020).  
1004 Construction of a human cell landscape at single-cell level. *Nature* 581, 303-309. [10.1038/s41586-020-  
1005 2157-4](https://doi.org/10.1038/s41586-020-2157-4).
- 1006

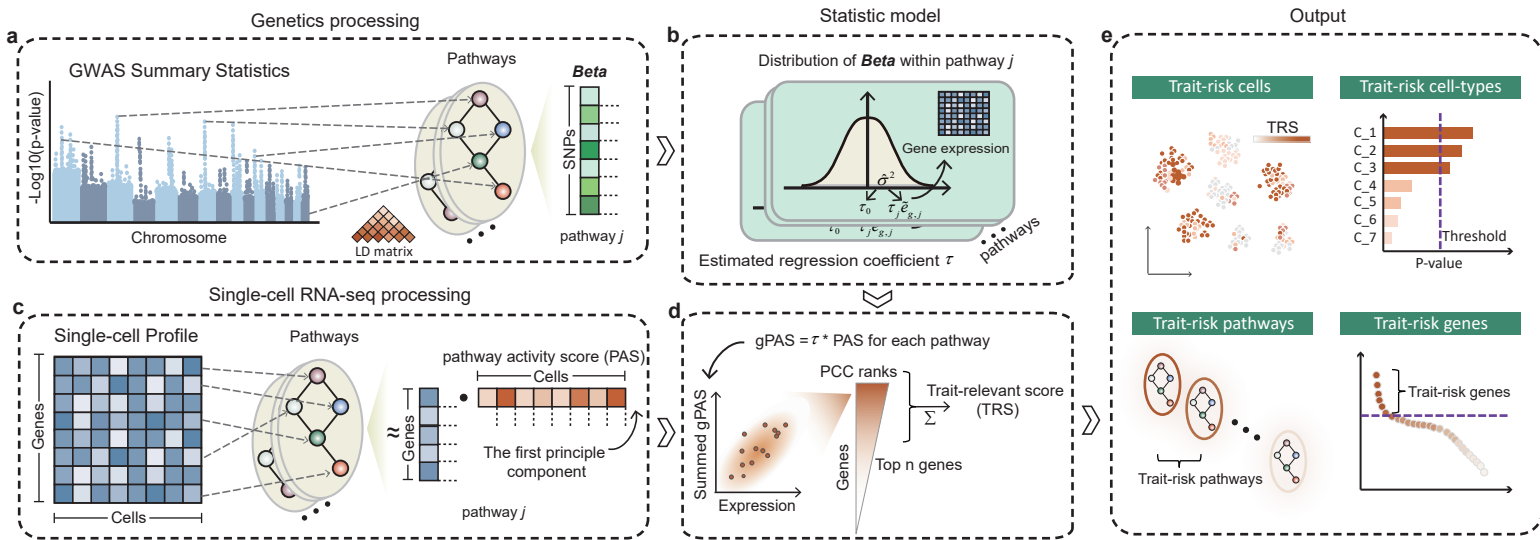
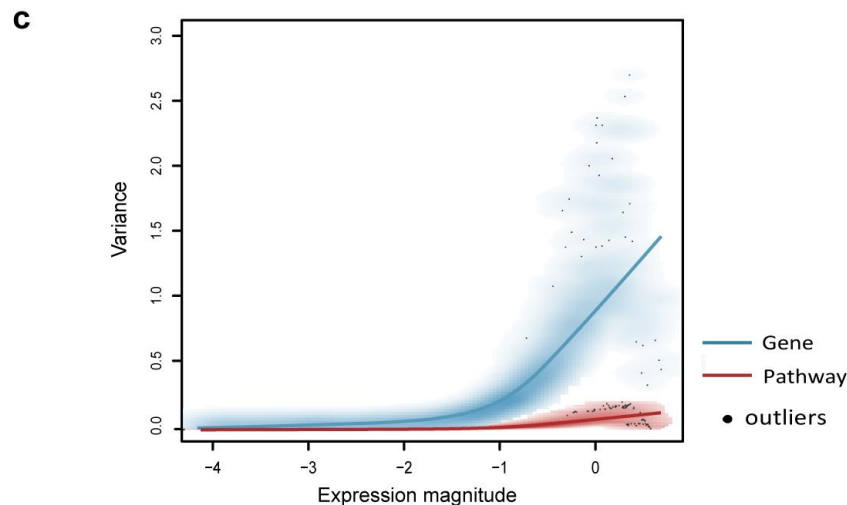
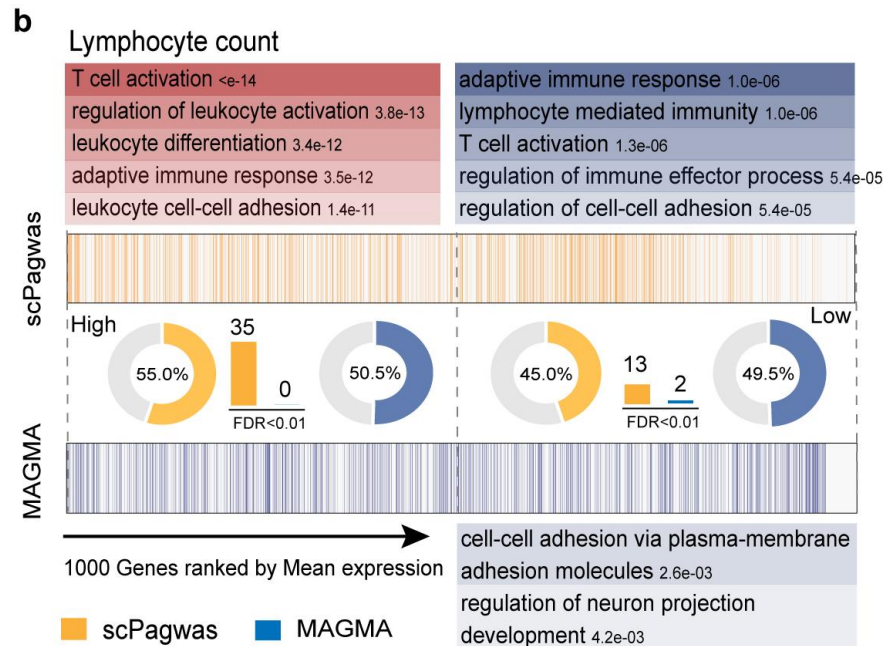
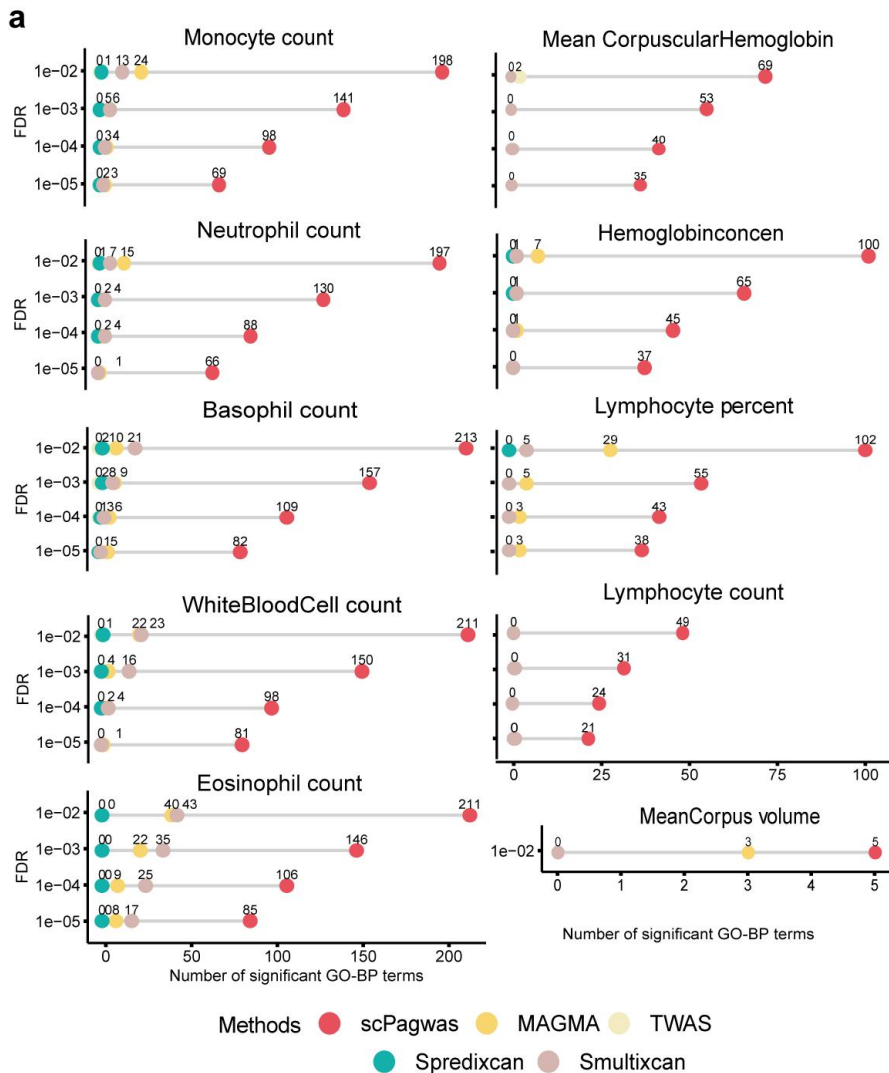
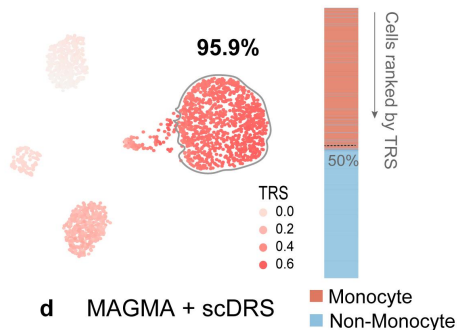
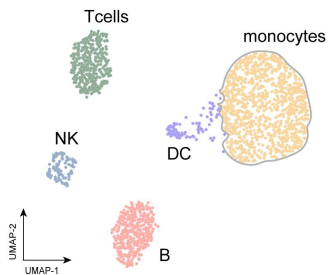
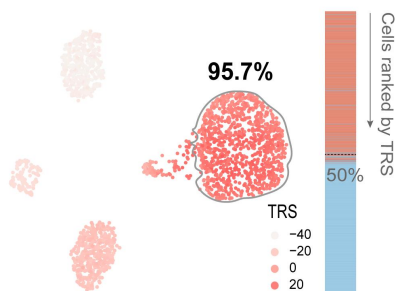
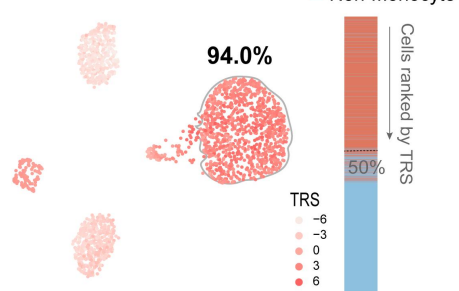
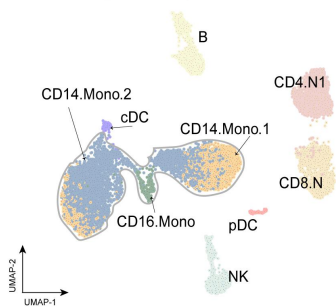
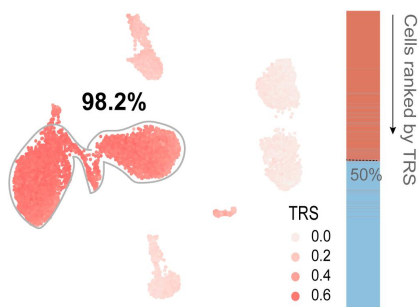
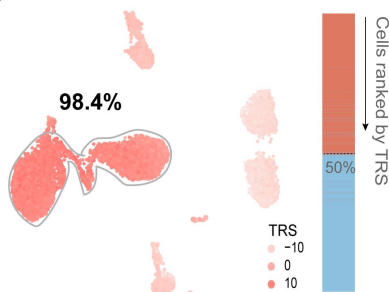
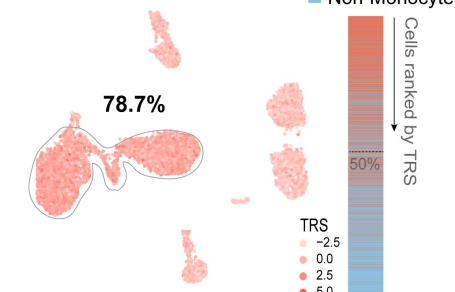
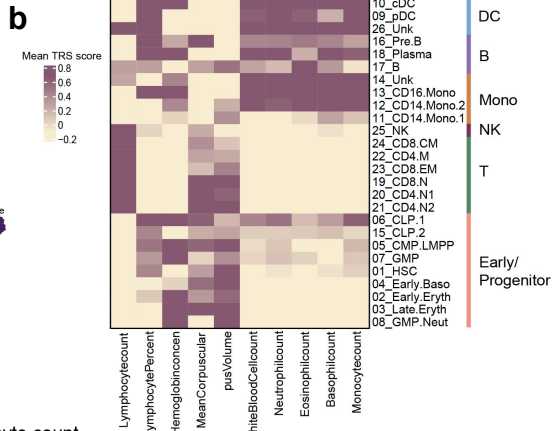
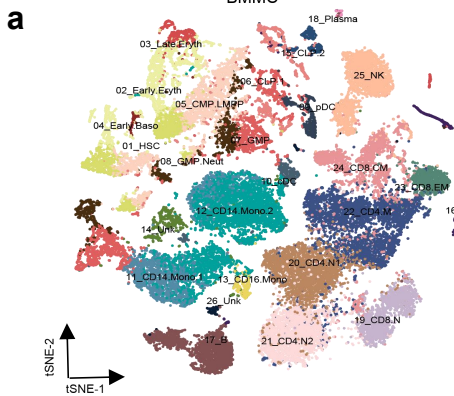


Figure 1

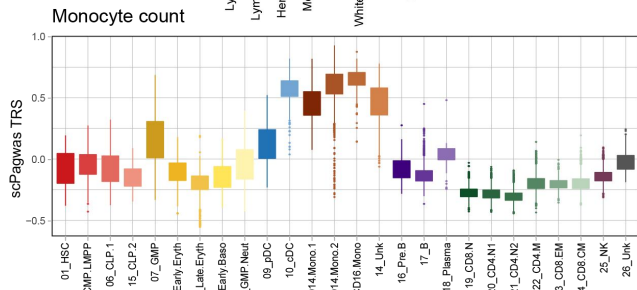
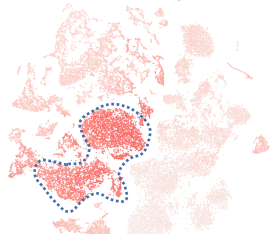


**a** Simulated single cell groundtruth data **b** scPagwas + Seurat**c** scPagwas + scDRS**d** MAGMA + scDRS**e** Real single cell groundtruth data**f** scPagwas + Seurat**g** scPagwas + scDRS**h** MAGMA + scDRS

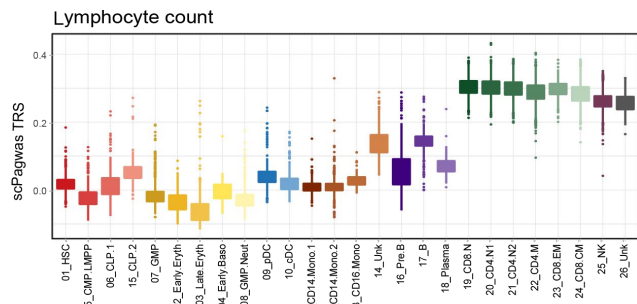
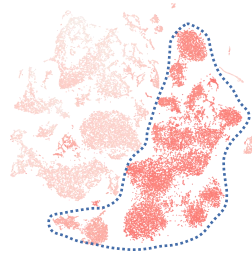
**BMMC**



**c** scPagwas + Seurat



**d** scPagwas + Seurat



**e** scPagwas + Seurat

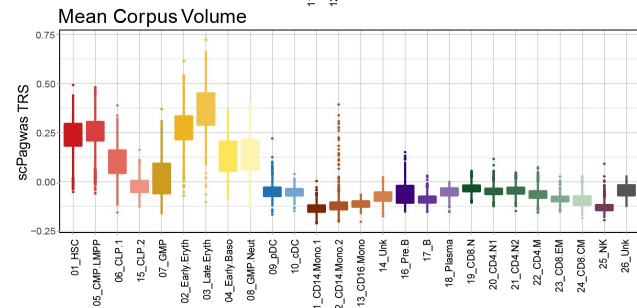
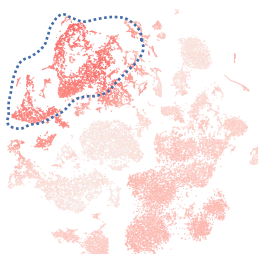


Figure 4

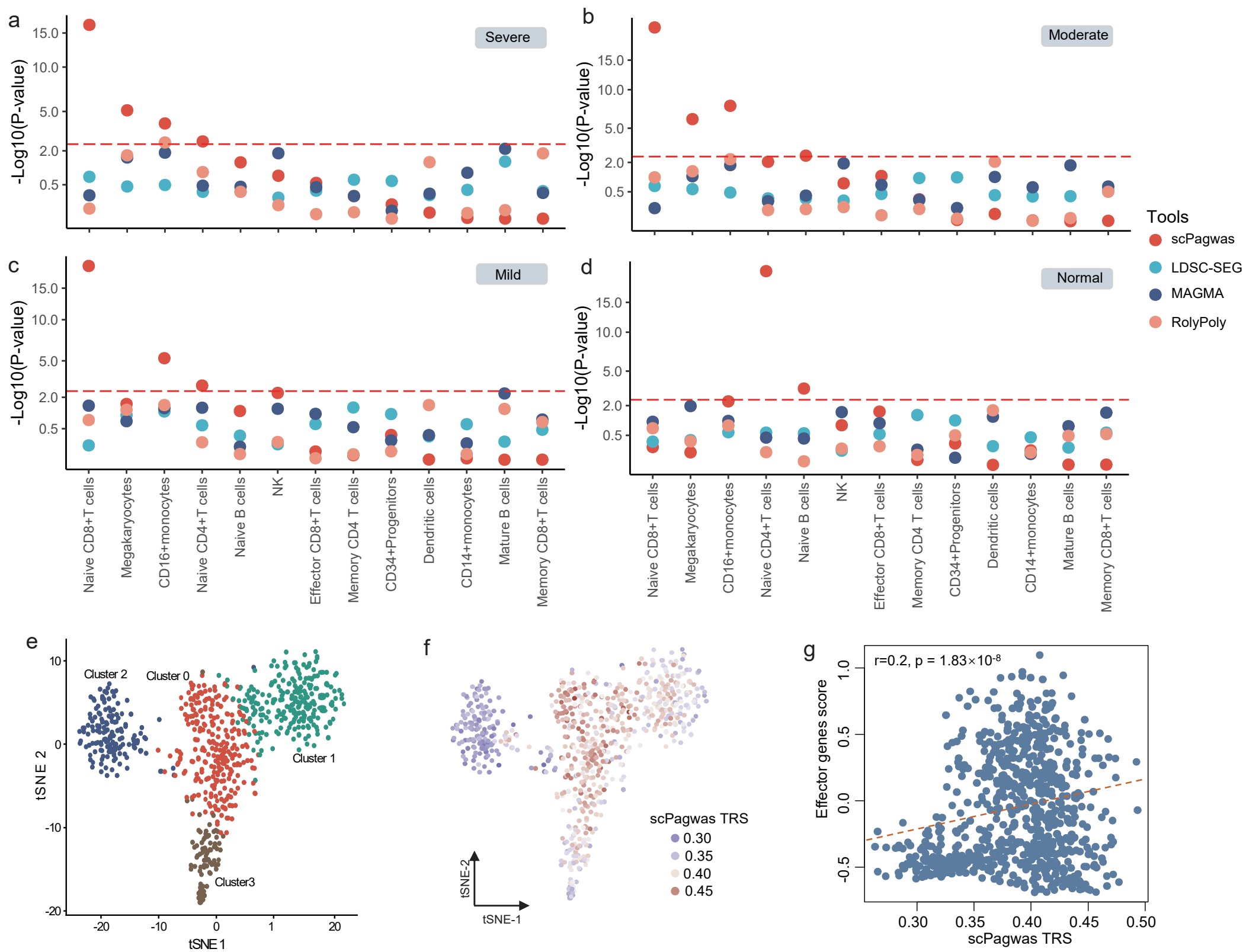


Figure 5

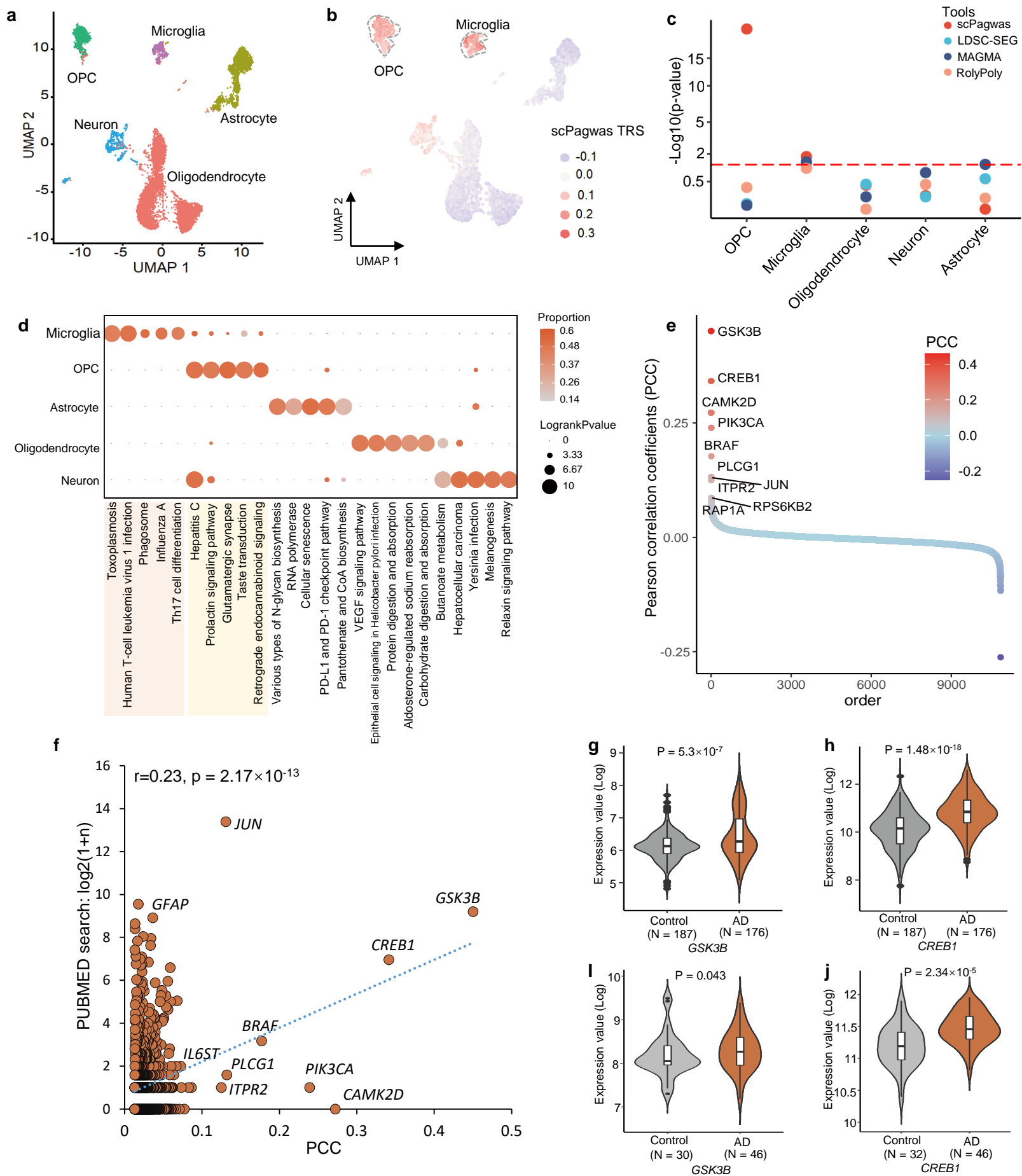


Figure 6