# Comprehensive evaluation of human brain gene expression deconvolution methods

Gavin J Sutton[1] and Irina Voineagu[1,*]

Author affiliations:

1. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

*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, similarly 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. For example, the human brain is unique in its transcriptomic diversity, and in the complexity of its cellularity, yet 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.

We evaluate 22 transcriptome deconvolution approaches, covering all main classes: 3 partial deconvolution methods, each applied with 6 different categories of cell-type signature data, 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. In contrast, cell-type signature only mildly influences deconvolution of pancreas transcriptome data, highlighting the importance of tissue-specific benchmarking. **(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. **(d)** We demonstrate that the impact of cellular composition differences on differential expression analyses is tissue-specific, and more pronounced for brain than for pancreas.

To facilitate wider implementation of correction for cellular composition, we develop a novel brain cell-type signature, *MultiBrain*, which integrates single-cell, immuno-panned, and single-nucleus datasets. We demonstrate that it achieves improved deconvolution accuracy over existing reference signatures. Deconvolution of transcriptome data from autism cases and controls using *MultiBrain* identified cell-type composition changes replicable across studies, and highlighted novel genes dysregulated in autism.

# Keywords

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

# Introduction

Human tissues are complex 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, none of these approaches are feasible and cost effective for human 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 a disease group. Furthermore, sorting-based techniques require *a priori* knowledge of the cell-type or sub-type of interest. Single-cell sequencing approaches, although able to identify cell-types without prior selection, provide sparse data, with only a subset of genes detected per cell[14]. In addition, the transcriptome complexity measured by scRNA-seq is limited, largely lacking measurements of non-coding RNAs and splicing isoforms[15]. Therefore, our understanding of the genetic regulation of gene expression in human tissues, its variation during development and aging, and its dysregulation in disease, relies on transcriptome data derived from bulk tissue samples[16,17].

Given the considerable effect of cellular composition on gene expression data[3,18,19] and the continuing need to utilise bulk tissue samples, an accumulating number of methods have been developed to deconvolve gene expression data, *i.e.* to estimate the cellular composition of a tissue sample from its gene expression profile (reviewed in Avila Cobos *et al.*[1]).

In general, deconvolution methods model gene expression data from a tissue sample (vector $X$) as the sum of gene expression levels in the cell-types of which it's comprised ("signature" expression matrix, $S$), weighted by the proportion of each cell-type in the sample (vector $P$), formalized as $X \sim S*P$.

Deconvolution methods fall into two broad categories:

(i) ***Partial or supervised deconvolution***[6,20–28] estimates the proportion of cell-types in a sample based on experimentally measured gene expression values from pure cell-types, *i.e.* determines P knowing X and S.

It is worth noting that the signature expression data (S) often comes from a different source than the bulk tissue data (X), and thus an intrinsic assumption of most partial deconvolution methods is that gene expression in a given cell-type is the same regardless of the source of cells (thus genetic background and environmental conditions including culture conditions are ignored)[1,28].

The most frequently employed methods for partial deconvolution are Non-negative Least Squares (NLS; *i.e.* optimising $X \sim S*P$ using a least-squares approach where P should be non-negative), and Support Vector Regression (SVR). DeconRNASeq[21] and CIBERSORT[22] are commonly-used examples of these two approaches, respectively.

A simplified version of partial deconvolution consists of calculating an enrichment score, rather than a proportion, for each cell-type (*e.g.* xCell[29], or BrainInABlender[7]). While this approach is intuitive, it has several limitations: its accuracy is harder to assess (as one cannot calculate error measures or goodness-of-fit), and its biological interpretation is often unclear since the scale of enrichment scores is variable.

(ii) ***Complete or reference-free/unsupervised deconvolution*** consists of estimating both the proportion of cell-types and cell-type specific expression, *i.e.* determining both P and S knowing X[30–34]. This is an under-determined problem, which requires biologically motivated constraints.

Partial deconvolution is the most widely used approach, since unlike complete deconvolution, it is not an under-determined problem. Regardless of the method used, the estimated proportions of cell-types across samples are used for two main types of downstream analyses: as a co-variate to correct gene expression analyses for cell-type composition[35,36], and as a variable of interest to determine factors that influence cell-type composition[3,6,7,16,37–40].

Deconvolution is 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 for transcriptome data have been developed for or assessed on blood/immune and tumour samples[22,29,41], 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. This question is particularly important as transcriptome deconvolution begins to be carried out in large-scale datasets of multiple tissues (*e.g.* the Genotype-Tiessue Expression Project (GTEx[42])).

We begin to address this question focussing on the human brain. It is worth noting that the main biological factors that influence brain cellular composition (*e.g.* brain 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 specific signatures are often obtained from tissue of a different brain region[43], species[44,45], and/or a different developmental stage[7] than the bulk brain samples. Alternatively, cells cultured *in vitro* have been used[29]. 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[46]. 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[5,7,45,47–49], a comprehensive comparative assessment of the performance of deconvolution methods on brain transcriptome data is currently lacking. At the same time, accumulating evidence supports the importance of correcting brain transcriptome data for confounding effects of cellular composition. During human brain development, 92% of the differentially expressed regions correlate with the proportion of neural progenitor cells[19], suggesting that they reflect changes in brain cellularity. In a mouse model of Alzheimer disease, transcriptome changes were shown to be driven primarily by cellular composition, rather than transcriptional regulation[39]. In human brain from schizophrenia and bipolar cases, differentially expressed genes correlated with cellular composition, and few significant changes were observed after correction for cellular composition[43]. However, many studies do not account for cellular composition effects[17,50–52], while those that do use widely different methods[9,16,19,35,36,39,43,44,47,53,54]. Therefore, a comparative assessment of human brain transcriptome deconvolution accuracy is an essential stepping-stone toward a wider implementation of correction for cellular composition and a consensus on best practice.

Here, we performed an extensive unbiased evaluation of brain transcriptome deconvolution by assessing the performance of seven algorithms (three partial deconvolution, two enrichment, and two complete deconvolution methods). The partial deconvolution methods were each combined with 6 types of signature data that differed in biological properties (cultured cells, immuno-purified cells, cross-species, and single-cell and single-nucleus data), or technical factors (RNA sequencing technology). These analyses were carried out on *in silico* mixtures of single-cell transcriptomes, mixtures of RNA extracted from pure populations of neurons and glial cells, as well as large-scale brain transcriptome data from the GTEx[17] and PsychEncode[16,51] consortia. Finally, the tissue-specificity of our key observations was evaluated by comparison with transcriptome data from human pancreas[55].

These analyses showed that cell-type signature data was the most important parameter for brain transcriptome deconvolution, but of mild consequence for deconvolution of pancreas transcriptomes, highlighting the importance of tissue-specific benchmarking. 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 CIB (implementing SVR) and DRS (implementing NLS) outperform complete deconvolution methods on human brain data. In turn, the performance of complete deconvolution methods depended on the variability of cell-type composition in the bulk transcriptome data.

We also found that the impact of cellular composition differences on differential expression analyses was tissue-specific, with stronger effects observed on brain than pancreas data, indicating that correction for cellular composition is particularly important for tissues with complex cellular composition, that include transcriptionally distinct cell-types, such as the neurons and glia of the brain.

Finally, we developed a novel brain cell-type signature, *MultiBrain*, by integrating single-cell, immuno-panned, and single-nucleus data from multiple individuals, which outperformed existing signature datasets on large-scale brain transcriptome data (GTEx[17], Parikshak *et al.*[51]). Deconvolution of transcriptome data from autism cases and controls using *MultiBrain* identified cell-type composition changes replicable across datasets, and highlighted novel genes dysregulated in ASD.

# Results

To benchmark transcriptome deconvolution for brain data, we selected widely employed methods from each category of deconvolution approaches, and where possible we 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[22], DeconRNASeq[21] (DRS) which implements the non-negative least square approach employed by the PsychENCODE consortium[16], and dtangle[23]. For enrichment-based methods we selected xCell[29], which has been recently applied by the GTEx consortium[42], and BrainInABlender[7], which was specifically developed for brain. Among complete deconvolution methods, we included Linseed[30], which extends previous methods[31,34], and the co-expression-based approach developed for brain data by Kelley *et al.*[5].

### ***Assessment of cell type composition estimates for brain gene expression data across multiple methods.***

To assess deconvolution accuracy of partial deconvolution and enrichment methods, we used simulated data based on single-cell transcriptomes from Darmanis *et al.*[13]; Figure 1A. These data include a total of 297 adult human brain cells from 5 brain cell-types: neurons, astrocytes, oligodendrocytes, microglia, and endothelia. 100 mixtures were simulated as the average expression of 100 randomly-sampled cells (Methods). Cell-type signature data, denoted as the SC signature, were also generated as the average of expression within each cell-type (Methods).

We estimated cell-type proportions in the 100 simulated mixtures mixture using CIB, DRS and dtangle, and estimated cell-type enrichment scores using xCell and Blender. For the enrichment methods, scores can be calculated only for cell-types included in their built-in signatures (*i.e.* Blender: all 5 cell-types; xCell: only neurons and astrocytes). Accuracy was assessed using Pearson correlation coefficients (*r*) between true and estimated proportions (or enrichment scores), as well as normalised mean absolute errors (*nmae*). *Nmae* was calculated for the deconvolution methods only, as enrichment scores are not directly comparable with proportions.

Deconvolution accuracy was very high for CIB (*r*: 0.92-0.97), and similarly high for DRS (*r*: 0.89-0.94) and Blender (*r*: 0.79-0.92). dtangle also performed well (*r*: 0.68-0.84). xCell, however, performed rather poorly, with an accuracy of *r*=0.63 for neurons and *r*=-0.04 for astrocytes (Figure 1B-C). The negative correlation obtained for astrocytes was particularly surprising. We reasoned that the poor performance of xCell in this analysis could have resulted from the fact that it utilises signature data from a distinct source than the mixtures, while for the deconvolution methods we had used signature and mixture data derived from the same dataset. Consistent with this notion, the signature data built into Blender, which also performed well, includes the Darmanis *et al.* single-cell dataset[7]. These results prompted us to further investigate the effect of varying reference signature data on deconvolution accuracy. We also noted that for all deconvolution methods, errors were higher for the cell-types with lower abundance, *i.e.* oligodendrocytes, microglia and endothelia (Figure 1B).

We next assessed deconvolution accuracy on *in vitro* RNA mixtures. We extracted RNA from cultured neurons and astrocytes, mixed them in known proportions (0:1, 0.4:0.6, 0.45:0.55, 0.5:0.5, 1:0), and carried out RNA-seq on the mixed as well as pure RNA samples (Figure 2A). The mixing proportions were chosen so that we could assess whether differences in cell-type proportions as low as 0.05 could be accurately detected.

As observed for the simulated single-cell-based data, when the signature was derived from the same source as the mixtures, the deconvolution accuracy was high: *nmae* = 0.035, 0.043, and 0.11 for CIB, DRS, and dtangle, respectively (Figure 2B); note that due to the low number of samples in the RNA mixture data (n=5), *r* is less informative than *nmae.*

### *The biological properties of the cell-type signature data strongly influences cell type composition estimates*

To investigate the effect of cell-type signatures, we built five additional cell-type signature data: RNA-seq of cells immuno-purified from adult human brain tissue using cell-type specific antibodies[56], designated as IP; CAGE data from the FANTOM5 consortium from cultured human neurons and astrocytes[57], designated F5; RNA-seq of cells immuno-purified from the mouse brain[58], designated MM; single-nucleus droplet-based RNA-seq from human prefrontal-cortex[59], designated LK; and single-nucleus SmartSeq from the human middle temporal gyrus[60], designated CA.

Across all deconvolution algorithms, we found that accuracy was strongly affected by the reference signature, both in the RNA mixture experiment (Figure 3A; Supplementary Figure 1), and in the simulated single-cell mixtures (Figure 3B-C; Supplementary Figure 2). In both cases, very high accuracy was achieved when the source of mixture samples and reference signature data were matched. Interestingly, among the non-matched signatures, for the RNA mixture experiment the F5 signature performed best (*nmae*=0.11), while in the single-cell mixture experiment the CA signature was most accurate (average *nmae* across cell types: matched = 0.06, CA = 0.27, IP = 0.51, MM = 0.75, F5 = 1.02, LK = 1.04).

These data suggest that, for brain deconvolution, it is essential for the biological source of the signature and mixtures to be closely matched. The RNA mixtures, which contained RNA from *in vitro* cultured neurons and astrocytes, were best deconvolved using the F5 signature, which is also derived from *in vitro* cultured cells. On the other hand, the single-cell mixtures, which contained cells directly isolated from adult brain, were poorly deconvolved by F5 but best deconvolved by the CA signature, which was derived from single-nuclei isolated from adult human brain by immuno-panning. These data also show that the effect of sequencing technology (CAGE vs. RNA-seq) is less pronounced than that of the biological factors discussed above: the F5 signature performed the best in the RNA mixture experiment, despite the fact that it was generated using CAGE while the mixture data was generated by standard Illumina RNA-seq. Furthermore, the single-nucleus based CA signature data performed very well for deconvolving single-cell mixtures. However, it is worth noting that the droplet-based single-nucleus signature (LK), where the data is sparse, performed poorly.

It had previously been shown that, in blood transcriptomes, cell-type signature had a stronger impact on deconvolution accuracy than algorithm[61]. Since we observed a similar phenomenon on brain tissue data, we wanted to determine whether this was a general property of transcriptome deconvolution for any tissue. To this end, we simulated a dataset using single-cell RNA-seq from freshly-isolated pancreas alpha and beta cells[55] (Methods). Mixtures and cell-type signature data were simulated by random sampling, as with the brain data (Methods). Deconvolution was carried out using either the matched signature generated from the same single-cell data, bulk RNA-seq of freshly-isolated cells[11,12], or cultured cells[12]. Interestingly, signature only weakly influenced deconvolution accuracy of pancreas mixtures, which was instead largely driven by algorithm (Supplementary Figure 3), with CIB performing rather poorly compared to DRS and dtangle. These data highlight the importance of tissue-specific benchmarking of deconvolution methods.

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

Since we observed a strong effect of the choice of reference signature data on the brain deconvolution outcome, and recent studies have proposed reference-free approaches to cell-type composition[30–32], we assessed the performance of two such methods on brain data. Linseed, a complete deconvolution algorithm[30], proposes to identify cell-type specific genes by representing the expression vector of each gene as a point in N-dimensional space (where N is the number of samples). The points represented by all genes form a (K-1)-dimensional simplex, where K is the number of cell-types in the mixture. Cell-type specific genes are represented by points located in the corners of the simplex. Unlike similar previous methods[31], where the K parameter had to be (arbitrarily) specified, Linseed proposes to use singular-value-decomposition (SVD) to determine K from the mixture data.

An alternative approach employs co-expression networks[62] to identify modules of co-expressed genes enriched for specific cell-type markers, and then uses the module eigengene values as cell-type enrichment scores[5]. We abbreviate this method as Coex throughout the manuscript.

When applying Linseed to the two benchmarking datasets, we found that the SVD approach did not correctly identify the number of cell types in the mixture (Methods). With the correct number of cell types specified, Linseed performed extremely well on the RNA mixtures ($r$=1, *nmae*=0.01; Supplementary Figure 4), but very poorly on the single-cell mixture dataset (max $r$=0.06 for neurons, max $r$=0.05 for astrocytes, Figure 4A).

Since Linseed relies on the detection of genes represented by points with “extreme” positions in the K-1 dimensional simplex (*i.e.* corners), its ability to detect cell-type specific genes from a mixture dataset will depend on whether these mixtures contain a wide range of cell-type proportions. Thus, we hypothesized that the difference in Linseed’s performance between the two datasets likely results from the wider distribution of cell-type proportions in the RNA mixtures (neuronal proportions: 0-100%), than in single-cell mixtures generated by random sampling (neuronal proportions: 39-59%). To test this hypothesis, we generated an alternative simulated dataset from the same single cell data, which contained 100 random samples with a gradient of neuronal proportions varying from 0-50% (Methods). The performance of Linseed improved markedly on this dataset (max *r*=0.77 for neurons, max *r*=0.49 for astrocytes, Figure 4B), but remained below that of deconvolution methods that employ an appropriate reference signature dataset.

The use of co-expression networks for estimating cell-type composition (Methods) was only possible for the single-cell mixture dataset, as the number of samples in the RNA mixture dataset was too low for reliable network construction. This approach performed well for estimating the proportion of the most abundant cell-types (neurons and astrocytes: *r*=0.83 and 0.92, respectively), but poorly for the less abundant cell-types (oligodendrocytes *r*=0.34, microglia *r*=0.61, and endothelia *r*=0.37); Figure 4C. Since the co-expression network approach also relies on gene expression co-variation driven by differences in cell-type proportions, it also performed better on the simulated dataset with a wider range of cell-type proportions, than the dataset with a narrow range of cell-type proportions (Figure 4D).

These data suggest that complete deconvolution methods less effective than partial deconvolution methods, particularly since the performance of these algorithms is related to the variance in cellular composition of the dataset.

### *Assessment of the interplay between cell-type composition and differential gene expression analyses.*

Since brain eQTL data as well as differential expression analyses typically identify effects of low magnitude, small differences in cellular composition need to be accurately detected. Therefore, we were interested in investigating the following questions. Firstly, how much should cell-type composition differ between two groups of brain samples to lead to false positive results in differential expression analyses? Secondly, does the inclusion of composition estimates as covariates lead to effective correction?

We used brain single-cell data[13] to generate randomly sampled datasets, with each dataset containing two groups of 50 samples (reference group A, and test group B). The proportion of neurons in group B was either higher or lower than in group A by a value varying between 0% and 10% (Methods; Figure 5A). We then carried out standard differential expression analysis comparing group B to group A, using a linear model with and without correction for cellular composition. The correction was carried out using either the known true proportion of neurons, or the estimated proportions (Methods). False-positives were defined as genes identified as differentially expressed at $p < 0.05$ after multiple testing correction[63]. We found that at < 5% difference in proportion of neurons, there were fewer than 10 false-positives, while a difference of 5-10% difference in neuronal proportions led to hundreds to thousands of false-positives (Figure 5B). False-positive detection was eliminated by the inclusion of either true or estimated proportions as a covariate (Figure 5B). As expected, the cell-type marker enrichment of false-positive genes was concordant with the confounding difference in neuronal proportions between the two groups (Figure 5C): neuronal markers were enriched among downregulated genes when the neuronal proportion was lower in group B, but enriched among upregulated genes when the neuronal proportion was higher.

To determine whether these observations were generalisable across tissues, we simulated a similar dataset using single-cell RNA-seq data of pancreatic alpha and beta cells[55] (Supplementary Figure 5). Overall the results recapitulated the trends seen in the brain: false-positives began to be observed at ~5% difference in cellular composition; however, the number of false-positives induced by cellular composition was lower than observed for brain: ~100 false-positives detected at a ±10% difference (Supplementary Figure 5A). This is likely explained by the fact that pancreatic alpha and beta cells are highly transcriptionally similar (*rho*=0.89), while neurons and glia show less transcriptional similarity (*rho*=0.43-0.68, Supplementary Figure 5C-D).

Taken together these data demonstrate that small changes in sample composition between groups induce false positive results in differential expression analyses, and the magnitude of such confounding effects varies across tissues. Our results underscore the need to quantify and control for cellular composition effects, particularly in tissues composed of transcriptionally dissimilar cell types, such as the neurons and glia of the brain.

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

Large-scale human brain transcriptome datasets are beginning to accumulate, both for control individuals (*e.g.* GTEx consortium[17]) and individuals with psychiatric disorders (*e.g.* PsychENCODE consortium[16]). We evaluated the performance of brain gene expression deconvolution focussing on a dataset of control individuals (GTEx data, n=1,671 samples; Methods), and a dataset of autism spectrum disorder (ASD) cases and controls (Parikshak *et al.*[51], n=251 samples; Methods). The GTEx data included samples from cerebellum (CB; n=309), cerebral cortex (CTX; n=408), subcortical regions (sCTX; n=863) and spinal cord (SP; n=91); the Parikshak *et al.* dataset included samples from CB (n=84) and CTX (n=167).

We applied to both datasets all combinations of 3 partial deconvolution methods and 6 cell-type signatures, the two enrichment methods, and Coex as a complete deconvolution method (Supplementary Table 1).

Given that genetic background influences transcript levels, and each of the human brain-derived signatures (SC, IP, CA, LK) included data from a small number of individuals (n<12), we developed a new brain signature dataset by merging the SC, IP, and CA data (Methods; LK was not included due to its poor performance in the benchmarking analyses). We refer to this new cell-type signature dataset as *MultiBrain*.

Given that true cell type composition was not known, the accuracy of composition estimates was evaluated using goodness-of-fit[61]: the Pearson correlation between measured gene expression and reconstructed gene expression values (Methods). Note that this measure can only be applied to partial deconvolution methods, since enrichment scores do not sum to one. Consistent with the results on simulated data, we found that cell-type signature data had a stronger impact on accuracy than the choice of algorithm (Figure 6A, Supplementary Figure 6).

The brain-derived cell-type signatures performed well, whether it was bulk RNA-seq data (IP), single-nucleus Smart-seq data (CA), or single-cell data (SC); the median goodness-of-fit values ranged between ~0.55-0.65 using CIB on CTX samples (Figure 6A-B; Supplementary Figure 6). In contrast, as we previously observed for simulated mixtures, the cultured-cell-derived F5 and the droplet-based single-nucleus LK signatures performed worst (Figure 6A-B; Supplementary Figure 6-7).

Combining cell-type signature data from multiple sources further increased the goodness-of-fit. Indeed, the best deconvolution outcome was achieved by using *MultiBrain* and CIB. This observation was replicable across both datasets (Figure 6A-B, Supplementary Figure 6-7; median CTX goodness-of-fit: 0.70 for Parikshak, 0.65 for GTEx).

We also noted that deconvolution for CB showed lower goodness of fit than CTX in both datasets (Supplementary Figure 6-7), consistent with the fact that all cell-type signatures were derived from CTX.

When the biological and technical differences between the bulk data and cell-type signatures are eliminated, as is the case of our *in silico* mixtures of single-cell data, goodness-of-fit averaged ~0.95 (Supplementary Figure 8).

Overall, these data demonstrate that cell-type signature data is critical for accurate deconvolution of brain transcriptomes, and provide a novel signature dataset that outperforms exiting reference signatures. Our results also suggest that further development of signature data, with broader coverage of brain regions and wider genetic background is warranted.

Notably, cell-type composition estimates were highly correlated across partial deconvolution methods applied with an appropriate signature, but less so for the F5 signature, enrichment methods and Coex (Figure 7; Supplementary Figure 9). Neuronal composition estimates showed Spearman correlation > 0.9 (GTEx) and > 0.7 (Parikshak) across all pairwise comparisons (except for those generated with the F5 signature and Coex). Astrocyte estimates showed similarly high correlations, with the exception of those generated with xCell, Blender and the F5 signature, further emphasizing the importance of cell-type signature data (Figure 7). However, composition estimates for lowly-abundant cell-types such as microglia and endothelia showed lower but highly variable correlation coefficients across methods (range: -0.2-0.9; Supplementary Figure 9).

Despite the high correlation of composition estimates for abundant cell types, when considering the absolute values of estimated cell-type composition, we found that algorithm choice had a substantial impact (Supplementary Figures 10-19). For example, across signatures, CIB consistently estimated a higher proportion of neurons than either DRS or dtangle (Supplementary Figures 10 and 15). These data suggest that studies where absolute values of composition estimates are required, such as QTL for cell-type composition, need careful benchmarking of the choice of deconvolution algorithm.

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[51]. Cell-type proportion estimates (CIB/*MultiBrain*), showed significantly higher astrocyte proportions in ASD CTX samples compared to controls (difference in means: 7.7%, p=1x$10^{-4}$, Wilcoxon rank sum test; Figure 8A; Supplementary Table 2). This result recapitulates recent single-nucleus data from ASD brain validated by immunohistochemistry[64], which showed higher proportion of astrocytes in ASD CTX samples. There were also significantly higher proportions of endothelia (0.77%, p=0.009) and microglia (0.4%, p=0.006; Wilcoxon rank sum test), although the overall proportion of these cell-types 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); see Methods. Astrocyte proportions, as well as proportions of any other cell-types that were not significantly correlated with it (*i.e.* microglia and oligodendrocytes), were included as covariates.

CD analyses identified 713 down- and 1885 up-regulated genes. In contrast, when correcting for composition estimates in CI analyses, we identified only 52 down- and 47 up-regulated genes (Figure 8B). Of these, 30 down- and 41 up-regulated genes overlapped between CI and CD analyses (Figure 8B). Thus, 22 down-regulated and 6 up-regulated genes were uncovered by the CI analysis. Conversely, 683 down-regulated and 1844 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 3). The CD upregulated genes were enriched for immune and inflammatory genes (Supplementary Table 3) as well as astrocyte markers ($p=4.5x10^{-4}$), consistent with higher astrocyte proportions in ASD samples.

Novel DE genes uncovered by correction for composition were defined as those significant in the CI analysis but in neither the CD analysis nor the initial Parikshak *et al.* study (6 up-regulated, 21 down-regulated; Supplementary Table 3). Notably, the top up-regulated novel gene, *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[64]. In addition, *CXXC4* was identified as the top associated gene in a GWAS meta-analysis of schizophrenia and ASD[65]. 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 the first comprehensive benchmarking of deconvolution methods on brain transcriptome data.

Cell-type signature data was the most important parameter in brain transcriptome deconvolution, having a stronger impact than the choice of method in all cases studied: simulated single-cell mixture data, RNA mixtures of known composition, and large-scale post-mortem transcriptome data. A similar observation has been previously made on deconvolution of blood microarray data [61]. Although crucial for brain and blood transcriptome deconvolution, cell-type signature had a weak impact for deconvolution of transcriptome data from pancreas. Our results thus underscore the importance of tissue-specific benchmarking of transcriptome deconvolution.

Unlike the results on blood data, where the microarray platform was the main factor driving differences between cell-type signature datasets[61], we found that for brain transcriptomes, biological factors outweighed technical factors. Matching the biological context of the brain signature and mixture data (*e.g.* brain region; *in vitro* cultured cells *vs.* cells isolated from brain) significantly improved the deconvolution accuracy (Figure X). Our results thus suggest caution when using black-box methods with built-in signatures[7,29] and suggest that a single-cell, single-nucleus or immuno-purifying-based signature is appropriate for deconvolving bulk brain tissue data, but signatures derived from *in vitro* cultured cells would be more appropriate when estimating cell-type proportions in *in vitro* differentiation experiments.

We also demonstrate that differences as low as 5-10% in cellular composition between sample groups can lead to the detection of false-positive differentially expressed genes (Figure 5A). However, the number of false-positive genes detected (at the same confounding difference in cellular composition between groups) was higher for brain than for pancreas (Supplementary Figure 5A), indicating that correction is particularly important for tissues with high complexity of cellular composition.

To streamline deconvolution analyses on brain data, we developed a novel signature dataset, *MultiBrain*, which integrates high-quality single-cell[13], immuno-panned[56], and single-nucleus data[60] and outperformed existing individual datasets on both the GTEx and Parikshak *et al*. data. Our study also provides a framework for further development of cell-type signature datasets for human brain, by demonstrating the utility of expanding the genetic background and the representation of brain regions in cell-type signature data.

Using *MultiBrain* to deconvolve gene expression data from CTX of ASD cases and controls[51], we obtained cell-type composition estimates (Figure 8) consistent with single-nucleus data from an independent study of ASD cases and controls[64], supporting the high accuracy of deconvolution with *MultiBrain* observed in benchmarking analyses.

It is worth noting that in both brain datasets, and across all deconvolution methods, there was a wide range of estimated cell-type proportions in any given brain region (Supplementary Figures 10-19). This is consistent with data from the PsychENCODE consortium[16], which used an NLS-based approach (similar to the one implemented in DRS) and reported a similarly wide range of proportion of neurons across 1867 dorsolateral prefrontal cortex samples: 2-54%. (http://resource.psychencode.org, PEC_DER-24_Cell-Fractions-Normalised). Such a wide range is also observed in brain methylome deconvolution[66] (0-50%) and likely reflects technical variability in dissection rather than biological inter-individual variability.

We hope that our study, carrying out the first comparative assessment of human brain transcriptome deconvolution accuracy, will provide a stepping-stone toward a wider implementation of correction for cellular composition in transcriptomics, and a consensus on best practice.

# Methods

## Datasets accessed

- **Brain tissue gene expression datasets (**Supplementary Table 4)

**Bulk brain gene expression data from Parikshak *et al.***[51] were obtained from Github (https://github.com/dhglab/Genome-wide-changes-in-lncRNA-alternative-splicing-and-cortical-patterning-in-autism/releases). Exon-level count data was obtained for 251 post-mortem samples (rRNA-depleted), including frontal cortex, temporal cortex, and cerebellar vermis samples from 48 ASD and 49 control individuals, aged 2-67 (Supplementary Table 4; see Parikshak *et al*. (2016) for complete metadata).

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

Outlier samples removed in the Parikshak *et al.* study were also removed from our analyses, leaving 121 ASD (43 frontal cortex, 39 temporal cortex, 39 cerebellum) and 126 control (45 frontal cortex, 36 temporal cortex, 45 cerebellum) samples; Supplementary Table 4.

**Bulk brain gene expression data from GTEx**[17] were obtained as read counts from the 2016-01-05 release (V7) at https://gtexportal.org/home/datasets. Counts were RPKM normalised as above. A minimum expression threshold was set at > 1 RPKM in at least 88 samples (*i.e.* the number of samples in the least-represented brain region).

- **Cell-type specific gene expression datasets and generation of cell-type signatures** (Supplementary Table 5)

***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/[57]. Tag-per-million normalised CAGE peak expression levels were aggregated by sum at gene level. Data from cultured neuron (n=3) and astrocyte (n=3) samples (Supplementary Table 5) were averaged to generate the F5 neuron and astrocyte signatures. A minimum expression threshold was set at > 1 tag-per-million in at least one cell-type.

***IP** (immuno-purified)*: RNA-seq data from cells immunopurified from human adult brain tissue extracted during surgery were obtained from Zhang *et al.* 2016[56]. FPKM-level data were accessed from Table S4 of Zhang *et al*. for neurons (n=1), astrocytes (n=12), oligodendrocytes (n=5), microglia (n=3), endothelia (n=2) (Supplementary Table 5). Cell-types derived from foetal brain were excluded (*i.e.*, foetal astrocytes). Samples of the same cell-type were averaged to generate the IP signature. A minimum expression threshold was set at > 1 FPKM in at least one of the five cell-types in the final signature matrix.

***MM** (Mus musculus)*: RNA-seq data from immunopurified mouse brain tissue was obtained from Zhang *et al.* 2014[58]. FPKM-level data were accessed from https://web.stanford.edu/group/barres_lab/brain_rnaseq.html, in which biological replicates of cell-type transcriptomes (neurons , astrocytes, oligodendrocytes, microglia, and endothelia were already aggregated across samples. Mouse genes were mapped to human orthologues using Gene ID homology information from http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt.

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

***SC** (Single-cell):* Brain single-cell gene expression data generated by Darmanis *et al.* (2015)[13] were downloaded as count-level data from https://github.com/VCCRI/CIDR-examples/tree/master/Brain[67]. Data were RPKM normalised as above. To generate the SC signature, expression was averaged across samples of each cell-type (*i.e.* astrocyte (n = 62), neuron (161), microglia (16), mature oligodendrocyte (38), or endothelia (20); Supplementary Table 5). Cell-types derived from foetal brain (quiescent neurons and replicating neurons) were excluded. Oligodendrocyte precursor cells were also excluded, for consistency with the IP and MM signatures. A minimum expression threshold was set at > 1 RPKM in at least one of the five cell-types in the final signature matrix.

***LK** (Lake)*: Gene expression data for 10,319 frontal cortex nuclei were accessed from Lake *et al*. 2018[59]. Nuclei with < 1000 unique molecular identifiers (UMIs, *i.e.* unique transcript counts) were excluded. Expression values were normalised to UMIs-per-million. To generate the LK signature, expression was average across nuclei of each cell-type: astrocytes (106), neurons (3795), oligodendrocytes (107), and microglia (24). Endothelia were excluded, as even when all (7) nuclei were pooled less than half of genes were expressed above zero. For consistency with other signatures, oligodendrocyte precursor cells and pericytes were excluded.

***CA** (Cell Atlas):* Exon-count-level expression data for 14,328 nuclei from the middle temporal gyrus were acquired from the Human Cell Atlas[60]. Data were RPKM-normalised as above. The CA signature was generated as the average expression within nuclei of each cell-type, including 291 astrocytes, 14,689 neurons, 313 oligodendrocytes, and 9 endothelia. Notably, all subtypes of neuron were included in the neuronal signature, for consistency with other signatures' neuronal definition. A minimum expression threshold of > 1 RPKM in at least one of the five cell-types was set. Oligodendrocyte precursor and unlabelled cells were also excluded.

***MultiBrain***: The *MultiBrain* signature of expression in neurons, astrocytes, oligodendrocytes, microglia, and endothelia was generated as their respective average RPKM expression values across the SC, IP, and CA signatures. All three signatures were quantile normalised[68] together prior to averaging. Only genes expressed in all three signatures were included.

For all cell-type signatures, only protein-coding genes were included.

## RNA-seq data generated in the present study

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

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

Neuronal differentiation of human neural progenitors stably transfected with pLRC-GFP was carried out for 2 weeks as previously described[69]. RNA extraction was carried out using a Qiagen miRNeasy kit, with on-column DNase digestion. RNA from differentiated neurons was kindly provided by Dr. Brent Fogel (UCLA)[69].

RNA mixtures were generated by mixing neuronal and astrocyte RNA in mass ratios of 40:60, 45:55, 50:50 neuron:astrocyte (n=1 for each ratio). In addition, a pure neuronal RNA sample and pure astrocyte RNA samples (n=3) were also included (Supplementary Table 6).

Library preparation using the Illumina TruSeq Stranded kit (http://www.illumina.com/products/truseq_stranded_total_rna_library_prep_kit.html) and sequencing on a NextSeq 500 Illumina sequencer were carried out at the UNSW Ramaciotti Centre for Genomics, generating 75 bp paired-end reads (Supplementary Table 6). Sequencing reads were mapped to the human genome (hg19) using STAR[70] with the following parameters: --outSJfilterOverhangMin 5 5 5 5 --alignSJoverhangMin 5 --alignSJDBoverhangMin 5 --outFilterMultimapNmax 1 --outFilterScoreMin 1 --outFilterMatchNmin 1 --outFilterMismatchNmax 2 --chimSegmentMin 5 --chimScoreMin 15 --chimScoreSeparation 10 --chimJunctionOverhangMin 5.

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

***IH** (in house)* cell-type signature data includes the RPKM-normalised data for the neurons sample, and averaged RPKM-normalised data across the 3 astrocyte samples. Data was thresholded for a minimum of 1 rpkm in at least one of the two cell types.

***RNA mixture data*** consists of RPKM-normalised data from the three RNA mixture samples (40:60, 45:55, 50:50 neuron:astrocyte ratios). Genes expressed at < 2 RPKM in at least one sample were filtered out.

## Generation of *in silico* mixture data

**Randomly sampled single-cell mixtures** were generated by sampling 100 single-cells from the Darmanis *et al.* dataset[13], and averaging the RPKM-normalised expression data of the sampled cells. All sampling was performed without replacement. Only cells classified as one of neurons, astrocytes, oligodendrocytes, microglia, or endothelia were included. A final dataset of 100 samples was generated.

**Gradient single-cell mixtures** were generated as the average expression of *n* randomly-sampled neurons (where *n* was a randomly-selected integer from 1-50), and 50 non-neuronal cells. All sampling was performed without replacement. A final dataset of 100 samples was generated.

**Simulated single-cell mixtures for differential expression analyses** were generated as datasets of 100 samples, split into groups A and B of 50 samples each. Each sample in group A (the reference group) was generated as the average expression of randomly selected *n* neuronal and 100-*n* non-neuronal cells, where *n* was a randomly selected integer between 40-60; this range was selected for consistence with known neuron:glia ratios in the brain[4]. Samples in group B (test group) were generated as for group A, except the range of neuronal proportion from which *n* was sampled was increased or decreased by 0-10 with a step of 1. All sampling was performed without replacement. Overall, we simulated 21 levels of difference in neuronal proportion between the two groups (-10 to 10 with a step of 1), and for each level we simulated 20 datasets, for a total of 420 simulated datasets. Differential expression analyses for group B vs. group A were carried out on each dataset as described in the "Differential Expression" section below.

## Estimation of cellular composition

***Note:*** Since cell-types differ in their total RNA content, transcriptome deconvolution estimates proportions of RNA from each cell type, rather than proportions of cells *per se*[30]. However, it is the proportion of RNA, not the proportion of cells, that is informative for gene expression analyses: gene expression measured in a tissue sample reflects gene expression levels measured in individual cell-types, weighted by the proportion of RNA (not the proportion of cells) from each cell-type. See Supplementary Figure 20 for an illustrative example. Thus, for simplicity, throughout the manuscript we use the term "cellular composition" to refer to a sample's composition of RNA from different cell-types.

Cell-type composition was estimated using three partial deconvolution methods (**DeconRNASeq**[21], **dtangle**[23], and **CIBERSORT**[22]), two enrichment methods with in-built signatures (**BrainInABlender**[7] and **xCell**[29]), and two complete deconvolution methods: **Linseed**[30], and a co-expression based approach proposed by Kelley *et al.*[5] (referred to as **Coex**).

All algorithms were run in R v3.6. All data used for deconvolution were RPKM-normalised expression values without log2 transformation[71] unless noted below.

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

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

**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[23].

**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[57], and ENCODE[72]. xCell generates a "Raw" and a "Transformed" enrichment score. We used the Transformed score as a measure of enrichment.

**Coex** was carried out by constructing co-expression networks using the *blockwiseModules* function from the WGCNA R package[62,73], 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 for a given cell-type. 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 SC signature data. Cell-type enrichment scores were defined as the CTM's eigengene values (*i.e.*, first principal component values of genes included in the CTM) according to Kelley *et al.*[5].

**Linseed** v0.99.2 was run using the R package from https://github.com/ctlab/LinSeed. We first used the SVD approach to determine the number of cell types in the mixture data. For the RNA mixtures which consisted of 2 cell types, the estimated k was 3 (Supplementary Figure 21). For the single-cell mixtures, which consisted of 5 cell types, the estimated k was more than 10 (Supplementary Figure 21). Therefore we used the known k value for RNA (2) and single-cell (5) mixtures, and k=5 for bulk brain data from Parikshak *et al.* and GTEx.

## Pancreas analyses

### Acquisition and pre-processing of pancreas alpha and beta signature data (Supplementary Figure 5)

In total, four signatures of pancreas alpha and beta cells were generated. Genes were excluded if they were not protein-coding, or if they were expressed at < 1 transcript-per-million (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.*[55]. 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 signature of alpha and beta expression was generated as the average of 998 alpha and 348 beta cells, respectively.

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

***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 comprised of isolated alpha and beta cells subjected to 1-week of culturing plus transduction with a GFP expression vector (average of 4 and 6 replicates, respectively).

### *In silico* mixture data

All *in silico* simulated mixtures for mixtures of pancreas alpha and beta cells were generated as per above, but using alpha and beta single-cell transcriptomes from Enge *et al.*[55]

## Goodness-of-fit

Goodness-of-fit was evaluated as the concordance between each sample's predicted and observed expression. For each sample, expression values were predicted using the following formula:

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

Here, $j$ denotes a cell-type, $s_j$ is the vector of gene expression in cell-type $j$ (found in the signature matrix), and $p_j$ is the estimated proportion of cell-type $j$ in the sample. In all applications, $n = 5$ as the proportion of five cell-types was estimated.

The concordance between the predicted and observed expression vectors was evaluated using a Pearson correlation coefficient on log2-transformed values offset by +0.5.

## Differential expression analyses

### Differential expression analyses of simulated data

Differential expression between group A and group B in simulated single cell mixtures was assessed using a linear model in R, on log2-transformed RPKM values offset by +0.5. Three models were used: a model with no covariates; a model adjusting for true neuronal proportion; and a model adjusting for estimated neuronal proportion. Neuronal proportion estimates were generated using the SC signature and DeconRNASeq. (Note that here we deconvolved 420 datasets with 100 simulated samples each, and CIB requires around 30 min per 100 samples, while DeconRNAseq is substantially faster). Multiple testing correction was conducted using the Benjamini-Hochberg approach[63].

### Differential expression analyses of ASD and control samples

Differential expression (DE) was carried out on count-level expression data. The same samples used by Parikshak *et al.*[51] for DE were included  in our analyses: 106 samples (43 ASD, 63 controls; Supplementary Table 4). Differential expression was carried out using a Wald test with Benjamini-Hochberg correction for multiple testing as implemented in DESeq2[74]. Composition-dependent DE adjusted for the following covariates: Age, Sex, Sequencing Batch, Brain Bank, Region, RIN, and the first two principal components of sequencing metadata, per Parikshak *et al.*[51]. Composition-independent DE used the same model, but with estimated astrocyte proportions from CIB/*MultiBrain* added, plus any cell-type proportions with which this didn't significantly correlate (*i.e.,* oligodendrocyte and microglial proportions; Supplementary Figure 22) to minimise co-linearity.

When composition estimates to be used in ASD *vs.* control analyses were generated, genes previously identified as DE in ASD by at least two studies[50–52] were removed from the signature (Supplementary Information Code).

## Other analyses

Gene ontology (GO) and pathway enrichment analyses were conducted using gProfiler2[75] in R, setting exclude_iea=TRUE and all other parameters as default. P-values were BH-corrected[63], Only results from GO, KEGG, Reactome, Human Phenotype, and Wikipathways were reported, with filtering performed after multiple-testing correction.

Cell-type marker enrichment analyses were performed by one-sided Fisher's Exact Test against 100 markers per cell-type from the SC signature. Markers were defined using the find_markers function from dtangle[23], setting marker_method = "diff".

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

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

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

## References

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

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

3. Glastonbury, C. A., Couto Alves, A., El-Sayed Moustafa, J. S. & Small, K. S. Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. *Am. J. Hum. Genet.* (2019).

doi:10.1016/j.ajhg.2019.03.025

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

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

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

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

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

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

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

11. Blodgett, D. M. *et al.* Novel Observations From Next-Generation RNA Sequencing of Highly Purified Human Adult and Fetal Islet Cell Subsets. *Diabetes* **64**, 3172–81 (2015).

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

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

14. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).

15. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

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

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

18. Murillo, O. D. *et al.* exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids. *Cell* **177**, 463-477.e15 (2019).

19. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* **18**, 154–161 (2015).

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

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

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

23. 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

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

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

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

27. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).

28. 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).

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

30. 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).

31. 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).

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

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

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

35. Lin, P. *et al.* Transcriptome analysis of human brain tissue identifies reduced expression of complement complex C1Q Genes in Rett syndrome. *BMC Genomics* **17**, 427 (2016).

36. Dillman, A. A. *et al.* Transcriptomic profiling of the human brain reveals that altered synaptic gene expression is associated with chronological aging. *Sci. Rep.* **7**, 16890 (2017).

37. Sarkisyan, D. *et al.* Damaged reward areas in human alcoholics: neuronal proportion decline and astrocyte activation. *Acta Neuropathol.* **133**, 485–487 (2017).

38. Yu, Q. & He, Z. Comprehensive investigation of temporal and autism-associated cell type composition-dependent and independent gene expression changes in human brains. *Sci. Rep.* **7**, 4121 (2017).

39. Srinivasan, K. *et al.* Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat. Commun.* **7**, 11295 (2016).

40. Kong, Y., Rastogi, D., Seoighe, C., Greally, J. M. & Suzuki, M. Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data. *PLoS One* **14**, e0215987 (2019).

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

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

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

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

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

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

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

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

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

50. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).

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

52. Liu, X. *et al.* Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism. *PLoS Biol.* **14**, (2016).

53. Collado-Torres, L. *et al.* Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron* **0**, (2019).

54. Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L. M. & Luthi-Carter, R. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods* **8**, 945 (2011).

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

56. 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).

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

58. 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).

59. 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).

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

61. 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).

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

63. 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).

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

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

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

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

68. Bolstad, B. M., Irizarry, R. ., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

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

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

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

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

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

74. 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).

75. 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: A.** Simulation design. DRS: DeconRNASeq. CIB: CIBERSORT. Blender: BrainInABlender. **B.** 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. **C.** Scatterplots of estimated proportion (or enrichment score) and true proportion for each cell-type. nmae: normalised mean absolute error. *nmae* and *r* values are those displayed in the barplots in B. *Red dotted line*: x=y. Grey line: regression line.

**Figure 2. A.** Outline of RNA mixtures and the in-house (IH) signature. **B.** Scatterplots of estimated and true proportions of neurons using CIB, DRS and dtangle, combined with the matching IH signature. **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).

**Figure 3. A.** Scatterplots of estimated and true proportions of neurons in the RNA mixture samples using CIB and varying the reference signature. *Matched*: refers to the IH signature, as both the mixture and signature derive from the same RNA extractions. *SC*: single-cell. *IP*: immuno-panned. *MM*: mouse brain. *F5*: FANTOM5. *LK*: Lake. *CA*: Cell Atlas. Note that deconvolution using CIB/LK was unable to run (Methods). **B.** Scatterplots of estimated and true proportions of all 5 cell-types in the single-cell mixture data. Deconvolution was carried out using CIB and varying the reference signature. *Matched*: refers to the SC signature, as both the mixture and signature derive from the same single-cell data. **C.** Barplots of *nmae* (top) and *r* (bottom) for data presented in B. *Red dotted line*: y=x. *Grey line*: regression line.

**Figure 4. A-B.** Signature-free deconvolution using Linseed. *Left*: correlation matrix between all Linseed identified cell-types and true cell-type proportions. *Right*: scatterplot of true neuronal proportion and the Linseed cell-type with the highest correlation with known neuronal proportions. Linseed1 – Linseed5: cell-types identified by Linseed. *A.* Mixtures simulated by random sampling. *B.* Mixtures simulated with a wide neuronal proportion range between 0 and 50%. *Neu*: neurons. *Ast*: Astrocytes. *Oli*: Oligodendrocytes. *Mic*: Microglia. *End*: Endothelia. **C-D.** Scatterplots of marker-enriched co-expression module eigengene and true proportions of the corresponding cell type. C. Mixtures simulated by random sampling. D. Mixtures simulated with a wide neuronal proportion range between 0 and 50%.

**Figure 5. A.** Outline of data simulation and analysis. **B.** Scatterplot of the number of false-positives (y-axis) *vs.* the simulated difference in neuronal proportion between two groups of samples (x-axis). Each data point represents a simulated dataset with two groups of samples, n=50 samples per group. **C.** Cell-type marker enrichment of false-positive genes. *Left*: Enrichment of false positives obtained when simulated neuronal proportion was lower in the test than reference group. *Right*: Enrichment of false positives obtained when neuronal proportion was higher in the test than reference group. Odds ratios were only calculated in simulations with > 10 false-positives.

**Figure 6.** Cell-type composition estimates in the bulk brain transcriptome resources. **A.** Heatmaps of the median correlation between measured expression and that reconstructed from deconvolution (goodness-of-fit (*r*)). Samples from all regions were included. In each cell, the number in parentheses is its rank, with lower ranks denoting better performance. Colouration is based upon this rank. **B-C.** Violin plots of goodness-of-fit using CIB and varying the signature. *B*. Data from GTEx (n=309 in cerebellum (CB); n=408 in cerebral cortex (CTX); n=863 in subcortical regions (sCTX); and n=91 for spinal cord (SP)). *C*. Data from Parikshak

*et al.* (n=84 in CB; n=167 in CTX). *Dotted horizontal line*: *r*=0.5. The bottom, middle, and top of the internal white boxes mark the first, second, and third quantiles, respectively.

**Figure 7. A-B.** Heatmaps of Spearman correlation between cell-type estimates (proportions or enrichment scores) across methods. *Left:* Pairwise correlations in neuronal composition. *Right:* Pairwise correlations in astrocytic composition. *A*. GTEx data. *B*. Parikshak *et al.* data.

**Figure 8. A.** Violin plots of cell-type composition in ASD and control (CTL) samples from Parikshak *et al.*. Composition was estimated using CIB/*MultiBrain* signature. *: p<0.05. **: p<0.01. **B.** Venn diagrams displaying the overlap of differentially-expressed genes (DEGs) in composition-dependent (CD) and composition-independent (CI) analyses (FDR < 0.05). *Left*: down-regulated. *Right*: up-regulated.

## Supplementary Figure Legends

**Supplementary Figure 1.** Scatterplots of estimated and true proportions of neurons in the RNA mixture data, obtained with either DRS (A) or dtangle (B). Titles refer to the signature used for deconvolution. *Matched* refers to the IH signature, as both the mixture and signature derive from the same RNA extractions. *Dotted red line*: y=x. *Black line*: regression line.

**Supplementary Figure 2.** Scatterplots of estimated and true proportions of all 5 cell types in the single-cell mixture data. Deconvolution was carried out using either DRS (A) or dtangle (B) varying the reference signature. Titles refer to the signature used for deconvolution. *Matched* refers to the SC signature, as both the mixture and signature derive from the same single-cell data. The LK signature lacked endothelia, while the F5 signature lacked oligodendrocytes, microglia, and endothelia. *Dotted red line*: y=x.

**Supplementary Figure 3**. Scatterplots of estimated and true proportions of alpha cells in *in silico* single-cell mixtures of pancreas alpha and beta cells. *Dotted red line*: y=x. *Grey line*: regression line. Titles denote the algorithm and signature combination used in deconvolution. *Matched*: the EN signature derived from the same single-cells used to generate the mixtures from Enge *et al.* 2017. *BA*: a bulk RNA-seq signature from sorted fresh pancreas tissue from Blodgett *et al.* 2015. *FS*: a bulk RNA-seq signature from sorted fresh pancreas tissue from Furuyama *et al.* 2019. *FG*: a bulk RNA-seq signature from sorted pancreas tissue with 1-week of culture and transduction with a GFP-expression vector from Furuyama *et al.* 2019.

**Supplementary Figure 4.** Scatterplot of proportions estimated by Linseed in the RNA mixtures of neurons and astrocytes setting number of cell types *k*=2. *Black line*: regression line. *Red dotted line*: y=x.

**Supplementary Figure 5.** Effect of composition differences between groups on differential expression in *in silico* mixtures of pancreas cell-types. **A.** Scatterplot of the number of false-positives (y-axis) versus the simulated difference in alpha cell proportion between two groups of samples (x-axis). Each data point represents a simulated dataset with two groups of samples, n=50 samples per group. **B.** Cell-type marker enrichment in false-positive genes. *Left*: Enrichment of false-positives obtained when simulated alpha proportion was lower in the test *vs.* reference group. *Right*: Enrichment of false-positives obtained when alpha proportion was higher in the test *vs.* reference group. **C.** Scatterplot of log gene expression in alpha and beta cells. **D.** Heatmap of spearman correlations between brain cell-types in the SC signature.

**Supplementary Figure 6.** Heatmaps of the median correlation between measured expression and that reconstructed from deconvolution **(**goodness-of-fit (*r*)). In each cell, the number in parentheses is its rank, with lower ranks denoting better performance. Colouration is based upon this rank. *A*. Samples from GTEx. *B*. Samples from Parikshak *et al*. *CTX*: cortical. *CB*: cerebellar. *sCTX*: sub-cortical. *SP*: spinal cord.

**Supplementary Figure 7.** Violin plots of the goodness-of-fit (r), using cell-type proportion estimates generated by different algorithm and signature combinations. Within each plot, samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively. **A.** GTEx dataset. **B.** Parikshak dataset.

**Supplementary Figure 8.** Violin plots of goodness-of-fit in *in silico* mixtures of brain cells. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively. Mixtures are the same as in Figure 1.

**Supplementary Figure 9.** Heatmaps of Spearman correlation between cell-type estimates (proportions or enrichment scores) across methods. *Left column*: GTEx. *Right column*: Parikshak *et al.*. *DTA*: dtangle. Correlations were calculated between cortical samples only.

**Supplementary Figure 10.** Violin plots of neuronal composition estimates for the GTEx dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. Dotted horizontal line: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 11**. Violin plots of astrocytic composition estimates for the GTEx dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. Dotted horizontal line: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 12.** Violin plots of oligodendrocytic composition estimates for the GTEx dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. Dotted horizontal line: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 13.** Violin plots of microglia composition estimates for the GTEx dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. Dotted horizontal line: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 14.** Violin plots of endothelial composition estimates for the GTEx dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex; *sCTX*-subcortical regions; *SP*-spinal cord. Dotted horizontal line: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 15.** Violin plots of neuronal composition estimates for the Parikshak dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 16** Violin plots of astrocytic composition estimates for the Parikshak dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 17.** Violin plots of oligodendrocytic composition estimates for the Parikshak dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 18.** Violin plots of microglial composition estimates for the Parikshak dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 19.** Violin plots of endothelial composition estimates for the Parikshak dataset as a function of algorithm and, where applicable, signature. Samples are grouped by brain region: *CB*-cerebellum; *CTX*-cortex. *Dotted horizontal line*: y=0.5. The bottom, middle, and top of the white boxes mark the first, second, and third quantiles, respectively.

**Supplementary Figure 20.** Hypothetical example showing that reconstructing gene expression data is accurate when using RNA proportions but not cell-type proportions. *Top panel, left:* Assume 2 cell-types (X and Y) express genes A-D at known numbers of copies per cell, but differ in the total amount of RNA per cell, such that Y cells contain double the RNA of X cells. Numbers are for illustrative purposes only. *Top panel, right*: X and Y's measured expression in an RNA-seq signature captures the relative expression of transcripts, as would be the case after normalising for sequencing depth. *Bottom panel:* Assume RNA is extracted and sequenced from a pool of 10X and 10Y cells. Reconstructed gene expression data matches measured expression when using RNA proportions but not cell-type proportions.

**Supplementary Figure 21. A.** Cumulative variance explained (y-axis) vs. singular value decomposition (SVD) dimension in the Linseed analysis of single-cell mixture data. *Top*: random mixtures of 5 cell-types. *Bottom*: gradient mixtures of 5 cell-types. **B.** Cumulative variance explained (y-axis) vs. SVD dimension in the Linseed analysis of RNA mixtures of neurons and astrocytes. Linseed proposes to estimate the number of cell types in the mixture as the SVD dimension for which the cumulative variance plateaus. Note that this value is > 10 in A and 3 in B.

**Supplementary Figure 22.** Heatmaps of correlations in cell-type composition estimates in samples used in ASD *vs.* control analyses. Composition was estimated using CIB/*MultiBrain.*

## 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.** Estimated cellular composition in brain transcriptomes from the GTEx and Parikshak *et al.* datasets.

**Supplementary Table 2.** Influence of deconvolution approach on ASD-related changes in cell-type composition from Parikshak *et al.*

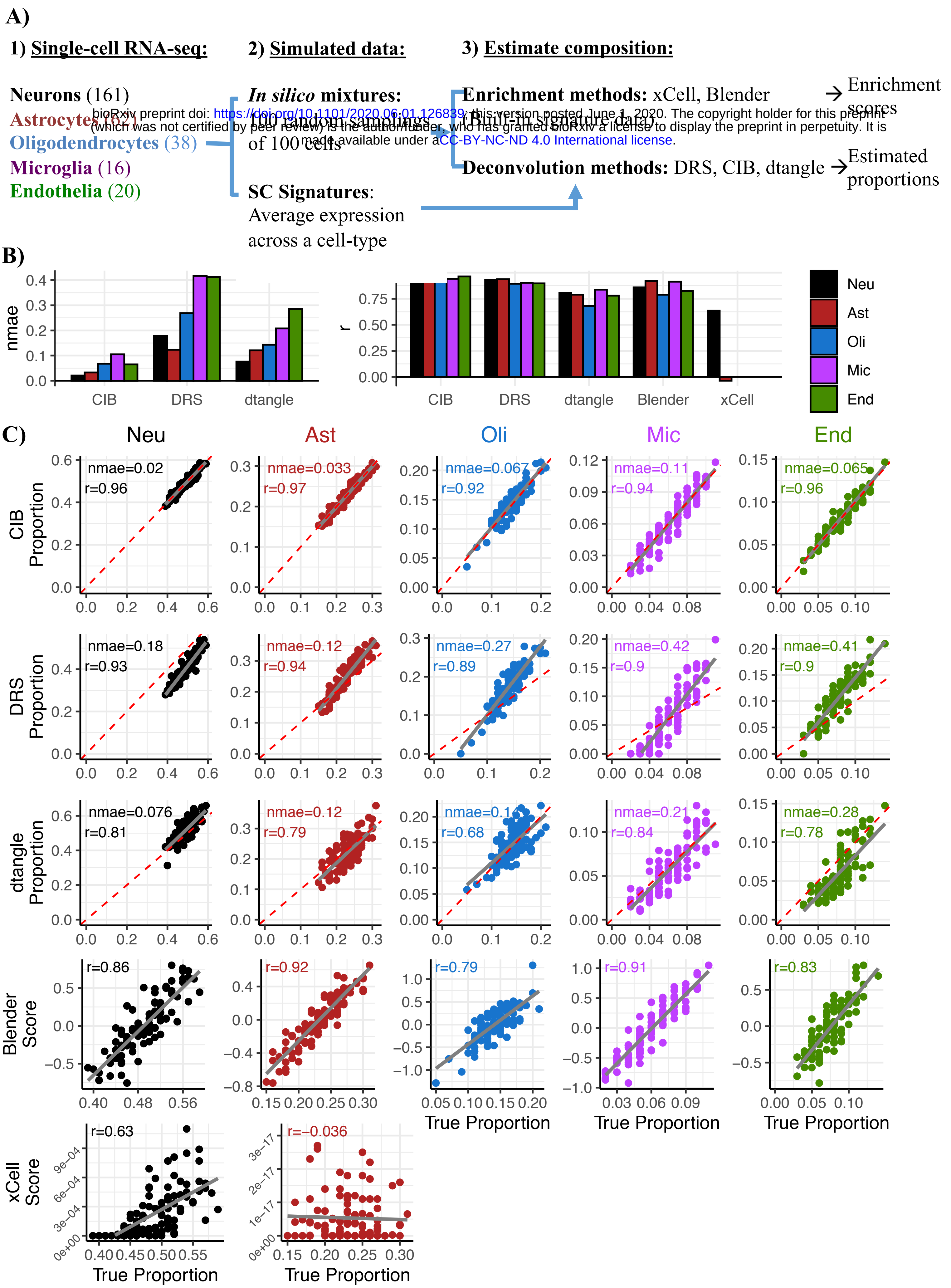
**Supplementary Table 3.** Differentially expression analysis results for ASD samples vs. controls for composition-dependent (CD) and composition-independent (CI) analyses. DEGs: genes significant at FDR< 0.05. GO: gene ontology terms significant at FDR< 0.05.

**Supplementary Table 4.** List of datasets accessed and the samples included in the present study from each dataset.

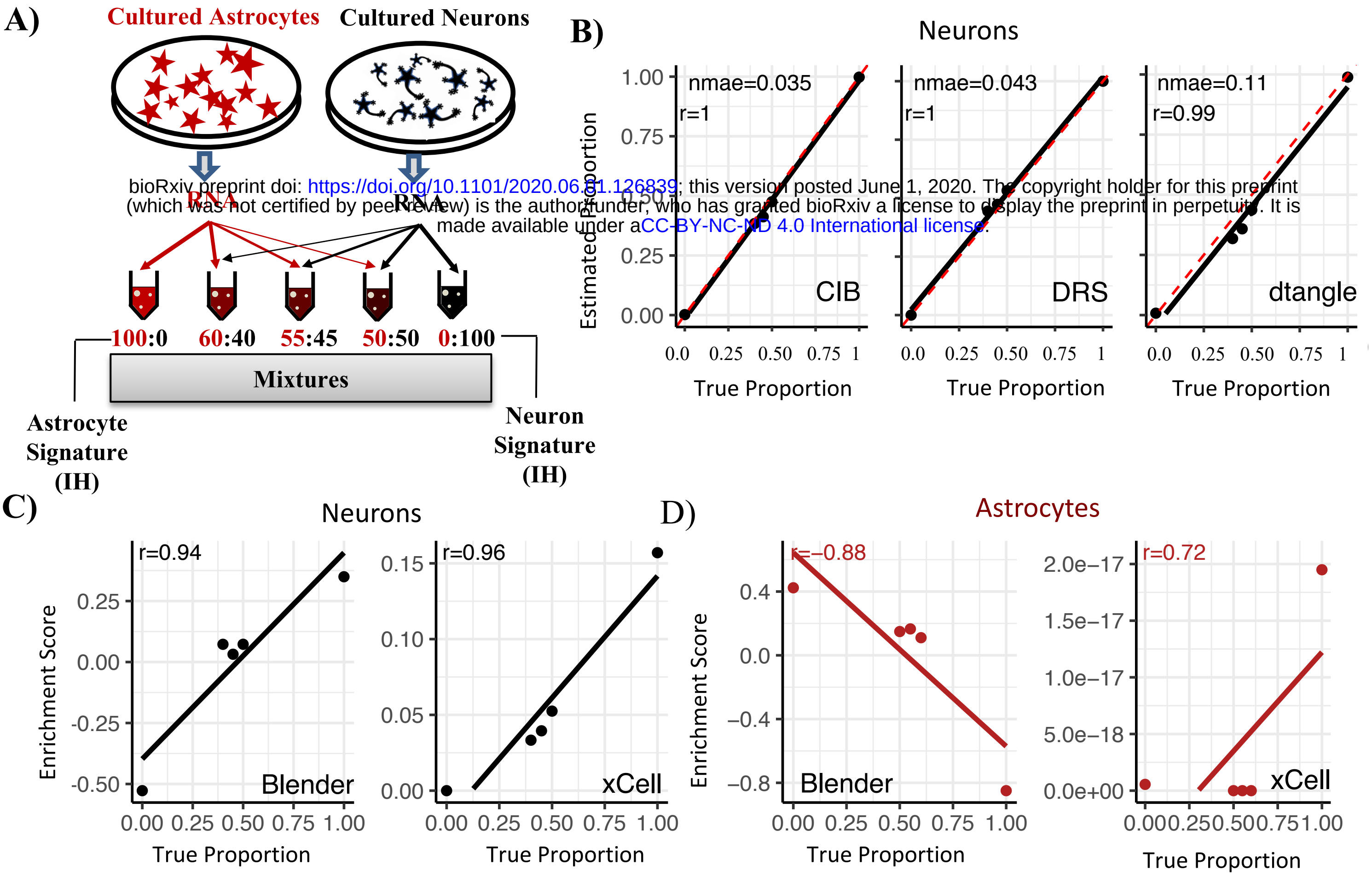
**Supplementary Table 5.** Cell-type specific gene expression signature data. Expression values are normalized and filtered as described in Methods.

**Supplementary Table 6.** Summary of RNA-seq data generated in the present study.

# Figure 1


A)
1) Single-cell RNA-seq:
2) Simulated data:
3) Estimate composition:
Neurons (161)
Astrocytes (62)
Oligodendrocytes (38)
Microglia (16)
Endothelia (20)
In silico mixtures:
100 random samplings
of 100 cells
SC Signatures:
Average expression
across a cell-type
Enrichment methods: xCell, Blender
(Built-in signature data)
Deconvolution methods: DRS, CIB, dtangle
Enrichment scores
Estimated proportions
B)
nmae
0.0 0.1 0.2 0.3 0.4
CIB DRS dtangle
r
0.00 0.25 0.50 0.75
CIB DRS dtangle Blender xCell
Neu Ast Oli Mic End
C)
Neu Ast Oli Mic End
CIB Proportion
nmae=0.02 r=0.96
nmae=0.033 r=0.97
nmae=0.067 r=0.92
nmae=0.11 r=0.94
nmae=0.065 r=0.96
DRS Proportion
nmae=0.18 r=0.93
nmae=0.12 r=0.94
nmae=0.27 r=0.89
nmae=0.42 r=0.9
nmae=0.41 r=0.9
dtangle Proportion
nmae=0.076 r=0.81
nmae=0.12 r=0.79
nmae=0.14 r=0.68
nmae=0.21 r=0.84
nmae=0.28 r=0.78
Blender Score
r=0.86
r=0.92
r=0.79
r=0.91
r=0.83
xCell Score
r=0.63
r=−0.036
True Proportion

# Figure 2



A)
Cultured Astrocytes
Cultured Neurons
RNA
RNA
100:0
60:40
55:45
50:50
0:100
Mixtures
Astrocyte Signature (IH)
Neuron Signature (IH)
B)
Neurons
nmae=0.035
r=1
CIB
nmae=0.043
r=1
DRS
nmae=0.11
r=0.99
dtangle
Estimated Proportion
True Proportion
0.00
0.25
0.50
0.75
1.00
0.0
0.25
0.50
0.75
1
C)
Neurons
r=0.94
Blender
r=0.96
xCell
Enrichment Score
True Proportion
-0.50
-0.25
0.00
0.25
0.00
0.05
0.10
0.15
D)
Astrocytes
r=−0.88
Blender
r=0.72
xCell
Enrichment Score
True Proportion
−0.8
−0.4
0.0
0.4
0.0e+00
5.0e−18
1.0e−17
1.5e−17
2.0e−17

# Figure 3

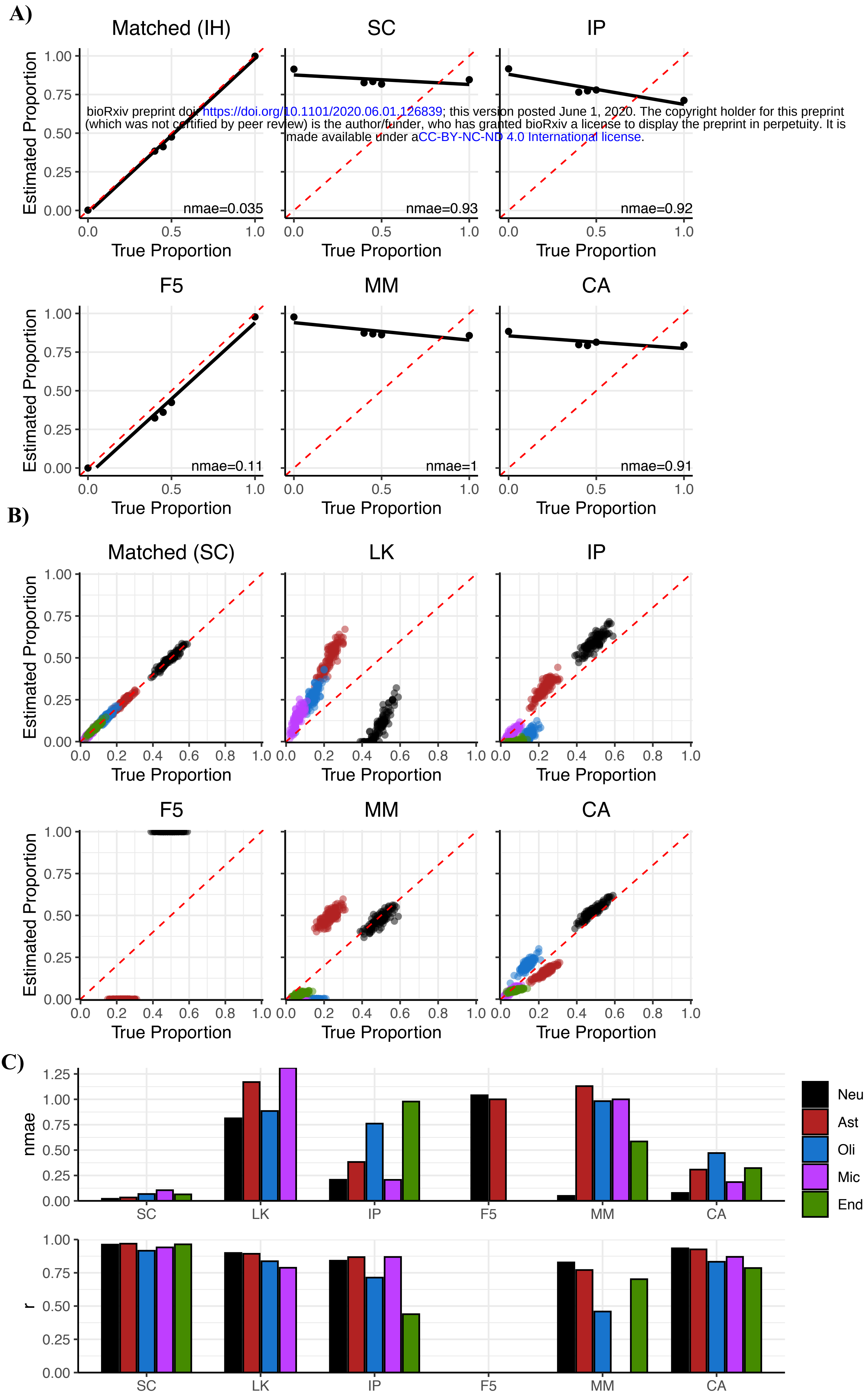
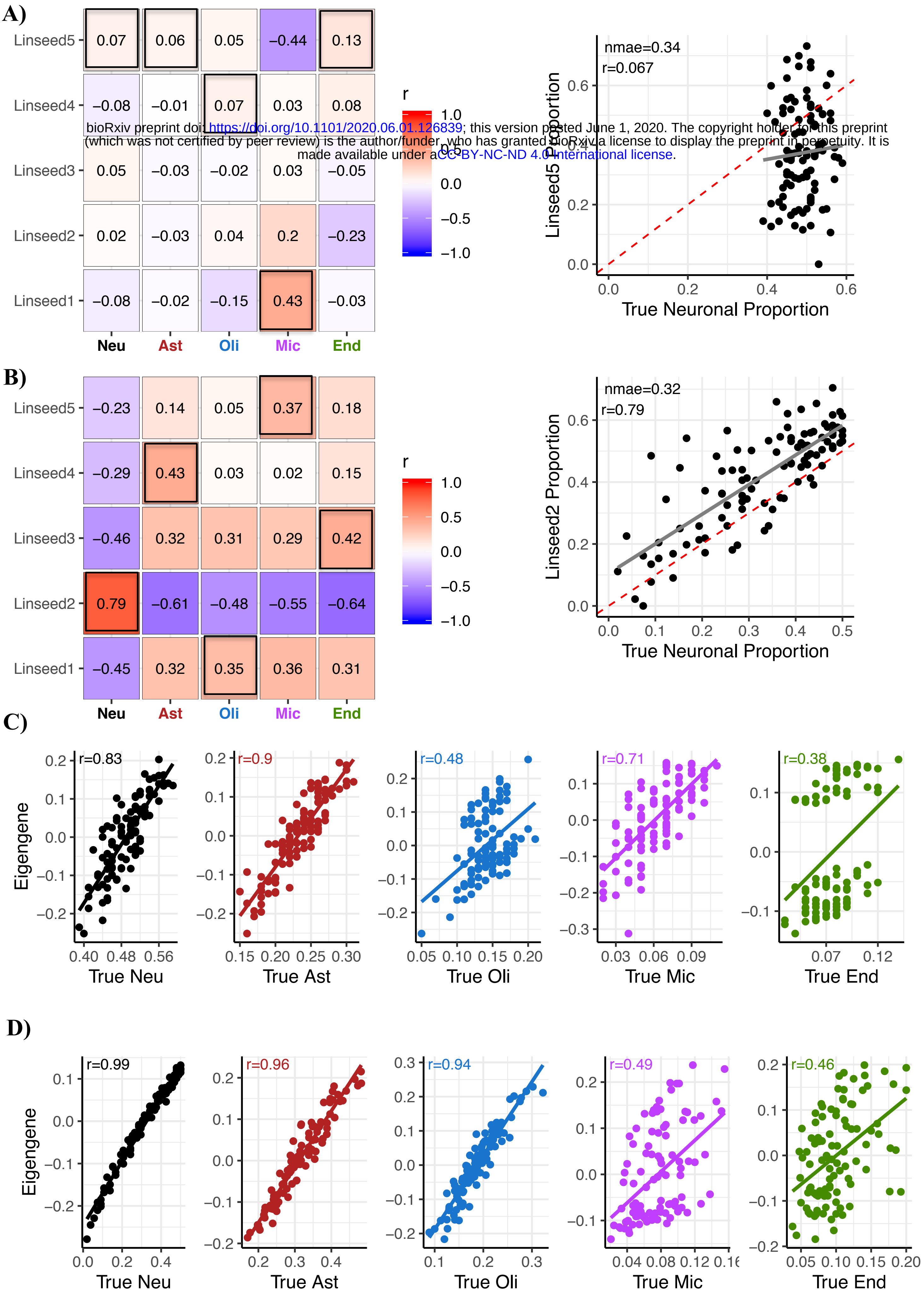
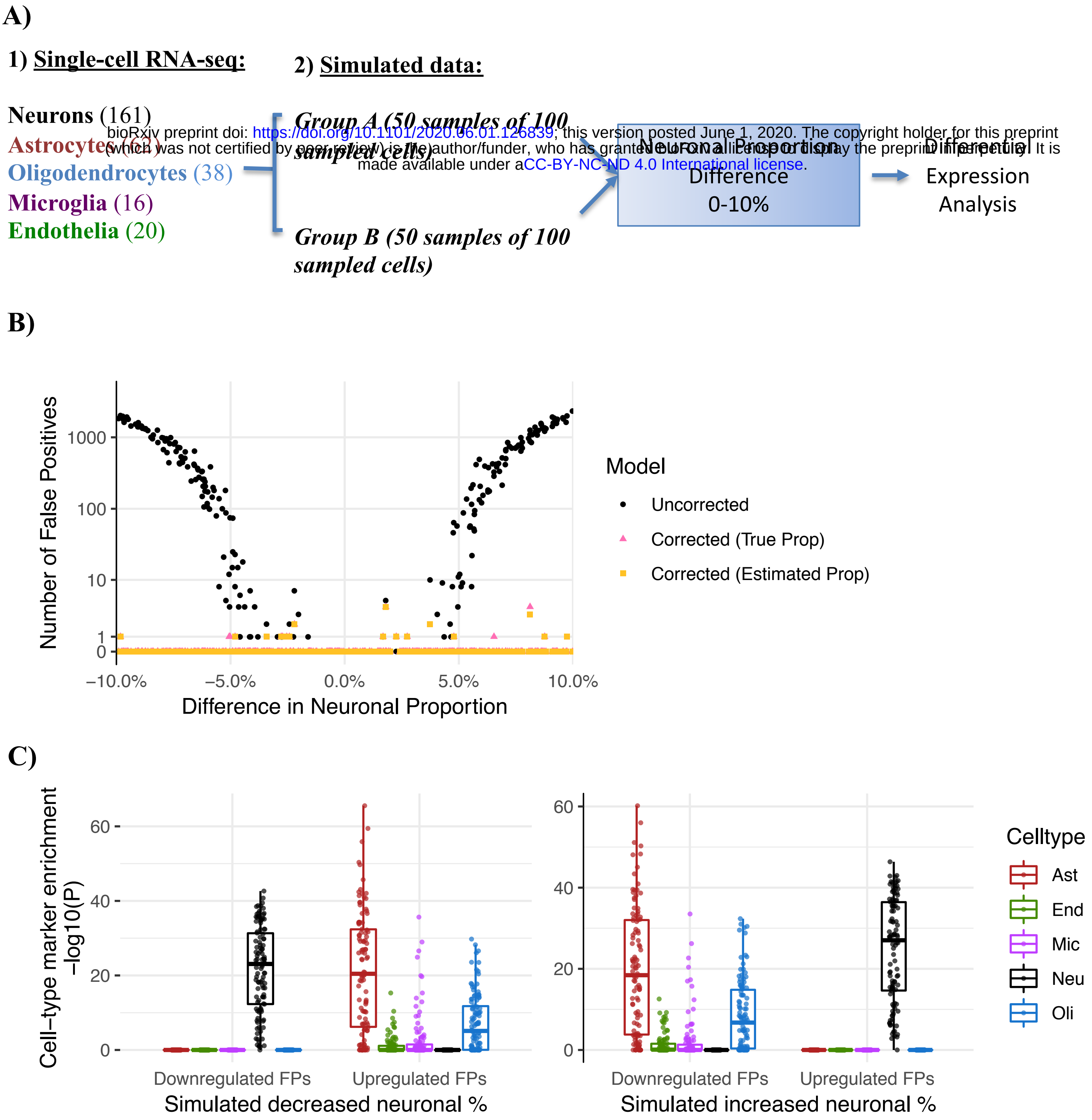

A)
Matched (IH)
SC
IP
F5
MM
CA
Estimated Proportion
True Proportion
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
nmae=0.035
nmae=0.93
nmae=0.92
nmae=0.11
nmae=1
nmae=0.91
B)
Matched (SC)
LK
IP
F5
MM
CA
Estimated Proportion
True Proportion
0.0
0.2
0.4
0.6
0.8
1.0
C)
nmae
1.25
1.00
0.75
0.50
0.25
0.00
SC
LK
IP
F5
MM
CA
r
Neu
Ast
Oli
Mic
End

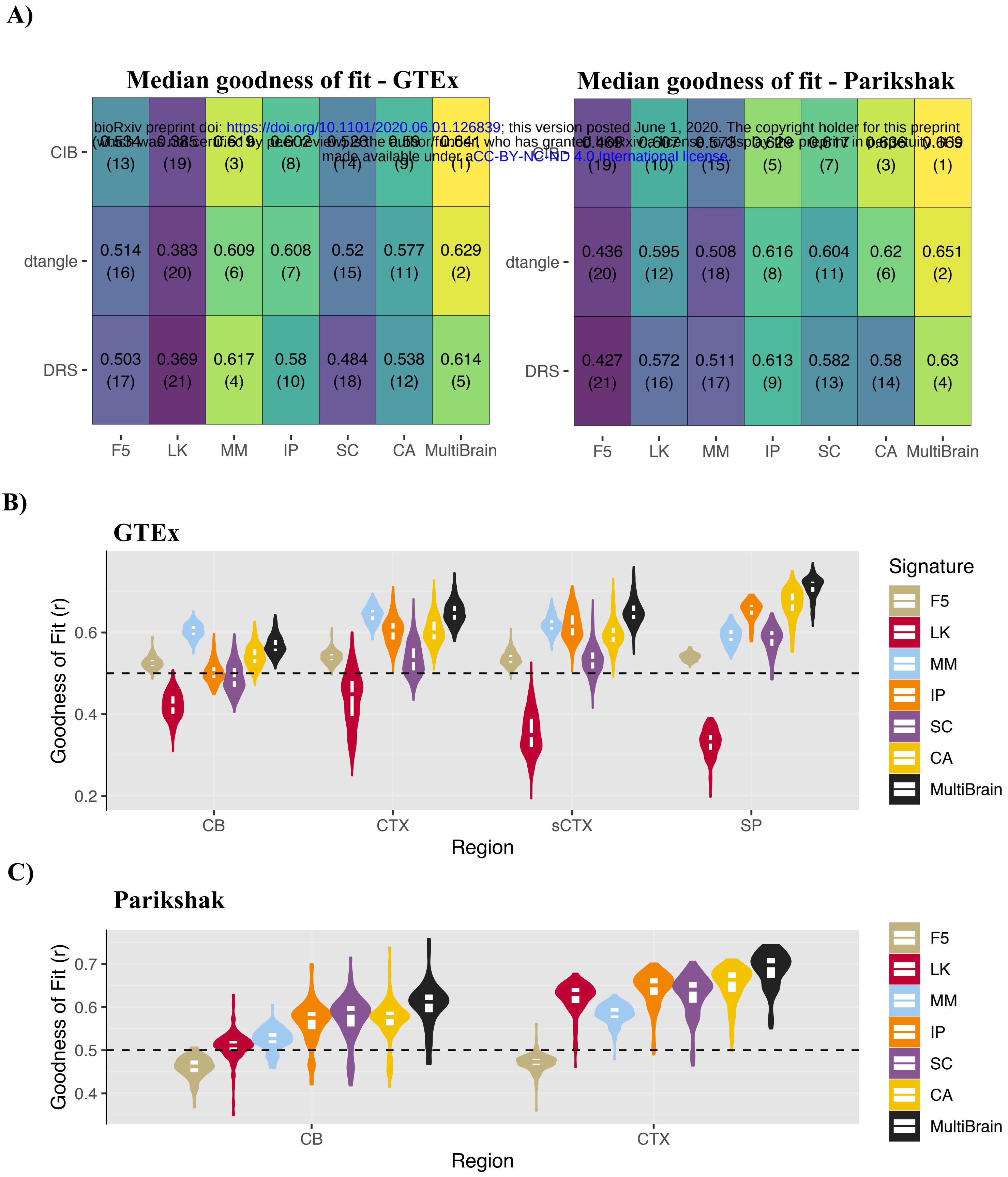
**Figure 4**


A)
Linseed5 0.07 0.06 0.05 −0.44 0.13
Linseed4 −0.08 −0.01 0.07 0.03 0.08
Linseed3 0.05 −0.03 −0.02 0.03 −0.05
Linseed2 0.02 −0.03 0.04 0.2 −0.23
Linseed1 −0.08 −0.02 −0.15 0.43 −0.03
Neu Ast Oli Mic End
r
1.0
0.5
0.0
−0.5
−1.0
nmae=0.34
r=0.067
Linseed5 Proportion
True Neuronal Proportion
0.0 0.2 0.4 0.6
B)
Linseed5 −0.23 0.14 0.05 0.37 0.18
Linseed4 −0.29 0.43 0.03 0.02 0.15
Linseed3 −0.46 0.32 0.31 0.29 0.42
Linseed2 0.79 −0.61 −0.48 −0.55 −0.64
Linseed1 −0.45 0.32 0.35 0.36 0.31
Neu Ast Oli Mic End
r
1.0
0.5
0.0
−0.5
−1.0
nmae=0.32
r=0.79
Linseed2 Proportion
True Neuronal Proportion
0.0 0.1 0.2 0.3 0.4 0.5
C)
Eigengene
r=0.83
True Neu
0.40 0.48 0.56
r=0.9
True Ast
0.15 0.20 0.25 0.30
r=0.48
True Oli
0.05 0.10 0.15 0.20
r=0.71
True Mic
0.03 0.06 0.09
r=0.38
True End
0.07 0.12
D)
Eigengene
r=0.99
True Neu
0.0 0.2 0.4
r=0.96
True Ast
0.2 0.3 0.4
r=0.94
True Oli
0.1 0.2 0.3
r=0.49
True Mic
0.04 0.08 0.12 0.16
r=0.46
True End
0.05 0.10 0.15 0.20

# Figure 5


A)
1) Single-cell RNA-seq:
Neurons (161)
Astrocytes (62)
Oligodendrocytes (38)
Microglia (16)
Endothelia (20)
2) Simulated data:
Group A (50 samples of 100 sampled cells)
Group B (50 samples of 100 sampled cells)
Neuronal Proportion Difference 0-10%
Differential Expression Analysis
B)
Number of False Positives
1000
100
10
1
0
−10.0%
−5.0%
0.0%
5.0%
10.0%
Difference in Neuronal Proportion
Model
Uncorrected
Corrected (True Prop)
Corrected (Estimated Prop)
C)
Cell−type marker enrichment −log10(P)
60
40
20
0
Downregulated FPs
Upregulated FPs
Simulated decreased neuronal %
Simulated increased neuronal %
Celltype
Ast
End
Mic
Neu
Oli

# Figure 6


A)
Median goodness of fit - GTEx
CIB
dtangle
DRS
0.534 (13)
0.385 (19)
0.619 (3)
0.602 (8)
0.528 (14)
0.59 (9)
0.641 (1)
0.514 (16)
0.383 (20)
0.609 (6)
0.608 (7)
0.52 (15)
0.577 (11)
0.629 (2)
0.503 (17)
0.369 (21)
0.617 (4)
0.58 (10)
0.484 (18)
0.538 (12)
0.614 (5)
F5
LK
MM
IP
SC
CA
MultiBrain
Median goodness of fit - Parikshak
CIB
dtangle
DRS
0.469 (19)
0.607 (10)
0.578 (15)
0.629 (5)
0.617 (7)
0.636 (3)
0.669 (1)
0.436 (20)
0.595 (12)
0.508 (18)
0.616 (8)
0.604 (11)
0.62 (6)
0.651 (2)
0.427 (21)
0.572 (16)
0.511 (17)
0.613 (9)
0.582 (13)
0.58 (14)
0.63 (4)
F5
LK
MM
IP
SC
CA
MultiBrain
B)
GTEx
Goodness of Fit (r)
0.2
0.4
0.6
CB
CTX
sCTX
SP
Region
Signature
F5
LK
MM
IP
SC
CA
MultiBrain
C)
Parikshak
Goodness of Fit (r)
0.4
0.5
0.6
0.7
CB
CTX
Region
F5
LK
MM
IP
SC
CA
MultiBrain

## Figure 7

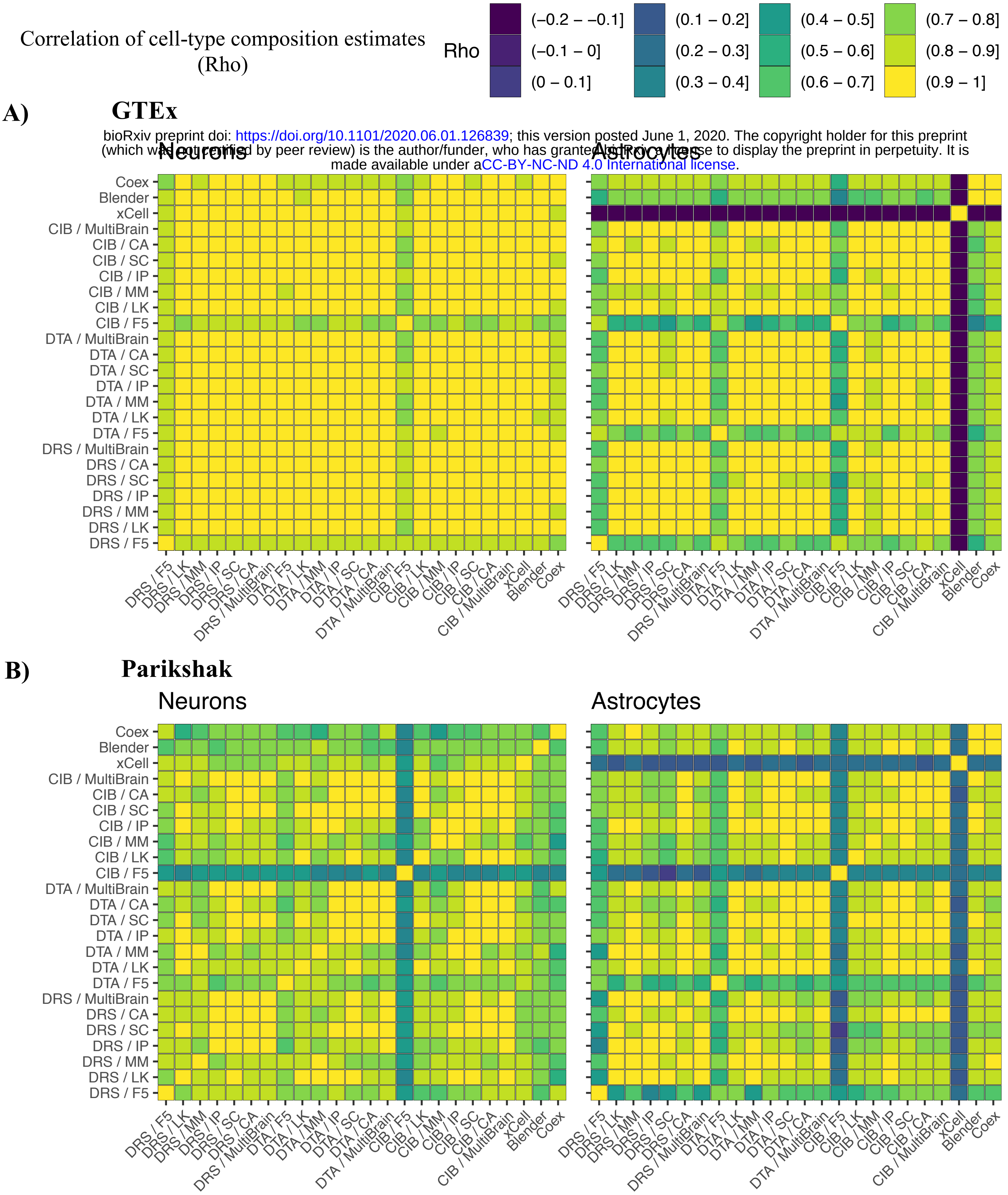

Correlation of cell-type composition estimates (Rho)
Rho
(−0.2 – −0.1]
(−0.1 – 0]
(0 – 0.1]
(0.1 – 0.2]
(0.2 – 0.3]
(0.3 – 0.4]
(0.4 – 0.5]
(0.5 – 0.6]
(0.6 – 0.7]
(0.7 – 0.8]
(0.8 – 0.9]
(0.9 – 1]
A)
GTEx
Neurons
Astrocytes
B)
Parikshak
Neurons
Astrocytes
Coex
Blender
xCell
CIB / MultiBrain
CIB / CA
CIB / SC
CIB / IP
CIB / MM
CIB / LK
CIB / F5
DTA / MultiBrain
DTA / CA
DTA / SC
DTA / IP
DTA / MM
DTA / LK
DTA / F5
DRS / MultiBrain
DRS / CA
DRS / SC
DRS / IP
DRS / MM
DRS / LK
DRS / F5
DRS / F5
DRS / LK
DRS / MM
DRS / IP
DRS / SC
DRS / CA
DRS / MultiBrain
DTA / F5
DTA / LK
DTA / MM
DTA / IP
DTA / SC
DTA / CA
DTA / MultiBrain
CIB / F5
CIB / LK
CIB / MM
CIB / IP
CIB / SC
CIB / CA
CIB / MultiBrain
xCell
Blender
Coex

Figure 8

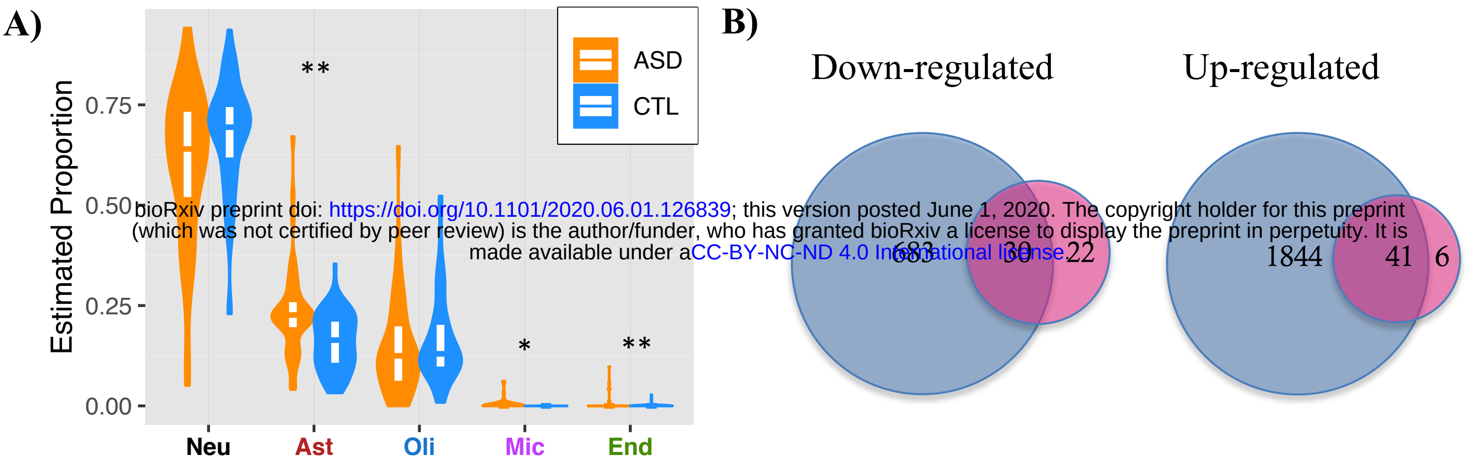

A)
Estimated Proportion
0.75
0.50
0.25
0.00
**
*
**
ASD
CTL
Neu
Ast
Oli
Mic
End
B)
Down-regulated
Up-regulated
683
30
22
1844
41
6

**Table 1**

| Algorithm | Class | Signature | Foundation | Output | Citation |
|---|---|---|---|---|---|
| DeconRNASeq* | Deconvolution | User-specified | Non-negative least squares | Proportions | Gong *et al.* (2013) |
| CIBERSORT* | Deconvolution | User-specified | Support vector regression | Proportions | Newman *et al.* (2015) |
| dtangle | Deconvolution | User-specified | Linear mixing model | Proportions | Hunt *et al.* (2019) |
| Linseed | Deconvolution | None | Simplex topology | Proportions of unlabelled cell-types** | Zaitsev *et al.* (2019) |
| BrainInABlender* | Enrichment | In-built (human and mouse brain) | Average scaled expression of marker genes | Enrichment | Hagenauer *et al.* (2018) |
| xCell | Enrichment | In-built (cultured human brain cells) | Gene set enrichment analysis | Enrichment | Aran *et al.* (2017) |
| Coex | Enrichment | None | Weighted gene co-expression network analysis | Enrichment for unlabelled cell-types** | Kelley *et al.* (2018) |