Comprehensive evaluation of deconvolution methods for human brain gene expression

Gavin J Sutton[1], Daniel Poppe[2,3], Rebecca K Simmons[2,3], Kieran Walsh[1], Urwah Nawaz[4], Ryan Lister[2,3], Johann A Gagnon-Bartsch[5], and Irina Voineagu[1,*]

Author affiliations:

1. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia
2. Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, the University of Western Australia, PO Box 7214, 6 Verdun Street, Nedlands, Western Australia, 6009, Australia.
3. Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Perth WA 6009, Australia
4. Adelaide Medical School, Robinson Research Institute, University of Adelaide, Adelaide, SA, Australia.
5. Department of Statistics, University of Michigan, 1085 South University Ave, Ann Arbor, MI 48109, USA.

*Corresponding author:

Assoc. Prof. Irina Voineagu

School of Biotechnology and Biomolecular Sciences

University of New South Wales

Kensington, Sydney NSW 2052 Australia

Phone: +61 (02) 9385 2029

Email: i.voineagu@unsw.edu.au

Abstract

Gene expression measurements, similar to DNA methylation and proteomic measurements, are influenced by the cellular composition of the sample analysed. Deconvolution of bulk transcriptome data aims to estimate the cellular composition of a sample from its gene expression data, which in turn can be used to correct for composition differences across samples. Although a multitude of deconvolution methods have been developed, it is unclear whether their performance is consistent across tissues with different complexities of cellular composition. The human brain is unique in its transcriptomic diversity, expressing the highest diversity of alternative splicing isoforms and non-coding RNAs. It comprises a complex mixture of cell-types including transcriptionally similar sub-types of neurons, which undergo gene expression changes in response to neuronal activity. However, a comprehensive assessment of the accuracy of transcriptome deconvolution methods on human brain data is currently lacking.

Here we carry out the first comprehensive comparative evaluation of the accuracy of deconvolution methods for human brain transcriptome data, and assess the tissue-specificity of our key observations by comparison with transcriptome data from human pancreas and heart.

We evaluate 8 transcriptome deconvolution approaches, covering all main classes: 4 partial deconvolution methods, each applied with 9 different cell-type signatures, 2 enrichment methods, and 2 complete deconvolution methods. We test the accuracy of cell-type estimates using *in silico* mixtures of single-cell RNA-seq data, mixtures of neuronal and glial RNA, as well as nearly 2,000 human brain samples.

Our results bring several important insights into the performance of transcriptome deconvolution: **(a)** We find that cell-type signature data has a stronger impact on brain deconvolution accuracy than the choice of method. **(b)** We demonstrate that biological factors influencing brain cell-type signature data (*e.g.* brain region, *in vitro* cell culturing), have stronger effects on the deconvolution outcome than technical factors (*e.g.* RNA sequencing platform). **(c)** We find that partial deconvolution methods outperform complete deconvolution methods on human brain data. To facilitate wider implementation of correction for cellular composition, we develop a webtool that implements the best performing methods, and is available at https://voineagulab.shinyapps.io/BrainDeconvShiny/ .

Keywords

Deconvolution; RNA-seq; Autism Spectrum Disorder; Benchmarking; Cellular Composition

Introduction

Human tissues are mosaics of cell-types and subtypes, which are diverse in their functionalities and express distinct sets of genes. Consequently, gene expression measurements in any tissue sample are the result of two main factors: gene expression levels within constituent cell-types, and the relative abundance of these cell-types in the sample[1,2]. The relative abundance of cell-types (*i.e.* cellular composition) in turn depends on both biological[3–6] and technical factors[7].

To circumvent the confounding effect of cellular composition, gene expression measurements could in principle be carried out by experimentally isolating individual cell-types by laser capture micro-dissection[8,9], cell sorting[10–12], or single-cell RNA-seq (scRNA-seq)[13]. In practice, these approaches are limited in feasibility and cost effectiveness for human brain transcriptome studies that require large sample sizes (hundreds to thousands of samples), such as eQTL studies or gene expression studies aiming to identify low-magnitude changes in disease samples.

Several methods for *in silico* deconvolution have been developed to estimate the cellular composition of a tissue sample from its gene expression profile (reviewed in Avila Cobos *et al.*[1]). *In silico* deconvolution offers the opportunity to leverage scRNA-seq data to obtain deeper insights into bulk tissue transcriptomes generated through large-scale studies such as GTEx[14], PsychEncode[15], the Common Mind Consortium[16], and BrainSpan[17].

Deconvolution methods fall into two main categories: partial deconvolution (including enrichment approaches), and complete deconvolution (see Methods), and are conceptually similar for any tissue and any type of molecular data (transcriptome, methylome, proteome, *etc.*). However, the complexity of cellular composition, and the transcriptome similarity across cell-types varies widely across tissues. Most deconvolution methods have been developed for, or assessed on, blood/immune and tumour samples[18–20], with limited assessment of their performance across tissues[2]. Therefore, an important outstanding question is whether transcriptome deconvolution methods perform equally well for any tissue.

We begin to address this question focussing on the human brain. The main biological factors that influence the cellular composition of brain samples (*e.g.* region, developmental stage, age[4,5]), and the technical factors involved (*e.g.* dissection protocol[7]) are distinct from those influencing cellular composition in blood. Furthermore, pure populations of cells from adult human brain are challenging to obtain, unlike blood or tumour cells. As a result, cell-type signature data are often obtained from a different brain region[21], species[22,23], and/or a different developmental stage[7] than the bulk brain samples. Alternatively, cells cultured *in vitro* have been used[19]. Whether such choices influence the accuracy of brain cell-type composition estimates is unknown. In addition, gene expression changes in most psychiatric disorders, similarly to effect-sizes of common variants, are of low magnitude[24]. Therefore, to serve as useful co-variates, cell-type composition estimates need to discriminate small differences in cellular composition[3]. While a few studies have proposed methods focussed on brain tissue[5,7,23,25–27], a comprehensive comparative assessment of the performance of deconvolution methods on brain transcriptome data is currently lacking.

Here, we performed a comprehensive evaluation of brain transcriptome deconvolution by assessing the performance of eight algorithms (four partial deconvolution, two enrichment, and two complete deconvolution methods). The partial deconvolution methods were each combined with nine types of cell-type signature data that differed in biological properties (cultured cells, immuno-purified cells, cross-species) or technical factors affecting RNA sequencing: single-nucleus RNA-seq (snRNA-seq), single-cell RNA-seq (scRNA-seq), bulk RNA-seq or CAGE-seq. These analyses were carried out on *in silico* mixtures of single-cell and single-nucleus transcriptomes, pure immuno-panned cell-types, mixtures of RNA extracted from pure populations of neurons and glial cells, as well as large-scale brain transcriptome data from the GTEx[14] and PsychENCODE[15,28] consortia.

Our results showed that cell-type signature data was the most important parameter for brain transcriptome deconvolution. The main biological factors influencing brain cell-type signature data, and consequently the deconvolution accuracy, were brain region and *in vitro* cell culturing. These factors had stronger effects on the deconvolution outcome than the sequencing platform (Illumina RNA-seq *vs.* Cap Analysis of Gene Expression (CAGE)). We also demonstrate that partial deconvolution methods, particularly CIBERSORT (implementing support vector regression), outperform complete deconvolution methods on human brain data. We also assessed the best approaches for correcting cell-type composition differences in differential expression analyses, and determined the magnitude of cell-type composition differences that can be effectively corrected for. Finally, we deconvolved large-scale gene expression data from GTEx and PsychENCODE consortia, and highlight the importance of

assessing deconvolution accuracy on each brain dataset; for this purpose we developed a user-friendly web tool that implements the best performing methods identified in this benchmarking study (https://voineagulab.shinyapps.io/BrainDeconvShiny/).

Results

To benchmark deconvolution methods for brain transcriptome data, we selected widely employed methods, and where possible included methods developed for brain data (Table 1). For partial deconvolution, we selected: CIBERSORT (*cib*), a highly-cited deconvolution method initially optimised for immune cell-types[18]; DeconRNASeq[29] (*drs*) which implements the non-negative least squares approach employed by the PsychENCODE consortium[15]; MuSiC[30](*music*), which is a single-cell-based deconvolution approach accounting for individual- and cell-specific expression variability in the signature; and dtangle[31]. For enrichment-based methods we selected xCell[19], which has been recently applied by the GTEx consortium[32], and BrainInABlender[7] (*blender*), which was specifically developed for brain-derived data. Among complete deconvolution methods, we included Linseed[33], which extends previous methods[34,35], and the co-expression-based approach developed for brain data by Kelley et al.[5] (*coex*).

Assessment of deconvolution accuracy across methods.

To assess deconvolution accuracy, we simulated data with known cell-type proportions using three adult human brain datasets: two snRNA-seq datasets, Velmeshev *et al.*[36] (VL: 24,646 nuclei, 10X Chromium, Fig.1) and Hodge *et al.* from the Human Cell Atlas[37] (CA: 11,314 nuclei, Smart-seq2, Supplementary Fig.1); as well as a scRNA-seq dataset Darmanis *et al.*[13] (DM: 297 cells, Smart-seq, Supplementary Fig.2). For each dataset, 100 mixtures were simulated as the average expression of 500 randomly-sampled nuclei (VL, CA; Methods) or 100 randomly-sampled cells (DM; Methods). Cell-type signatures were generated as the average expression within each cell-type (Methods).

We first estimated cell-type proportions in these mixtures using *cib*, *drs*, *dtangle*, and *music*, and enrichment scores using *xCell* and *blender*, evaluating 6 major brain cell-types: neurons, astrocytes, oligodendrocytes, oligodendrocyte precursor cells (OPCs), microglia, and endothelia. Focusing on mixtures generated from the largest dataset (VL), we found that deconvolution accuracy was very high for *cib* (mean r across cell-types = 0.87), *music* (0.82), and *dtangle* (0.87), but lower for *drs* (0.50) (Fig.1B, left, and Supplementary Fig.3,4). For the two enrichment algorithms, *blender*'s accuracy was moderately high but inconsistent across cell-types, while *xCell* poorly estimated cell-type abundance (r = -0.06 and 0.02 for neurons and astrocytes, respectively); Fig.1C, and Supplementary Fig.4. These observations were replicated in both the CA- (Supplementary Fig.1,5,6) and DM-based simulations (Supplementary Fig.2), suggesting that (a) deconvolution of major brain cell-types is accurate across a range of partial deconvolution algorithms, with *cib* generally performing best and (b) enrichment methods are less accurate than partial deconvolution methods, with *xCell* showing particularly low accuracy.

We next assessed deconvolution accuracy on five *in vitro* RNA mixture samples of known composition (Supplementary Figure 7) and twenty-one RNA samples from pure populations of cells immuno-panned with cell-type-specific antibodies[38]; Supplementary Figure 8. In both cases, the deconvolution accuracy was very high when the signature was derived from the same source as the mixtures. For RNA mixtures the normalised mean absolute error was 0.035, 0.043, and 0.11 for *cib, drs,* and *dtangle,* respectively (Supplementary Figure 7). For RNA extracted from sorted cells, the immuno-panned cell-type was identified on average as 96.3%, 93.0%, and 92.6% abundant by *cib, drs,* and *dtangle,* respectively (Supplementary Figure 8).

- **Deconvolution of cellular subtypes**

We next explored how including cellular sub-types affected deconvolution accuracy for brain data. First, we used broad neuronal sub-types, *i.e.* excitatory and inhibitory neurons (Fig.1B middle, Supplementary Fig.3,9), and found that deconvolution accuracy was high ($r > 0.8$ for all algorithms), with *cib* performing best ($r = 0.94$ and 0.95 for excitatory and inhibitory, respectively). The accuracy for the other cell-types was largely unaffected by neurons being sub-classified (Fig.1B, middle, Supplementary Fig.3,9). This result was replicated in the CA-based simulations (Supplementary Fig.1,5,10).

When including all cell sub-types detected in the VL dataset (11 neuronal and 2 astrocyte sub-types), deconvolution with *cib* remained accurate ($r > 0.8$) for most cell populations (Fig.1B, right, Supplementary Fig.3,11,12). However, two main factors led to a reduction in accuracy for certain cell-types: low abundance of the cell-type (< 2%) and high collinearity (gene expression correlation with another cell-type *rho* > 0.95); Supplementary Fig.13,14. This observation was replicated in the CA dataset with most cell sub-types being accurately deconvolved ($r > 0.8$; Supplementary Fig.1,5,15,16) and collinearity being the main factor that led to reduced accuracy (Supplementary Fig.17,18).

- **Deconvolution accuracy when a cell-type is missing from the reference signature data**

We also explored how deconvolution was affected when a cell-type was missing from the signature. We removed one cell-type at a time from the VL-derived mixtures, and found that when an abundant cell-type was missing (Neurons, 87.4%), the deconvolution accuracy was substantially reduced (mean r was reduced from 0.85 to 0.41, and normalised mean absolute error increased from 0.33 to 10.3). However, when lowly-abundant cell-types were missing from the signature, the effect on deconvolution was minimal (Supplementary Figure 19). We then tested the effect of removing a sub-type of neurons, excitatory or inhibitory neurons, which are highly correlated in expression (*rho*=0.92). Deconvolution accuracy was reduced to a lesser extent then when all neurons were missing: r was reduced from 0.87 to 0.71 when excitatory neurons were missing, and from 0.86 to 0.76 when inhibitory neurons were missing (Supplementary Figure 19).

The biological properties of cell-type signature data strongly influence deconvolution

We hypothesised that the brain cell-type signature data could have a major impact on the deconvolution outcome. This has been previously reported for deconvolution of blood transcriptomes[39], and is supported by our observation that xCell performed significantly worse than the other deconvolution methods (noting that its signature data is built-in).

To investigate how the properties of the signature data influence the deconvolution outcome, we deconvolved the human brain snRNA-seq mixtures (VL) using cell-type signature data from several datasets (Methods; Fig.2) with different sequencing methods and various sources of brain tissue: human brain snRNA-seq (CA[37], NG[40], LK[41]); scRNA-seq from the human (DM[13]) or mouse (TS[42]) brain; bulk RNA-seq of immuno-panned cells from the human (IP[38]) or mouse brain (MM[43]); or CAGE-seq from cultured human brain cells (F5[44]).

We found that the choice of cell-type signature data strongly affected the deconvolution accuracy. Using data from cultured brain cells (F5) dramatically reduced the accuracy (Fig.2A-B). This likely explains the poor performance of xCell, which has F5 as the main built-in signature. Using signature data from the mouse brain (TS, MM) also reduced the deconvolution accuracy (Fig.2A-B). These observations were consistent across deconvolution algorithms (Supplementary Fig. 20) and were replicated when deconvolving *in silico* mixtures based on

215 the CA and DM data (Supplementary Fig.21,22), as well as with deconvolution of broad
216 neuronal subtypes (Supplementary Fig.23,24). Conversely, when deconvolving RNA mixtures
217 of known composition from cultured cells (Methods), using the cultured-cell F5 signature data
218 performed the best despite the difference in sequencing technology (CAGE-seq *vs*. RNA-seq;
219 Supplementary Fig.7).

220 Overall, these data demonstrate that the biological properties of the cell-type signature data
221 strongly impact the deconvolution accuracy, having a more pronounced effect than the
222 sequencing methods, and highlight *in vitro* culturing of brain cells as an important biological
223 factor.

224 • **The effect of compartment specific genes on deconvolution accuracy**

225 Since most single-cell data from the adult human brain are generated using single-nucleus
226 RNA-seq, while bulk RNA-seq is based on total RNA, we next investigated whether
227 compartment-specific genes (*i.e*. those either enriched or depleted from the nucleus) influence
228 the outcome of deconvolution. For this purpose, we generated paired bulk RNA-seq and
229 nuclear RNA-seq from five frozen brain tissue samples (Methods), as well as snRNA-seq from
230 the same brain samples. We identified compartment-specific genes as those differentially-
231 expressed between the nuclear and total bulk RNA-seq (FDR < 0.05, |FC| > 1.3; Supplementary
232 Table 1). We then carried out several deconvolution analyses with and without filtering-out the
233 compartment-specific genes.

234 Firstly, we deconvolved the twenty-one bulk RNA-seq samples from sorted brain cells[38], where
235 true cell-type composition is known (i.e. each sample is expected to be a nearly-pure cell-type,
236 with some experimental variability of the sorting efficiency). We deconvolved these data with
237 either the identical cell-type signature (derived from the sorted dataset; IP), an scRNA-seq
238 signature (DM), and four snRNA-seq signatures (VL, CA, NG, and LK) ; Supplementary Table
239 2. When using the IP signature, the immuno-panned cell-type was estimated as > 80% abundant
240 in all samples. Thus we assessed the proportion of samples in which the sorted cell-type was
241 correctly identified (i.e estimated proportion > 80%) using the scRNA-seq and snRNA-seq
242 signatures, with or without filtering out compartment-specific genes. We found that the
243 snRNA-seq signatures performed well even prior to filtering out compartment-specific genes,
244 correctly identifying the sorted cell-type in an average of 86% of samples (71%-95%). As
245 expected, the single-cell-based signature (Supplementary Fig.25) performed somewhat better,
246 identifying the sorted cell-type in an average of 90% of samples. Removing compartment-
247 specific genes further improved the outcome for snRNA-seq signatures: the sorted cell-type
248 was correctly identified in an average of 88% of samples (86%-95%); Supplementary Fig.25,
249 eliminating the difference between the scRNA-seq and snRNA-seq signatures.

250 To increase the complexity of the deconvolution task, we asked how accurately the five whole-
251 tissue samples were deconvolved when using snRNA-seq data from the same individuals, as
252 compared to a whole-cell based signature (Supplementary Fig.25). In this case, if
253 compartment-specific genes were not removed from the cell-type signature, the correlation
254 between cell-type proportions estimated using the snRNA-seq signature and the whole-cell
255 signature was modest (r=0.27). However, the correlation improved substantially by filtering
256 out compartment-specific genes (r=0.98) suggesting that this filtering approach should be
257 considered when using snRNA-seq-based cell-type signatures.

258 ***Reference-free complete deconvolution methods are less effective on brain gene***
259 ***expression data than partial deconvolution methods.***

260 Since we observed a strong effect of the choice of reference signature data on the brain
261 deconvolution outcome, and recent studies have proposed reference-free approaches to cell-

262 type composition[33,34,45], we assessed the performance of two such methods on brain data.
263 Linseed, a complete deconvolution algorithm[33], proposes to identify cell-type specific genes
264 by representing the expression vector of each gene as a point in N-dimensional space (where
265 N is the number of samples). It also proposes using singular-value-decomposition (SVD) to
266 determine the number of cell-types from the mixture data. An alternative approach, *coex*[46],
267 employs co-expression networks to identify modules of co-expressed genes enriched for
268 specific cell-type markers, and then uses the module eigengene values as cell-type enrichment
269 scores[5].

270 When applying Linseed to *in silico* mixtures generated by random sampling from the three
271 benchmarking datasets (VL, CA, DM), we found that the SVD approach did not correctly
272 identify the number of cell-types in the mixture (Methods; Supplementary Fig.26). With the
273 correct number of cell-types specified, Linseed performed less accurately than the partial
274 deconvolution methods, with $r > 0.8$ achieved for only two cell-types for VL and CA, and none
275 of the cell-types for DM (Fig.3B, Supplementary Fig.27). On the RNA mixtures however,
276 Linseed performed very accurately (r=1; Supplementary Fig.28). Since Linseed relies on the
277 detection of genes represented by points with "extreme" positions in the k-1 dimensional
278 simplex, we hypothesized that the difference in its performance between the datasets likely
279 results from the wider distribution of cell-type proportions in the RNA mixtures (neuronal
280 proportions: 0-100%), than in the mixtures generated by random sampling from real brain
281 datasets. To test this hypothesis, we generated mixtures with a broad range of cell-type
282 abundances using VL and CA (Methods; Fig.3A,C). The performance of Linseed improved
283 markedly on both datasets using these controlled *in silico* mixtures (Fig.3D, Supplementary
284 Fig.29,30), with the SVD approach also better identifying the number of cell-types in the
285 mixture (Methods).

286 We found that *coex* also performed significantly less accurately than the partial deconvolution
287 methods on the randomly-sampled mixtures (Fig.3B, Supplementary Fig.27). Since the co-
288 expression network approach also relies on gene expression co-variation driven by differences
289 in cell-type proportions, its performance improved on simulations with a wider range of cell-
290 type proportions, but did not achieve accurate deconvolution for all cell-types (Fig.3D,
291 Supplementary Fig.29,30).

292 These data suggest that complete deconvolution methods are less effective than partial
293 deconvolution methods, particularly since the performance of these algorithms is related to the
294 variance in cellular composition of the dataset, which is not known *a priori*.

295 ***Assessment of the interplay between cell-type composition and differential gene (DE)***
296 ***expression analyses.***

297 We next investigated how cellular composition influences DE analyses, in particular: (i) how
298 much should cell-type composition differ between two groups of brain samples to lead to false
299 positive results in DE analyses, and (ii) what is the best approach for correcting cell-type
300 composition differences in DE analyses?

301 We used the CA dataset[37] (Smart-seq2, high coverage per gene) to generate simulated data for
302 two-group DE analyses. Each dataset contained two groups of 50 samples (group A and group
303 B). The proportion of one of the cell-types (excitatory neurons) was simulated as either higher
304 or lower in group B than in group A by 0% to 40% (Methods; Fig.4). We then carried out DE
305 comparing group B to group A, using either a linear model (LM) or a generalised LM as
306 implemented in DESeq2[47] with and without correction for cellular composition. False-positives
307 driven by cellular composition were defined as genes differentially expressed at a false
308 discovery rate (FDR) < 0.05 (Methods).

309 We found that without correction, differences in cellular composition of less than 5% between
310 the sample groups led to fewer than 10 false-positive DE genes. However, above 5% the
311 number of false-positive genes increased steeply with the difference in cellular composition,
312 reaching > 10,000 at a 20% difference in cellular composition (Fig.4A). Inclusion of excitatory
313 neuron proportions as a covariate in the LM effectively eliminated false-positive genes
314 (Fig.4A). We didn't observe any additional benefit when using a spline matrix as covariate,
315 while quadratic regression was less effective than linear regression at larger composition
316 confounds (Fig. 4A). We also found that including cellular composition estimates in DESeq2
317 was similarly effective at eliminating false-positives (Fig.4A). As expected, markers of
318 excitatory neurons were enriched among downregulated genes when the proportion of this cell-
319 type was reduced in the test group, but enriched among upregulated genes when the proportion
320 was increased (Fig.4B).

321 We next investigated the more challenging case where there are true differences in gene
322 expression between the two groups, in addition to differences in cell-type composition. To this
323 end, we simulated data with differences in cell-type composition as above, while also
324 introducing gene expression changes in a set of 200 genes, of which 100 are markers of
325 excitatory neurons and 100 are non-marker genes (Methods). Several sets of simulations were
326 generated with a mean expression difference between groups of 1.1-, 1.3-, 1.5- or 2-fold.

327 To quantify how effectively cellular composition was corrected for, we calculated
328 discriminatory power as the fraction of the 200 perturbed genes that were in the top 200 most
329 significant DE genes. This measure rewards true-positives while penalising false-positives. We
330 found that without correction, the discriminatory power decreased with the magnitude of cell-
331 type composition difference between the two groups (Fig.4C; uncorrected). Correction for cell-
332 type composition was effective at restoring discriminatory power for gene expression
333 differences of 1.5 fold when composition differences were up to 12.5% (Fig.4D; corrected).
334 As expected, for expression differences of lower magnitude (1.1 and 1.3) the effective
335 correction range was narrower (6.3% and 6.9% respectively), while for stronger expression
336 differences (2-fold) the effective correction range was wider (25.7%); Fig.4D, Supplementary
337 Fig.31. All correction approaches performed similarly in this analysis, with the exception of
338 spline regression which we found to be less effective (Fig.4C,D, Supplementary Fig.31).

339 We also investigated whether the cell-type where differential expression occurs can be
340 uncovered through deconvolution. To this end, we used CIBERSORTx[48], which takes a bulk
341 mixture and estimates gene expression values in each cell-type present in the signature data;
342 these cell-type-specific expression values can then be used to carry out cell-type specific DE
343 analyses. We thus simulated data with 1.5-fold change in expression specific to a given cell-
344 type, with or without superimposed differences in cell-type composition between the two
345 groups (Supplementary Fig.32), and tested whether genes were identified as DE in the correct
346 cell-type. As above, the 1.5-fold expression difference was simulated for 200 genes, of which
347 100 are markers of the perturbed cell-type and 100 are non-marker genes.

348 The expression difference was first simulated in excitatory neurons. In the absence of
349 confounding cell-type composition differences between the two groups, more than 95% of the
350 perturbed excitatory marker genes were detected as DE in the correct cell-type (excitatory
351 neurons), while for the non-marker genes ~45% were detected as DE in excitatory neurons and
352 another ~30% were incorrectly detected as DE in inhibitory neurons; Fig. 4E. The false-
353 positive rate (*i.e.* the fraction of non-perturbed genes detected as DE) was 0% (Supplementary
354 Fig.32).

355 When a composition difference was superimposed (~10% increase in excitatory neurons;
356 Methods), the true-positive rate was unchanged except for upregulated marker genes, where it

357 was reduced, likely due to the fact that the composition change and the expression change were
358 confounded (both variables were higher in group B *vs*. group A). The false-positive rate was
359 less than 12% in all cell-types, thus drastically reduced relative to no correction for cell-type
360 composition (32%), but higher than when correcting for composition differences in a standard
361 linear model (0%). Similar results were observed when the gene expression change was
362 modelled in inhibitory neurons (Supplementary Fig.32).

363 Overall, these results suggest that using cell-type specific gene expression for DE analyses is
364 effective at detecting DE genes in the right cell-type when the gene expression and composition
365 changes are not confounded, but this comes at the expense of a moderate increase in false-
366 positives.

367 ***Cell-type composition estimates in large-scale human brain transcriptome data.***

368 We next evaluated the performance of brain gene expression deconvolution on large-scale
369 datasets, focussing on a dataset of control individuals (GTEx[14], n=1,671 samples; Methods),
370 and a dataset of autism spectrum disorder (ASD) cases and controls (PsychENCODE;
371 Parikshak *et al*.[28], n=251 samples; Methods). The GTEx data included samples from
372 cerebellum (n=309), cerebral cortex (n=408), subcortical regions (n=863) and spinal cord
373 (n=91); the Parikshak *et al*. dataset included samples from cerebellum (n=84) and cerebral
374 cortex (n=167).

375 We assessed all combinations of 4 partial deconvolution methods and 9 cell-type signatures,
376 the two enrichment methods, and *coex* as a complete deconvolution method (Supplementary
377 Table 3). We also generated an additional signature (MultiBrain) by merging CA, IP, DM, NG,
378 and VL signatures derived from cortex, reasoning that this approach will average-out inter-
379 individual and technical differences, as previously proposed[39].

380 The accuracy of composition estimates was first evaluated on cortical samples using goodness-
381 of-fit, *i.e*. the Pearson correlation between measured gene expression and reconstructed gene
382 expression values (Methods). Consistent with the results on simulated data, we found that cell-
383 type signature data had a stronger impact on accuracy than the choice of algorithm
384 (Supplementary Fig.33). Although there was some variation between the two datasets, the CA
385 and MultiBrain signatures performed consistently well, while the cultured-cell-derived F5 and
386 the single-nucleus LK signatures performed worst (Fig.5A,B). Cerebellar samples showed
387 lower goodness of fit than cortical samples in both datasets (Supplementary Fig.34,35),
388 consistent with the fact that all cell-type signatures were derived from cerebral cortex. When
389 the biological and technical differences between the bulk data and cell-type signatures are
390 eliminated, as is the case of our *in silico* mixtures of single-cell data, goodness-of-fit averaged
391 ~0.95 (Supplementary Fig.36). Further details on the deconvolution results of the GTEx and
392 Parikshak *et al*. data (correlation and absolute values of cell-type abundance estimates) are
393 included in Supplementary Note.

394 While the important role of signature data has been previously reported[39], here we uncovered
395 the effect of novel biological factors that affect deconvolution accuracy, in particular *in-vitro*
396 culturing and brain region. Since *in-vitro* culturing may be relevant to other tissues as well, we
397 investigated whether using cultured-cell-derived or tissue-derived signature data influences the
398 accuracy of deconvolution for two additional tissues in GTEx: pancreas and heart. We found
399 that deconvolution accuracy for left heart ventricle and arterial appendage samples was
400 significantly reduced when using signature data from cultured cells, while the deconvolution
401 accuracy for pancreas data was mildly reduced (Supplementary Fig.37). These data suggest
402 that the influence of biological factors on cell-type signature data vary across tissues, indicating
403 that tissue-specific benchmarking of deconvolution approaches is warranted.

Finally, we applied the results of the cell-type composition analyses to get further insights into genes differentially expressed in brain tissue samples from ASD cases[28]. Cell-type proportion estimates (CIB/*MultiBrain*), showed significantly higher astrocyte proportions in ASD cortex samples compared to controls (difference in means: 7.2%, p=0.0002, Wilcoxon rank sum test; Fig.5C). This result recapitulates recent single-nucleus data from ASD brain validated by immunohistochemistry[36], which showed higher proportion of astrocytes in ASD cortex samples. There were also significantly higher proportions of microglia (0.7%, p=0.003; Wilcoxon rank sum test), although the overall abundance of microglia was low.

We next carried out differential expression (DE) analyses either without correction for cellular composition (composition-dependent; CD) or including cell-type proportion estimates from CIB/*MultiBrain* in the model (composition-independent; CI); Methods. Astrocyte, oligodendrocyte, and microglial proportions were included as covariates. CD analyses identified 713 down- and 1885 up-regulated genes (Fig.5D). In contrast, when correcting for composition estimates in our CI analyses, we identified only 46 down- and 21 up-regulated genes (Fig.5D). Of these, 20 down- and 18 up-regulated genes overlapped between CI and CD analyses (Fig.5D). Thus, 26 down-regulated and 3 up-regulated genes were uncovered by the CI analysis, and have not previously been reported as DE in ASD[28]. Conversely, 693 down-regulated and 1867 up-regulated genes were identified in the CD analysis only, and thus likely reflect differences in cellular composition between the ASD and control samples, rather than gene expression dysregulation (Supplementary Table 4). The CD upregulated genes were enriched for immune and inflammatory genes (Supplementary Table 4) as well as astrocyte and microglial markers (p=2.5x10^{-11} and 4.7x10^{-36}, respectively), consistent with their higher proportions in ASD samples. Notably, one of the top up-regulated novel genes, *CXXC4,* which encodes a protein involved in Wnt signalling, has also been identified as upregulated in ASD CTX layer 4 neurons by single-nucleus RNA-seq[36]. In addition, *CXXC4* was identified as the top associated gene in a GWAS meta-analysis of schizophrenia and ASD[49]. These data indicate that correction for cellular composition can identify novel, disease-relevant gene expression changes.

Discussion

Here, we began to address the question of tissue-specificity in transcriptome deconvolution, by carrying out a comprehensive benchmarking of deconvolution methods on brain transcriptome data. We assessed eight deconvolution methods, as well as multiple parameters of deconvolution: the biological and technical properties of the cell-type signature data; the effect of deconvolving brain cell sub-types; the effect of missing cell-types in the signature data; and the effect of nuclear-enriched or depleted transcripts on deconvolution using snRNA-seq signatures. We also investigated how effectively cell-type composition differences can be corrected in DE analyses.

It has previously been shown that cell-type signature data has a strong effect on deconvolution accuracy[39,50]. In blood, the microarray platform was the main factor driving differences between cell-type signature datasets[39]. For deconvolution of solid tumours, accurate estimation of immune cell-type composition required tumour-derived cell-type signatures, rather than blood-derived signatures[50]. For brain transcriptomes, we found that cell-type signature data had a stronger impact than the choice of method in all cases studied: simulated single-cell mixture data, RNA mixtures of known composition, immuno-panned cells, and large-scale post-mortem transcriptome data. We also found that for brain transcriptomes, biological factors outweighed technical factors, and among biological factors *in vitro* culturing (Supplementary Fig.33-35) and brain region (cortex *vs*. cerebellum) (Supplementary Fig.33-35) had the strongest impact. *In vitro* culturing also affected the performance of deconvolution for other

452 tissues (heart and pancreas), but to different extents (Supplementary Fig.37), highlighting the
453 importance of tissue-specific benchmarking.

454 We found that snRNA-seq derived cell-type signatures performed well, particularly the Human
455 Cell Atlas data (CA), which has high-coverage, while low sequencing depth (LK) led to
456 reduced accuracy. Removing compartment-specific genes from the snRNA-seq signatures
457 improved the deconvolution accuracy (Supplementary Fig.25).

458 Another factor known to influence deconvolution accuracy is the absence of cell-types present
459 in mixtures from the signature data[51,52]. Consistent with previous results[51,52], we found that if
460 an abundant brain cell-type was missing from the signature data, the deconvolution accuracy
461 was reduced, particularly for cell-types highly correlated with the missing cell-type
462 (Supplementary Fig.19). The absence of a lowly-abundant cell-type, such as microglia and
463 endothelia, had a minimal impact on deconvolution accuracy, suggesting that signature datasets
464 missing these cell types can be used in deconvolution of brain data.

465 Since neuronal sub-types are highly similar in gene expression profiles, we investigated how
466 different deconvolution methods handled co-linearity in brain transcriptome data. We found
467 that CIBERSORT best handled co-linearity, and deconvolution of brain cell sub-types was
468 accurate provided that they were not lowly abundant (<2%) or highly collinear with other cell-
469 types (*rho* > 0.95); (Supplementary Fig.14,18).

470 It was previously shown that semi-supervised and unsupervised complete deconvolution
471 methods underperform relative to supervised (*i.e.* partial) deconvolution methods[51,52]. Our
472 results support these observations, and we further determine that the range of cell-type
473 composition across samples in the bulk dataset is a major factor influencing the performance
474 of complete deconvolution methods (Figure 3, Supplementary Fig.27-30).

475 When assessing the interplay between cellular composition and DE analyses, we found that
476 false-positives are induced in DE analyses by as low as 5-10% difference in cell-type
477 composition (Figure 5A). Inclusion of cell-type composition estimates as covariates effectively
478 eliminated composition-induced false-positive genes, and restored discriminatory power for
479 gene expression differences of 2-fold when cell-type composition differences were up to ~25%
480 (Figure 5C,D, Supplementary Fig.31).

481 The deconvolution of large GTEx data and PsychENCODE data showed that the best-
482 performing signature may differ across datasets, and thus it is worth assessing goodness-of-fit
483 for multiple signatures when deconvolving brain gene expression data. Notably, in both
484 datasets, and across all deconvolution methods, there was a wide range of estimated cell-type
485 proportions in any given brain region (Supplementary Note). This is consistent with data from
486 the PsychENCODE consortium[15], which used an NLS-based approach (similar to the one
487 implemented in *drs*) and reported a similarly wide range of proportion of neurons across 1867
488 dorsolateral prefrontal cortex samples: 2-54%. (http://resource.psychencode.org, PEC_DER-
489 24_Cell-Fractions-Normalised). Such a wide range is also observed in brain methylome
490 deconvolution[53] (0-50%) and likely reflects technical variability in dissection rather than
491 biological inter-individual variability.

492 Overall, for deconvolution of brain transcriptome data we recommend that **(a)** CIBERSORT
493 and either dTangle or MuSiC are good choices of methods **(b)** cell-type signature data should
494 be well matched to the bulk samples, in terms of *in vitro* culture state and brain region, **(c)**
495 cellular sub-types should only be included in deconvolution if they are > 2% abundant and <
496 95% correlated with other cell-types/sub-types, **(d)** when using snRNA-seq based signatures,
497 removal of nuclear-specific genes (Supplementary Table 1) from the signature should be

considered, **(e)** only attempt to use reference-free deconvolution methods if the bulk dataset is known to have a wide range of cell-type compositions.

To facilitate the choice of cell-type signature data, we provide the ten cell-type signatures compiled here as an R package (https://github.com/Voineagulab/brainyR), and developed a web tool which implements the best performing algorithms and all the cell-type signatures, as well as calculation of goodness-of-fit, in a user-friendly format, available at: https://voineagulab.shinyapps.io/BrainDeconvShiny/.

506 # Methods

507 **Datasets accessed and pre-processing**

508 - **Brain RNA-seq datasets**

509 **Bulk brain gene expression data from Parikshak *et al.***[28] were obtained from Github
510 (https://github.com/dhglab/Genome-wide-changes-in-lncRNA-alternative-splicing-and-
511 cortical-patterning-in-autism/releases). Exon-level count data was obtained for 251 post-
512 mortem samples (rRNA-depleted), including frontal cortex, temporal cortex, and cerebellar
513 vermis samples from 48 ASD and 49 control individuals, aged 2-67 (Supplementary Table 5;
514 see Parikshak *et al*. (2016) for complete metadata).

515 Gene-level normalised data was generated by aggregating exon counts followed by reads per
516 kilobase per million reads (RPKM) normalisation using the total exonic length of each gene
517 (Ensembl V19 (hg19) assembly). A minimum expression threshold was then set at > 1 RPKM
518 in at least 40 samples (*i.e.*, half of the number of samples in the least-represented region).

519 Outlier samples removed in the Parikshak *et al.* study were also removed from our analyses,
520 leaving 121 ASD (43 frontal cortex, 39 temporal cortex, 39 cerebellum) and 126 control (45
521 frontal cortex, 36 temporal cortex, 45 cerebellum) samples; Supplementary Table 5.

522 **Bulk brain gene expression data from GTEx**[14] were obtained as gene-level read counts from
523 the 2016-01-05 release (V7) at https://gtexportal.org/home/datasets. Counts were RPKM
524 normalised as above. A minimum expression threshold was set at > 1 RPKM in at least 88
525 samples (*i.e.* the number of samples in the least-represented brain region).

526 - **Brain cell-type-specific gene expression datasets and generation of cell-type**
527 **signatures**

528 Information about samples used and final expression values are available in in Supplementary
529 Tables 5 and 6, respectively. Metrics of signature similarity are presented in Supplementary
530 Figures 38-39.

531 ***F5*** *(FANTOM5)*: Cap Analysis of Gene Expression (CAGE) data for robust CAGE peaks was
532 obtained from the FANTOM5 consortium: http://fantom.gsc.riken.jp/5/data/[44]. Tag-per-
533 million normalised CAGE peak expression levels were aggregated by sum at gene level. Data
534 from cultured neuron (n=3) and astrocyte (n=3) samples were averaged to generate the F5
535 neuron and astrocyte signatures. A minimum expression threshold was set at > 1 tag-per-
536 million in at least one cell-type.

537 ***IP*** *(immuno-purified)*: RNA-seq data from cells immunopurified from human adult brain tissue
538 extracted during surgery were obtained from Zhang *et al.* 2016[38]. FPKM-level data were
539 accessed from Table S4 of Zhang *et al*. for neurons (n=1), astrocytes (n=12), oligodendrocytes
540 (n=5), microglia (n=3), endothelia (n=2). Cell-types derived from foetal brain were excluded
541 (*i.e.*, foetal astrocytes). Samples of the same cell-type were averaged to generate the IP
542 signature. A minimum expression threshold was set at > 1 FPKM in at least one of the five
543 cell-types in the final signature matrix.

544 ***MM*** *(Mus musculus)*: RNA-seq data from immunopurified mouse brain tissue was obtained
545 from Zhang *et al.* 2014[43]. FPKM-level data were accessed from
546 https://web.stanford.edu/group/barres_lab/brain_rnaseq.html, in which biological replicates of
547 cell-type transcriptomes (neurons, astrocytes, oligodendrocytes, microglia, and endothelia
548 were already aggregated across samples. Mouse genes were mapped to human orthologues
549 using Gene ID homology information from
550 http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt.

551 Expression data from oligodendrocyte precursors and newly-formed oligodendrocytes were
552 excluded. A minimum expression threshold was set at > 1 FPKM in at least one of the five
553 cell-types in the final signature matrix.

554 ***DM*** *(Darmanis):* Human brain single-cell gene expression data from the middle temporal gyrus
555 generated by Darmanis *et al.* (2015)[13] were downloaded as count-level data from
556 https://github.com/VCCRI/CIDR-examples/tree/master/Brain[54]. To generate the DM
557 signature, RPKM or counts-per-million (CPM) expression was averaged across samples of
558 each cell-type (*i.e.* astrocyte (n = 62), neuron (161), microglia (16), mature oligodendrocytes
559 (38), oligodendrocyte precursor cells (OPCs) (18), or endothelia (20). Cell-types derived from
560 foetal brain (quiescent neurons and replicating neurons) were excluded. A minimum expression
561 threshold was set at > 1 RPKM or CPM in at least one cell-type in the final signature matrix.

562 ***LK*** *(Lake)*: Gene expression data for 10,319 human adult frontal cortex nuclei were accessed
563 from Lake *et al*. 2018[41]. Seurat[55] was used to pre-process raw count expression data, removing
564 nuclei with 1) fewer than 1000 counts or 2) 200 expressed genes, or 3) >5% of counts attributed
565 to mitochondrial genes, or 4) a number of reads >99.5th percentile of its dataset. Only 3930
566 nuclei passed these QC criteria. To generate the LK signature, RPKM or CPM values were
567 averaged across nuclei of each cell-type: astrocytes (97), excitatory neurons (2611), inhibitory
568 neurons (1051), oligodendrocytes (96), OPCs (46), and microglia (22). An expression profile
569 for neurons was also generated, as the average of all excitatory and inhibitory nuclei. A
570 minimum expression threshold of > 1 RPKM or CPM in at least one cell-type was required.
571 Note that endothelia were excluded for having fewer than 10 nuclei (7).

572 ***VL*** *(Velmeshev)*: 10X Chromium for single-nucleus data from the post-mortem adult human
573 brain were accessed Velmeshev *et al.*[36]. Only nuclei from control prefrontal cortex samples
574 were included. Seurat processing, cell-type aggregation, and thresholding were performed as
575 described above in LK After filtering, 24,556 nuclei remained, classified as astrocytes (2229),
576 excitatory neurons (9718), inhibitory neurons (4238), oligodendrocytes (4721), OPCs (2677),
577 microglia (450), and endothelia (523).

578 ***CA*** *(Cell Atlas):* Count-level exon expression data for NeuN+ sorted adult nuclei from the
579 middle temporal gyrus were acquired from the Human Cell Atlas[37]. Seurat processing, cell-
580 type aggregation, and thresholding were performed as described above in LK. After filtering,
581 15,524 nuclei remained, classified as astrocytes (291), excitatory neurons (10492), inhibitory
582 neurons (4118), oligodendrocytes (313), OPCs (238), microglia (63). Endothelia were
583 excluded for having fewer than 10 representatives (9).

584 ***NG*** *(Nagy)*: 10X Chromium single-nucleus expression data from the adult human post-mortem
585 human prefrontal cortex were accessed from Nagy *et al.*[40] . Only nuclei from control samples
586 were included. Seurat processing, cell-type aggregation, and thresholding were performed as
587 described above in LK. After filtering, 23,168 nuclei remained, classified as astrocytes (1195),
588 excitatory neurons (14624), inhibitory neurons (5940), oligodendrocytes (757), OPCs (505),
589 microglia (85), and endothelia (62).

590 ***TS*** *(Tasic)*: Exon-level SmartSeq2 single-cell expression data from the adult mouse cortex
591 were accessed from Tasic *et al.*[42]. Only cells from the Anterior Lateral Motor Cortex were
592 included. Further, cells labelled by the authors as low quality or with no class were excluded.
593 Seurat processing, cell-type aggregation, and thresholding were performed as described above
594 in LK. After filtering, 8075 nuclei remained, classified as astrocytes (195), excitatory neurons
595 (3851), inhibitory neurons (3767), oligodendrocytes (69), OPCs (24), microglia (80), and
596 endothelia (89).

MB (*Multibrain*): this composite signature was generated by quantile-normalising and averaging the RPKM-level expression of the CA, IP, DM, NG, and VL signatures for five cell-types (neurons, astrocytes, oligodendrocytes, microglia, and endothelia). All signatures are cortical in origin but represent a range of purification protocols (scRNA-seq by SmartSeq (DM), snRNA-seq by 10X (VL, NG), snRNA-seq by SmartSeq (CA), and immuno-panning (IP))

- **Heart and pancreas RNA-seq datasets**

Bulk gene expression data from GTEx[14] for pancreas (n=268) and heart (n=310 and 417 atrial appendage and left ventricle, respectively) were processed as per the GTEx brain samples, except the pancreas samples were normalised to the level of transcripts-per-million (TPM)

Cell-type-specific RNA-seq data from pancreas alpha and beta cells were obtained from three studies as described below. For each dataset, genes were excluded if they were not protein-coding, or if they were expressed at < 1 TPM in both cell-types.

***EN (Enge)*:** count-level expression data for single-cells from freshly-isolated, FACS-sorted human pancreas were acquired from Enge *et al.*[56]. Data were normalised to the level of transcripts-per-million (TPM), using the total exonic length of each gene per the Ensembl V19 (hg19) assembly. The expression signature of alpha and beta cells was generated as the average of 998 alpha and 348 beta cells.

BL (Blodgett): TPM-level expression data for bulk RNA-seq from freshly-isolated, FACS-sorted alpha and beta cells from human pancreas were acquired from Blodgett *et al.*[11]. The expression signature of alpha and beta cells was generated as the average of 7 adult alpha-cell and 7 adult beta-cell bulk RNA-seq samples.

FS* and *FG* *(Furuyama): count-level expression data for human pancreas alpha and beta cells were acquired from Furuyama *et al.*[12]. After TPM normalisation, the **FS** (Furuyama Sorted) signature was constructed from freshly-isolated, FACS sorted alpha and beta cells (average of 5 replicates each), while the **FG** (Furuyama GFP) signature consists of isolated alpha and beta cells subjected to 1-week of culturing. These cells had been transduced with a GFP expression vector for imaging purposes[12] (average of 4 and 6 replicates, respectively).

Cell-type-specific RNA-seq data from heart were accessed from three publicly available datasets, containing cardiomyocytes (CM), cardiac endothelia (EC), cardiac fibroblasts (FC), and smooth muscle cells (SMC). For each dataset genes were excluded if they were not protein-coding, or if they were expressed at < 1RPKM across all four cell-types.

F5 *(FANTOM5):* Cap Analysis of Gene Expression (CAGE) data for robust CAGE peaks was obtained from the FANTOM5 consortium: http://fantom.gsc.riken.jp/5/data/[44]. Tag-per-million normalised CAGE peak expression levels were aggregated by sum at gene level. n=3, 4, 6, and 3 for CM, EC, FC, and SMC, respectively.

EN *(ENCODE):* FPKM-level RNA-seq data for cultured primary cells were accessed from the ENCODE consortium[57]; n=2 for all cell-types.

SC *(Single-cell):* single-cell RNA-seq data from freshly-isolated tissue samples were accessed from Wang *et al.* (2020)[58] (GSE109816). Only left atrial samples were used. Cell-type specific expression was generated as the average RPKM of all cells in each classification. n=1934, 1111, 257, and 427 for CM, EC, FC, and SMC, respectively.

640 **RNA-seq data generated in the present study and data pre-processing**

641 - **RNA Mixture Experiment**

642 Total RNA was extracted from human primary astrocytes and from neurons derived from
643 human foetal neural progenitors.

644 Human primary astrocytes (Lonza, #CC-2565) stably expressing GFP from pCMV6-AC-GFP
645 had been generated by selection with G418 (Thermo Fisher Scientific, #10231027) at
646 800µg/ml. Cells were cultured in RPMI GlutaMAX™ (Thermo Fisher Scientific, #35050061)
647 supplemented with 10% foetal bovine serum, 1% streptomycin (10,000 µg/ml), 1% penicillin
648 (10,000 units/ml) and 1% Fungizone (2.5 µg/ml) and seeded into 6-well tissue culture plates at
649 a density of 0.5×10^6 cells 24 hours prior to RNA extraction. Total RNA was extracted using
650 TRIzol® reagent and a Qiagen miRNeasy kit and treated with 1 µl DNase I (Thermo Fisher
651 Scientific, #AM2238) per 10 µg of RNA.

652 Neuronal differentiation of human neural progenitors stably transfected with pLRC-GFP was
653 carried out for 2 weeks as previously described[59]. RNA extraction was carried out using a
654 Qiagen miRNeasy kit, with on-column DNase digestion. RNA from differentiated neurons was
655 kindly provided by Dr. Brent Fogel (UCLA)[59].

656 RNA mixtures were generated by mixing neuronal and astrocyte RNA in mass ratios of 40:60,
657 45:55, 50:50 neuron:astrocyte (n=1 for each ratio). In addition, a pure neuronal RNA sample
658 and pure astrocyte RNA samples (n=3) were also included (Supplementary Table 7).

659 Library preparation using the Illumina TruSeq Stranded kit
660 (http://www.illumina.com/products/truseq_stranded_total_rna_library_prep_kit.html) and
661 sequencing on a NextSeq 500 Illumina sequencer were carried out at the UNSW Ramaciotti
662 Centre for Genomics, generating 75 bp paired-end reads (Supplementary Table 7). Sequencing
663 reads were mapped to the human genome (hg19) using STAR v2.5.2b[60] with the following
664 parameters: --outSJfilterOverhangMin 5 5 5 5 --alignSJoverhangMin 5 --
665 alignSJDBoverhangMin 5 --outFilterMultimapNmax 1 --outFilterScoreMin 1 --
666 outFilterMatchNmin 1 --outFilterMismatchNmax 2 --chimSegmentMin 5 --chimScoreMin
667 15 --chimScoreSeparation 10 --chimJunctionOverhangMin 5.

668 Gene counts for GENCODE V19 annotated genes were obtained from the STAR output and
669 RPKM-normalised.

670 ***IH (in house) cell-type signature*** data includes the RPKM-normalised data from the the pure
671 neurons and astrocyte samples (averaged across the 3 replicates for astrocytes). Data was
672 thresholded for a minimum of 1 RPKM in at least one cell-type.

673 ***RNA mixture data*** consists of RPKM-normalised data from the three RNA mixture samples.
674 Genes expressed at < 2 RPKM in at least one sample were filtered out.

675 - **Bulk RNA-seq data generated from brain tissue**

676 Brain tissue samples were obtained from the NICHD Brain and Tissue Bank, and included
677 frontal cortex samples (BA9/10) from 2 control, 2 ASD, and 1 Fragile-X premutation carrier
678 individuals. For each brain sample, frozen tissue was pulverised using a CellCrusher
679 (https://cellcrusher.com/) and the tissue was then divided for nuclear RNA extraction and RNA
680 extraction from bulk tissue.

681 **Nuclei Isolation**

682 Around 30 mg of tissue was lysed in 2.5 ml lysis buffer (10 mM Tris-HCl, 3 mM MgCl2, 10
683 mM NaCl, 0.005% NP40) for 17 minutes on ice. After lysis, 2.5 ml of ice-cold PBS was added

684 to the sample and tissue was homogenized using a Pasteur pipette until no large chunks were
685 visible. Tissue was then filtered through a 30 µm strainer and centrifuged at 500 g for 5 minutes
686 at 4°C. Supernatant was removed and the pellet was resuspended in 400 µl PBS with 1% BSA
687 and DAPI. DAPI-positive singlet nuclei were sorted using a BD Influx with a 70 µm nozzle at
688 20 PSI to collect approx. 100,000 nuclei per sample.

689 **Bulk RNA extraction and library generation**

690 To extract RNA from bulk tissue, the Qiagen mini RNA prep kit was used following the
691 manufacturer's instructions, including a DNAse treatment step. From sorted nuclei, RNA was
692 extracted by a hot Trizol extraction method. Nuclei were washed in PBS and resuspended in
693 Trizol at 65°C and incubated on a shaker at 1,300 rpm for 15 min. RNA was enriched using a
694 guanidinium HCl buffer and silica-coated magnetic beads with a DNAse I treatment step. RNA
695 amounts and quality were assessed on a TapeStation using RNA Screen Tape (Agilent), and
696 20-100 ng of total RNA was used per replicate to generate RNA-seq libraries. ERCC ExFold
697 RNA Spike-In mixes (Thermo Scientific) were added as internal control. Libraries were
698 prepared using the TruSeq Stranded mRNA library prep kit (Illumina), using TruSeq RNA
699 unique dual index adapters. Libraries were quantified by qPCR on a CFX96/C1000 cycler
700 (Biorad) and sequenced on a Novaseq 6000 (Illumina) for 2x 53 bp as paired-end, generating
701 around 25 M reads per sample.

702 Sequencing reads were mapped to the human genome (hg38) using STAR v2.5.2b[60] with the
703 following parameters: --outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15 --
704 alignSJDBoverhangMin 15 --outFilterMultimapNmax 1 --outFilterScoreMin 1 --
705 outFilterMatchNmin 1 --outFilterMismatchNmax 2 --chimSegmentMin 15 --chimScoreMin
706 15 --chimScoreSeparation 10 --chimJunctionOverhangMin 15 --bamRemoveDuplicatesType
707 UniqueIdenticalNotMulti. Note that nuclear samples were mapped to a pre-mRNA hg38
708 transcriptome.

709 Gene counts for GENCODE V19 annotated genes were obtained from the STAR output and
710 RPKM-normalised.

711 Nuclear enrichment was confirmed using the expression of the nuclear-specific transcript
712 MALAT1 (22.1-fold enrichment in nuclear samples, $p = 6.7x10^{-5}$, t-test; Supplementary Table
713 8)

714 - **Single-nucleus RNA-seq data generated from bulk brain tissue**

715 snRNA-seq data were generated from the same five brain samples described in the previous
716 section, but from a different chunk of the dissection.

717 **Nuclei Isolation**

718 Around 30 mg of tissue was lysed in 400 µl of lysis buffer (10 mM Tris-HCl, 3 mM MgCl2,
719 10 mM NaCl, 0.005% NP40) in 1.5 ml tubes and broken down with a pellet pestle. Tissue was
720 dissociated by passing through a polished silanized Pasteur pipette 3-4 times, then incubated
721 on ice for 10 minutes. Dissociation was repeated at 5 minutes and 10 minutes. After incubation,
722 the dissociated tissue was added to 2.5 ml of wash buffer in a 15 ml falcon tube. The sample
723 was then passed through a 30 µm strainer into a 50 ml falcon tube and centrifuged for 5 minutes
724 at 500 x g at 4°C in a swinging bucket centrifuge. Following centrifugation, the supernatant
725 was removed and sample resuspended in 100 µl of wash buffer (PBS with 1% BSA) for every
726 30 mg of tissue used. Using only 100 µl of the resuspended sample, 180 µl of a 1.8 M sucrose
727 solution (made with Sigma Nuclei Pure Prep kit) was added and homogenized using a P1000
728 pipette. In a 2 ml Eppendorf tube, 1 ml of a 1.3 M sucrose solution with 1% BSA was placed.
729 280 µl of the nuclei suspension mixed with sucrose was slowly layered on top of the 1.3 M

730 sucrose solution. The sucrose gradient was centrifuged for 10 minutes at 3,000 x g at 4°C in a
731 swinging bucket centrifuge. After centrifugation, the debris from the top of the sucrose gradient
732 were removed by soaking a Kimwipe from the top of the tube, slowly lowering it together with
733 the sinking meniscus until a volume of less than 100 µl remained, which was removed with a
734 pipette. Nuclei were resuspended in 20-50 µl of wash buffer and 10 µl of the suspension was
735 stained with Trypan Blue to count for concentration.

736 **Library generation and sequencing**

737 The 10X Genomics 3' v2 and v3 single cell expression kit was used to generate single nuclei
738 RNA-seq libraries. Using 16,000 nuclei in total to aim for 10,000 nuclei recovery, the standard
739 protocol was used according to manufacturer instruction with the following alterations: 17 PCR
740 cycles in total for cDNA amplification and 13 PCR cycles in total for library amplification.
741 Libraries were then sequenced on a Novaseq 6000 (Illumina) generating around 100 M reads
742 per sample.

743 **Sequence Alignment and UMI counting**

744 A pre-mRNA transcriptome was built using the Cell Ranger mkref command and default
745 parameters starting with the refdata-cellranger-GRCh38-1.2.0 transcriptome as per the
746 instructions provided by 10X Genomics. Reads were demultiplexed by sample index using the
747 Cell Ranger mkfastq command. Fastq files were aligned to the custom transcriptome, cell
748 barcodes were demultiplexed, and UMIs corresponding to genes were counted using the cell
749 ranger count command using default parameters. Cell Ranger version 2.1.0 was used for all
750 steps.

751 **Data preprocessing**

752 Cell Ranger output was pre-processed using Seurat v3[55]. Filtering-out criteria for nuclei: < 500
753 counts, or < 200 expressed genes, or >20% of counts attributed to mitochondrial genes, or total
754 number of reads in the top 99.5th percentile of its dataset. UMI counts were log2-transformed
755 and normalised for library size and mitochondrial percentage, and finally scaled. Nuclei from
756 all individuals were then integrated using canonical correlation analysis in Seurat, setting the
757 numbers of dimensions to be 30.

758 After retransformaing and renormalising data, clustering was performed using tSNE[61] on the
759 top 35 principal components of the 2000 most variable genes, with the resolution parameter set
760 to 1.5 (Supplementary Fig.40). Clusters were annotated using SingleR[62] to transfer cell-type
761 annotation labels from the NG signature.

762 **Signature generation**

763 A separate cell-type specific signature was generated from each of the five individuals. This
764 was calculated as the average RPKM of each individual's cells within each cluster. Only cell-
765 types represented in all individuals were used (Neurons, Astrocytes, and Oligodendrocytes).

766

767 **<u>Simulated datasets</u>**

768 - **Simulations for assessing deconvolution accuracy**

769 **Randomly sampled single-nucleus mixtures** were generated using data from VL[36] and the
770 CA[37] datasets.

771 Simulated data was generated separately from each dataset. Seurat v3 was used to pre-process
772 raw count expression data, removing nuclei with 1) fewer than 1000 counts or 200 expressed
773 genes, 2) >5% of counts attributed to mitochondrial genes, or 3) a number of reads >99.5th

774 percentile of its dataset. In addition, cells assigned to a cell-type or cell-subtype with fewer
775 than 200 cells were excluded. Next, the dataset was randomly split into two: half was used to
776 generate cell-type signatures and half for simulated mixture. One hundred mixtures were
777 simulated by summing the counts of 500 randomly-sampled single nuclei. Random sampling
778 was performed without replacement.

779 **Randomly sampled single-cell mixtures** were generated using single-cells from the Darmanis
780 *et al.* dataset[13], using a method largely as above but with three key differences: first, only cells
781 classified as one of neurons, astrocytes, oligodendrocytes, OPCs, microglia, or endothelia were
782 included, without regard for the number of representatives (*i.e.,* non-hybrid cells from adult
783 samples); second, the number of cells aggregated per mixture was only 100, owing to the lower
784 number of total cells (285); and finally, the dataset was not randomly split into two for mixture
785 and signature generation.

786 We confirmed that the single-nucleus and single-cell simulated mixtures, had similar
787 expression distributions to data from bulk brain tissue, and were not zero-inflated
788 (Supplementary Fig.41).

789 **Single-nucleus mixtures with a wide range of cell-type compositions** were generated using
790 single nuclei from the VL and CA datasets. 100 mixtures were simulated. To obtain a defined
791 range of cell-type proportions in the mixture, for each cell-type j we randomly sampled without
792 replacement between 1 and n_j nuclei where $n_j = (n/k)/(s_j/\min(s))$ where n=500, the chosen
793 number of cells per mixture; k is the number of cell-types in each dataset; s is the vector of
794 total library sizes for all k cell-types; s_j is the total library size for cell-type j.

795 If more than 500 total nuclei were randomly-sampled by this approach, then a random subset
796 of 500 was kept; conversely, if fewer than 500 nuclei were initially sampled, then additional
797 nuclei were randomly-sampled from any cell-type until 500 was reached. Mixtures were
798 simulated by summing the counts of these single nuclei followed by counts-per-million
799 normalisation.

800 • **Simulations for assessing the interaction between composition and differential**
801 **expression**

802 **Simulated data with cell-type composition differences between sample groups**

803 Single-nucleus mixtures for DE analyses were generated using snRNA-seq data from the CA
804 dataset. Nuclei were classified as one of Excitatory, Inhibitory, Oligodendrocyte, OPC, or
805 Astrocyte. Each simulation was created as a dataset of 100 samples, split into groups A and B
806 of 50 samples each. Each sample in group A (the reference group) was generated as the
807 summed expression of randomly selected n excitatory neurons and 500-n non-excitatory cells,
808 where n was a randomly selected integer from [200-300] so that the simulated proportion of
809 excitatory neurons varies between 40-60%). Samples in group B (test group) were generated
810 as per group A, except n was sampled from [200+k, 300+k] for increased proportions or [200-
811 k, 300-k] for decreased proportions where k varied from 0 to 195 with a step of 5. All sampling
812 was performed without replacement. Differential expression analyses for group B *vs.* group A
813 were performed on each dataset as described in the "Differential Expression" section below.

814 **Simulated data with cell-type composition and gene expression differences between**
815 **sample groups**

816 The expression of expression of 200 genes was altered by 1.1-, 1.3-, 1.5-, or 2-fold in the above
817 simulated mixtures, in group A samples only. The 200 genes selected for perturbation included
818 the top 100 excitatory neuron marker genes and 100 randomly-selected non-marker genes. Half
819 of each set was randomly assigned to be upregulated or downregulated.

820 To simulate cell-type-specific expression differences, the expression alteration was introduced
821 only to nuclei from the cell-type-of-interest (*i.e.* excitatory or inhibitory neurons) prior to
822 aggregation.

823 **Estimation of cellular composition**

824 - ***Overview of deconvolution methods***

825 In general, deconvolution methods model gene expression data from a tissue sample (vector X)
826 as the sum of gene expression levels in the cell-types of which it's comprised ("signature"
827 expression matrix, S), weighted by the proportion of each cell-type in the sample (vector P),
828 formalized as $X \sim S*P$. Deconvolution methods fall into two broad categories – partial and
829 complete – as described below.

830 ***Partial or supervised deconvolution***[6,18,29,31,63–68] estimates the proportion of cell-types in a
831 sample based on experimentally measured gene expression values from pure cell-types, *i.e.*
832 determines P knowing X and S.

833 It is worth noting that the signature expression data (S) often comes from a different source
834 than the bulk tissue data (X), and thus an intrinsic assumption of most partial deconvolution
835 methods is that gene expression in a given cell-type is the same regardless of the source of cells
836 (thus genetic background and environmental conditions including culture conditions are
837 ignored)[1,68]. The most frequently employed methods for partial deconvolution are Non-
838 negative Least Squares (*i.e.* optimising $X \sim S*P$ using a least-squares approach where P should
839 be non-negative) (*e.g.* DeconRNASeq[29]), and Support Vector Regression (*e.g.*
840 CIBERSORT[18]).

841 A simplified version of partial deconvolution consists of calculating an enrichment score,
842 rather than a proportion, for each cell-type (*e.g.* xCell[19], or BrainInABlender[7]). While this
843 approach is intuitive, it has several limitations: its accuracy is harder to assess (as one cannot
844 calculate error measures or goodness-of-fit), and its biological interpretation is often unclear
845 since the scale of enrichment scores is variable.

846 ***Complete or reference-free/unsupervised deconvolution*** consists of estimating both the
847 proportion of cell-types and cell-type specific expression, *i.e.* determining both P and S
848 knowing X[33–35,45,69]. This is an under-determined problem, which requires biologically
849 motivated constraints.

850 - **Deconvolution methods used**

851 Cell-type composition was estimated using four partial deconvolution methods
852 (**DeconRNASeq**[29], **dtangle**[31], **MuSiC**[30], and **CIBERSORT**[18]), two enrichment methods with
853 in-built signatures (**BrainInABlender**[7] and **xCell**[19]), and two complete deconvolution
854 methods: **Linseed**[33], and a co-expression based approach proposed by Kelley *et al.*[5] (referred
855 to as **Coex**).

856 All algorithms were run in R v3.6. All data used for deconvolution were RPKM-normalised
857 expression values without log2 transformation[70] unless noted below.

858 **CIBERSORT v1.04** was run using the *CIBERSORT* R package obtained from
859 https://cibersort.stanford.edu with default parameters.

860 **DeconRNASeq v1.26** was run using the *DeconRNASeq* Bioconductor R package with default
861 parameters.

862 **Music** v0.1.1 was run using the music_prop() function from R package available at
863 https://github.com/xuranw/MuSiC. Raw count data was used as input for both signatures and

mixtures. Only single-cell- or single-nucleus-derived signatures were used; their individual cells/nuclei were not aggregated, metadata about the individual-of-origin was included as well as predefined cell-type labels.

dtangle v0.3.1 was run using the *dtangle* CRAN R package. Cell-type markers were selected as the top 1% of markers using its *find_markers* function with method="diff". Data was log2 transformed with an offset 0.5, as recommended[31].

BrainInABlender v0.9 was run using the R package obtained from https://github.com/hagenaue/BrainInABlender using default parameters. Cell-type signature data built into BrainInABlender is derived from numerous resources of brain cell-type specific expression, including human data from Darmanis *et al.*[13], and various mouse datasets (full list in Hagenauer *et al.,* 2018). Both publication-specific indices and an averaged index are generated; we used the averaged index as the enrichment score in all analyses.

xCell v1.1.0 was run using the R package from https://github.com/dviraran/xCell using default parameters with the built-in signature data. Cell-type signature data for neurons and astrocytes are built in xCell, and are derived from *in vitro* cultured data from FANTOM5[44], and ENCODE[57]. xCell generates a "Raw" and a "Transformed" enrichment score; we used the latter as a measure of enrichment.

Coex was carried out by constructing co-expression networks using the *blockwiseModules* function from the WGCNA R package[46,71], with the following parameters: deepSplit = 4, minModuleSize = 150, mergeCutHeight = 0.2, detectCutHeight = 0.9999, corType = "bicor", networkType = "signed", pamStage = FALSE, pamRespectsDendro = TRUE, maxBlockSize = 30000. The beta power was selected for each network so that the scale-free topology fit r^2 was > 0.8 and median connectivity < 100 (Supplementary Information Code). Genes were assigned to the module with the highest kME (correlation with the module eigengene), provided $kME > 0.5$, and $p < 0.05$ (BH-corrected Student's t-test). Co-expression networks were built on log2 transformed RPKM values, offset by 0.5.

A cell-type module (CTM) was defined as the module most significantly enriched for the top 100 markers of a given cell-type, requiring an enrichment p-value $<10^{-5}$ and odds ratio >5.

Enrichment was assessed using a one-sided Fisher's Exact Test. Cell-type markers were defined using the *find_markers* function in the dtangle R package applied to the matching cell-type signature data for simulations, and MB for GTEx and Parikshak. Cell-type enrichment scores were defined as the CTM's eigengene values (*i.e.*, first principal component values of genes included in the CTM), as per Kelley *et al.*[5].

Linseed v0.99.2 was run using the R package from https://github.com/ctlab/LinSeed. We used a collinearity threshold of p=0.01 to filter genes. Output was transformed to sum-to-one.

We also tested the SVD approach to determine the number of cell-types in the mixture data, which involves looking for the plateau (Supplementary Figure 26). For the VL-based simulations, with 7 cell-types, the estimated k was greater than 10 for random and 7 for wide-range mixtures. For the CA-based simulations, with 5 cell-types, the estimated k was ~5-7 for random and 5 for wide-range mixtures. For the RNA mixtures which consisted of 2 cell-types, the estimated k was 3. For the DM mixtures, which consisted of 5 cell-types, the estimated k was more than 10. Therefore, we used the known k value for all mixtures.

- **Deconvolution of specific datasets**

Parikshak, GTEx and RNA mixture data were deconvolved using RPKM-normalised signatures and mixtures, while for single-cell and single-nucleus simulated datasets, signatures

and mixtures were CPM-normalised, if raw count data was available (otherwise the normalised data available from the original publication was used).

Assessment of deconvolution accuracy

For simulated datasets, deconvolution accuracy was assessed by two measures: (i) Pearson correlation between true and estimated proportions and (ii) normalised mean absolute error (nmae) calculated as mean error divided by the mean of true proportions, where error is the per-sample absolute difference between estimate and true proportion. Note that nmae can only be calculated when estimates are bounded between 0-1 *i.e.* are proportions rather than relative enrichment scores like Blender's or xCell's output.

For datasets without a ground truth, such as bulk brain samples, goodness-of-fit was evaluated as the Pearson correlation for each sample's reconstructed and observed expression, log2-transformed with an offset of +0.5.

First, observed expression and cell-type signature data were quantile normalised. Then, for each sample, reconstructed expression values were calculated using the following formula:

$$\sum_{j=1}^{n} s_j . p_j$$

where, j denotes a cell-type, s_j is the vector of gene expression in cell-type j (from the signature matrix), and p_j is the estimated proportion of cell-type j in the sample, and n is the number of cell-types.

Differential expression (DE) analyses

Differential expression analyses in simulated data

DE between group A and group B in simulated single-nucleus mixtures was assessed using either a linear model on log2-transformed CPM values offset by +0.5 as implemented in the *lm* function in R, or a generalised linear model implemented in DESeq2[47] on count data.

Excitatory neuron proportions were included as covariates in the model either as linear term, a quadratic term, or after conversion to a spline matrix using the bs() function from the R *splines* package, setting degree = 3 and knots at its 25th, 50th, and 75th percentiles.

Multiple testing correction was conducted using the Benjamini-Hochberg approach[72].

Cell-type marker enrichment analyses were performed by one-sided Fisher's exact test for 100 markers per cell-type using the CA cell-type signature. Markers were defined using the *find_markers* function from dtangle[31], setting marker_method = "diff". Only simulations where the number of false positive genes was >100 were tested for cell-type enrichment, to ensure adequate power for the test.

Discriminatory power was calculated as the fraction of the 200 true perturbed genes that were in the top 200 most significant genes by p-value. No significance threshold was applied to DE p-values.

Cell-type-specific DE analyses

Cell-type-specific expression was first extracted using the high-resolution algorithm of CIBERSORTx[48] webtool at https://cibersortx.stanford.edu/ with default settings using the CA signature. Due to the webtool's computational constraints, a subset of 1000 genes was analysed, including the 100 perturbed cell-type marker genes, the 100 perturbed non-marker genes, the top 100 markers for each of the four other cell-types, and 400 randomly-selected non-marker genes. The resultant cell-type-specific expression data was used for DE analyses,

951 using a linear model on log2 transformed data offset by +0.5. Multiple testing correction was
952 conducted using the Benjamini-Hochberg approach, adjusting for the size of the full
953 transcriptome rather than that of the smaller subset[72].

954 **DE analyses of ASD and control samples**

955 DE was carried out using DESeq2 v1.22.2[47] on count-level expression data. The same samples
956 used by Parikshak *et al.*[28] for DE were included in our analyses: 106 samples (43 ASD, 63
957 controls; Supplementary Table 5). Differential expression was carried out using a Wald test
958 with Benjamini-Hochberg correction for multiple testing as implemented in DESeq2[47].
959 Composition-dependent DE adjusted for the following covariates: Age, Sex, Sequencing
960 Batch, Brain Bank, Region, RIN, and the first two principal components of sequencing
961 metadata, per Parikshak *et al.*[28]. Composition-independent DE used the same covariates as
962 above, but adding the estimated proportions of astrocytes and any other cell-types not
963 significantly correlated with astrocyte proportions ($p<0.05$, Pearson correlation test) *i.e.,*
964 oligodendrocyte and microglia (Supplementary Figure 42), to minimise co-linearity.

965 **Determination of compartment-specific genes**

966 Compartment-specific genes were identified using the RNA-seq data generated from bulk brain
967 tissue (5 total RNA and 5 nuclear RNA samples) pre-processed as described above, and log2-
968 transformed with an offset of 0.5. Compartment specific genes were identified as genes DE
969 between groups using a linear model as implemented in the *lm* function in R (absolute fold-
970 change > 1.3 and a Benjamini-Hochberg-adjusted[72] p-value < 0.05).

971 **<u>Other analyses</u>**

972 Gene ontology (GO) and pathway enrichment analyses were conducted using gProfiler2 v0.2[73]
973 in R, setting exclude_iea=TRUE and all other parameters as default. P-values were BH-
974 corrected[72]. Only results from GO, KEGG, Reactome, Human Phenotype, and Wikipathways
975 were reported, with filtering performed after multiple-testing correction.

976 For all set enrichment analyses, the background was set to the relevant list of all expressed
977 genes.

978 **<u>Note on cell-type proportions vs. RNA proportions.</u>**

979 Since cell-types differ in their total RNA content, transcriptome deconvolution estimates
980 proportions of RNA from each cell-type, rather than proportions of cells *per se*[33]. It is important
981 to note that bulk RNA-seq sequences a mixture of RNA molecules (primarily protein coding,
982 after poly-A selection or ribo-depletion), and thus the goal of transcriptome deconvolution is
983 in fact to estimate the proportion of the sequenced RNA molecules coming from a given cell-
984 type (pRNA), rather than the proportion of cells. *A-priori,* pRNA (not pCt) should be relevant
985 for reconstruction of gene expression data, and thus useful as a co-variate in differential
986 expression analyses. To test this hypothesis, we deconvolved pseudo-bulk data from the
987 Velmeshev *et al.* snRNA-seq dataset (10 individuals), where we know both pCt and pRNA
988 (calculated as the proportion of RNA-seq reads from each cell-type), and found that
989 deconvolution estimates perfectly correlate with pRNA but less so with pCt (Supplementary
990 Figure 43), consistent with previous data[33]. Note that pRNA and pCt are themselves correlated
991 in this dataset (r=0.86). We then assessed goodness-of-fit for these pseudo-bulk data using
992 either pRNA or pCt to reconstruct gene expression. We found that goodness-of-fit was always
993 higher when using pRNA (Supplementary Figure 43). These data demonstrate that pRNA, the
994 output of transcriptome deconvolution, is the appropriate measure to use for re-constructing
995 gene expression data and thus as a co-variate in DE analyses. For simplicity, we refer to pRNA
996 as "cell-type proportions" throughout the manuscript.

997 **Data availability**

998 The sequencing data generated in the present study is available on GEO/SRA accession number
999 GSE175772.

1000 **Code availability**

1001 Data analysis code is available at: https://github.com/Voineagulab/BrainCellularComposition.

1002 All brain cell-type signatures are available as an R package:
1003 https://github.com/Voineagulab/brainyR.

1004 Deconvolution with the top performing algorithms is implemented at:
1005 https://voineagulab.shinyapps.io/BrainDeconvShiny/.

1006 **Acknowledgements:** This work was supported by an ARC Future Fellowship and a UNSW
1007 Scientia Fellowship to I.V. and an RTP PhD Scholarship to G.J.S.

1008

1009 References

1010 1. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational
1011 deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**,
1012 1969–1979 (2018).

1013 2. Mohammadi, S., Zuckerman, N. S., Goldsmith, A. & Grama, A. A Critical Survey of
1014 Deconvolution Methods for Separating Cell Types in Complex Tissues. *Proc. IEEE* **105**,
1015 340–366 (2017).

1016 3. Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-Type
1017 Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals
1018 Disease-Relevant Cell-Specific eQTLs. *Am. J. Hum. Genet.* (2019).
1019 doi:10.1016/j.ajhg.2019.03.025

1020 4. Pelvig, D. P., Pakkenberg, H., Stark, A. K. & Pakkenberg, B. Neocortical glial cell
1021 numbers in human brains. *Neurobiol. Aging* **29**, 1754–1762 (2008).

1022 5. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among
1023 intact tissue samples reveals the core transcriptional features of human CNS cell classes.
1024 *Nat. Neurosci.* **21**, 265397 (2018).

1025 6. Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data.
1026 *Nat. Methods* **16**, 327–332 (2019).

1027 7. Hagenauer, M. H. *et al.* Inference of cell type content from human brain transcriptomic
1028 datasets illuminates the effects of age, manner of death, dissection, and psychiatric
1029 diagnosis. *PLoS One* **13**, 89391 (2018).

1030 8. Yang, L. *et al.* Transcriptomic Landscape of von Economo Neurons in Human Anterior
1031 Cingulate Cortex Revealed by Microdissected-Cell RNA Sequencing. *Cereb. Cortex* **29**,
1032 838–851 (2019).

1033 9. Kuhn, A. *et al.* Cell population-specific expression analysis of human cerebellum. *BMC*
1034 *Genomics* **13**, 610 (2012).

1035 10. Mendizabal, I. *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain.
1036 *Genome Biol.* **20**, 135 (2019).

1037 11. Blodgett, D. M. *et al.* Novel Observations From Next-Generation RNA Sequencing of

1038 Highly Purified Human Adult and Fetal Islet Cell Subsets. *Diabetes* **64**, 3172–81 (2015).

1039 12. Furuyama, K. *et al*. Diabetes relief in mice by glucose-sensing insulin-secreting human
1040 α-cells. *Nature* **567**, 43–48 (2019).

1041 13. Darmanis, S. *et al*. A survey of human brain transcriptome diversity at the single cell
1042 level. *Proc. Natl. Acad. Sci.* **112**, 7285–7290 (2015).

1043 14. Consortium, Gte. Genetic effects on gene expression across human tissues. *Nature* **550**,
1044 204–213 (2017).

1045 15. Wang, D. *et al*. Comprehensive functional genomic resource and integrative model for
1046 the human brain. *Science (80-.).* **362**, eaat8464 (2018).

1047 16. Hoffman, G. E. *et al*. CommonMind Consortium provides transcriptomic and
1048 epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **6**, 1–14 (2019).

1049 17. Miller, J. A. *et al*. Transcriptional landscape of the prenatal human brain. *Nature* **508**,
1050 199–206 (2014).

1051 18. Newman, A. M. *et al*. Robust enumeration of cell subsets from tissue expression
1052 profiles. *Nat. Methods* **12**, 453–7 (2015).

1053 19. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular
1054 heterogeneity landscape. *Genome Biol.* **18**, (2017).

1055 20. Sturm, G. *et al*. Comprehensive evaluation of transcriptome-based cell-type
1056 quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

1057 21. Ramaker, R. C. *et al*. Post-mortem molecular profiling of three psychiatric disorders.
1058 *Genome Med.* **9**, 72 (2017).

1059 22. Xu, X., Nehorai, A. & Dougherty, J. D. Cell type-specific analysis of human brain
1060 transcriptome data to predict alterations in cellular composition. *Syst. Biomed.* **1**, 151–
1061 160 (2013).

1062 23. Mancarci, B. O. *et al*. Cross-laboratory analysis of brain cell type transcriptomes with
1063 applications to interpretation of bulk tissue data. *eNeuro* **4**, ENEURO-0212 (2017).

1064 24. Grove, J. *et al*. Identification of common genetic risk variants for autism spectrum
1065 disorder. *Nat. Genet.* **51**, 431–444 (2019).

1066 25. Li, Z. *et al*. Genetic variants associated with Alzheimer's disease confer different
1067 cerebral cortex cell-type population structure. *Genome Med.* **10**, 43 (2018).

1068 26. McCoy, M. J. *et al*. LONGO: an R package for interactive gene length dependent
1069 analysis for neuronal identity. *Bioinformatics* **34**, i422–i428 (2018).

1070 27. Wang, J., Devlin, B. & Roeder, K. Using multiple measurements of tissue to estimate
1071 subject- and cell-type-specific gene expression. *Bioinformatics* (2019).
1072 doi:10.1093/bioinformatics/btz619

1073 28. Parikshak, N. N. *et al*. Genome-wide changes in lncRNA, splicing, and regional gene
1074 expression patterns in autism. *Nature* **540**, 423–427 (2016).

1075 29. Gong, T. & Szustakowski, J. D. DeconRNASeq: A statistical framework for
1076 deconvolution of heterogeneous tissue samples based on mRNA-Seq data.
1077 *Bioinformatics* **29**, 1083–1085 (2013).

1078 30. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type
1079 deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**,

380 (2019).

31. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* 290262 (2018). doi:10.1093/bioinformatics/bty926

32. Aguet, F. *et al*. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *BioRxiv* 787903 (2019).

33. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* **10**, 2209 (2019).

34. Zhu, Y., Wang, N., Miller, D. J. & Wang, Y. Convex analysis of mixtures for separating non-negative well-grounded sources. *Sci. Rep.* **6**, 38350 (2016).

35. Wang, N. *et al*. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6**, (2016).

36. Velmeshev, D. *et al*. Single-cell genomics identifies cell type–specific molecular changes in autism. *Science (80-.).* **364**, 685–689 (2019).

37. Hodge, R. D. *et al*. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

38. Zhang, Y. *et al*. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).

39. Vallania, F. *et al*. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat. Commun.* **9**, 4735 (2018).

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

41. Lake, B. B. *et al*. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

42. Tasic, B. *et al*. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

43. Zhang, Y. *et al*. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–47 (2014).

44. Forrest, A. R. R. *et al*. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

45. Wang, N. *et al*. UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics* **31**, 137–139 (2015).

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

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

48. Newman, A. M. *et al*. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

1122 49. Reay, W. R. & Cairns, M. J. Pairwise common variant meta-analyses of schizophrenia
1123 with other psychiatric disorders reveals shared and distinct gene and gene-set
1124 associations. *BioRxiv* 725614 (2019).

1125 50. Schelker, M. *et al*. Estimation of immune cell content in tumour tissue using single-cell
1126 RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).

1127 51. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K.
1128 Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat.*
1129 *Commun.* **11**, 1–14 (2020).

1130 52. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic
1131 testing environments. *Genome Biol.* **22**, 1–23 (2021).

1132 53. Guintivano, J., Aryee, M. J. & Kaminsky, Z. A. A cell epigenotype specific model for
1133 the correction of brain cellular heterogeneity bias and its application to age, brain region
1134 and major depression. *Epigenetics* **8**, 290–302 (2013).

1135 54. Lin, P., Troup, M. & Ho, J. W. K. CIDR: Ultrafast and accurate clustering through
1136 imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).

1137 55. Stuart, T. *et al*. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-
1138 1902.e21 (2019).

1139 56. Enge, M. *et al*. Single-Cell Analysis of Human Pancreas Reveals Transcriptional
1140 Signatures of Aging and Somatic Mutation Patterns. *Cell* **171**, 321-330.e14 (2017).

1141 57. Djebali, S. *et al*. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

1142 58. Wang, L. *et al*. Single-cell reconstruction of the adult human heart during heart failure
1143 and recovery reveals the cellular landscape underlying cardiac function. *Nat. Cell Biol.*
1144 **22**, 108–119 (2020).

1145 59. Fogel, B. L. *et al*. RBFOX1 regulates both splicing and transcriptional networks in
1146 human neuronal development. *Hum. Mol. Genet.* **21**, 4171–4186 (2012).

1147 60. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
1148 (2013).

1149 61. Van Der Maaten, L. & Hinton, G. *Visualizing Data using t-SNE. Journal of Machine*
1150 *Learning Research* **9**, (2008).

1151 62. Aran, D. *et al*. Reference-based analysis of lung single-cell sequencing reveals a
1152 transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

1153 63. Du, R., Carey, V. & Weiss, S. deconvSeq: Deconvolution of Cell Mixture Distribution
1154 in Sequencing Data. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz444

1155 64. Tsoucas, D. *et al*. Accurate estimation of cell-type composition from gene expression
1156 data. *Nat. Commun.* **10**, 2975 (2019).

1157 65. Shen-Orr, S. S. *et al*. Cell type–specific gene expression differences in complex tissues.
1158 *Nat. Methods* **7**, 287 (2010).

1159 66. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F.
1160 Deconvolution of blood microarray data identifies cellular activation patterns in
1161 systemic lupus erythematosus. *PLoS One* **4**, e6098 (2009).

1162 67. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex
1163 tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89

(2013).

68. Qiao, W. *et al*. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Comput. Biol.* **8**, e1002838 (2012).

69. Li, Z. & Wu, H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.* **20**, 190 (2019).

70. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8 (2012).

71. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54 (2007).

72. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 289–300 (1995).

73. Raudvere, U. *et al*. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

Figure Legends

Figure 1. Deconvolution accuracy across methods. A. Simulation design. Single-nucleus RNA sequencing data was acquired from Velmeshev *et al*. and used to create *in silico* mixtures with known proportions. *Left:* Piechart displaying the composition of the dataset; n: number of cells per cell-type. For each cell type the number of sub-types is listed in between brackets. *Right:* analysis outline. *OPC*: oligodendrocyte precursor cells. *Neurons_Exc*, *Neurons*_Inh, and *Neurons*_NRGN: excitatory, inhibitory, and NRGN$^+$ neurons, respectively. *DRS*: DeconRNASeq. *CIB*: *CIBERSORT*. *Blender*: BrainInABlender. **B.** Barplots of Pearson correlation coefficients (*r*) between true and estimated cell-type proportions in 100 in silico mixtures. *Left:* cells are grouped by major cell-types; *middle:* excitatory and inhibitory neuron subtypes are included in the signature; *right:* all cell-subtype labels are used in the signature. **C.** Barplots of Pearson correlation coefficients (*r*) between true proportion and cell-type enrichment scores in 100 *in silico* mixtures. *Dotted lines*: $r = 0.8$.

Figure 2. Effect of signature choice on deconvolution accuracy. A. Scatterplots of true and CIBERSORT-estimated proportions in VL *in silico* mixtures, for nine different signatures. *Matched:* the signature and mixture were derived from the same dataset. *VL*: Velmeshev. *NG*: Nagy. *CA*: Human Cell Atlas. *LK*: Lake. *TS*: Tasic. *Dotted line:* y=x. **B.** Barplots of normalized mean absolute error (*nmae*) for all cell-types and signatures presented in A. *Ast*: astrocytes. *End*: endothelia. *Mic*: microglia. *Oli*: oligodendrocytes. *OPC*: oligodendrocyte precursor cells. *Exc*: excitatory neurons. *Inh*: inhibitory neurons. *Dotted line:* nmae = 1. **C.** Barplots of Pearson correlation (*r*) for all cell-types and signatures presented in A. *Dotted line*: $r = 0.8$.

Figure 3. Reference-free deconvolution. A. Violin plots of the distribution of true cell-type proportions in VL, CA, and DM *in silico* random simulations (left, middle, and right respectively). *Black horizontal bar*: median. *Ast*: astrocytes. *End*: endothelia. *Exc*: excitatory neurons. *Inh*: inhibitory neurons. *Neu:* neurons. *Oli*: oligodendrocytes. *OPC*: oligodendrocyte precursors. **B.** Heatmaps of Pearson correlation coefficients between estimated and true cell-type proportions for random simulations based on VL, CA, and DM. *Top:* Linseed; y-axis: Cell-types defined by Linseed; x-axis: true cell-type in simulated data. *Bottom:* Coex; for each cell type the true vs. estimated correlation coefficient is displayed for the coexpression module

assigned to that cell-type based on marker enrichment (Methods); zero values represent cases where no coexpression module was assigned to the corresponding cell-type. **C.** Violin plots of the distribution of cell-type abundances in simulations with wide cell-type ranges based on VL (left) and CA (right). **D.** Heatmaps of Pearson correlation coefficients between estimated and true cell-type proportions for simulations with wide cell-type ranges displayed in C, based on VL and CA. *Top:* Linseed. *Bottom:* Coex.

Figure 4. Effect of brain cell-type composition on differential expression (DE) analyses. A. Scatterplot of the number of false positive genes versus the simulated difference in excitatory neuron proportion between two groups of 50 samples. Each point represents a different simulated dataset. DE was assessed with either a linear model (LM) or DESeq2, with or without correction for composition, *Coloured lines:* local regression line. **B.** Cell-type marker enrichment within false positive genes. Each point represents a single simulated dataset. *Y-axis*: enrichment p-value (one-sided Fisher test); Methods. *FPs*: false positive genes. **C.** Scatterplot of the discriminatory power, *i.e.* fraction of the 200 perturbed genes in the top 200 most significantly differentially expressed genes (y-axis) versus simulated difference in excitatory neuron proportion between sample groups (x-axis) for simulated 1.5-fold expression difference. *Coloured lines*: local regression line. *Dotted line:* expected discriminatory power, *i.e.* 0.95 times the discriminatory power in the absence of cell-type composition differences between groups. **D.** Model robustness to cell-type composition differences across a range of fold-changes, quantified as the smallest composition change where discriminatory power fell below its expected value. **E.** Cell-type specific DE analysis using CIBERSORTx for simulated mixtures with 1.5-fold difference in gene expression between two groups, without (*Left*) or with (*Right*) a superimposed cell-type composition difference between the two groups. Gene expression was perturbed in excitatory neurons for 100 non-marker genes and 100 excitatory neuron marker genes. The cell type composition of simulated data is shown in Supplementary Fig. 32. Y-axis: fraction of perturbed genes significantly DE at FDR < 0.05; DE was carried out using a linear model (Methods).

Figure 5. Cell-type composition estimates in large-scale human brain transcriptome data. A-B. Goodness of Fit across signatures in cortex samples from the GTEx consortium (A) and Parikshak *et al*. (B). Deconvolution was performed using CIBERSORT. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. **C.** Composition estimates in ASD (n=43) and Control (n=63) cortical samples. Deconvolution was performed using CIBERSORT and the MultiBrain signature. *ASD*: autism spectrum disorder. *CTL*: control. *: p<0.01. ***: p<0.001. **D.** Venn diagrams of the overlap between composition-dependent and composition-dependent DE genes between ASD and CTL samples.

Supplementary Figure Legends

Supplementary Figure 1. Generation and deconvolution of a replication simulated dataset using single-nuclei from the CA dataset. **A.** Process for generating 100 in silico mixtures. **B.** Barplots show Pearson correlation coefficients (r) between true and deconvolution-estimated proportions in 100 *in silico* mixtures. The left column shows results when only major cell-type labels are used in the signature; the middle column shows results when a mix of major cell-type and cell-subtype labels are used in the signature; the right column shows results when only cell-subtype labels are used in the signature. *Dotted line*: *r*=0.8.

Supplementary Figure 2. Benchmarking deconvolution algorithms on simulated mixtures using data from Darmanis *et al.* **A.** Simulation design. **B.** Scatterplots of

estimated proportion (or enrichment score) and true proportion for each cell-type. *Ast*: astrocytes. *End*: endothelia. *Mic*: microglia. *Neu*: Neurons. *Oli*: oligodendrocytes. *Red dotted line*: y=x. Grey line: regression line. **C.** Barplots of normalised mean absolute error *(nmae*; left) and Pearson correlation coefficients between true and estimated proportions (*r*; right) based on 100 *in silico* mixtures.

Supplementary Figure 3. Barplots of normalised mean absolute error (*nmae*) between true and deconvolution-estimated proportions in 100 VL-derived *in silico* mixtures. *Nmae* was calculated as the average error divided by the average true value. **A.** Using only major cell-type labels in the signature. **B.** Using a mix of major cell-type and cell-subtype labels in the signature. **C.** Using all cell-subtype labels are the signature. *Dotted black line*: *nmae* = 1.

Supplementary Figure 4. Scatterplots of true and deconvolution-estimated proportions in 100 VL-derived *in silico* mixtures. Each row represents a different algorithm. The signature used only major cell-types. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient. Note that nmae was not calculated for xCell and Blender as their output is an enrichment score rather than a proportion

Supplementary Figure 5. Barplots of normalised mean absolute error (*nmae*) between true and deconvolution-estimated proportions in 100 CA-derived *in silico* mixtures. *Nmae* was calculated as the average error divided by the average true value. **A.** Using only major cell-type labels in the signature. **B.** Using a mix of major cell-type and cell-subtype labels in the signature. **C.** Using all cell-subtype labels are the signature. *Dotted black line*: *nmae* = 1.

Supplementary Figure 6. Scatterplots of true and deconvolution-estimated proportions in 100 CA-derived *in silico* mixtures. Each row represents a different algorithm. The signature used only major cell-types. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient. Note that nmae was not calculated for xCell and Blender as their output is an enrichment score rather than a proportion

Supplementary Figure 7. Deconvolving mixtures of RNA from cultured neurons and astrocytes. A. Outline of RNA mixtures and its corresponding in-house (IH) signature. **B.** Scatterplots of estimated and true proportions of neurons using CIB, DRS and DTA, combined with the matching IH signature. Note that the MUS algorithm was not used, as the algorithm is only compatible with single-cell-level data. **C.** Scatterplots of neuron enrichment scores obtained with Blender (left) and xCell (right). **D.** Scatterplots of astrocyte enrichment scores obtained with Blender (left) and xCell (right). **E**. Scatterplots of true versus estimated neuronal proportion when using CIB and mismatched signatures. All signatures contained just neuronal and astrocyte expression values.

Supplementary Figure 8. Estimated proportions in immuno-panned purified brain. *Thick horizontal line*: mean. *Neu*: neurons. *Ast*: astrocytes. *Oli*: oligodendrocytes. *Mic*: microglia. *CIB*: CIBERSORT. *DRS*: DeconRNASeq. *DTA*: dtangle. Note that MuSiC was not applied as it requires single-cell or –nucleus data for its signature.

Supplementary Figure 9. Scatterplots of true and deconvolution-estimated proportions in 100 VL-derived *in silico* mixtures. The signature used a range of cell-subtypes and major cell-types. **A.** CIBERSORT deconvolution. **B.** DeconRNASeq. **C.** dtangle. **D.** MuSiC. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient. *Neurons_Inh and Neurons_Exc*: Inhibitory and excitatory neurons, respectively.

Supplementary Figure 10. Scatterplots of true and deconvolution-estimated proportions in 100 CA-derived *in silico* mixtures. Each row represents a different algorithm. The signature used a range of cell-subtypes and major cell-types. **A.** CIBERSORT deconvolution. **B.** DeconRNASeq. **C.** dtangle. **D.** Music. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient. *Neurons_Inh and Neurons_Exc*: Inhibitory and excitatory neurons, respectively.

Supplementary Figure 11. Scatterplots of true and deconvolution-estimated proportions in 100 VL-based *in silico* mixtures. The signature used all cell-subtypes from the original publication by Velmeshev *et al.* (2019). **A.** CIBERSORT deconvolution. **B.** DeconRNASeq. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient.

Supplementary Figure 12. Scatterplots of true and deconvolution-estimated proportions in 100 VL-based *in silico* mixtures. The signature used all cell-subtypes from the original publication by Velmeshev *et al.* (2019). **A.** dtangle deconvolution. **B.** MuSiC. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient.

Supplementary Figure 13. Heatmap of Spearman correlations between cell-subtypes in the VL dataset. Labels are taken from the original publication**.** Numbers in brackets on the right axis labels indicate the number of nuclei in that class.

Supplementary Figure 14. Effect of cell-type abundance and collinearity on deconvolution accuracy in VL-based simulations. Each point represents a cell-subtype in the Velmeshev dataset. Points are labelled by text indicating the cell-subtype classification. Colours represent a binary code for good and poor deconvolution performance. *Rho*: Spearman correlation coefficient. *X axis*: mean abundance across the 100 simulated mixtures. *Y axis*: the highest correlation a cell-subtype has to any of the other cell-subtypes in the dataset, indicating collinearity. Note that “o” and “a” are partially overlapping at x=0.5, y = 0.9, as are “k” and “I” at x=15 and y = 0.96.

Supplementary Figure 15. Scatterplots of true and deconvolution-estimated proportions in 100 CA-based *in silico* mixtures. The signature used all cell-subtypes from the original publication by Hodge *et al.* (2019). **A.** CIBERSORT deconvolution. **B.** DeconRNASeq. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient.

Supplementary Figure 16. Scatterplots of true and deconvolution-estimated proportions in 100 CA *in silico* mixtures. The signature used all cell-subtypes from the original publication by Hodge *et al.* (2019). **A.** dtangle deconvolution. **B.** MuSiC deconvolution. *Solid black line*: regression line. *Nmae*: normalised mean absolute error. *r*: Pearson correlation coefficient.

Supplementary Figure 17. Heatmap of Spearman correlations between cell-subtypes in the CA dataset. Labels are taken from the original publication. Numbers in brackets on the right axis labels indicate the number of nuclei in that class.

1356 **Supplementary Figure 18. Effect of cell-type abundance and collinearity on**
1357 **deconvolution accuracy in CA-based simulations.** Each point represents a cell-subtype in
1358 the CA dataset. Points are labelled by text indicating the cell-subtype classification. Colours
1359 represent a binary code for good and poor deconvolution performance. *Rho*: Spearman
1360 correlation coefficient. *X axis*: mean abundance across the 100 simulated mixtures. *Y axis*: the
1361 highest correlation a cell-subtype has to any of the other cell-subtypes in the dataset,
1362 indicating collinearity. Note that the following labels are partially overlapping: "l", "g", and
1363 "o" at x=3, y=0.87; and "b", "f", and "h" at x=7, y=0.95
1364
1365 **Supplementary Figure 19. Effect of removing cell-types or cell-subtypes from the**
1366 **signature matrix.** For each cell-type, its mean abundance in the mixtures is shown in
1367 backets, and scatterplots display the deconvolution accuracy when all cell types are present in
1368 the signature (x-axis) *vs.* when the cell type is absent from the signature (y-axis). Accuracy is
1369 measured as either *r* correlation coefficient (left panel) or normalised mean absolute error
1370 (right panel). Calculations of mean *r* and mean NMAE for the x-axis label do not include the
1371 absent cell-type, and thus differ across plots. *Dotted red line:* y = x.
1372
1373 **Supplementary Figure 20. Effect of varying the signature in VL-based simulated**
1374 **mixtures**. Heatmaps of Pearson correlation (*r*; **A.**) and normalised mean absolute error
1375 (nmae; **B.**) for estimated versus true proportion when varying the reference signature. The
1376 mixtures are 100 *in silico* VL simulations. Signatures only included a pan-neuronal
1377 expression profile, rather than excitatory or inhibitory sub-types. Blank squares indicate that
1378 the cell-type was not present in the signature, and thus no statistic was calculated. Grey
1379 squares indicate NA, indicating that the cell-type was present in the signature but the statistic
1380 could not be calculated; for *r*, this means there was no variance in the composition estimates,
1381 typically meaning all 100 samples' estimates were 0 or 1. For more details about signature
1382 characteristics, see methods.
1383
1384 **Supplementary Figure 21. Effect of varying the signature in CA-based simulated**
1385 **mixtures**. Heatmaps of Pearson correlation (*r*; **A.**) and normalised mean absolute error
1386 (nmae; **B.**) for estimated versus true proportion when varying the reference signature. The
1387 mixtures are 100 *in silico* CA simulations. Signatures only included a pan-neuronal
1388 expression profile, rather than excitatory or inhibitory sub-types. Blank squares indicate that
1389 the cell-type was not present in the signature, and thus no statistic was calculated. Grey
1390 squares indicate NA, indicating that the cell-type was present in the signature but the statistic
1391 could not be calculated; for *r*, this means there was no variance in the composition estimates,
1392 typically meaning all 100 samples' estimates were 0 or 1. For more details about signature
1393 characteristics, see methods.
1394
1395 **Supplementary Figure 22. Effect of varying the signature while including neuronal**
1396 **subtypes in DM-based simulated mixtures.** Heatmaps of Pearson correlation (*r*; top panel)
1397 and normalised mean absolute error (nmae; bottom panel) for estimated versus true
1398 proportion when varying the reference signature. Signatures only included a pan-neuronal
1399 expression profile, rather than excitatory or inhibitory sub-types. *Neu*: Neurons. *Ast*:
1400 Astrocytes. *Oli*: Oligodendrocytes. *Mic*: Microglia. *End*: Endothelia. For more details about
1401 signature characteristics, see methods.
1402
1403 **Supplementary Figure 23. Effect of varying the signature while including neuronal**
1404 **subtypes in VL-based simulated mixtures.** Heatmaps of Pearson correlation (*r*; **A.**) and
1405 normalised mean absolute error (nmae; **B.**) for estimated versus true proportion when varying

1406 the reference signature. The mixtures are 100 *in silico* VL simulations. All signatures here
1407 include information about the broad neuronal subtype (excitatory or inhibitory). Blank
1408 squares indicate that the cell-type was not present in the signature, and thus no statistic was
1409 calculated. Grey squares indicate NA, indicating that the cell-type was present in the
1410 signature but the statistic could not be calculated; for *r*, this means there was no variance in
1411 the composition estimates, typically meaning all 100 samples' estimates were 0 or 1. For
1412 more details about signature characteristics, see methods.
1413
1414 **Supplementary Figure 24. Effect of varying the signature on CA-based simulated**
1415 **mixtures.** Heatmaps of Pearson correlation (*r*; **A.**) and normalised mean absolute error
1416 (nmae; **B.**) for estimated versus true proportion when varying the reference signature. The
1417 mixtures are 100 *in silico* CA simulations. All signatures here include information about the
1418 broad neuronal subtype (excitatory or inhibitory). Blank squares indicate that the cell-type
1419 was not present in the signature, and thus no statistic was calculated. Grey squares indicate
1420 NA, indicating that the cell-type was present in the signature but the statistic could not be
1421 calculated; for *r*, this means there was no variance in the composition estimates, typically
1422 meaning all 100 samples' estimates were 0 or 1. For more details about signature
1423 characteristics, see methods.
1424
1425 **Supplementary Figure 25. The role of compartment-specific genes when using snRNA-**
1426 **seq signatures. A and B:** Estimated proportions for pure samples of immuno-panned cell-
1427 types, using whole-cell and snRNA-seq signatures. (A) all genes were included in the
1428 signature. (B) compartment-specific genes were filtered-out. *Thick horizontal line*: mean.
1429 *Dotted vertical line*: separates whole-cell from snRNA-seq signatures. *Neu*: neurons. *Ast*:
1430 astrocytes. *Oli*: oligodendrocytes. *Mic*: microglia. **C.** Scatterplot of estimated proportions for
1431 bulk brain samples using the RNA-seq signature, IP (x-axis) or the snRNA-seq signature
1432 derived from the same individual (y-axis). *Left*: all genes were included in the signature;
1433 *right*: compartment-specific genes were filtered-out. *Individual*: NICHD brain bank id
1434 number. Cell-type proportions were estimated using CIBERSORT.
1435
1436 **Supplementary Figure 26. Proportion of variance in gene expression explained by**
1437 **singular-value decomposition.** Linseed proposes that the saturation point of the curve (*i.e.*
1438 the number of linearly-independent components that contribute to the mixture) is its number
1439 of constituent cell-types. **A-B.** VL-derived mixtures based on random sampling (A) and those
1440 simulated with a wide range and variance in cell-type composition (B). **C-D.** CA mixtures
1441 based on random sampling (C) and and those simulated with a wide range and variance in
1442 cell-type composition (D). **E.** DM mixtures based on random sampling. **F.** RNA mixtures.
1443
1444 **Supplementary Figure 27. Scatterplots of reference-free deconvolution estimates versus**
1445 **true proportion**. **A-C.** Linseed. Plots are only shown for inferred cell-types correlated to a
1446 true cell-type at $r > 0.5$. A. VL-derived random simulations. B. CA-derived random
1447 simulations. C. DM-derived random simulations. *Dotted line*: $y = x$. **D-F.** Coex. Plots are
1448 only shown for inferred cell-types that were assigned to a true cell-type through marker
1449 enrichment analysis (Fisher Test, $p<1x10^{-5}$, odds ratio > 5; Methods) . D. VL-derived random
1450 simulations. E. CA-derived random simulations. F. DM-derived random simulations.
1451
1452 **Supplementary Figure 28.** Scatterplot of proportions estimated by Linseed in the RNA
1453 mixtures of neurons and astrocytes setting number of cell types *k*=2. *Black line*: regression
1454 line. *Red dotted line*: y=x.
1455

1456 **Supplementary Figure 29.** Scatterplots of reference-free deconvolution estimates versus
1457 true proportion in simulations with increased cell-type variance. Simulations were based on
1458 VL single-nuclei. **A.** Linseed. **B.** Coex.
1459
1460 **Supplementary Figure 30.** Scatterplots of reference-free deconvolution estimates versus
1461 true proportion in simulations with increased cell-type variance. Simulations were based on
1462 CA single-nuclei. **A.** Linseed. **B.** Coex.
1463
1464 **Supplementary Figure 31. Interplay between confounds in excitatory cell-type**
1465 **proportion and differential expression for true up- or down-regulated genes**. **A.**
1466 Scatterplots show the relationship between the confound in excitatory proportion within a
1467 simulation, and the discriminatory ability (fraction of known perturbed genes in the 200
1468 genes with the smallest p-value) (left), the true positive rate for marker genes of excitatory
1469 neurons (middle), and the true positive rate for non-marker genes (right). *Top-left*: true fold-
1470 change of 1.1. *Top-right*: true fold-change of 1.3. *Bottom-left*: true fold-change of 2.
1471 *Coloured lines*: local regression line. *Dotted line in the left* panel: 0.95 times the
1472 discriminatory ability for LM when x = 0. **B.** Model robustness to composition confounds.
1473 Barplots show the smallest composition decrease where discriminatory ability fell below 0.95
1474 of the baseline (*i.e.* that from an uncorrected linear model on a simulation with no
1475 composition confound). Labels are per A.
1476
1477 **Supplementary Figure 32. Cell-type-specific differential expression analysis using**
1478 **CIBERSORTx. A.** Composition distribution of simulated datasets. *Left*: simulated data
1479 without a composition difference between the two groups. *Right*: simulated data with a
1480 composition difference between the two groups. Each group contained 50 samples. **B.** Gene
1481 expression was perturbed 1.5-fold in Group B in inhibitory neurons for 100 non-marker
1482 genes plus 100 inhibitory neuronal marker genes. CIBERSORTx was used to extract cell-
1483 type-specific expression (Methods), with a linear model then run to assess differential
1484 expression in each cell-type. The plot displays the fraction of the true perturbed genes with an
1485 FDR < 0.05. Note that the fraction was calculated using only the subset of perturbed genes
1486 which were detected in the given cell-type. **C and D.** False positive rate across the different
1487 simulations when expression was perturbed in either excitatory neurons (C) or inhibitory
1488 neurons (D).
1489
1490 **Supplementary Figure 33. Median goodness-of-fit in large bulk brain RNA-seq datasets**
1491 **across signatures and algorithms. A.** GTEx. **B.** Parikshak *et al.*. Each panel aggregates
1492 results from samples from a given region. Rows represent signatures, and columns represent
1493 algorithms. Within each cell, the number of top is the median goodness-of-fit, while the
1494 number in parentheses below is its rank across all algorithm/signature combinations. Colours
1495 represent rank, ranging from purple (worst performance and high rank) to yellow (best
1496 performance and low rank)
1497
1498 **Supplementary Figure 34. Violin plots of goodness of fit across signatures and regions**
1499 **in GTEx data.** The top, middle, and bottom of the white internal boxes mark the 75th, 50th,
1500 and 25th percentiles, respectively. Note that the data presented in the top-left panel
1501 (CIB/Cortex) was also shown in Figure 5A.
1502
1503 **Supplementary Figure 35. Violin plots of goodness of fit across signatures and regions**
1504 **in the Parikshak dataset.** The top, middle, and bottom of the white internal boxes mark the

1505 75th, 50th, and 25th percentiles, respectively. Note that the data presented in the top-left
1506 panel (CIB/Cortex) was also shown in Figure 5B.
1507
1508 **Supplementary Figure 36. Goodness of fit across signatures in simulated data.** Each
1509 point represents one of the hundred mixtures per simulated dataset. **A.** VL-based simulation.
1510 **B.** CA-based simulation. **C.** DM-based simulation. Note that the order of signatures along the
1511 x-axis is based on median goodness-of-fit in that panel, and therefore differs between panels.
1512
1513 **Supplementary Figure 37. Violin plots of goodness-of-fit in two non-brain tissues in the**
1514 **GTEx dataset**. **A.** Pancreas samples. **B.** Heart left ventricle samples. **C.** Heart atrial
1515 appendage samples. *Fresh:* signature derived from freshly-processed human tissue.
1516 *Cultured:* signature derived from cultured cells. The bottom, middle, and top of the white
1517 boxes mark the first, second, and third quantiles, respectively.
1518
1519 **Supplementary Figure 38. Heatmaps of Spearman correlations across signatures**. **A.**
1520 Across nine brain signatures. *Top left*: Neurons. *Top middle*: Astrocytes. *Top right*: legend.
1521 *Bottom left*: Oligodendrocytes. *Bottom middle*: Microglia. *Bottom right*: Endothelia. **B.**
1522 Across four pancreas signatures. *Left*: alpha cells. R*ight*: beta cells. Legend is per the top
1523 right panel of A. **C.** Across three heart signatures. *From left to right*: cardiomyocytes, smooth
1524 muscle cells, fibroblasts, and endothelial cells. Legend is per the top right panel of A.
1525
1526 **Supplementary Figure 39. Heatmaps of intersection between the top 100 cell-type**
1527 **marker genes across signatures**. **A.** Across nine brain signatures. **B.** Across four pancreas
1528 signatures. Legend is per the bottom right panel of A. **C.** Across three heart signatures. *CM*:
1529 cardiomyocytes. *SMC*: smooth muscle cells. *FB*: fibroblasts. *EC*: endothelial cells. Legend is
1530 per the bottom right panel of A.
1531
1532 **Supplementary Figure 40. tSNE dimensionality reduction plot of snRNA-seq data**
1533 **generated as part of the present study**. Nuclei were annotated using the SingleR package to
1534 transfer labels from the NG signature. *Ast*: astrocytes. *End*: endothelia. *Exc*: excitatory
1535 neurons. *Inh*: inhibitory neurons. *Oli*: oligodendrocytes. *OPC*: oligodendrocyte precursor
1536 cells.
1537
1538 **Supplementary Figure 41. Distribution of gene expression values in real and simulated**
1539 **brain mixtures.** *Brain*: bulk brain RNA-seq from Parikshak et al. (2016). *Simulations*: *in*
1540 *silico* mixtures simulated from the corresponding dataset. *DM*: scRNA-seq from Darmanis *et*
1541 *al.*. *CA*: snRNA-seq data from the Human Cell Atlas. *VL*: snRNA-seq data from Velmeshev
1542 *et al.* (2019). Simulations contained 500 cells/mixtures for CA and VL, and 100 cells/mixture
1543 for DM. *Nuclei, Cells*: single-nuclei and single-cells from the corresponding dataset. Ten
1544 samples were randomly for each plot to minimise overplotting. Note: DM simulation is in
1545 units of RPKM rather than CPM.
1546
1547 **Supplementary Figure 42. Heatmap of Pearson correlations for cell-ty[e proportion in**
1548 **samples used for ASD analyses**. Proportions were estimated using CIBERSORT and the
1549 Multibrain signature.
1550
1551 **Supplementary Figure 43. The relationship between deconvolution estimates, RNA**
1552 **proportions (pRNA) and cell-type proportion (pCt). A.** Boxplots of RNA content per cell,
1553 reflected in the number of unique molecular identifiers (UMIs) per-cell across cell types in
1554 the VL dataset. *Ast:* astrocytes. *End:* Endothelia. *Exc:* Excitatory Neurons. Inh: Inhibitory

Neurons. *Mic*: Microglia. *Oli*: Oligodendrocytes. *OPC*: Oligodendrocyte Precursor Cells. **B.** Scatterplot of pCt vs. pRNA in pseudo-bulk samples from 10 individuals. **C.** Scatterplot of Estimated proportion (y-axis) versus true pCt (left) or pRNA (right)**. D.** Scatterplot of goodness-of-fit when reconstructing gene expression using pCt or pRNA. *Dotted black lines:* y=x

Supplementary Figure 44. Heatmap of correlations in neuronal estimates across signatures and algorithms in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. Black squares represent NA, where the cell-type estimate had a variance of 0 (typically all estimates being all 0 or all 1). *Dotted black line:* y=0.5.

Supplementary Figure 45. Heatmap of correlations in astrocyte estimates across signatures and algorithms in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. Black squares represent NA, where the cell-type estimate had a variance of 0 (typically all estimates being all 0 or all 1). *Dotted black line:* y=0.5.

Supplementary Figure 46. Heatmap of correlations in oligodendrocyte estimates across signatures and algorithms in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. Black squares represent NA, where the cell-type estimate had a variance of 0 (typically all estimates being all 0 or all 1). *Dotted black line:* y=0.5.

Supplementary Figure 47. Heatmap of correlations in microglial estimates across signatures and algorithms in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. Black squares represent NA, where the cell-type estimate had a variance of 0 (typically all estimates being all 0 or all 1). *Dotted black line:* y=0.5.

Supplementary Figure 48. Heatmap of correlations in endothelial estimates across signatures and algorithms in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. Black squares represent NA, where the cell-type estimate had a variance of 0 (typically all estimates being all 0 or all 1). *Dotted black line:* y=0.5.

Supplementary Figure 49. Distribution of neuronal deconvolutions estimates in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. *Dotted black line:* y=0.5.

Supplementary Figure 50. Distribution of astrocyte deconvolutions estimates in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. *Dotted black line:* y=0.5.

Supplementary Figure 51. Distribution of oligodendrocyte deconvolutions estimates in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. *Dotted black line:* y=0.5.

Supplementary Figure 52. Distribution of microglial deconvolutions estimates in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. *Dotted black line:* y=0.5.

Supplementary Figure 53. Distribution of endothelial deconvolutions estimates in bulk brain datasets. A. GTEx. **B.** Parikshak *et al.*. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. *Dotted black line:* y=0.5.

Tables

Table 1. Description of algorithms benchmarked in this study. *: For brevity, DeconRNASeq, CIBERSORT, and BrainInABlender will be referred to in-text as DRS, CIB, and Blender, respectively. **: the identities of unlabeled cell-types were inferred through cell-type marker enrichment (Methods)

Supplementary Tables

Supplementary Table 1. Differential expression analysis comparing nuclear and whole-cell brain tissue preparations.

Supplementary Table 2. Composition estimates in pure immunopanned brain cell-types using either all genes or stable genes. All signatures contained Neurons, Astrocytes, Oligodendrocytes, and Microglia. The estimates shown are those for the corresponding pure cell-type only. The column name shows the algorithm and signature combination used. *Above_0.8*: percentage of samples in which the composition estimate is > 0.8.

Supplementary Table 3. Composition estimates in GTEx and Parikshak *et al*. bulk brain transcriptomes, across signatures and methods

Supplementary Table 4. Differentially expression analysis results for ASD samples vs. controls for composition-dependent (CD) and composition-independent (CI) analyses. DEGs: genes significant at FDR< 0.05. GO: gene ontology terms significant at FDR< 0.05.

Supplementary Table 5. List of datasets accessed and the samples included in the present study from each dataset.

Supplementary Table 6. Cell-type specific gene expression signature data. Expression values are normalised and filtered as described in Methods. Each tab shows expression for a different signature.

Supplementary Table 7. Summary of RNA-seq data generated from mixtures of RNA from cultured cells.

Supplementary Table 8. Summary of RNA-seq and snRNA-seq data generated for nuclear versus whole-cell comparisons.

Figure 1

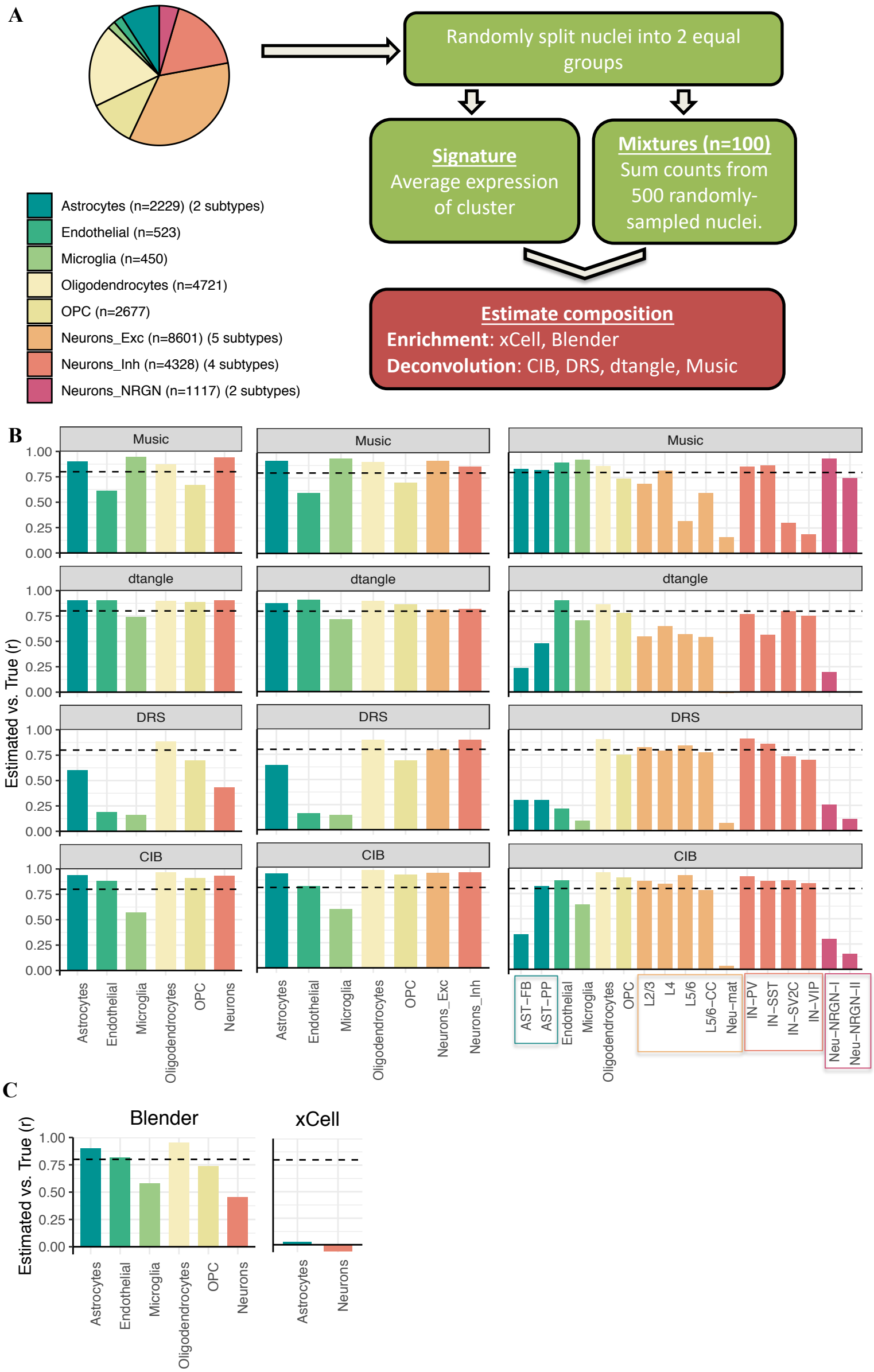

Figure 1. Deconvolution accuracy across methods. A. Simulation design. Single-nucleus RNA sequencing data was acquired from Velmeshev *et al*. and used to create *in silico* mixtures with known proportions. *Left:* Piechart displaying the composition of the dataset; n: number of cells per cell-type. For each cell type the number of sub-types is listed in between brackets. *Right:* analysis outline. *OPC*: oligodendrocyte precursor cells. *Neurons_Exc*, *Neurons*_Inh, and *Neurons*_NRGN: excitatory, inhibitory, and NRGN^{+} neurons, respectively. *DRS*: DeconRNASeq. *CIB*: *CIBERSORT*. *Blender*: BrainInABlender. **B.** Barplots of Pearson correlation coefficients (*r*) between true and estimated cell-type proportions in 100 in silico mixtures. *Left:* cells are grouped by major cell-types; *middle:* excitatory and inhibitory neuron subtypes are included in the signature; *right:* all cell-subtype labels are used in the signature. **C.** Barplots of Pearson correlation coefficients (*r*) between true proportion and cell-type enrichment scores in 100 *in silico* mixtures. *Dotted lines*: $r = 0.8$.

Figure 2

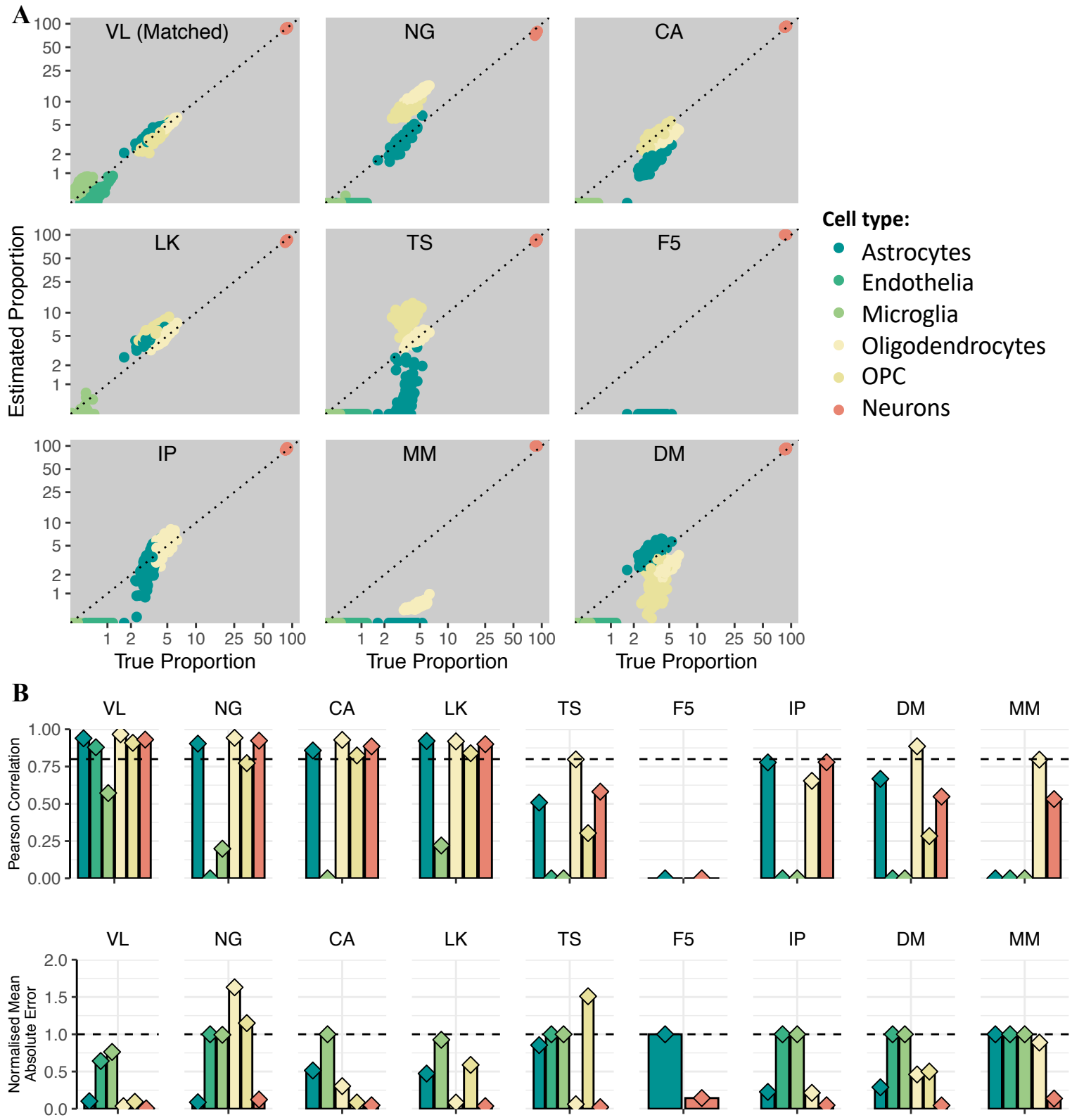

Figure 2. Effect of signature choice on deconvolution accuracy. A. Scatterplots of true and CIBERSORT-estimated proportions in VL *in silico* mixtures, for nine different signatures. *Matched:* the signature and mixture were derived from the same dataset. *VL*: Velmeshev. *NG*: Nagy. *CA*: Human Cell Atlas. *LK*: Lake. *TS*: Tasic. *Dotted line:* y=x. **B.** Barplots of normalized mean absolute error (*nmae*) for all cell-types and signatures presented in A. *Ast*: astrocytes. *End*: endothelia. *Mic*: microglia. *Oli*: oligodendrocytes. *OPC*: oligodendrocyte precursor cells. *Exc*: excitatory neurons. *Inh*: inhibitory neurons. *Dotted line:* nmae = 1. **C.** Barplots of Pearson correlation (*r*) for all cell-types and signatures presented in A. *Dotted line*: $r = 0.8$.

Figure 3

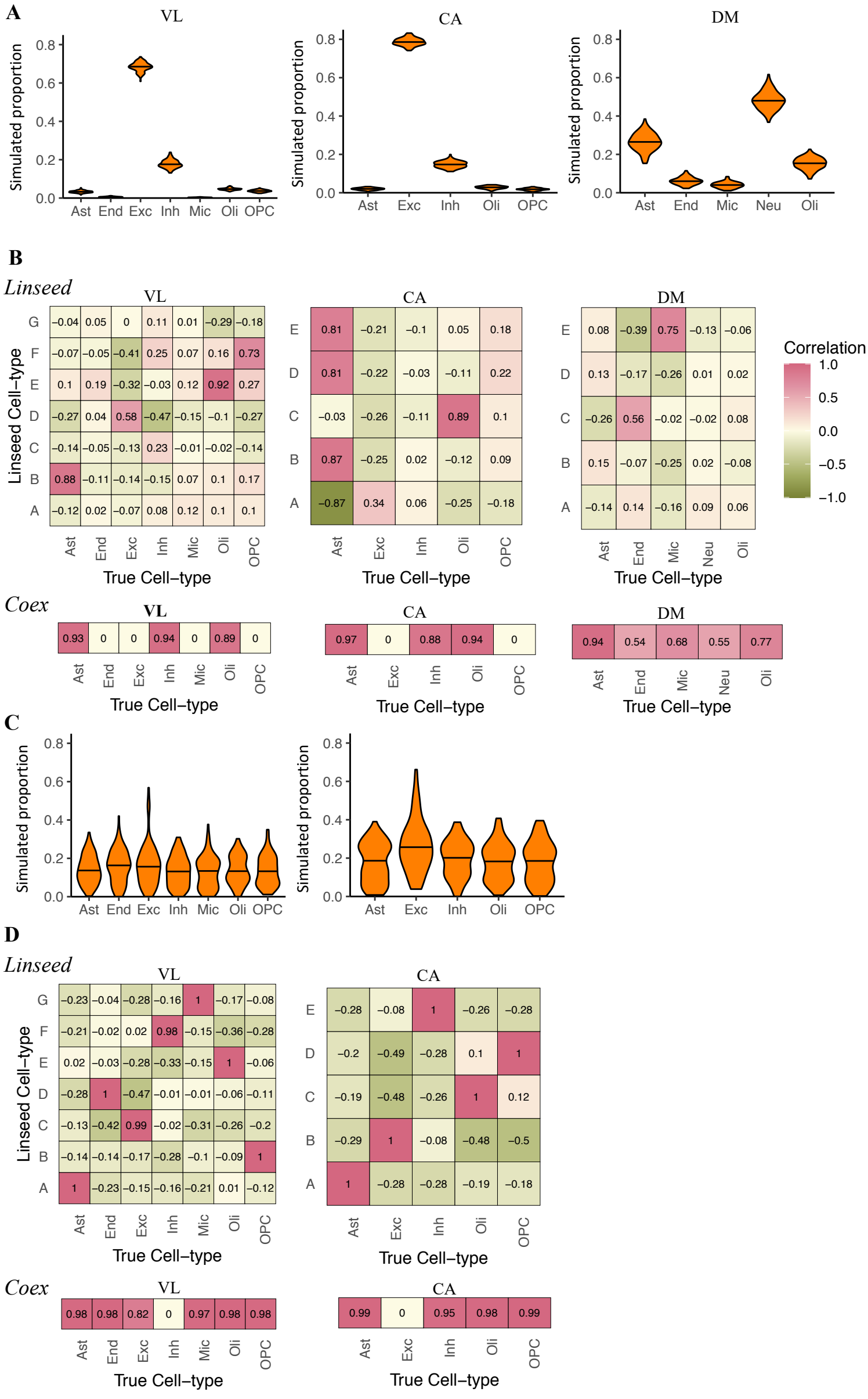

Figure 3. Reference-free deconvolution. A. Violin plots of the distribution of true cell-type proportions in VL, CA, and DM *in silico* random simulations (left, middle, and right respectively). *Black horizontal bar*: median. *Ast*: astrocytes. *End*: endothelia. *Exc*: excitatory neurons. *Inh*: inhibitory neurons. *Neu:* neurons. *Oli*: oligodendrocytes. *OPC*: oligodendrocyte precursors. **B.** Heatmaps of Pearson correlation coefficients between estimated and true cell-type proportions for random simulations based on VL, CA, and DM. *Top:* Linseed; y-axis: Cell-types defined by Linseed; x-axis: true cell-type in simulated data. *Bottom:* Coex; for each cell type the true vs. estimated correlation coefficient is displayed for the coexpression module assigned to that cell-type based on marker enrichment (Methods); zero values represent cases where no coexpression module was assigned to the corresponding cell-type. **C.** Violin plots of the distribution of cell-type abundances in simulations with wide cell-type ranges based on VL (left) and CA (right). **D.** Heatmaps of Pearson correlation coefficients between estimated and true cell-type proportions for simulations with wide cell-type ranges displayed in C, based on VL and CA. *Top:* Linseed. *Bottom:* Coex.

Figure 4

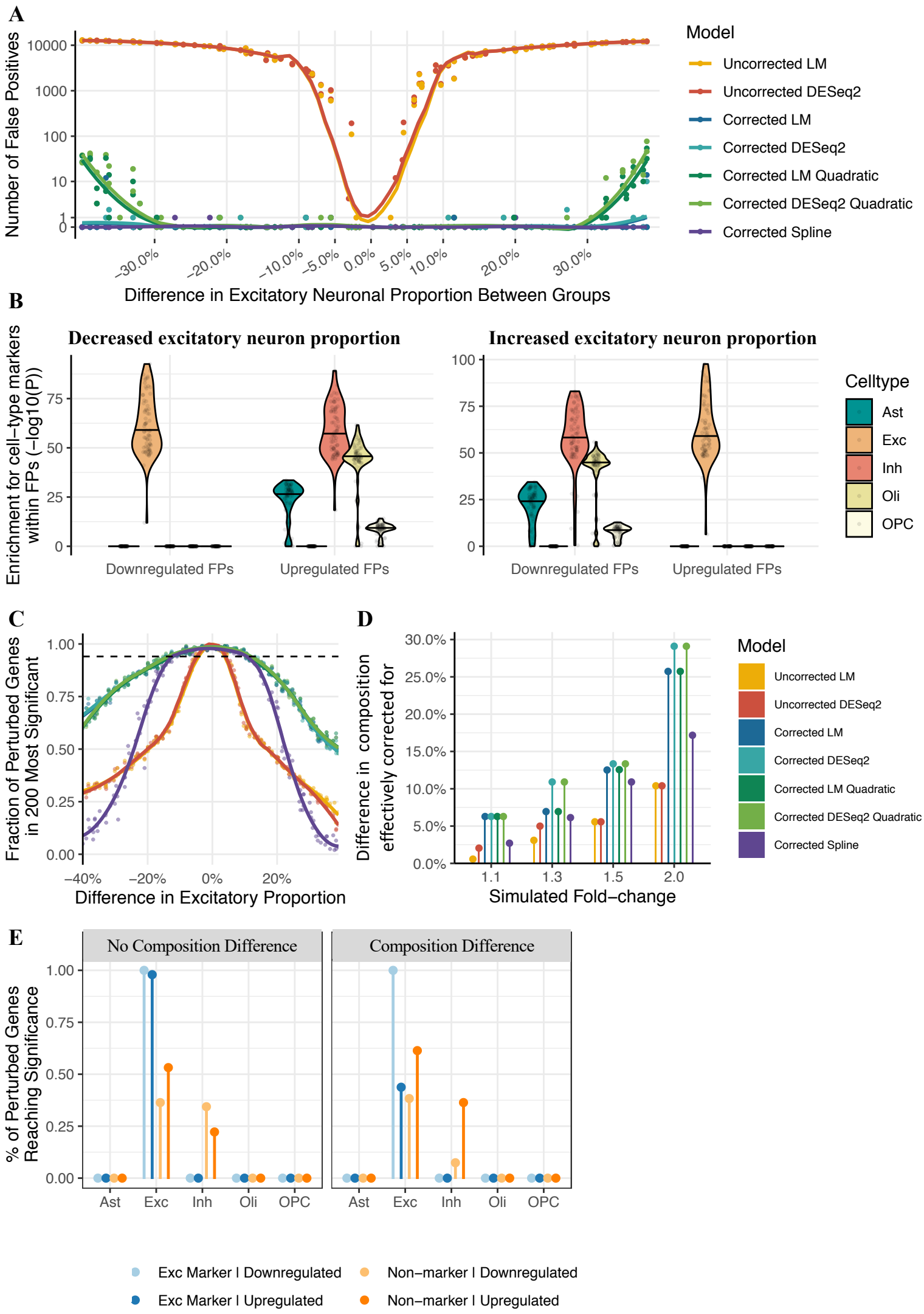

Figure 4. Effect of brain cell-type composition on differential expression (DE) analyses. A. Scatterplot of the number of false positive genes versus the simulated difference in excitatory neuron proportion between two groups of 50 samples. Each point represents a different simulated dataset. DE was assessed with either a linear model (LM) or DESeq2, with or without correction for composition, *Coloured lines:* local regression line. **B.** Cell-type marker enrichment within false positive genes. Each point represents a single simulated dataset. *Y-axis*: enrichment p-value (one-sided Fisher test); Methods. *FPs*: false positive genes. **C.** Scatterplot of the discriminatory power, *i.e.* fraction of the 200 perturbed genes in the top 200 most significantly differentially expressed genes (y-axis) versus simulated difference in excitatory neuron proportion between sample groups (x-axis) for simulated 1.5-fold expression difference. *Coloured lines*: local regression line. *Dotted line:* expected discriminatory power, *i.e.* 0.95 times the discriminatory power in the absence of cell-type composition differences between groups. **D.** Model robustness to cell-type composition differences across a range of fold-changes, quantified as the smallest composition change where discriminatory power fell below its expected value. **E.** Cell-type specific DE analysis using CIBERSORTx for simulated mixtures with 1.5-fold difference in gene expression between two groups, without (*Left*) or with (*Right*) a superimposed cell-type composition difference between the two groups. Gene expression was perturbed in excitatory neurons for 100 non-marker genes and 100 excitatory neuron marker genes. The cell type composition of simulated data is shown in Supplementary Fig. 32. Y-axis: fraction of perturbed genes significantly DE at FDR < 0.05; DE was carried out using a linear model (Methods).

Figure 5

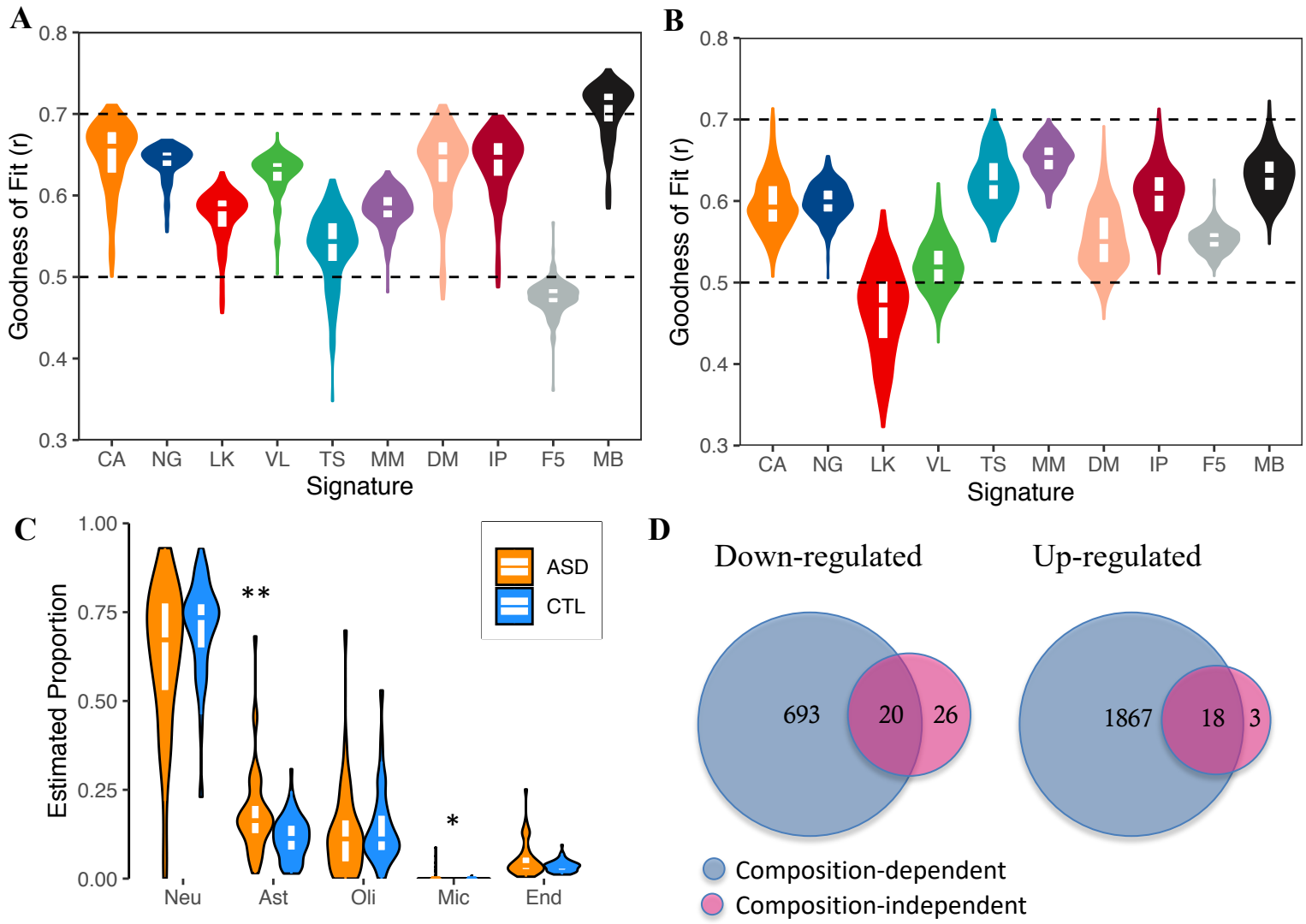

Figure 5. Cell-type composition estimates in large-scale human brain transcriptome data. A-B. Goodness of Fit across signatures in cortex samples from the GTEx consortium (A) and Parikshak *et al*. (B). Deconvolution was performed using CIBERSORT. The top, middle, and bottom of the white internal boxes mark the 75th, 50th, and 25th percentiles, respectively. **C.** Composition estimates in ASD (n=43) and Control (n=63) cortical samples. Deconvolution was performed using CIBERSORT and the MultiBrain signature. *ASD*: autism spectrum disorder. *CTL*: control. *: p<0.01. ***: p<0.001. **D.** Venn diagrams of the overlap between composition-dependent and composition-dependent DE genes between ASD and CTL samples.