

1 **Evaluating performance and applications of sample-wise cell**
2 **deconvolution methods on human brain transcriptomic data**

3

4 **Short title: Evaluating sample-wise cell deconvolution**

5

6 Rujia Dai¹, Tianyao Chu², Ming Zhang², Xuan Wang², Alexandre Jourdon³,
7 Feinan Wu³, Jessica Mariani³, Flora M. Vaccarino^{3,4}, Donghoon Lee⁵, John F.
8 Fullard⁵, Gabriel E. Hoffman⁵, Panos Roussos⁵, Yue Wang⁶, Xusheng Wang⁷,
9 Dalila Pinto⁸, Sidney H. Wang⁹, Chunling Zhang¹⁰, PsychENCODE consortium,
10 Chao Chen^{2*}, Chunyu Liu^{1,2,10*}

11

12 ¹Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA

13 ²Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of
14 Life Sciences, Central South University, Changsha, China

15 ³Child Study Center, Yale University, New Haven, CT, USA

16 ⁴Department of Neuroscience, Yale University, New Haven, CT, USA

17 ⁵Center for Disease Neurogenomics, Departments of Psychiatry and Genetics and
18 Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

19 ⁶Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and
20 State University, VA, USA

21 ⁷ Department of Biology, University of North Dakota, Grand Forks, ND, USA

22 ⁸ Department of Psychiatry, Department of Genetics and Genomic Sciences, Mindich
23 Child Health and Development Institute, and Icahn Genomics Institute for Data

24 Science and Genomic Technology, Seaver Autism Center, Icahn School of Medicine at
25 Mount Sinai, New York, NY, USA.

26 ⁹Center for Human Genetics, The Brown foundation Institute of Molecular Medicine,
27 The University of Texas Health Science Center at Houston, Houston, TX, USA

28 ¹⁰Department of Neuroscience & Physiology, SUNY Upstate Medical University,
29 Syracuse, NY, USA

30

31 **Abstract**

32 Sample-wise deconvolution methods have been developed to estimate cell-
33 type proportions and gene expressions in bulk-tissue samples. However, the
34 performance of these methods and their biological applications has not been
35 evaluated, particularly on human brain transcriptomic data. Here, nine
36 deconvolution methods were evaluated with sample-matched data from bulk-
37 tissue RNAseq, single-cell/nuclei (sc/sn) RNAseq, and immunohistochemistry.

38 A total of 1,130,767 nuclei/cells from 149 adult postmortem brains and 72
39 organoid samples were used. The results showed the best performance of
40 dtangle for estimating cell proportions and bMIND for estimating sample-wise
41 cell-type gene expression. For eight brain cell types, 25,273 cell-type eQTLs
42 were identified with deconvoluted expressions (decon-eQTLs). The results
43 showed that decon-eQTLs explained more schizophrenia GWAS heritability
44 than bulk-tissue or single-cell eQTLs alone. Differential gene expression
45 associated with multiple phenotypes were also examined using the

46 deconvoluted data. Our findings, which were replicated in bulk-tissue RNAseq
47 and sc/snRNAseq data, provided new insights into the biological applications
48 of deconvoluted data.

49

50 **Introduction**

51 Brain transcriptome is essential for studying brain biology and related disorders,
52 but important cell type information can be obscured when bulk tissue is used
53 for the data production. Several brain projects have generated valuable
54 transcriptomic resources from human brains, such as GTEx(1),
55 PsychENCODE(2), CommonMind(3), Brainspan(4), and ROSMAP(5).
56 However, most of the existing transcriptomes are from bulk tissue, which are
57 mixtures of many different cell types, and gene regulatory mechanisms are
58 known to vary across brain cell types, obscuring the cellular mechanisms
59 underlying bulk tissue expression changes.

60

61 Sorted-cell RNAseq and single-cell/nuclei (sc/sn) RNAseq(6, 7) offer solutions
62 for profiling brain transcriptome at the cell-type resolution but with several
63 limitations. Cell sorting relies on marker genes, which are not always available.
64 The specificity of such marker genes is frequently a concern. A combination of
65 several marker genes only can sort limited cell types. The data from
66 sc/snRNAseq is sparse due to the limited RNA input from each cell(8).
67 Sc/snRNAseq data suffer from the large number of zero values, which are

68 called “dropout events”. Moreover, it is challenging to discriminate between two
69 possible causes of dropouts: biologically true zero expression and technical
70 random missing data(9). The presence of dropouts may result in potential
71 problems in gene expression quantification. Another limitation is the high cost
72 of sc/snRNAseq. Even though multiplexing methods have been developed to
73 simultaneously profile cells from numerous samples(10), using sc/snRNAseq
74 in large-scale studies typically requiring hundreds of subjects, such as disease
75 association and expression quantitative trait loci (eQTL) mapping(11), can be
76 cost-prohibitive.

77

78 Computational algorithms for cell deconvolution have been developed to
79 estimate cell proportions. These algorithms can be classified into two types:
80 supervised deconvolution uses prior information from a cell-type reference data
81 to facilitate the estimation of the cell proportions of each cell types in bulk-tissue
82 samples, while unsupervised does not need a reference. This study focused on
83 evaluating supervised deconvolution methods.

84

85 The performance of methods for estimating cell proportions has been
86 previously evaluated (12-16). Studies have evaluated the accuracy of
87 estimated cell proportions with data from the brain and other tissues. Methods
88 like DSA(17), OLS, CIBERSORT(18), dtangle(19), and MuSiC(20) showed
89 good performance in these evaluations. Additionally, the effect of cell type

90 marker gene selection, covariates, data transformation and normalization, and
91 cell subtypes on cell deconvolution has been evaluated, which provided
92 guidelines for data processing before cell deconvolution.

93

94 Estimated cell proportions have been used for cell-type studies but with
95 limitations. Cell proportions were used to represent cell types for case-control
96 comparison(21, 22). However, proportional changes in cell types are just one
97 aspect of possible changes. Disease-related changes also involve cell-type-
98 specific gene expressions(23-26). Cell proportions have also been used to map
99 eQTLs associated with cell types, which are called cell-type interaction
100 eQTLs(27) (ieQTLs). The genetic regulators that were associated with gene
101 expression when the cell proportion varied were mapped. ieQTL has two major
102 limitations. Firstly, ieQTLs are not necessarily specific to cell types. They may
103 refer to other cell types with positive or negative correlation with the cell
104 proportion of the target cell type. Secondly, the power of ieQTL mapping is low.
105 Less than 50 ieQTLs were reported for neurons with 15,201 samples (27),
106 which is much less than what standard eQTL can map on bulk tissues.
107 Therefore, there is still the need to discover expression changes associated
108 with diseases and eQTLs from cell-type expression data.

109

110 Cell-type gene expressions can be deconvoluted from bulk-tissue expression
111 data. Methods have been developed to estimate cell-type expressions for each

112 sample, such as bMIND(28), swCAM(29), and TCA(30). We call this sample-
113 wise deconvolution of gene expression. These methods use expression
114 references from sc/snRNAseq or sorted-cell expressions. For example, bMIND
115 used the Bayesian model and Markov Chain Monte Carlo to estimate
116 expression for each gene in the cell types of each sample. The cell-type
117 expressions of the individual samples will enable eQTL mapping and differential
118 expression analysis in cell types. The deconvoluted data can cover the majority
119 of genes in bulk tissue and is less sparse than sc/snRNAseq data. It makes the
120 large sample study of cell-type expression affordable since bulk tissue data is
121 either ready to use or can be generated at a relatively low cost.

122

123 The methods for estimating sample-wise cell-type expressions have been
124 partially evaluated with major blind spots. The performance of bMIND, TCA,
125 and swCAM has been evaluated in their original methodology papers. However,
126 these studies used artificially-constructed pseudo-bulk data other than bulk-
127 tissue data to benchmark their performance. Pseudo-bulks were constructed
128 by simulating cell proportions and multiplying these proportions with
129 expressions from sc/snRNAseq or sorted-cell expression data. Therefore,
130 pseudo-bulk data is less complex than data from real bulk tissue(31). The
131 differences among cell types in the pseudo-bulk are easier to be captured than
132 those in bulk-tissue data. The benchmark conclusion based on the pseudo-bulk
133 data may not apply to data from brain tissues. Head-to-head comparisons of all

134 these methods on brain data have not been conducted to date. The
135 downstream applications based on deconvoluted data, such as eQTL mapping
136 and differential expression, have also not been evaluated to showcase the
137 validity of deconvolution.

138

139 The current study aimed at evaluating the performance of algorithms for
140 sample-wise deconvoluting cell proportions and cell-type expressions, as well
141 as research applications based on the deconvoluted data. Specifically, we
142 evaluated six commonly-used deconvolution methods for estimating cell
143 proportions and three deconvolution methods for estimating the cell-type
144 expressions of individual samples. Data from bulk-tissue RNAseq,
145 sc/snRNAseq, and immunohistochemistry (IHC) of matched adult postmortem
146 brains and brain organoids were used for evaluation. Downstream analyses of
147 the deconvoluted results were also conducted, including their use in eQTL
148 mapping, schizophrenia (SCZ) GWAS heritability enrichment, differential
149 expression for Alzheimer's disease (AD), SCZ, and brain development in cell
150 types. Based on the evaluation, we recommended the best practice for brain
151 transcriptome deconvolution.

152

153 **Results**

154 **Benchmarking of sample-wise deconvolution methods with brain transcriptome**
155 **data**

156 To evaluate commonly-used deconvolution methods, we selected six methods
157 (DSA, OLS, CIBERSORT, dtangle, MuSiC, and Bisque(32)) for estimating cell
158 proportions and three methods (bMIND, swCAM, and TCA) for estimating cell-
159 type expressions (Fig. 1). Bulk-tissue RNAseq, snRNAseq, and IHC data from
160 ROSMAP were used as primary data for evaluation(33). Data from adult brains
161 in CommonMind (CMC)(34) and brain organoids(35) were used for
162 confirmation (Table 1). Cell proportions from IHC and sc/snRNAseq data were
163 used as ground truth for evaluating the accuracy of estimated cell proportions.
164 Gene expressions in sc/snRNAseq data were used as ground truth for
165 evaluating the accuracy of estimated cell-type expressions. The root-mean-
166 square error (RMSE) and Spearman correlation coefficient were used as
167 evaluation metrics. After method evaluation, eQTL mapping, GWAS heritability
168 enrichment, and differential expression analysis were performed on the cell-
169 type expressions estimated by the best performing method. To further evaluate
170 the quality of outputs of these deconvolution methods by actual applications,
171 the eQTLs, explained GWAS heritability, and phenotype-associated genes
172 derived from deconvoluted expressions were compared to corresponding
173 results based on sc/snRNAseq and bulk-tissue data.

174

175 **Evaluation of cell proportions estimated by deconvolution methods**

176 The overall performance of six deconvolution methods (DSA, OLS,
177 CIBERSORT, dtangle, MuSiC, and Bisque) for estimating cell proportions was

178 evaluated with ground truth from matched samples. To ensure the
179 deconvolution performance, the intersection of marker genes identified at the
180 individual-cell level and the pseudo-bulk level was used to guide deconvolution
181 (see details in methods). Using ROSMAP IHC data as ground truth (n of
182 samples=49), dtangle, and OLS showed lower RMSE than other methods (Fig.
183 2A). dtangle also showed relatively low RMSE in CMC (n=94) and brain
184 organoid (n=55) data. MuSiC and Bisque did not perform well, even though
185 they are designed to use sc/snRNAseq data as a reference. Using cell
186 proportions computed from sc/snRNAseq data as the ground truth, Bisque had
187 the lowest RMSE in all three datasets. The accuracy of deconvoluted cell
188 proportions in major cell types was better than that in minor cell types (Fig. 2B,
189 Fig. S1). The RMSE increased sharply when the cell proportion was below 5%,
190 such as in oligodendrocyte precursor cells (Opc), microglia, endothelial cells,
191 and pericytes in adult brains. Similar results were observed using Spearman
192 correlation as an evaluation metric (Fig. S2).

193

194 **Evaluation of sample-wise cell-type expressions estimated by deconvolution**
195 **methods**

196 The accuracy of sample-wise cell-type expressions deconvoluted by bMIND, swCAM,
197 and TCA was evaluated using ground truth generated from sc/snRNAseq expressions
198 of matched samples (n=35 for ROSMAP, n=94 for CMC, and n=55 for brain organoid).
199 Cell proportions estimated by Bisque and dtangle were selected as input for the three

200 methods, since they showed the best performance in the above evaluation for
201 estimating cell proportions. bMIND showed the best performance for estimating cell-
202 type expressions in all datasets, followed by swCAM (Fig. 3A). For bMIND, the
203 averaged correlation coefficient between estimated expression and sc/snRNAseq data
204 was 0.62 in ROSMAP data, 0.75 in CMC adult brain data, and 0.85 in brain organoid
205 data. We did not observe a substantial difference in performances for estimating
206 expressions of major and rare cell types (Fig. 3B). However, bMIND performed more
207 steadily and overall better in major cell types than in minor cell types. The deconvoluted
208 expressions by bMIND correlated with corresponding cell types in sample-matched
209 sc/snRNAseq data, and they were less correlated with unrelated cell types (Fig. S3).
210 A number of well-known marker genes were highly expressed in corresponding cell
211 types, thus indicating that the deconvoluted data have good cell-type specificity (Fig.
212 3C).
213

214 **Cell-type eQTL mapping with deconvoluted sample-wise expression data**
215 To identify SNPs that cis-regulate gene expression in specific cell types, cell-type eQTL
216 mapping was performed for the association between genotypes and deconvoluted
217 gene expression data of individual samples. The cell-type eQTLs identified with
218 deconvoluted gene expression data were named deconvolution eQTLs (decon-eQTL).
219 RNAseq data of 1,112 bulk-tissue samples of ROSMAP collection were deconvoluted.
220 Cell proportions and cell-type expressions were estimated with dtangle and bMIND,
221 respectively. Out of the 1,112 samples, 861 had genotype data and were used for

222 decon-eQTL mapping. The effect of SNPs within a 1-megabase window around the
223 transcription start site (TSS) of genes was tested. The numbers of input genes for
224 decon-eQTL mapping ranged from 8,521 to 12,418 across all cell types. The number
225 of input SNPs was 4,954,561 for all cell types. Effects of known and hidden covariates
226 on deconvoluted expressions were corrected. A total of 1,088,634 to 2,245,945 decon-
227 eQTLs were detected across eight cell types at a genome-wide significant level
228 (FDR<0.05). To identify the independent effect in SNPs, a permutation test was
229 performed for each gene. A total of 25,273 (4,541~ 8,149) independent decon-eQTLs
230 were identified at FDR<0.05 for eight cell types (Fig. 4B). As expected, eQTL SNPs
231 (eSNPs) were enriched around the TSS region of eQTL genes (eGenes) (Fig. S4). The
232 numbers of detected decon-eQTLs were positively correlated with the proportions of
233 cell types in the tissue (Fig. 4B). To test the robustness of identified decon-eQTLs,
234 sample IDs were randomly shuffled before the eQTL mapping. The absence of
235 significant eQTL in the shuffled data supported that the identified decon-eQTLs were
236 not due to random noise (Fig. S5).

237
238 Identified decon-eQTLs from ROSMAP data were replicated with another
239 deconvoluted data from BrainGVEX(36). The same deconvolution and eQTL mapping
240 procedures were performed on RNAseq data of 400 postmortem brain samples from
241 BrainGVEX to obtain the decon-eQTLs. Across all eight cell types, 3,479 to 5,718
242 independent eQTLs were identified in the deconvoluted data from BrainGVEX at
243 FDR<0.05. To measure the replication rate of ROSMAP decon-eQTLs in BrainGVEX

244 data, Pi1 statistic(37), which is the proportion of true eQTL associations in the
245 replication data, were calculated. The Pi1 of ROSEMAP decon-eQTLs in BrainGVEX
246 data was 0.59 ~ 0.74 for the matched cell types (Fig. 4C). Decon-eQTLs of
247 oligodendrocyte had relatively better replication than other cell types.

248

249 Cell-type eQTLs from snRNAseq data in Bryois et al.(38) were also used to replicate
250 our decon-eQTLs. This replication study had performed genotyping and snRNAseq on
251 192 cortical samples. A total of 7,607 independent eQTLs across eight cell types were
252 identified. Even though the replication data had less statistical power than our
253 deconvoluted data, 17%~57% of decon-eQTLs were replicated (Fig. 4D). eQTLs of
254 excitatory neurons ($\text{Pi1}=0.57$) had higher Pi1 values than other cell types (averaged
255 $\text{Pi1}=0.38$).

256

257 To illustrate the value of decon-eQTLs, we compared decon-eQTLs to bulk-tissue
258 eQTLs from ROSEMAP (Fig. 4E). Overall, decon-eQTLs had good replication in bulk-
259 tissue data, with $\text{Pi1}>0.95$. The eQTLs that were significant at the cell-type level but
260 insignificant at the bulk-tissue level were defined as cell-type-specific eQTLs. A total of
261 1,206 ~ 3,006 (24.3% ~ 36.89%) cell-type-specific eQTLs were identified in the
262 deconvoluted data. Cell-type-specific eQTLs had Pi1 values of 0.17~0.52 in single-cell
263 eQTLs, which were similar to Pi1 values of decon-eQTLs that were shared with bulk-
264 tissue eQTLs (Fig. S6). This demonstrated that a good proportion of eQTLs regulate
265 gene expressions in a cell-type-specific way, and they can be detected by decon-

266 eQTLs.

267

268 **Cell-type eQTLs enriched for the risk heritability in SCZ GWAS data**

269 To test whether cell-type eQTLs are enriched for genetic risk heritability of SCZ,
270 stratified linkage disequilibrium score regression (sLDSC)(39) was used to calculate
271 the heritability of SCZ GWAS mediated by decon-eQTLs. Single-cell eQTLs and bulk-
272 tissue eQTLs were also included for comparison. Decon-eQTLs explained more SCZ
273 GWAS heritability (averaged $h^2 = 37\%$) than single-cell eQTLs (averaged $h^2 = 6\%$) for
274 all cell types (Fig. 5A). Bulk-tissue eQTLs explained 49% of SCZ GWAS heritability.
275 Integrating decon-eQTLs and bulk-tissue eQTLs increased the explained heritability to
276 63%, whereas the integration of single-cell eQTLs only resulted in an increase of
277 heritability to 53%. The total proportion of explained heritability was correlated with the
278 proportions of each cell type. To control the effect of SNP numbers, heritability was
279 normalized by the number of decon-eQTLs, which was called enrichment. Decon-
280 eQTLs of all cell types were enriched for SCZ GWAS heritability (Fig. 5B, P value <0.05).
281 Decon-eQTLs of oligodendrocytes showed the strongest per-SNP enrichment across
282 all cell types. The SCZ GWAS heritability was only significantly enriched in single-cell
283 eQTLs from oligodendrocytes and excitatory neurons. Decon-eQTLs of most of the
284 cell types showed higher enrichment of SCZ GWAS heritability than bulk-tissue eQTLs
285 (Fig. 5B), indicating that some of the SCZ risk SNPs may affect gene expression in cell
286 type-specific ways. Deconvolution analyses uncovered more such cell-type-specific
287 regulations associated with the genetic risk of SCZ.

288

289 **Identification of gene expression changes associated with disease and brain**
290 **development within cell types**

291 To identify genes associated with various phenotypes in specific cell types, differential
292 gene expression analysis was conducted using the deconvoluted sample-wise
293 expression data. Associations with AD, SCZ, and brain development modeled by
294 organoids were tested in three deconvoluted datasets independently. More samples
295 were included in AD ($N_{AD}=743$, $N_{control}=367$) and SCZ ($N_{SCZ}=246$, $N_{control}=279$) data. For
296 AD and SCZ, the Wilcoxon signed-rank test was performed on the deconvoluted data.
297 For brain development, the linear regression model was used to test the correlation
298 between deconvoluted data and culture days of organoids ($N_{day0}=15$, $N_{day30}=22$,
299 $N_{day60}=18$). With a threshold of $FDR<0.05$, 4,419, 10,964, and 9,562 phenotypes-
300 associated genes (PAGs) were identified for AD, SCZ, and brain development,
301 respectively.

302

303 To test the reliability of PAGs identified from deconvoluted data, these PAGs were
304 compared to those identified from bulk-tissue and sc/snRNAseq data (Fig. 6). In total,
305 81%, 49%, and 89% of PAGs for AD, SCZ and brain development, respectively, were
306 replicated in bulk-tissue data. Among these PAGs, most of them (>95%) had the same
307 direction of expression changes in bulk-tissue data. For AD and SCZ, less than 15%
308 of PAGs overlapped with PAGs from snRNAseq data. However, 35% of development-
309 related PAGs could be replicated in scRNAseq data. The possible explanation for the

310 difference in replication rate in the three datasets was that the expression changes
311 associated with brain development were larger than the changes associated with AD
312 or SCZ (Fig. S7). The low replication rate with sc/snRNAseq data suggested that
313 sc/snRNAseq data was underpowered to detect PAGs of small effect size.

314

315 **Discussion**

316 Using matched samples of bulk-tissue RNAseq, IHC, and sc/snRNAseq data, the
317 performance of six methods for estimating cell proportions and three methods for
318 estimating sample-wise cell-type gene expression was systematically evaluated. The
319 transcriptome data used for evaluation were from adult brains and cultured brain
320 organoids, providing data representative from different states of cell maturity and
321 developmental processes. In addition, the results of eQTL mapping, SCZ GWAS
322 heritability enrichment, and differential expression analysis based on deconvoluted
323 data demonstrate the utility of deconvolution.

324

325 dtangle had better accuracy for estimating cell type proportions than other methods.
326 Previous studies have benchmarked the performance of deconvolution methods for
327 estimating cell type proportions with ground truth data created from simulated
328 proportions. In those studies, dtangle showed good performance in Sutton et al(15)
329 but poor performance in Avila Cobos et al.(12) One possible reason is the difference
330 of pseudo-bulk simulation between studies. Sutton et al. constructed pseudo-bulks
331 from 500 cells while Avila Cobos et al. used only 100 cells for each pseudo-bulk. Given

332 that sc/snRNAseq data are remarkably sparse, 100 cells may not be representative of
333 cell composition in bulk tissue. This inconsistency suggests the necessity of using real
334 ground truth data in benchmarking studies. By using bulk-tissue RNAseq and IHC data
335 from matched samples, dtangle was found to be the best deconvolution method for
336 estimating cell proportions in this study. The excellent performance of dtangle was
337 preserved in two replication data.

338

339 Deconvolution methods using sc/snRNAseq data as reference, such as Bisque and
340 MuSiC, did not outperform old methods using pooled-cell reference. Given that Bisque
341 learns prior information from the reference of sc/snRNAseq data, it is not surprising to
342 see that Bisque showed perfect performance when using cell proportions from
343 sc/snRNAseq data as ground truth. However, the cell proportions measured by single-
344 cell technologies can be easily biased by the sorting strategy(40, 41). The proportions
345 from single-cell data as prior reference and ground truth should be used with caution.

346

347 Cell-type expressions deconvoluted to individual samples from brain tissues were
348 further evaluated for the first time here for eQTL mapping and differential expression
349 analysis, which require sample-wise expression data. The development of sample-
350 wise deconvolution satisfies these needs. Sample-wise deconvolution can estimate
351 cell-type expression for each sample, without cell sorting and sc/snRNAseq. bMIND
352 was the best method for estimating cell-type expressions in our evaluation, since the
353 correlation coefficients between estimated expressions by bMIND and ground truth

354 were higher than other methods. Moreover, our evaluation showed that the
355 deconvoluted data by bMIND have good cell-type specificity. The deconvoluted
356 expressions by bMIND had a high correlation with matched cell types but a low
357 correlation with other cell types in the ground truth data (Fig. S3). The deconvoluted
358 cell types by bMIND expressed well-known marker genes. For example, NRGN and
359 GAD1 were highly expressed in excitatory and inhibitory neurons respectively, but
360 poorly expressed in glial cell types. These results indicated that bMIND is the best
361 method for generating cell-type-specific gene expression data for each sample directly
362 from bulk tissue data.

363

364 The deconvolution performance on rare cell types is in general poor and capturing rare
365 cell types is thus a challenge for cell deconvolution. The accuracy of deconvoluting cell
366 proportions decreased sharply when cell proportion was less than 5%. Similarly, the
367 accuracy of estimated gene expressions was low for rare cell types in brains, such as
368 endothelial cells (correlation coefficient between deconvoluted expressions and
369 ground truth = 0.33) and pericytes (correlation coefficient = 0.43). The low abundance
370 of rare cell types may be masked by dominant cell types in the bulk tissue. Rare cell
371 types may need to be studied using RNAseq of sorted cells, or high coverage
372 sc/snRNAseq, or techniques that enrich for rare cell types.

373

374 The most important benefit of cell deconvolution was that, more cell-type eQTLs were
375 identified using deconvoluted data than using single-cell data and bulk-tissue data.

376 Cell-type eQTLs have been generated with sc/snRNAseq data(42, 43). However, the
377 number of individuals profiled was typically limited(11). To date, a total of 7,607 eQTLs
378 have been identified in the largest single-cell eQTL study(38) comprising 192 human
379 brain samples. Besides the small sample size, the quality of single-cell data can be
380 potentially affected by poor expression quantification, with serious dropout issues and
381 high technical variability(9). Consequently, the eQTLs identified from sc/snRNAseq
382 data may be affected. In contrast, this study deconvoluted data from 861 human brain
383 and mapped 25,273 decon-eQTLs, far more than eQTLs identified from single-cell
384 studies. The union for decon-eQTLs of all cell types was more than bulk-tissue eQTLs
385 (n=9,148). A total of 24.3% ~ 36.89% of top decon-eQTLs were not detected by bulk-
386 eQTL mapping. This indicates that many cell-type-specific eQTLs are buried in bulk-
387 tissue data since the expression of diverse cell types is mixed. Overall, decon-eQTLs
388 could be replicated in bulk-tissue eQTLs (averaged $Pi1=0.99$) and single-cell eQTLs
389 (averaged $Pi1=0.38$), indicating the reliability of eQTLs identified in deconvoluted data.
390 Sample-wise deconvolution provides a valuable opportunity to study genetic
391 regulations in specific cell types with comparable power to bulk-tissue eQTL studies.
392

393 Decon-eQTLs explained SCZ GWAS heritability that was missed by single-cell and
394 bulk-tissue eQTLs. Nearly six times more SCZ GWAS heritability was explained by
395 decon-eQTLs than by single-cell eQTLs. Integrating decon-eQTLs and bulk-tissue
396 eQTLs explained 63% of SCZ GWAS heritability, which was 14% more than heritability
397 explained only by bulk-tissue eQTLs. These results suggested that SCZ GWAS risk

398 may be mediated by genetic regulations in specific cell types, and such an effect can
399 be captured by deconvoluted data.

400

401 Risk genes associated with SCZ GWAS can be revealed by decon-eQTL mapping.

402 Identification of genetically risk genes and pathways in specific cell types is an
403 essential application of decon-eQTLs. For example, we identified that the association
404 between rs12466331 and CALM2 was significant in excitatory neurons but not in bulk-

405 tissue data (Fig. S8A). CALM2, a gene encoding calmodulin, is highly expressed in
406 excitatory neurons (Fig. S8B) and has been found downregulated in the postmortem
407 brains of SCZ patients(44). Moreover, rs12466331 was colocalized with SCZ GWAS

408 risk locus rs144040771 (Fig. S8C). These data suggest that rs12466331 may regulate
409 the expression of CALM2 in excitatory neurons and that dysregulation of such pathway

410 may be associated with SCZ. Thus, mapping decon-eQTLs enabled the discovery of
411 the genetic risk of disease and helped identify their molecular mechanisms in specific
412 cell types.

413

414 Cell-type eQTLs mapping with deconvoluted data is an advanced alternative for ieQTL
415 mapping. ieQTLs are the results of interaction between genetic regulation and cell-
416 type enrichment, while decon-eQTLs are based on deconvoluted data, which are the
417 direct relationship between genotypes and cell type expression for each SNP-gene
418 pair. Both ieQTLs and decon-eQTLs were mapped with ROSMAP data in the current
419 study. Nearly twenty times more decon-eQTLs (n=27,339) were identified than ieQTLs

420 (n=1,822) for the same sample size. Moreover, decon-eQTL is more robust than ieQTL.

421 Compared to single-cell eQTLs, the replication rate of decon-eQTLs (averaged

422 $P_{i1}=0.38$) is clearly superior than ieQTLs (averaged $P_{i1}=0.16$, Fig.S9).

423

424 This study offers a practical guideline for conducting brain cell deconvolution. Using

425 dtangle to estimate cell proportions and bMIND to estimate cell-type expressions is

426 recommended. Rare cell types (proportion<5%) are not recommended to be included

427 in cell deconvolution analysis.

428

429 This study has several limitations. The results were based on the analysis of human

430 brain data with specific parameters tested. More tests may be needed to generalize

431 the conclusion to other tissues and situations. Unsupervised deconvolution methods

432 have not been evaluated by this study. This evaluation only focused on the major cell

433 types in brains, and the deconvolution performance of cell subtypes could be further

434 explored to validate our findings.

435

436 Conclusion

437 This study comprehensively evaluated the commonlu-used methods for sample-wise

438 deconvolution of cell proportions and cell type gene expressions. The downstream

439 analysis of eQTL mapping, GWAS heritability enrichment, and differential expression

440 was also evaluated. Our analysis is a crucial methodological foundation for other

441 studies where deconvolution can be used. A practical guideline is offered for a broad

442 community interested in cell-type-specific studies of brain functions and disorders
443 when only bulk-tissue transcriptome is available.

444

445 **Materials and Methods**

446 **Data processing**

447 Bulk-tissue RNAseq data. Three RNAseq data from brain tissues and brain organoids
448 were used (Table 1). TMM normalization(45) was applied to the raw counts data and
449 log-transformed counts per million reads mapped (CPM) were used. Gene with
450 $\log_2\text{CPM} > 0.1$ in at least 25% of samples were retained. Connectivity between samples
451 was calculated by weighted correlation network analysis (WGCNA)(46) and z-score
452 was normalized. Samples with z-score connectivity $< (-3)$ were labeled as outliers and
453 were removed from downstream analysis. Data were then quantile normalized with the
454 preprocessCore(47) package. The batch effect was corrected with combat in the sva
455 package(48).

456 Sc/snRNAseq data. The processed count matrix and metadata were used. The
457 ROSMAP snRNAseq data were downloaded from
458 <https://www.synapse.org/#/Synapse:syn18681734>. For CMC snRNAseq data, the
459 processing pipeline can be found in the Capstone paper (syn48958066). scRNAseq
460 data of brain organoids can be found in Jourdon et al(35).

461 IHC data. IHC data were downloaded from
462 <https://github.com/ellispatrick/CortexCellDeconv>. Cell proportions were normalized
463 according to the sum-to-1 constraint.

464 **Construction of references and pseudo-bulks**

465 Two types of references were used. For DSA, dtangle, OLS, and CIBERSORT, pooled-
466 cell reference is required. To build a pooled-cell reference, the count matrix was
467 averaged by cell types. The averaged counts matrix was normalized into CPM and
468 was log2-transformed. For MuSiC and Bisque, a single-cell reference was used, which
469 was the gene-by-cell count matrix. To build pseudo-bulks, the count matrix was
470 summed by cell types and by individuals.

471 **Marker gene identification**

472 Marker genes were identified at the cell level and pseudo-bulk level. One versus
473 second high strategy was used. For each gene, the expression difference between the
474 cell type with the highest expression and the cell type with the second highest
475 expression was calculated. At the cell level, marker genes were identified with
476 Seurat(49). Genes having a proportion of zero expression>15% in the target cell type
477 were removed. The Wilcoxon signed-rank test was used to test the expression
478 difference. Genes with $\text{log2FC}>1$ and FDR corrected p value<0.05 were defined as
479 marker genes at the cell level. At the pseudo-bulk level, marker genes were tested in
480 DESeq2(50). The likelihood ratio test was used to test the expression difference
481 between the two cell groups. Marker genes with $\text{log2FC}>2$ and FDR-corrected p-value
482 <0.05 were defined as marker genes at the pseudo-bulk level.

483 **Estimation of cell proportions**

484 Three inputs were required for all deconvolution methods: bulk tissue data, reference,
485 and marker genes. Batch-corrected data was used as input for bulk tissue data. The

486 intersected genes between marker genes at the cell level and pseudo-bulk level were
487 used as input of marker genes. For DSA, dtangle, OLS, and CIBERSORT, pooled-cell
488 reference was used. For MuSiC and Bisque, single-cell reference was used. The
489 genes that have no expression variation were removed from the reference.

490 **Evaluation of cell proportions**

491 Two ground truths were used to evaluate estimated cell proportions: cell proportions
492 from IHC data and cell proportions from sc/snRNAseq data. For one sample, cell
493 proportions from sc/snRNAseq were calculated by dividing the number of cells of one
494 specific cell type by the total number of cells. RMSE was used as an evaluation metric.
495 The formula of RMSE is:

$$496 \quad RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}}$$

497 y_i is the estimated cell proportion and \hat{y} is the ground truth. n is the number of cell
498 types in the sample-level evaluation, and n is the number of samples in the cell-type-
499 level evaluation.

500 **Estimation of cell-type expression**

501 Batch-corrected data was used as input for bulk tissue data. The proportion from DSA,
502 dtangle, OLS, CIBERSORT, MuSiC, and Bisque was used independently. Pooled-cell
503 reference was used as prior for bMIND and swCAM. For TCA, only bulk-tissue data
504 and cell proportions were used to estimate cell type expressions for each sample. Data
505 were transformed into a log scale for bMIND and TCA and a linear scale for swCAM.

506 **Evaluation of cell type expression**

507 To construct ground truth for evaluating estimated expressions, averaged counts by

508 cell types in each sample were calculated. Then the averaged cell type expressions
509 were normalized into CPM and were log2-transformed. Sample-to-sample spearman
510 correlation was tested between estimated expression and ground truth for each cell
511 type.

512 **Genotyping quality control**

513 The ROSMAP whole genome sequencing (WGS) dataset was downloaded from
514 <https://www.synapse.org/#/Synapse:syn11724057>. The data is already imputed. Only
515 individuals with both genotype and deconvolution results were retained for the eQTL
516 analysis. SNPs with minor allele frequency (MAF) <5% or deviating from Hardy–
517 Weinberg equilibrium ($P < 1 \times 10^{-6}$) were excluded. After quality control, we obtained
518 high-quality genotypes for ~4.9 million SNPs (MAF > 5%) in 861 individuals.

519 **eQTL mapping**

520 decon-eQTLs. To identify decon-eQTLs, we tested the associations between
521 genotypes and deconvoluted expressions. We mapped cis-eQTLs within a 1-Mb
522 window of the TSS of each gene using QTLtools(51). For each gene, QTLtools
523 performs permutations of the expression data and records the best p-value for each
524 SNP in the cis window after each permutation. We used estimated cell-type expression
525 by bMIND as phenotype data. Phenotype data of eight cell types were tested
526 independently. Quantile normalization was used for normalizing expression matrixes
527 before eQTL mapping. PEER was used to identify hidden covariates in the data(52).
528 8-35 PEER factors were included as covariates in eQTL mapping.
529 Bulk-tissue eQTLs. To map bulk-tissue eQTLs, the same eQTL mapping procedure

530 was performed on the bulk-tissue expression data. 33 PEER factors were included as
531 covariates in eQTL mapping.

532 **Replication of decon-eQTLs in BrainGVEX data**

533 To replicate decon-eQTLs, we deconvoluted RNAseq data from BrainGVEX(36) and
534 mapped eQTLs with the deconvoluted data. 430 brain samples with both genotypes
535 and RNAseq data were used. dtangle was used to estimate cell proportions, with the
536 marker genes and the reference from ROSEMAP sn/RNAseq data. Then, bulk-tissue
537 data were deconvoluted into cell-type expressions for eight major cell types with
538 bMIND. The same eQTL mapping process was performed on the deconvoluted data
539 to identify decon-eQTLs in BrainGVEX data.

540 The proportion of true associations (π_1) in the qvalue package(37) was used to
541 measure the replicate rate of significant decon-eQTLs in ROSEMAP data in decon-
542 eQTLs in BrainGVEX data. With the distribution of corresponding p values for the
543 overlapped eSNP-eGene pairs in two datasets, we calculated π_0 , i.e., the proportion
544 of true null associations based on distribution. Then, $\pi_1 = 1 - \pi_0$ estimated the lowest
545 bound for true-positive associations.

546 **Replication of decon-eQTLs in single-cell eQTLs**

547 To measure the replication rate of decon-eQTLs in the sc/snRNAseq dataset, we
548 downloaded cell-type eQTLs identified from the snRNAseq data of 192 individuals(38).
549 With the single-cell eQTLs as a reference, π_1 statistics were calculated for eight cell
550 types independently.

551 **SCZ heritability enrichment**

552 Stratified linkage disequilibrium score regression(39) (S-LDSC) was used to calculate
553 SCZ GWAS heritability enrichment in decon-eQTLs. GWAS summary statistics from
554 three published SCZ studies were downloaded(53-55). Conditional analysis was
555 performed on decon-eQTLs to select the top SNP for each gene ($r^2 > 0.2$ in 1000
556 Genomes European individuals(56)). Then script ldsc.py with the “--l2” parameter was
557 used to generate the gene-set-specific annotation and LD score files. Then ldsc.py
558 with the “--h²-cts” parameter was used to generate stratified heritability by decon-
559 eQTLs of eight cell types.

560 **Co-localization**

561 For each gene in decon-eQTLs, the co-localization between eSNP and SCZ GWAS
562 signals(57) was tested. The ‘coloc.abf’ function in the Coloc(58) package (version 5.1.0)
563 was used for testing. The threshold for significance is SNP.PP.H4>0.95.

564 **Differential expression analysis**

565 Differential expression analysis was performed on deconvoluted data and bulk-tissue
566 data to identify genes associated with AD, SCZ, and brain development. For AD and
567 SCZ, differential expression analysis was conducted in each cell type with the
568 Wilcoxon rank-sum test. For brain development, the linear regression model was used
569 to identify genes showing significant expression changes. The p values were corrected
570 by FDR. Genes with FDR q value <0.05 were identified as phenotype-associated
571 genes (PAGs).

572 To compare deconvoluted PAGs and PAGs from sc/snRNAseq data, PAGs for AD(23)
573 and SCZ(59) in cell types were downloaded. For brain development, PAGs were

574 identified in pseudo-bulk data. The linear regression model was used to identify PAGs
575 for each cell type independently. The p values were corrected by FDR.

576 **Data availability**

577 The source data described in this manuscript are available via the PsychENCODE
578 Knowledge Portal (<https://psychencode.synapse.org/>). The PsychENCODE
579 Knowledge Portal is a platform for accessing data, analyses, and tools generated
580 through grants funded by the National Institute of Mental Health (NIMH)
581 PsychENCODE Consortium. Data is available for general research use according to
582 the following requirements for data access and data attribution:
583 (<https://psychencode.synapse.org/DataAccess>). For access to content described in this
584 manuscript see: <https://www.synapse.org/#!Synapse:syn51072187/datasets/>. The eQTL
585 and PAG results can be accessed at
586 <https://www.synapse.org/#!Synapse:syn50908925>.

587

588 **Acknowledgment**

589 We thank Richard Kopp at SUNY Upstate Medical University for his help in polishing
590 words. We thank all the participants involved in the ROSEMAP and PsychENCODE
591 study for making the data available. This work was supported by NIH grants
592 U01MH122591, U01MH116489, R01MH110920, U01MH103340, R01MH126459 and
593 R01MH109648; the Simons Foundation 632742; the National Natural Science
594 Foundation of China 82022024 and 31970572; the science and technology innovation
595 Program of Hunan Province 2021RC4018 and 2021RC5027. Data were generated as

596 part of the PsychENCODE Consortium, supported by: U01DA048279, U01MH103339,
597 U01MH103340, U01MH103346, U01MH103365, U01MH103392, U01MH116438,
598 U01MH116441, U01MH116442, U01MH116488, U01MH116489, U01MH116492,
599 U01MH122590, U01MH122591, U01MH122592, U01MH122849, U01MH122678,
600 U01MH122681, U01MH116487, U01MH122509, R01MH094714, R01MH105472,
601 R01MH105898, R01MH109677, R01MH109715, R01MH110905, R01MH110920,
602 R01MH110921, R01MH110926, R01MH110927, R01MH110928, R01MH111721,
603 R01MH117291, R01MH117292, R01MH117293, R21MH102791, R21MH103877,
604 R21MH105853, R21MH105881, R21MH109956, R56MH114899, R56MH114901,
605 R56MH114911, R01MH125516, R01MH126459, R01MH129301, R01MH126393,
606 R01MH121521, R01MH116529, R01MH129817, R01MH117406, and P50MH106934
607 awarded to: Alexej Abyzov, Nadav Ahituv, Schahram Akbarian, Kristin Brennand,
608 Andrew Chess, Gregory Cooper, Gregory Crawford, Stella Dracheva, Peggy Farnham,
609 Michael Gandal, Mark Gerstein, Daniel Geschwind, Fernando Goes, Joachim F.
610 Hallmayer, Vahram Haroutunian, Thomas M. Hyde, Andrew Jaffe, Peng Jin, Manolis
611 Kellis, Joel Kleinman, James A. Knowles, Arnold Kriegstein, Chunyu Liu, Christopher
612 E. Mason, Keri Martinowich, Eran Mukamel, Richard Myers, Charles Nemeroff, Mette
613 Peters, Dalila Pinto, Katherine Pollard, Kerry Ressler, Panos Roussos, Stephan
614 Sanders, Nenad Sestan, Pamela Sklar, Michael P. Snyder, Matthew State, Jason Stein,
615 Patrick Sullivan, Alexander E. Urban, Flora Vaccarino, Stephen Warren, Daniel
616 Weinberger, Sherman Weissman, Zhiping Weng, Kevin White, A. Jeremy Willsey,
617 Hyejung Won, and Peter Zandi. The authors declare that they have no competing

618 interests.

619

620 **References**

- 621 1. G. T. Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
- 622 2. E. C. Psych *et al.*, The PsychENCODE project. *Nat Neurosci* **18**, 1707-1712 (2015).
- 623 3. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
- 624 4. J. A. Miller *et al.*, Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206 (2014).
- 625 5. D. A. Bennett *et al.*, Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis* **64**, S161-S189 (2018).
- 626 6. S. Darmanis *et al.*, A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).
- 627 7. Y. Zhang *et al.*, Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).
- 628 8. D. Lahnemann *et al.*, Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020).
- 629 9. P. Brennecke *et al.*, Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093-1095 (2013).
- 630 10. J. Cheng, J. Liao, X. Shao, X. Lu, X. Fan, Multiplexing Methods for Simultaneous Large-Scale Transcriptomic Profiling of Samples at Single-Cell Resolution. *Adv Sci (Weinh)* **8**, e2101229 (2021).
- 631 11. M. Maria, N. Pouyanfar, T. Ord, M. U. Kaikkonen, The Power of Single-Cell RNA Sequencing in eQTL Discovery. *Genes (Basel)* **13**, (2022).
- 632 12. F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, K. De Preter, Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun* **11**, 5650 (2020).
- 633 13. H. Jin, Z. Liu, A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol* **22**, 102 (2021).
- 634 14. B. B. Nadel *et al.*, Systematic evaluation of transcriptomics-based deconvolution methods and references using thousands of clinical samples. *Brief Bioinform* **22**, (2021).
- 635 15. G. J. Sutton *et al.*, Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat Commun* **13**, 1358 (2022).
- 636 16. E. Patrick *et al.*, Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput Biol* **16**, e1008120 (2020).
- 637 17. Y. Zhong, Y. W. Wan, K. Pang, L. M. Chow, Z. Liu, Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
- 638 18. A. M. Newman *et al.*, Robust enumeration of cell subsets from tissue expression

659 profiles. *Nat Methods* **12**, 453-457 (2015).

660 19. G. J. Hunt, S. Freytag, M. Bahlo, J. A. Gagnon-Bartsch, dtangle: accurate and robust
661 cell type deconvolution. *Bioinformatics* **35**, 2093-2099 (2019).

662 20. X. Wang, J. Park, K. Susztak, N. R. Zhang, M. Li, Bulk tissue cell type deconvolution
663 with multi-subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).

664 21. X. Wang *et al.*, Deciphering cellular transcriptional alterations in Alzheimer's disease
665 brains. *Mol Neurodegener* **15**, 38 (2020).

666 22. G. Pei *et al.*, Gene expression imputation and cell-type deconvolution in human brain
667 with spatiotemporal precision and its implications for brain-related disorders. *Genome*
668 *Res* **31**, 146-158 (2021).

669 23. H. Mathys *et al.*, Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**,
670 332-337 (2019).

671 24. D. Velmeshev *et al.*, Single-cell genomics identifies cell type-specific molecular
672 changes in autism. *Science* **364**, 685-689 (2019).

673 25. C. Nagy *et al.*, Single-nucleus transcriptomics of the prefrontal cortex in major
674 depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons.
675 *Nat Neurosci* **23**, 771-781 (2020).

676 26. U. Pfisterer *et al.*, Identification of epilepsy-associated neuronal subtypes and gene
677 expression underlying epileptogenesis. *Nat Commun* **11**, 5038 (2020).

678 27. S. Kim-Hellmuth *et al.*, Cell type-specific genetic regulation of gene expression across
679 human tissues. *Science* **369**, (2020).

680 28. J. Wang, K. Roeder, B. Devlin, Bayesian estimation of cell type-specific gene
681 expression with prior derived from single-cell data. *Genome Res* **31**, 1807-1818 (2021).

682 29. L. Chen *et al.*, swCAM: estimation of subtype-specific expressions in individual
683 samples with unsupervised sample-wise deconvolution. *Bioinformatics* **38**, 1403-1410
684 (2022).

685 30. E. Rahmani *et al.*, Cell-type-specific resolution epigenetics without the need for cell
686 sorting or single-cell biology. *Nat Commun* **10**, 3417 (2019).

687 31. T. P. Morris, I. R. White, M. J. Crowther, Using simulation studies to evaluate statistical
688 methods. *Stat Med* **38**, 2074-2102 (2019).

689 32. B. Jew *et al.*, Accurate estimation of cell composition in bulk expression through robust
690 integration of single-cell information. *Nat Commun* **11**, 1971 (2020).

691 33. P. L. De Jager *et al.*, A multi-omic atlas of the human frontal cortex for aging and
692 Alzheimer's disease research. *Sci Data* **5**, 180142 (2018).

693 34. G. E. Hoffman *et al.*, CommonMind Consortium provides transcriptomic and
694 epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data* **6**, 180 (2019).

695 35. A. Jourdon *et al.*, ASD modelling in organoids reveals imbalance of excitatory cortical
696 neuron subtypes during early neurogenesis. *bioRxiv*, 2022.2003.2019.484988 (2023).

697 36. D. Wang *et al.*, Comprehensive functional genomic resource and integrative model for
698 the human brain. *Science* **362**, (2018).

699 37. D. A. a. R. D. Bass JDSwcfAJ, qvalue: Q-value estimation for false discovery rate
700 control. *R package version* 2.2.2, (2015).

701 38. J. Bryois *et al.*, Cell-type-specific cis-eQTLs in eight human brain cell types identify
702 novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**, 1104-1112

703 703 (2022).

704 39. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-
705 wide association summary statistics. *Nat Genet* **47**, 1228-1235 (2015).

706 40. J. Rammohan *et al.*, Comparison of bias and resolvability in single-cell and single-
707 transcript methods. *Commun Biol* **4**, 659 (2021).

708 41. E. Denisenko *et al.*, Systematic assessment of tissue dissociation and storage biases
709 in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21**, 130 (2020).

710 42. D. Neavin *et al.*, Single cell eQTL analysis identifies cell type-specific genetic control of
711 gene expression in fibroblasts and reprogrammed induced pluripotent stem cells.
Genome Biol **22**, 76 (2021).

712 43. S. Yazar *et al.*, Single-cell eQTL mapping identifies cell type-specific genetic control of
713 autoimmune disease. *Science* **376**, eabf3041 (2022).

714 44. J. M. Nascimento, D. Martins-de-Souza, The proteome of schizophrenia. *NPJ Schizophr* **1**, 14003 (2015).

715 45. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression
716 analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

717 46. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network
718 analysis. *BMC Bioinformatics* **9**, 559 (2008).

719 47. B. B, A collection of pre-processing functions. *R package version 1.60.1*, (2022).

720 48. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The sva package for
721 removing batch effects and other unwanted variation in high-throughput experiments.
Bioinformatics **28**, 882-883 (2012).

722 49. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587
723 e3529 (2021).

724 50. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion
725 for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

726 51. O. Delaneau *et al.*, A complete tool set for molecular QTL discovery and analysis. *Nat
727 Commun* **8**, 15452 (2017).

728 52. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of
729 expression residuals (PEER) to obtain increased power and interpretability of gene
730 expression analyses. *Nat Protoc* **7**, 500-507 (2012).

731 53. A. F. Pardinas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant
732 genes and in regions under strong background selection. *Nat Genet* **50**, 381-389
733 (2018).

734 54. D. Bipolar, d. r. v. e. Schizophrenia Working Group of the Psychiatric Genomics
735 Consortium. Electronic address, D. Bipolar, C. Schizophrenia Working Group of the
736 Psychiatric Genomics, Genomic Dissection of Bipolar Disorder and Schizophrenia,
737 Including 28 Subphenotypes. *Cell* **173**, 1705-1715 e1716 (2018).

738 55. C. Schizophrenia Working Group of the Psychiatric Genomics, Biological insights from
739 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).

740 56. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**,
741 68-74 (2015).

742 57. V. Trubetskoy *et al.*, Mapping genomic loci implicates genes and synaptic biology in
743 schizophrenia. *Nature* **604**, 502-508 (2022).

744

745

746

747 58. C. Giambartolomei *et al.*, Bayesian test for colocalisation between pairs of genetic
748 association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

749 59. W. B. Ruzicka *et al.*, Single-cell multi-cohort dissection of the schizophrenia
750 transcriptome. *medRxiv*, 2022.2008.2031.22279406 (2022).

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

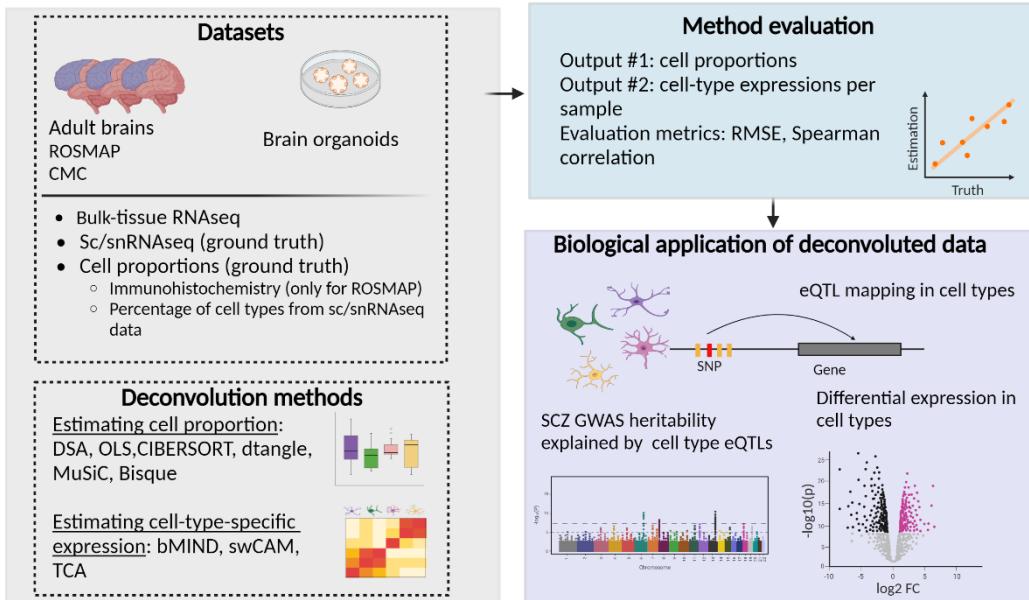
767

768

769

770

771 **Figures and Tables**

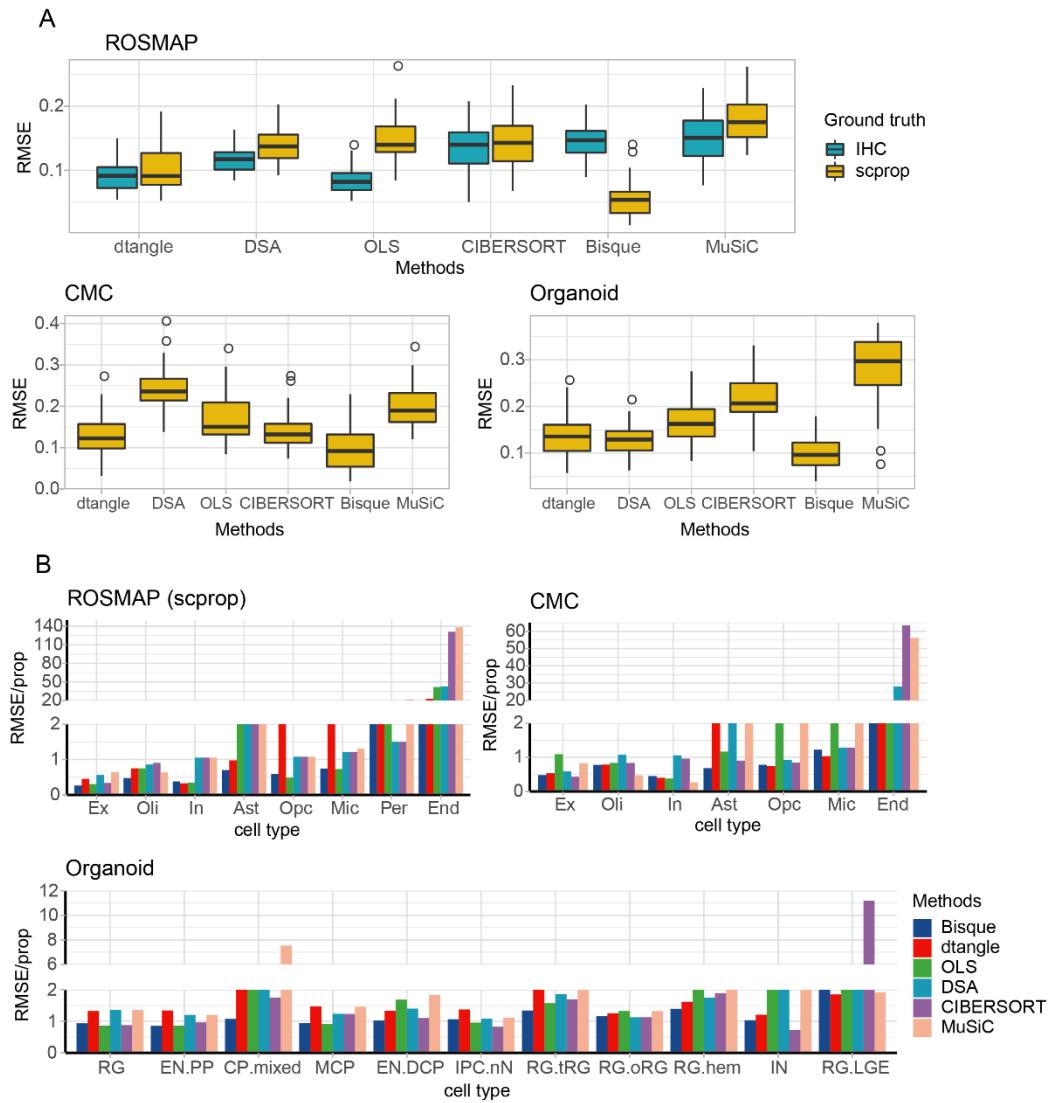


772

773 **Fig. 1. Study overview.**

774

775

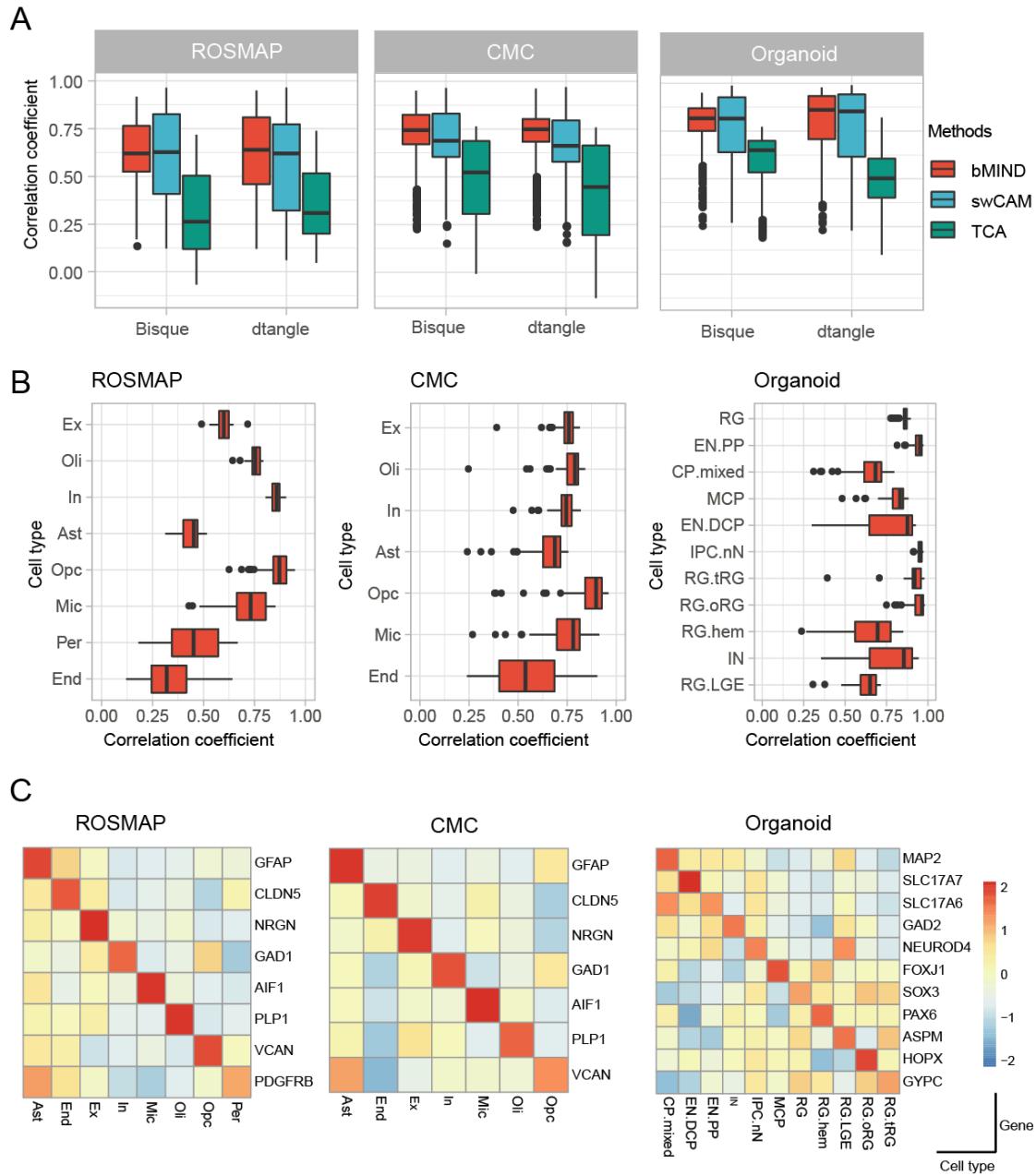


776

777 **Fig. 2. Assessment of cell proportions estimated by examined deconvolution methods.**

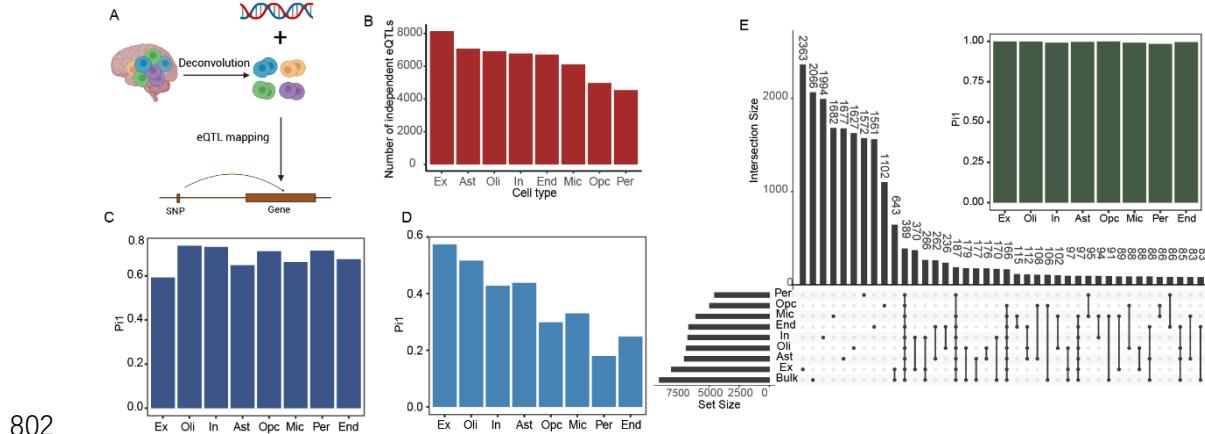
778 (A). Sample-level RMSE values between estimated cell proportions and ground truth. IHC:
779 immunohistochemistry; scprop: cell proportions calculated from sc/snRNAseq data, scprop =
780 the number of cells of specific cell type/number of total cells. (B). Cell-type-level RMSE values
781 between estimated cell proportions and ground truth data. RMSE values were normalized by
782 the value of cell proportions to make them comparable across cell types. Cell types were
783 ordered by cell proportions in a decreasing way. Ex: excitatory neurons, In: inhibitory neurons,
784 Ast: astrocytes, Opc: oligodendrocyte precursor cells, Mic: microglia, Per: pericytes, End:
785 endothelial cells; RG: radial glia, EN.PP: early born excitatory neurons of the pre-plate/subplate,
786 CP.mixed: cortical plate mixed neurons, MCP: medial cortical plate, EN-DCP: dorsal cortical
787 plate excitatory neurons, IPC-nN: intermediate progenitor cell or newborn neuron, RG.tRG:
788 truncated radial glia, RG.oRG: outer radial glia, RG.hem: radial glia in cortical hem, IN: inhibitory
789 neurons, RG-LGE: progenitors corresponding to a putative ventrolateral ganglionic eminence
790 fate.

791



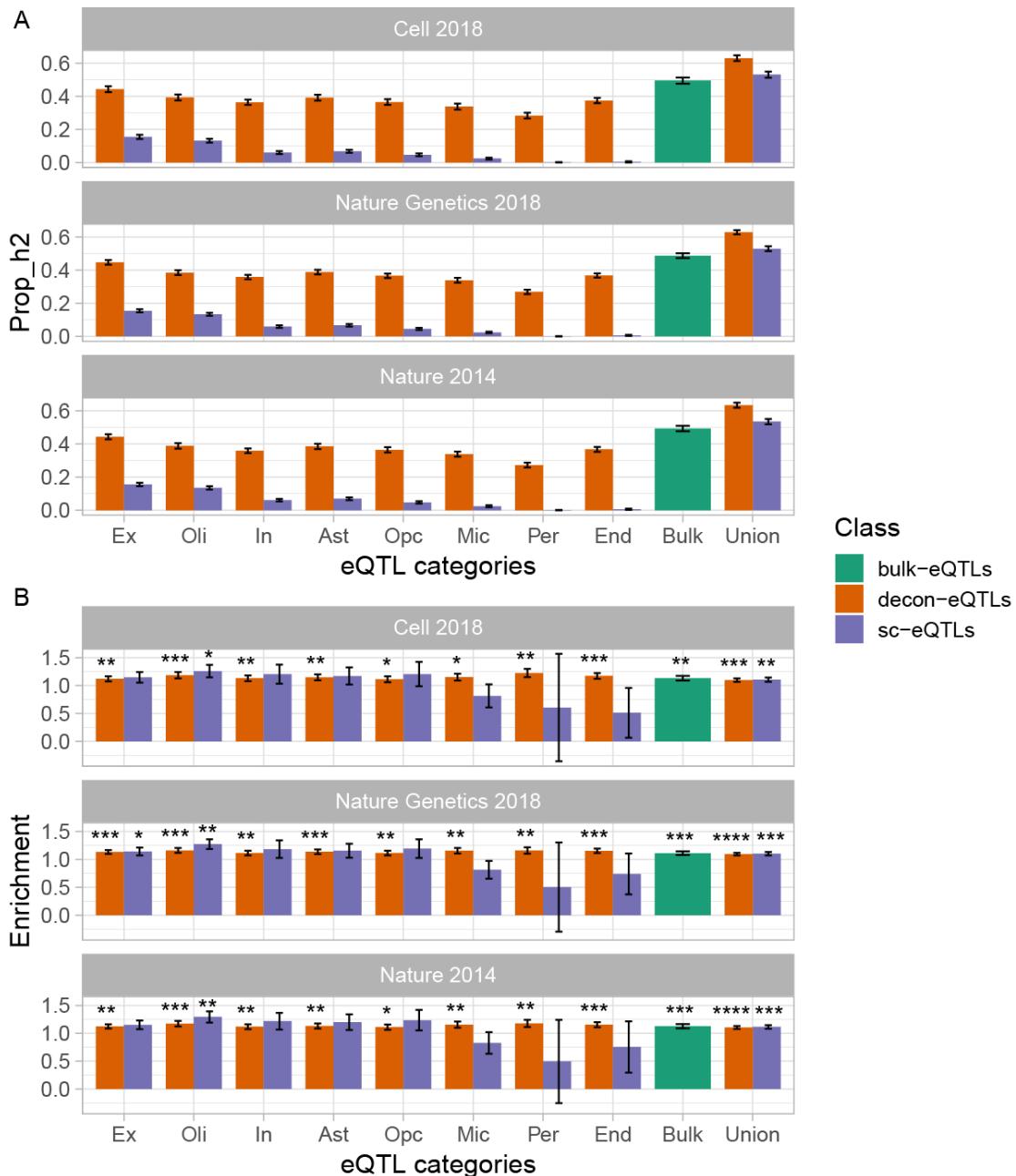
792

793 **Fig. 3. Assessment of sample-wise cell-type expressions deconvoluted from bulk-tissue**
 794 **data. (A)**. Overall assessment of methods for estimating cell-type expressions. Spearman
 795 correlations between deconvoluted data and sc/snRNAseq data from matched samples. The
 796 averaged expression by cell types was used as ground truth. **(B)**. Cell-type-level assessment
 797 of methods for estimating cell-type expressions. Correlations between deconvoluted data by
 798 bMIND and sc/snRNAseq data were calculated for each cell type. Cell proportions estimated
 799 by dtangle were used for input. Cell types on the y-axis were ordered by cell proportions
 800 computed from sc/snRNAseq. **(C)**. Assessment of cell type specificity in estimated expressions.
 801 The figure shows the expression of marker genes in deconvoluted data by bMIND.



802

803 **Fig. 4. Cell-type eQTL mapping based on deconvoluted sample-wise expression data. (A)**
804 Illustration of decon-eQTL mapping. **(B)** The number of decon-eQTLs identified in different cell
805 types at FDR < 0.05 in the permutation test. **(C)** Pi_1 statistics of decon-eQTLs in BrainGVEX
806 decon-eQTLs and **(D)** eQTLs from snRNAseq study (Bryois et al.). **(E)** Comparison of decon-
807 eQTLs and bulk-tissue eQTLs. The top barplot shows the Pi_1 values of decon-eQTLs in bulk-
808 tissue eQTLs. The bottom plot shows the intersections between decon-eQTLs and bulk-tissue
809 eQTLs, as well as intersections of decon-eQTLs across various cell types.



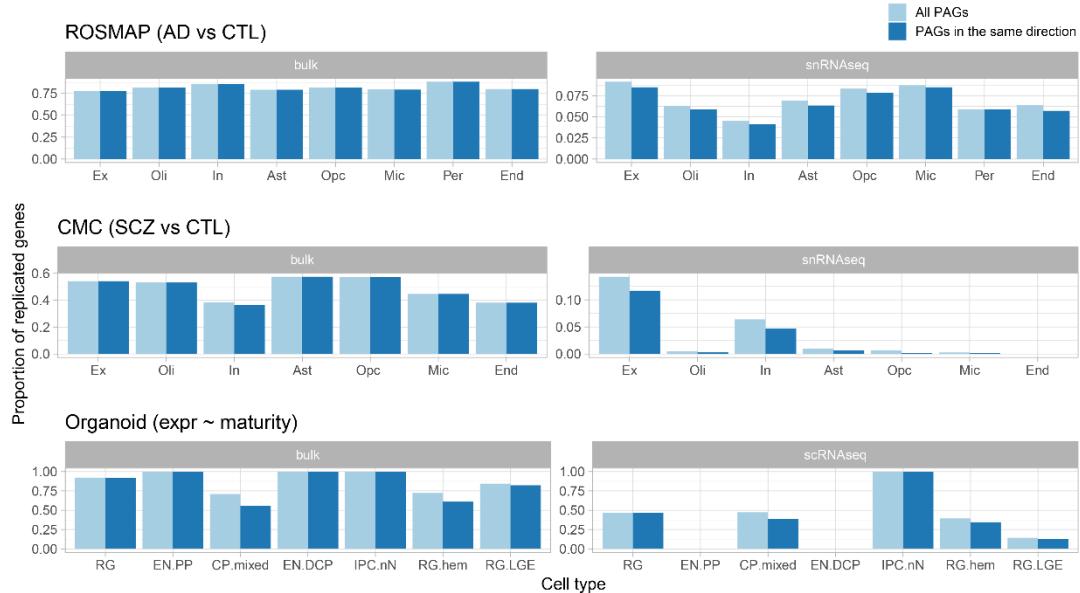
810

811

812

813

Fig. 5. SCZ GWAS heritability explained by cell-type eQTLs and bulk-tissue eQTLs. (A)
Total SCZ GWAS heritability (h^2) explained by eQTLs. **(B)** SCZ GWAS heritability enrichment in eQTLs. Enrichment = h^2 /number of SNPs in each eQTL category.



814

815 **Fig. 6. The proportion of phenotypes-associated genes (PAGs) replicated in bulk-**
816 **tissue data (left panel) and sc/snRNAseq data (right panel).** The light blue bar shows the
817 proportions of replicated PAGs with $FDR < 0.05$. The dark blue bar shows the proportions of
818 replicated PAGs with $FDR < 0.05$ and having the same direction of changes in replication data.
819 expr denotes expression. The maturity of organoids was measured by the days of cell culture.

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843 **Table 1 Datasets used for evaluation**

Study	Brain region	Data type	Sample size	Number of cells	Number of cell types	Number of genes
ROSMAP	PFC	Bulk-tissue RNAseq	1,112	-	-	17,128
	PFC	snRNAseq	48	69,611	8	17,926
	PFC	IHC	49	-	5	-
CMC	PFC	Bulk-tissue RNAseq	572	-	-	25,774
	PFC	snRNAseq	101	569,289	7	33,822
Brain organoid	-	Bulk-tissue RNAseq	130	-	-	20,125
	-	scRNAseq	72	490,844		33,538

844 *PFC: prefrontal cortex

845

846

847

848

849

850

851

852

853

854

855

856

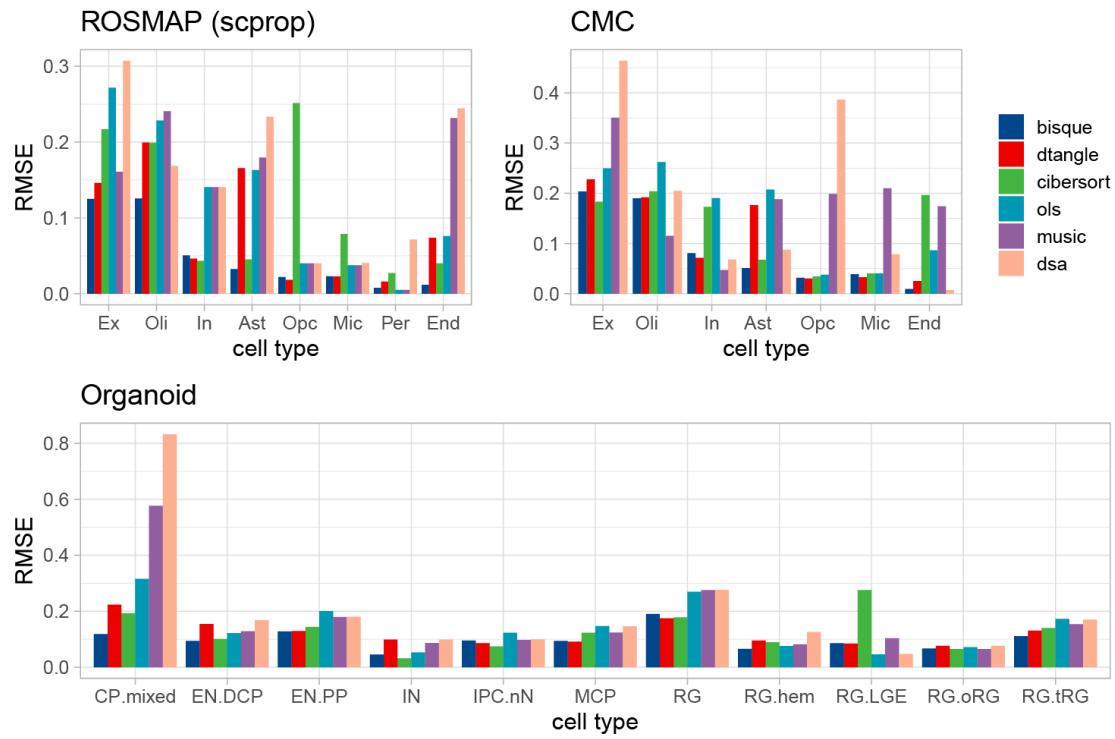
857

858

859

860 **Supplemental Figures**

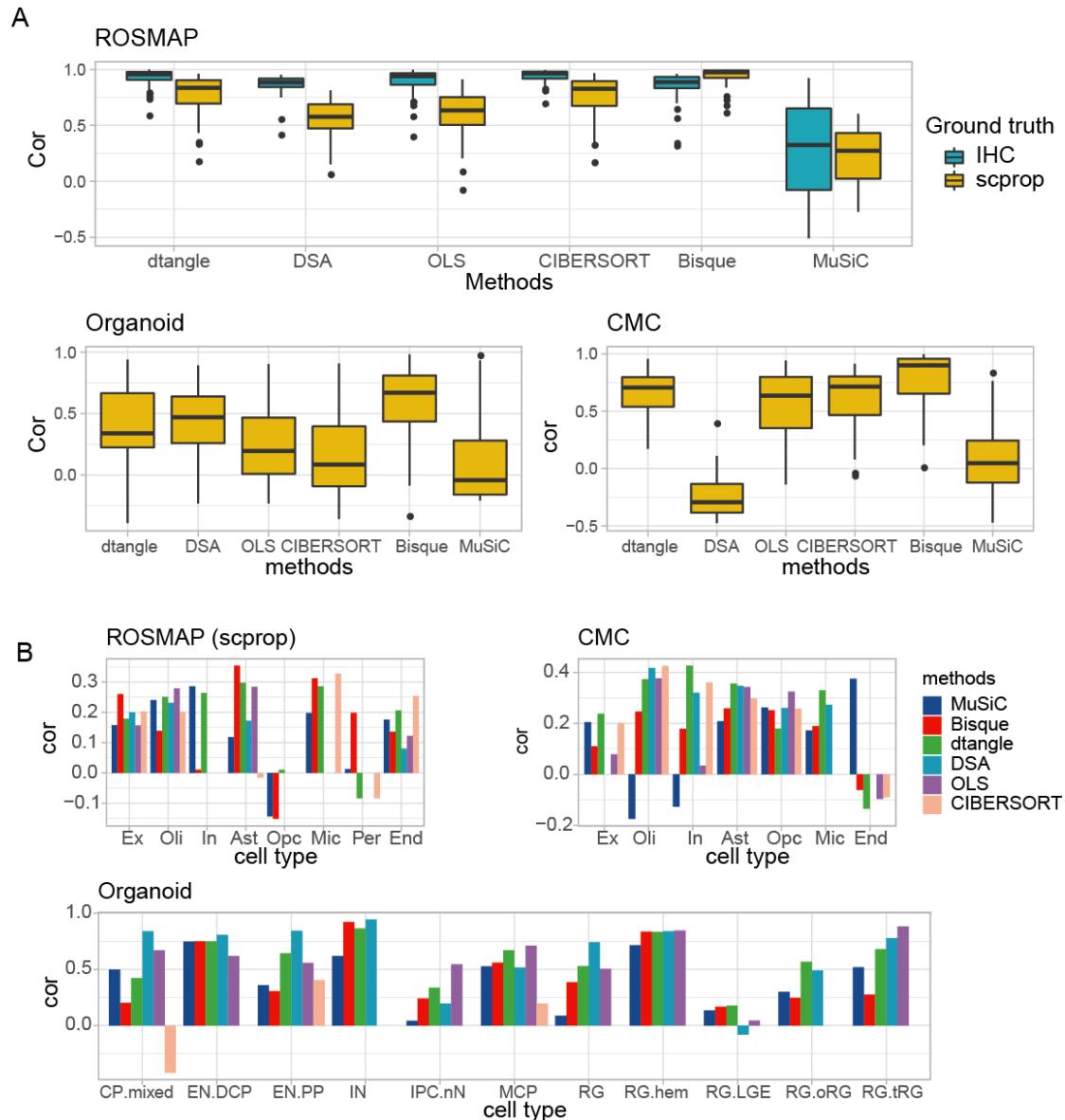
861



862

Fig. S1 Cell-type-level RMSE values between estimated cell proportions and ground truth.

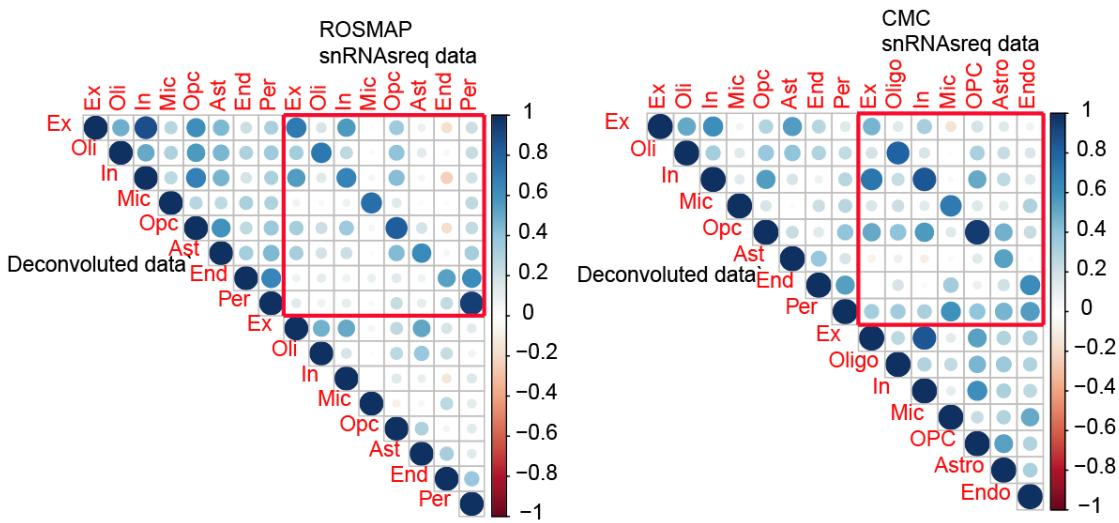
863 This is the full version of Fig. 2B in terms of RMSE values.



865

866 **Fig. S2 Assessment of cell proportions estimated by deconvolution methods based on**
 867 **Spearman correlation. (A).** The sample-level correlation coefficient between estimated cell
 868 proportions and ground truth. IHC: immunohistochemistry; scprop: cell proportions calculated
 869 from sc/snRNASeq data, scprop = the number of cells of specific cell type/number of total cells.
 870 **(B).** The cell-type-level correlation coefficient between estimated cell proportions and ground
 871 truth. Cell types were ordered by cell proportions in a decreasing way. Ex: excitatory neurons,
 872 In: inhibitory neurons, Ast: astrocytes, Opc: oligodendrocyte precursor cells, Mic: microglia, Per:
 873 pericytes, End: endothelial cells; RG: radial glia, EN.PP: early born excitatory neurons of the
 874 pre-plate/subplate, CP.mixed: cortical plate mixed neurons, MCP: medial cortical plate, EN-
 875 DCP: dorsal cortical plate excitatory neurons, IPC.nN: intermediate progenitor cell or newborn
 876 neuron, RG.tRG: truncated radial glia, RG.oRG: outer radial glia, RG.hem: radial glia in cortical
 877 hem, IN: inhibitory neurons, RG-LGE: progenitors corresponding to a putative ventrolateral
 878 ganglionic eminence fate.

879

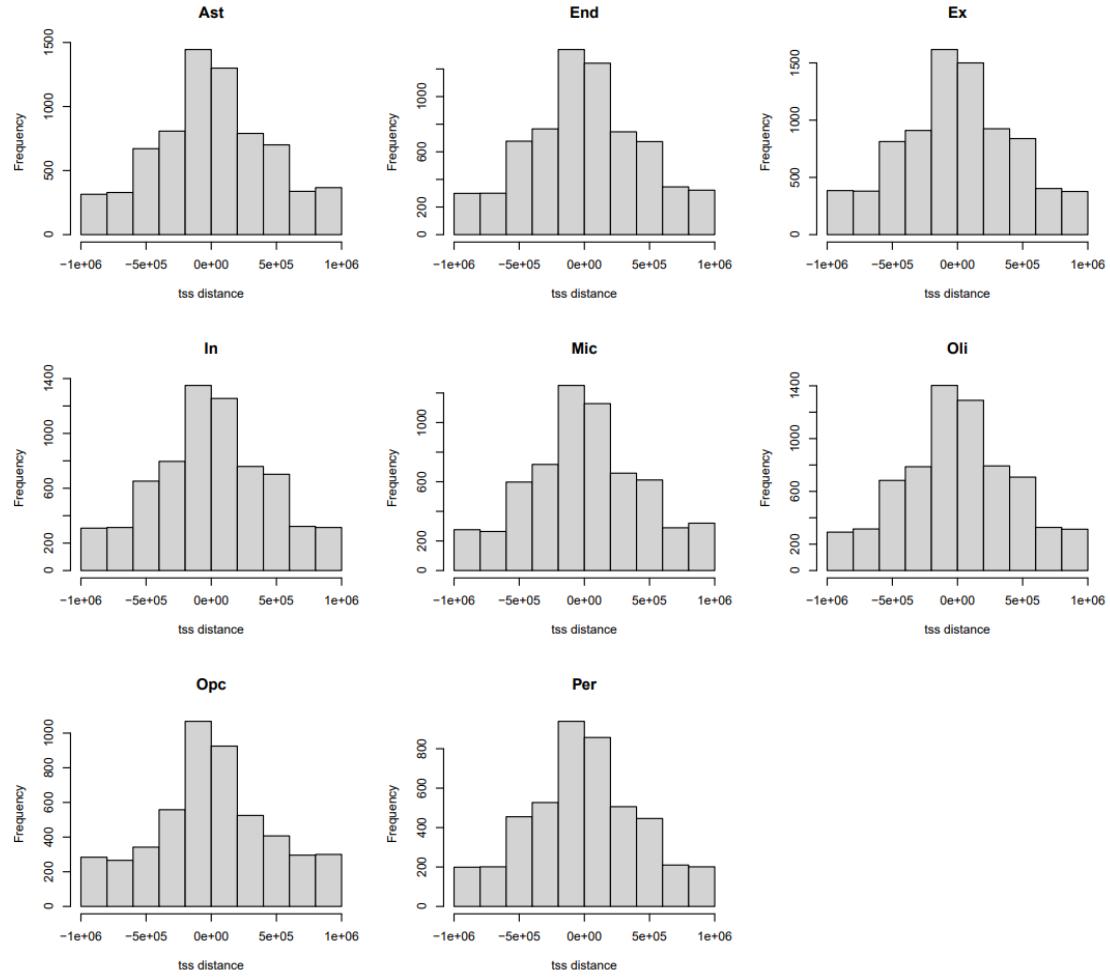


880

881 **Fig. S3. Correlations between deconvoluted expression (ROSMAP) and snRNAseq data**
882 **from ROSMAP and CMC.** For each cell type, averaged expressions across all samples were
883 used.

884

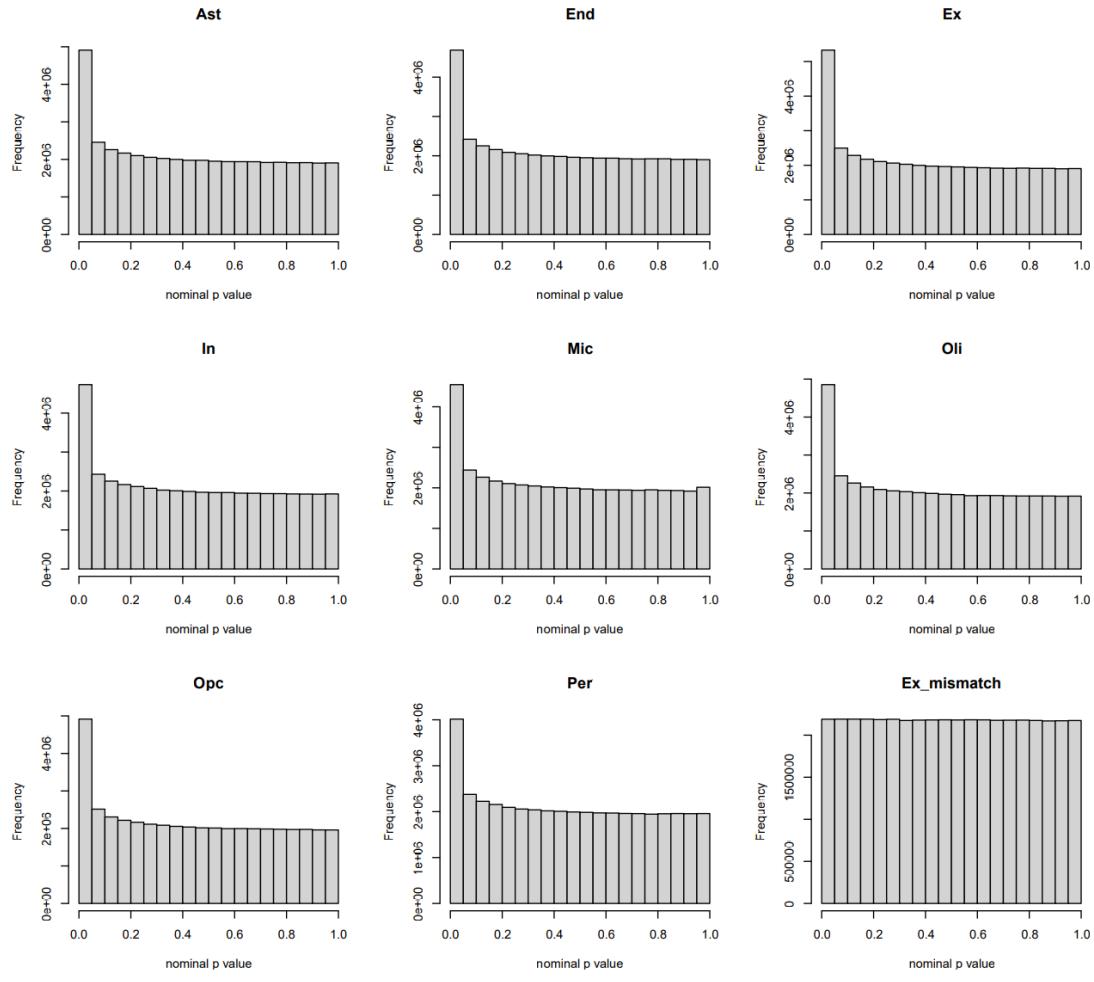
885



886

887

Fig. S4. Distance between transcription start sites (TSS) and eQTL SNPs (eSNPs).



888

889 **Fig. S5. Distribution of decon-eQTL p values.** Ex_mismatch represents eQTL
890 mapping results based on sample-shuffled data of deconvoluted excitatory neurons.

891

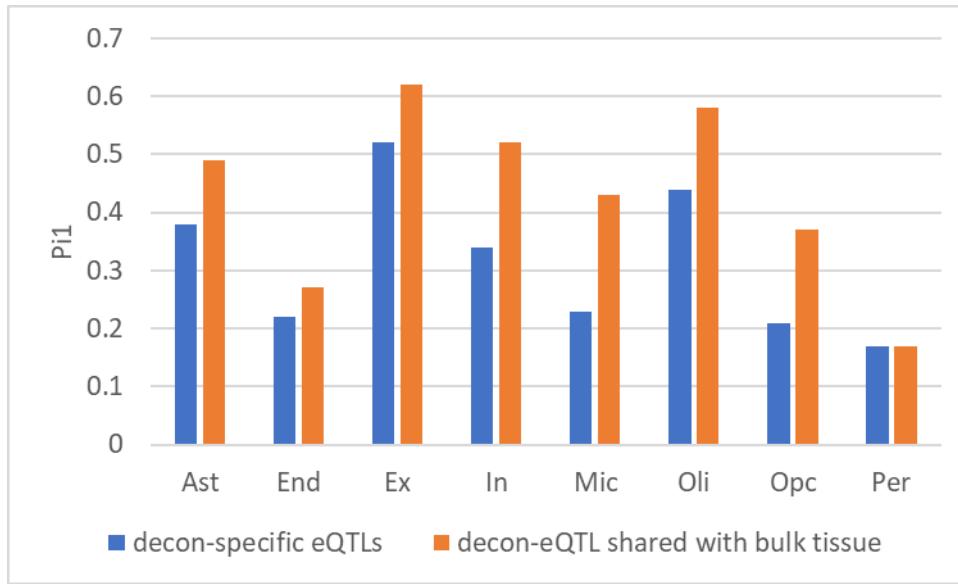
892

893

894

895

896



897
898 **Fig. S6. Replication of decon-eQTLs in single-cell eQTLs.** Wilcoxon signed-rank test was
899 used to test the difference in Pi1 of two eQTL classes. $P=0.15$.

900

901

902

903

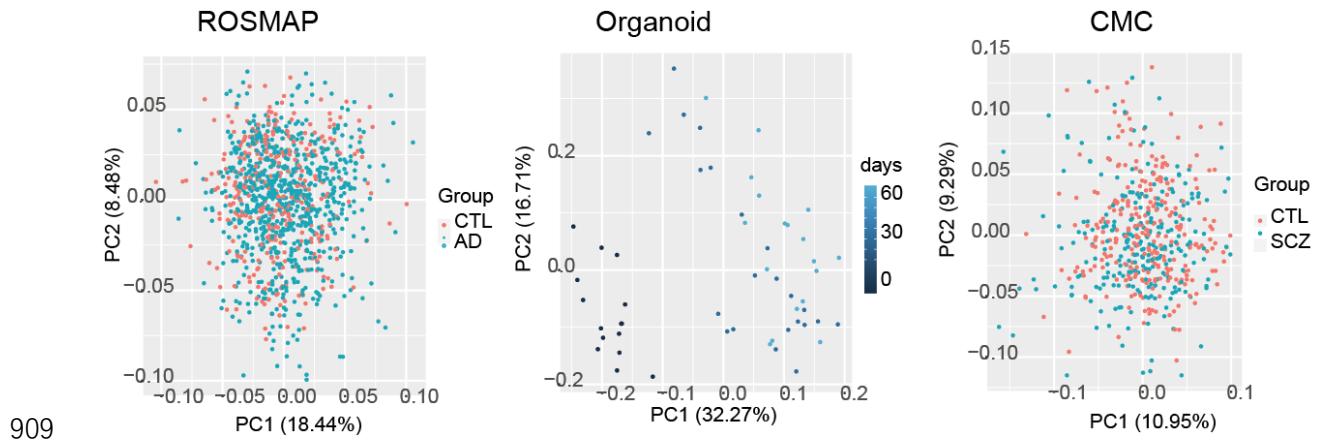
904

905

906

907

908



909

910 **Fig. S7.** PCA plot of samples in bulk-tissue datasets. Batch-corrected were used.

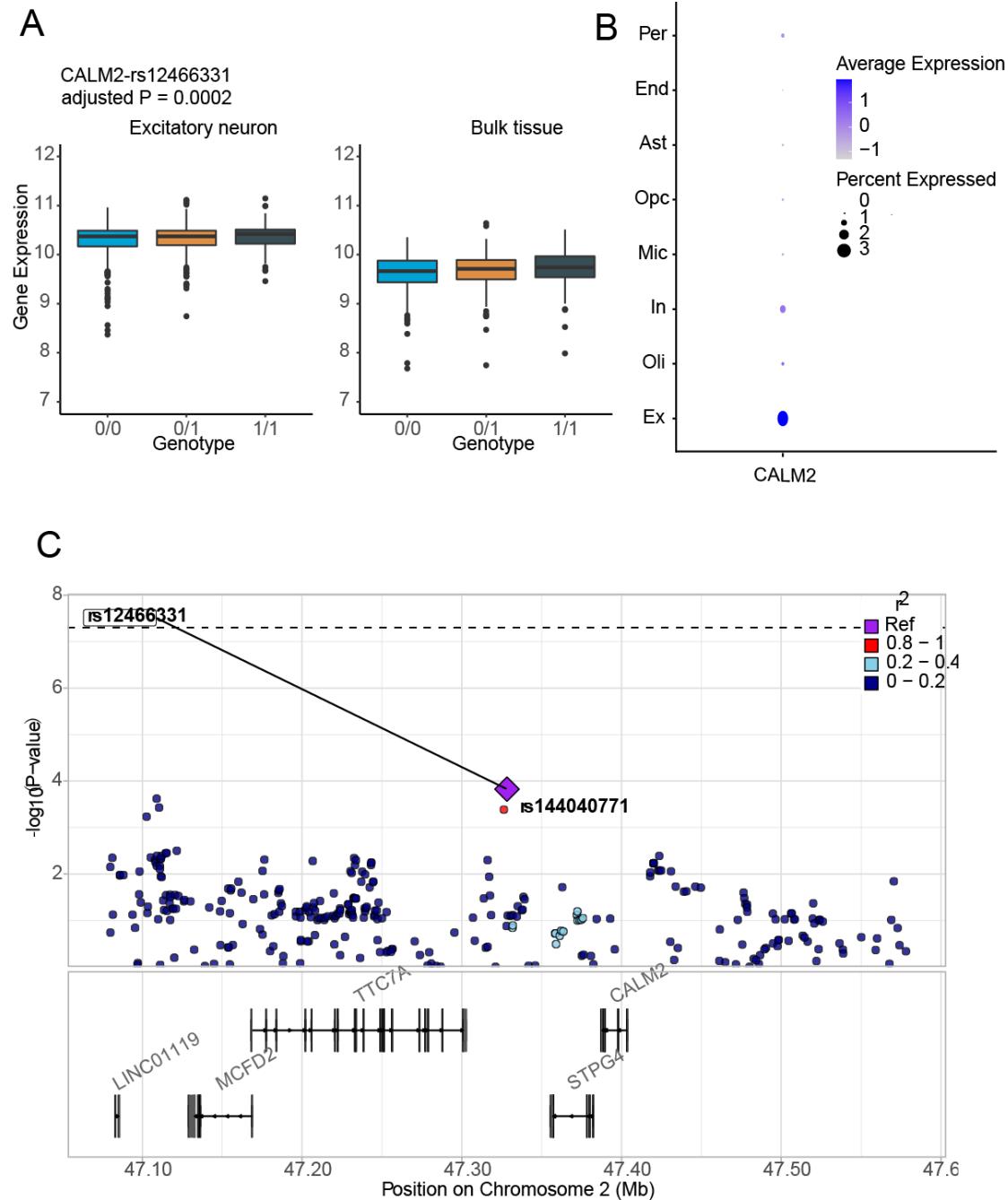
911

912

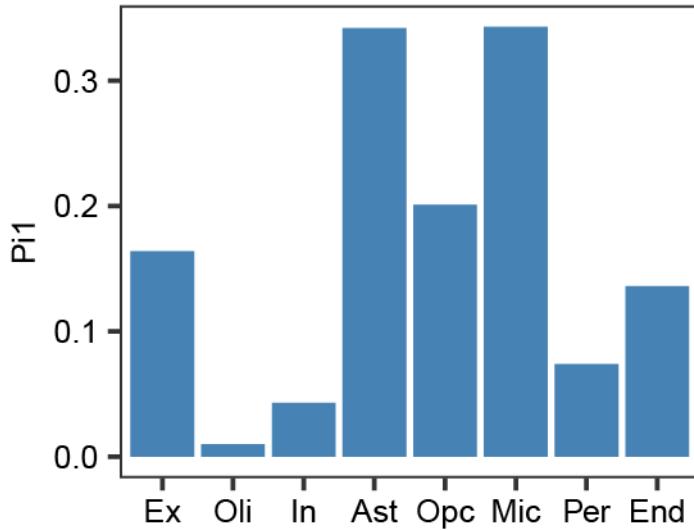
913

914

915



917 **Fig. S8. An example of cell-type specific eQTLs in excitatory neurons. (A)** Expression of
918 CALM2 in individuals with different genotypes. **(B)** Expression of CALM2 in ROSMAP
919 snRNAseq data. **(C)** Colocalization of eSNPs on CALM2 and SCZ GWAS risk locus.



920

921

Fig. S9. Replication of ieQTLs in single-cell eQTLs.