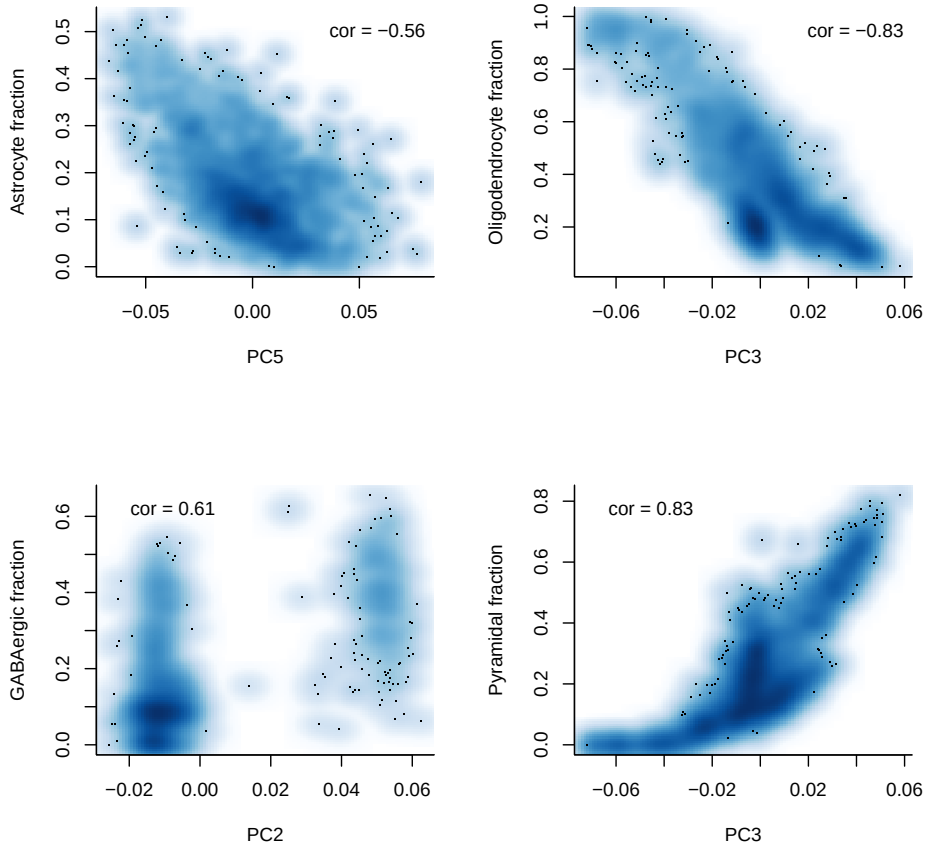
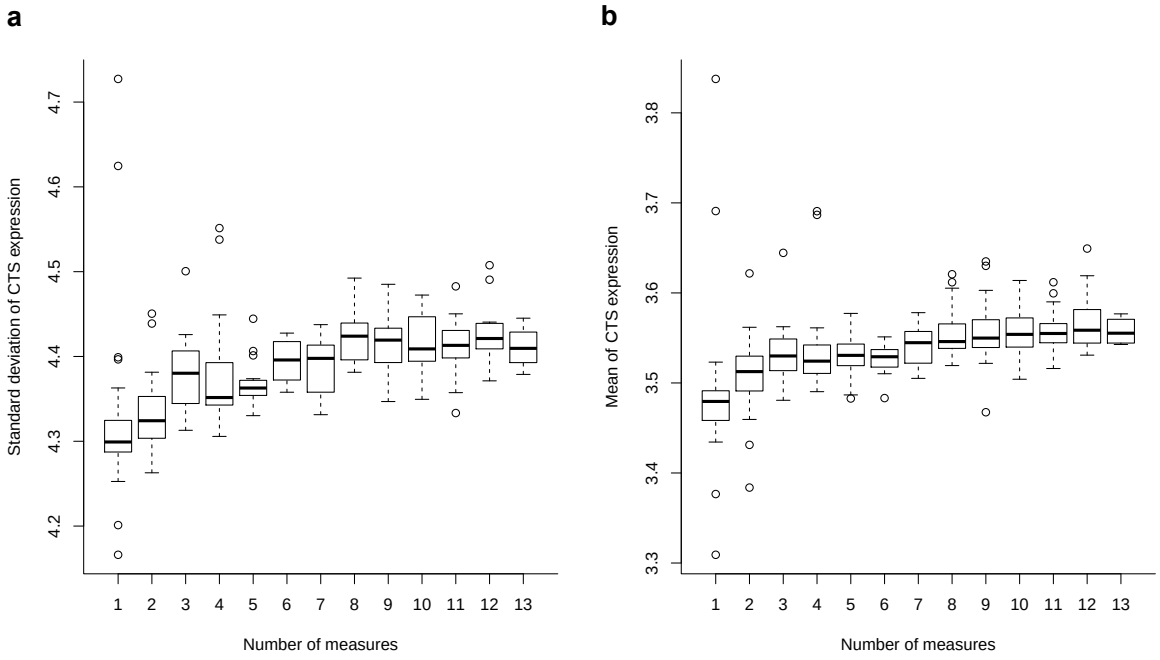
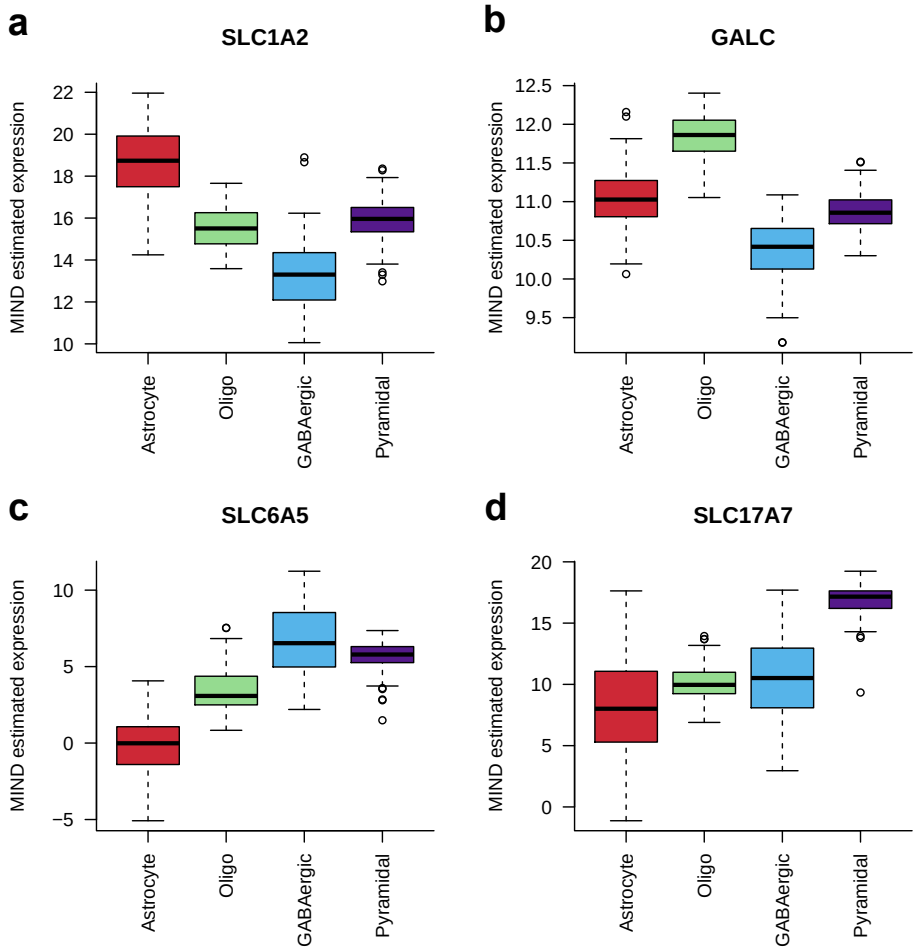
Supplementary Figure



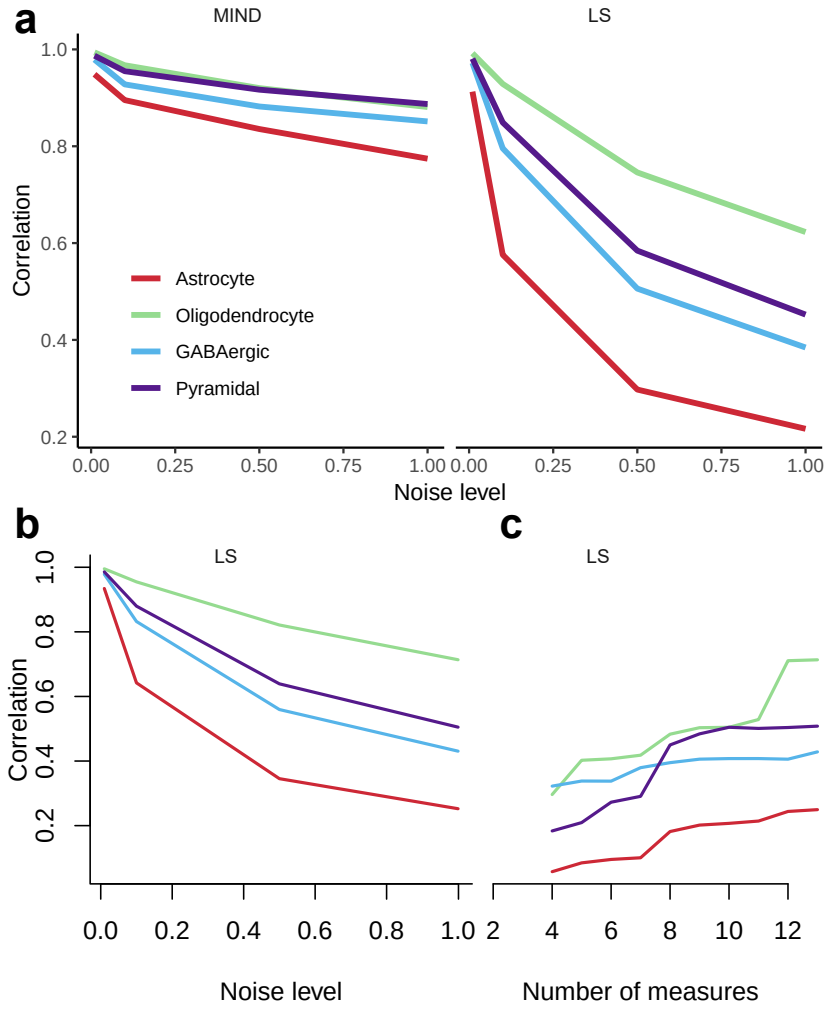
Supplementary Figure 1. Smoothed scatter plots of cell type fractions and principal components (PCs) for GTEx brain tissue samples. Correlations are shown on the figure. We used quantile normalized and scaled data to construct PCs. All genes were used and the top 10 PCs were computed. For each cell type, we chose the PC with the highest absolute correlation with the cell type fraction. The fraction of GABAergic neuron is associated with PC2, which differentiates cerebellum (where GABAergic neuron is prevalent) from other brain regions. Note that PCs have been used as a surrogate for cell type fractions[1].



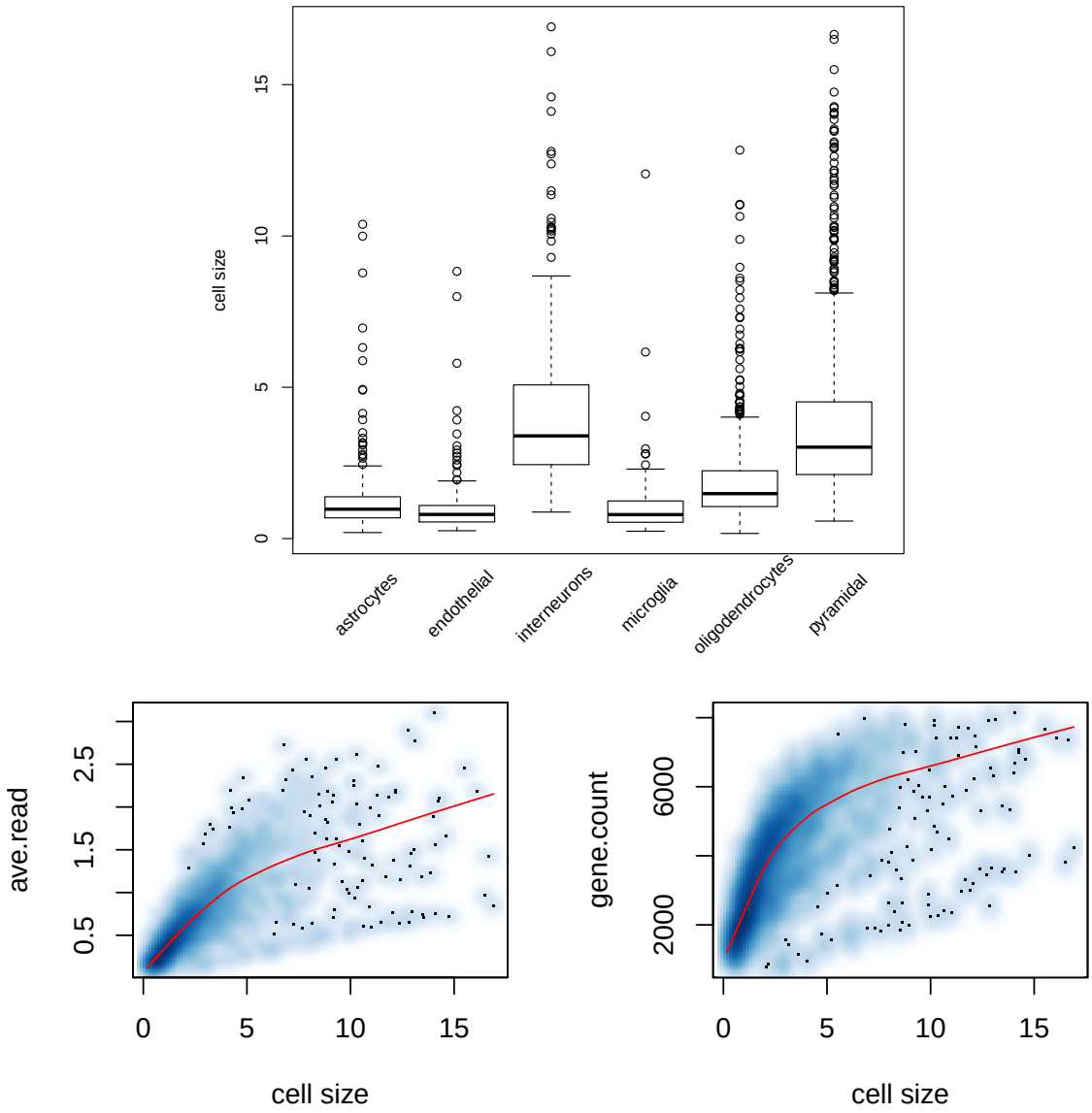
Supplementary Figure 2. The standard deviation (**a**) and mean (**b**) of the MIND estimated cell-type-specific (CTS) expression across all genes for each subject as a function of number of measures in GTEx brain data.



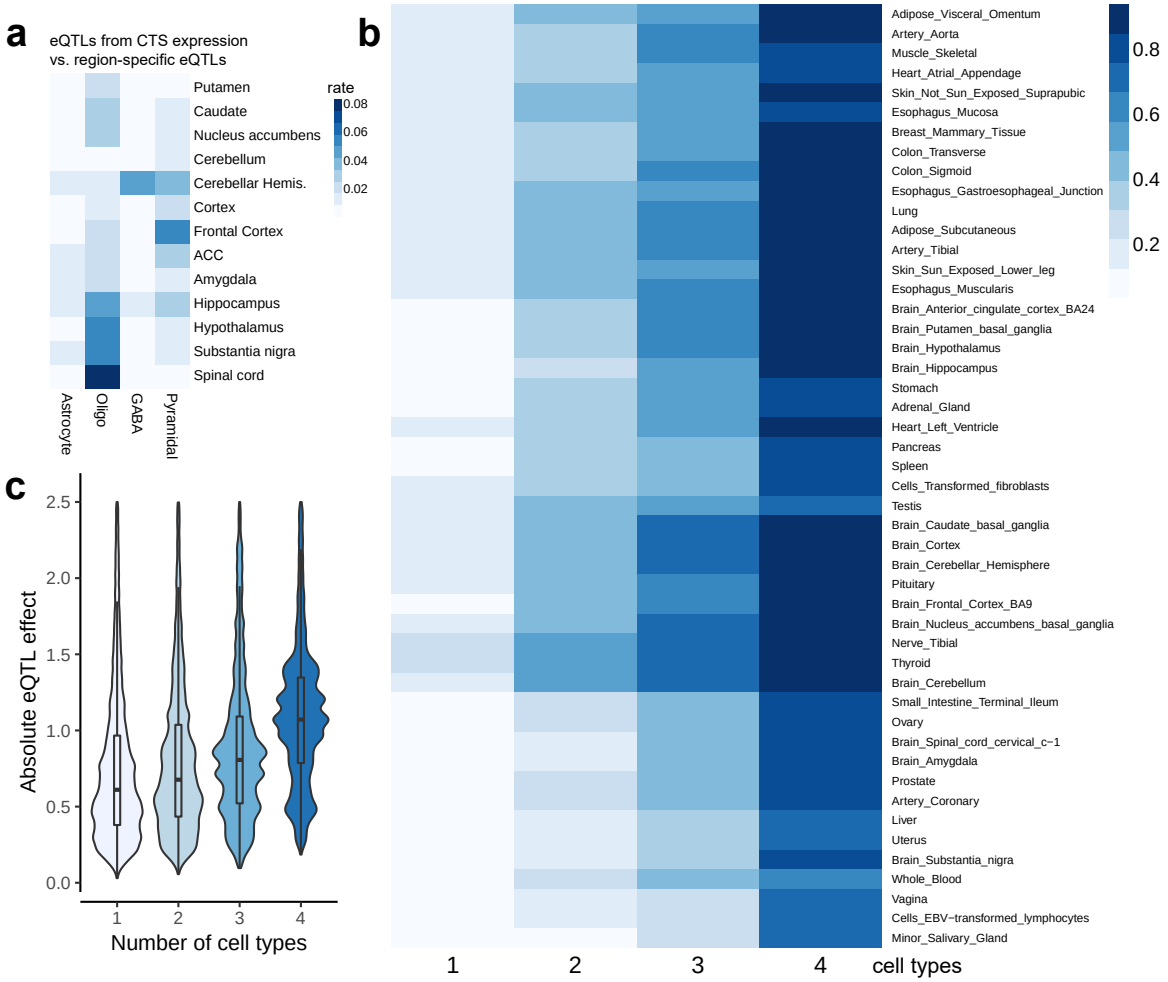
Supplementary Figure 3. The MIND deconvolved expression distinguishes cell types according to known marker genes for astrocyte (**a**), oligodendrocyte (**b**), GABAergic (**c**), and pyramidal neuron (**d**). The boxplots visualize the distribution of CTS expression for GTEx subjects with at least nine measures. For each marker gene, the cell type it marks matches with the cell type that has the maximum average deconvolved expression. Note that these marker genes are not used by CIBERSORT to estimate the cell type fractions.



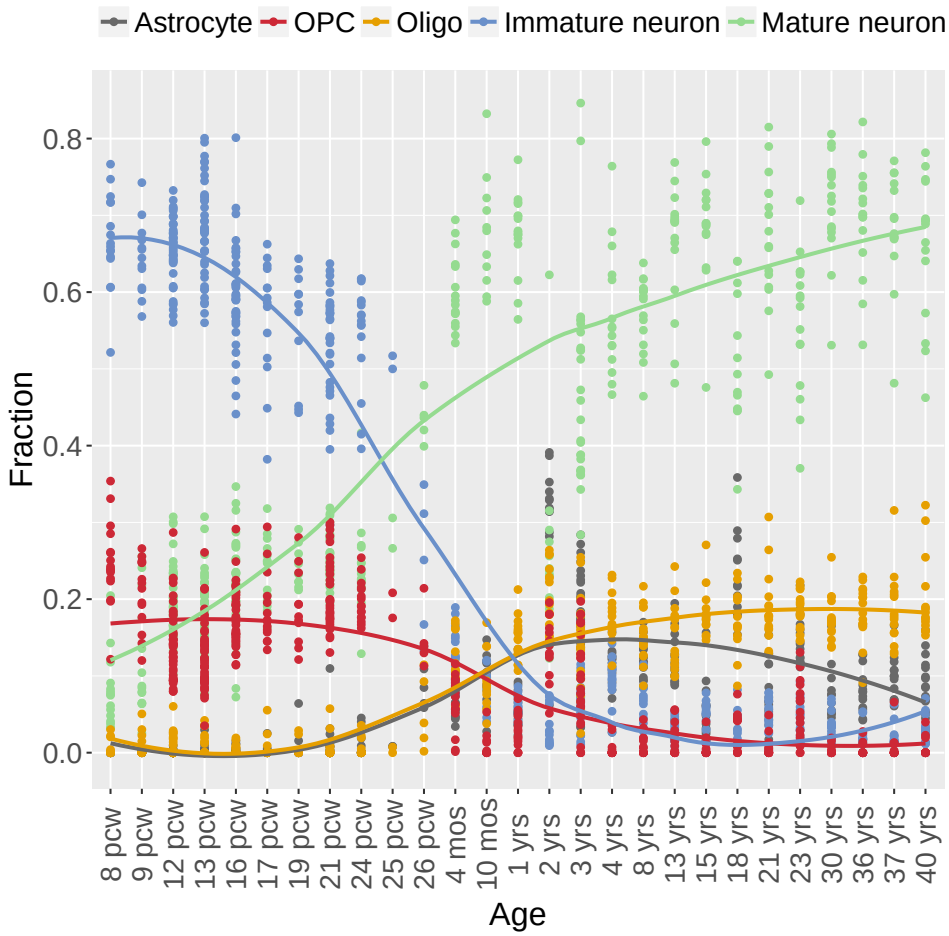
Supplementary Figure 4. (**a**) The impact of region-specific CTS expression ($\boldsymbol{A}_{ij}$) to MIND. We compute the correlation between the deconvolved and true $\boldsymbol{A}_{ij}$ for each cell type, comparing MIND with a least-squares-based (LS) method. (**b-c**) The performance of least-squares-based method under the same simulation setting as in **Fig. 3c,d**. The correlation between the measured and deconvolved expression for each cell type as a function of the noise level. (**c**) The correlation between the measured and deconvolved expression for each cell type as a function of the number of measures. We simulate cell mixture data using the measured CTS expression and the estimated cell type fractions from the GTEx data.



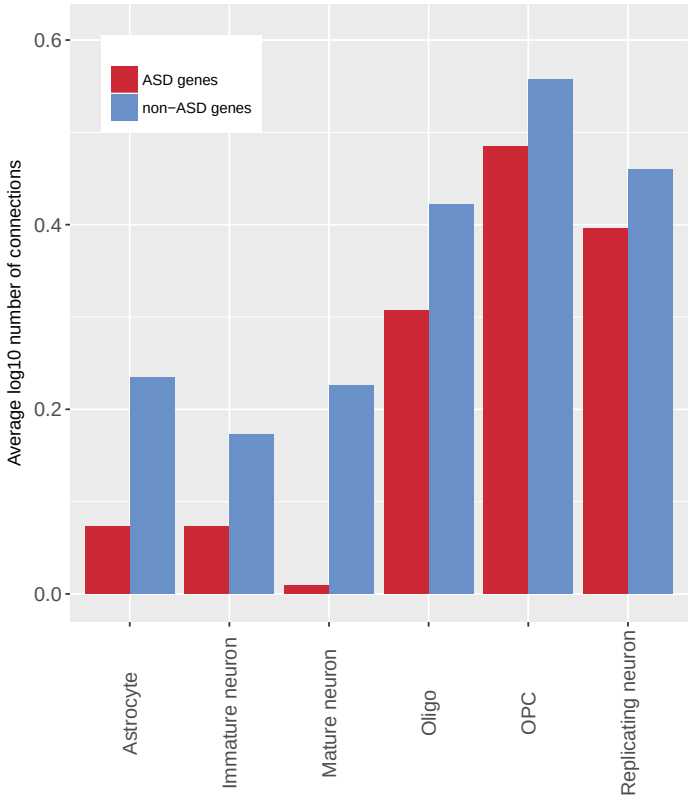
Supplementary Figure 5. The estimated cell size in the scRNA-seq data of Zeisel et al.[2]. Top: neurons (interneurons and pyramidal neurons) have larger cell sizes as compared to non-neurons. Bottom: the average read (left) and gene count with nonzero read (right) vs. cell size. Both have a correlation of 0.6-0.7. The red line is a smooth curve.



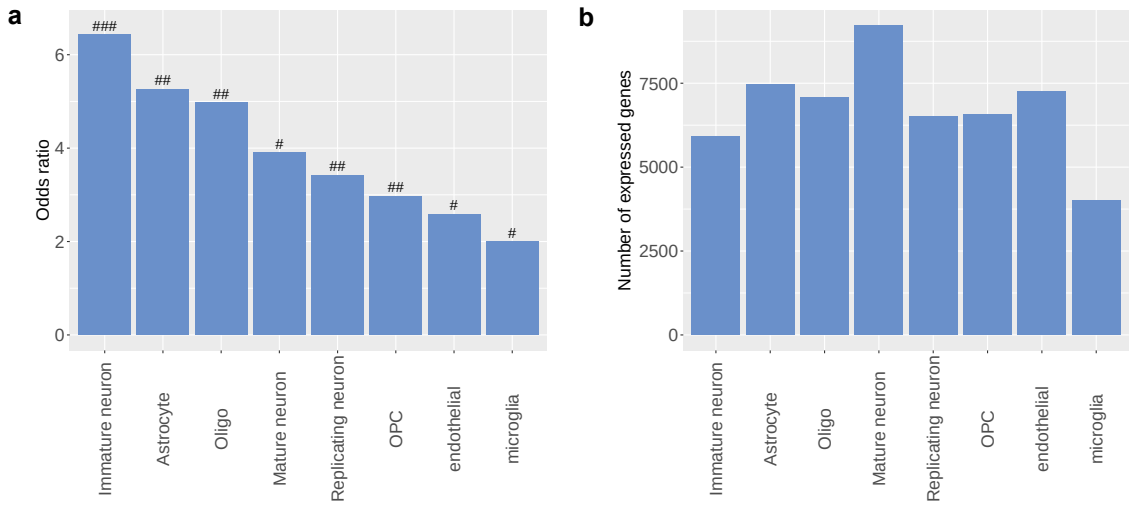
Supplementary Figure 6. Expression quantitative trait loci (eQTL) discovered from tissue-level or CTS gene expression. (**a**) Rate of correspondence between region-specific eQTLs from GTEx samples to eQTLs discovered by MIND from CTS gene expression. (**b**) Overlap between eQTLs appearing in multiple cell types and those in each GTEx tissue type. For eQTLs that appear in one, two, three, and four cell types, respectively, we calculate their probability of being identified in each tissue type. (**c**) Absolute eQTL effects as a function of the number of cell types in which they are identified.



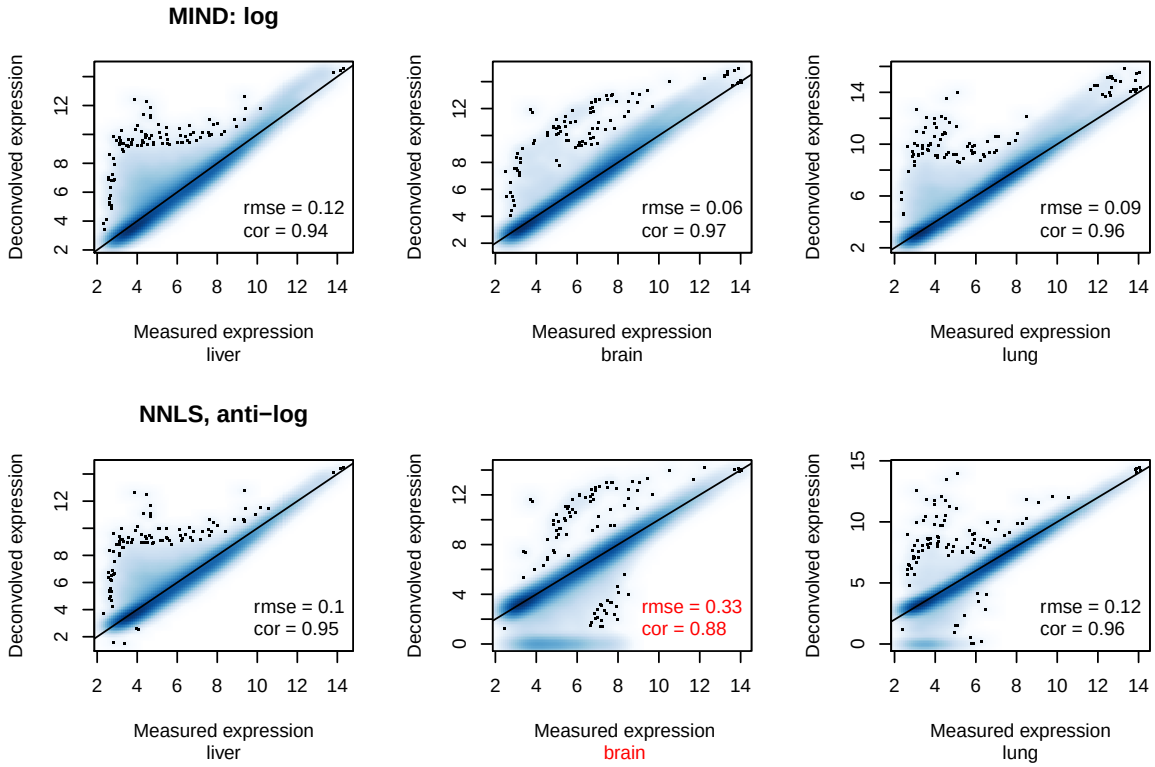
Supplementary Figure 7. The cell type compositions across the lifespan in human brains of BrainSpan data. The curves denote the smooth lines of estimated fractions (represented by dots). Microglia, endothelial cells, and fetal replicating neurons have deconvolved cell fractions close to zero and thus are not shown.



Supplementary Figure 8. The average number of connected genes for ASD genes and non-ASD genes in different cell types (in log10 scale) based on scRNA-seq data from Darmanis et al.[3]. Endothelial and microglia cells are excluded since the numbers of cells are small ($\leq$20).



Supplementary Figure 9. The enrichment analysis of ASD genes expressed in the scRNA-seq data from Darmanis et al.[3]. We focus on 11,215 genes that are expressed in at least 15% cells of one or more cell types. (**a**) The OR (odds ratio) assessing the association between being expressed and ASD genes. We test if OR = 1 for each cell type using Fisher's exact test. "#" denotes p-value $> 10^{-3}$, "##" denotes $10^{-5} \leq$ p-value $\leq 10^{-3}$, and "###" denotes p-value $< 10^{-5}$). (**b**) The number of genes expressed per cell type.



Supplementary Figure 10. Smoothed scatter plots comparing log and anti-log transformation in deconvolution using mixtures of tissue expression in Shen-Orr et al.[4]. There are three tissue types mixed: liver, brain, and lung. NNLS: non-negative least squares; rmse: root mean square error; cor: Pearson correlation.

Supplementary Table

Supplementary Table 1. Analysis of simulated data mimicking bulk gene expression data using MIND. For each simulation setting, we vary the true value of variance parameters, σ_e^2, σ_c^2, and $\sigma_c^{kk'}$, which denote the error variance, and the variance and covariance of CTS expression, respectively. We present the average estimates of variance parameters and the correlation between the estimated (est.) and true CTS expression. The correlation is calculated for each cell type. The results are based on 100 replications.

setting	true value			parameter estimate			correlation of est. and true CTS expression			
	σ_e^2	σ_c^2	$\sigma_c^{kk'}$	$\hat{\sigma}_e^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_c^{kk'}$	astrocyte	oligo	GABAergic	pyramidal
A	1	1	0.5	0.98	1.25	0.36	0.66	0.84	0.71	0.77
B	2	2	1.0	1.99	2.02	0.83	0.69	0.84	0.73	0.78
C	3	3	1.5	3.01	2.76	1.31	0.69	0.84	0.73	0.78
D	4	4	2.0	4.03	3.49	1.79	0.69	0.84	0.73	0.78
E	5	5	2.5	5.05	4.22	2.27	0.69	0.84	0.73	0.78

Supplementary Table 2. The correlation between the estimated fraction of each cell type and the expression fraction of the corresponding marker gene within each of the GTEx brain samples. Since Scheme 4 does not provide specific fractions for GABAergic and pyramidal neurons, we use all neurons and the correlations for the two cell types are italicized. The scheme with the highest correlation for each marker gene is in boldface.

Cell type	Gene	Scheme 1	Scheme 2	Scheme 3	Scheme 4
Astrocyte	SLC1A2	**0.81**	0.76	0.70	0.68
	AQP4	**0.62**	0.61	0.60	0.56
	FGFR3	**0.80**	**0.80**	0.77	0.74
	GJB6	**0.73**	0.71	0.67	0.64
Oligodendrocyte	MBP	0.76	**0.79**	0.78	**0.79**
	SOX10	**0.88**	0.85	0.83	0.84
	MAG	**0.78**	**0.78**	0.76	0.75
	MOG	**0.82**	0.81	0.79	0.79
GABAergic neuron	GAD1	0.33	0.35	**0.40**	***0.54***
	GAD2	0.41	0.40	**0.48**	*0.34*
	SLC32A1	0.48	0.46	**0.50**	*0.34*
Pyramidal neuron	SLC17A7	0.77	0.76	**0.84**	*0.61*
GABAergic+Pyramidal neuron	MYT1L	0.79	0.79	**0.85**	0.83

Supplementary Note

An EM algorithm for the multi-measure deconvolution model

The complete data log-likelihood is given by

$$\ell\left(\Sigma_c, \sigma_e^2\right) = const - \frac{p}{2}\sum_{i=1}^{n} T_i \log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\boldsymbol{X}_{ij} - \boldsymbol{W}_i\boldsymbol{A}_{ij}\right)'\left(\boldsymbol{X}_{ij} - \boldsymbol{W}_i\boldsymbol{A}_{ij}\right)$$
$$-\frac{1}{2}np\log|\Sigma_c| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\boldsymbol{A}_{ij}'\Sigma_c^{-1}\boldsymbol{A}_{ij}.$$

E-step

The E-step is to calculate the expected value of the above statistics given the observed data and the current parameter estimates ($\boldsymbol{\gamma}^{(t)} = (\Sigma_c^{(t)}, \sigma_e^{2(t)})$),

$$E\left(\ell\left(\Sigma_c, \sigma_e^2\right)|\boldsymbol{X}, \boldsymbol{\gamma}^{(t)}\right) = const - \frac{p}{2}\sum_{i=1}^{n} T_i \log(\sigma_e^2) - \frac{1}{2}np\log|\Sigma_c|$$
$$-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[E\left(\boldsymbol{e}_{ij}'|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right)E\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \gamma^{(t)}\right)\right)\right]$$
$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[\boldsymbol{A}_{ij}^{(t)\prime}\Sigma_c^{-1}\boldsymbol{A}_{ij}^{(t)} + \mathrm{tr}\left(\ \Sigma_c^{-1}\Sigma_{ij}^{(t)}\right)\right],$$

where

$$\boldsymbol{A}_{ij}^{(t)} = E\left(\boldsymbol{A}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \Sigma_c^{(t)}\boldsymbol{W}_i'\left(\boldsymbol{W}_i\Sigma_c^{(t)}\boldsymbol{W}_i' + \sigma_e^{2(t)}I_{T_i}\right)^{-1}\boldsymbol{X}_{ij} = \Sigma_{ij}^{(t)}\boldsymbol{W}_i'\boldsymbol{X}_{ij}/\sigma_e^{2(t)}$$

is the empirical Bayes estimate of $\boldsymbol{A}_{ij}$ and its covariance matrix is

$$\Sigma_{ij}^{(t)} = \mathrm{var}\left(\boldsymbol{A}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \Sigma_c^{(t)} - \Sigma_c^{(t)}\boldsymbol{W}_i'\left(\boldsymbol{W}_i\Sigma_c^{(t)}\boldsymbol{W}_i' + \sigma_e^{2(t)}I_{T_i}\right)^{-1}\boldsymbol{W}_i\Sigma_c^{(t)}$$
$$= \left(\boldsymbol{W}_i'\boldsymbol{W}_i/\sigma_e^{2(t)} + \left(\Sigma_c^{(t)}\right)^{-1}\right)^{-1}.$$

For the error term,

$E\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \sigma_e^{2(t)}\left(R_{ij}^{(t)}\right)^{-1}\boldsymbol{X}_{ij}$, $\mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \sigma_e^{2(t)}I_{T_i} - \sigma_e^{4(t)}\left(R_{ij}^{(t)}\right)^{-1}$,

$R_{ij}^{(t)} = \boldsymbol{W}_i\Sigma_c^{(t)}\boldsymbol{W}_i' + \sigma_e^{2(t)}I_{T_i}$.

M-step

In the M-step, we derive the estimate of the covariance matrix of random effects as

$$\Sigma_c^{(t+1)} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[\boldsymbol{A}_{ij}^{(t)}\boldsymbol{A}_{ij}^{(t)\prime} + \Sigma_{ij}^{(t)}\right].$$

The error variance estimate is

$$\sigma_e^{2(t+1)} = \frac{1}{p\sum_{i=1}^{n}T_i}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[E\left(\boldsymbol{e}_{ij}'|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right)E\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{X}_{ij}, \gamma^{(t)}\right)\right)\right].$$

log vs. anti-log transformation

Zhong and Liu[5] raised a concern about using log-transformed data in deconvolution. Shen-Orr et al.[6] provided convincing argument about using log-transformation in their response. In addition, Shannon et al.[7] showed more accurate results when using quantile normalized and log-transformed data to estimate cell type fractions.

Here we further address this issue using the same data as Zhong and Liu[5], i.e., the mixtures of tissue expression in liver, brain, and lung by Shen-Orr et al.[4]. There are 33 mixtures of the three tissues with known mixing fractions. We compare the measured and deconvolved expression, for MIND using log-transformed data and NNLS (non-negative least squares) using anti-log transformed data (**Supplementary Fig. 10**). In MIND, the problem is formulated as 33 measures from a single subject, and NNLS treats it as 33 samples. The goal is to estimate the expression for each of the three tissues. The two approaches are comparable in liver and lung, in terms of root mean square error (rmse) and correlation, but anti-log transformed data produce much worse results in brain, which is the focus of our paper. The reason is that NNLS with anti-log transformed data fails to accurately deconvolve some genes and forces 6% of deconvolved expression exactly as zero.

Sensitivity of estimating cell type fractions

To assess the sensitivity of estimating cell type fractions, we use four schemes of deconvolution via CIBERSORT with different numbers of NeuroExpresso samples as the reference to estimate the cell type fractions in GTEx brain data.

- Scheme 1: use 269 NeuroExpresso samples with 11 types of known neurotransmitter and 2 single-cell clusters of endothelial cell; deconvolve GTEx brain tissue into 12 cell types.
- Scheme 2: use 212 NeuroExpresso samples of three glial cell types (astrocyte, oligodendrocyte, and microglia) and four neuronal cell types (GABAergic, pyramidal, cholinergic, and glutamatergic); six oligodendrocyte samples in NeuroExpresso that may be contaminated as shown in hierarchical clustering are excluded. Deconvolve GTEx brain tissue into 7 cell types.
- Scheme 3: exclude cholinergic and glutamatergic neurons in Scheme 2; use 188 NeuroExpresso samples and deconvolve GTEx brain tissue into 5 cell types.
- Scheme 4: use 212 NeuroExpresso samples as in Scheme 2 and deconvolve GTEx brain tissue into 4 cell types, including three glial cell types and neuron.

To compare the performance of different deconvolution schemes, we calculate the correlation between the estimated fraction of each cell type and the expression fraction of the corresponding marker gene within each of the GTEx brain samples (**Supplementary Table 2**). The expression fraction of a marker gene within each tissue sample is calculated as the ratio of the expression of the marker gene over the sum of the expression of all genes[8].

We choose to use Scheme 3 to estimate cell type fractions in GTEx since it has better performance for the two neuronal subtypes (GABAergic and pyramidal) and comparable results for glial cells. The estimated fraction of microglia is ignorable and thus we exclude the three microglia samples in NeuroExpresso. Finally, we use 185 samples to deconvolve GTEx brain data into four cell types: astrocyte, oligodendrocyte, GABAergic, and pyramidal neurons.

References

1. Mancarci, B. O. *et al.* Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eneuro* ENEURO.0212–17.2017 (2017).

2. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* **347**, 1138–1142 (2015).

3. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112**, 7285–7290 (2015).

4. Shen-Orr, S. S. *et al.* Cell type–specific gene expression differences in complex tissues. *Nature Methods* **7**, 287–289 (2010).

5. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nature methods* **9**, 8 (2012).

6. Shen-Orr, S. S., Tibshirani, R. & Butte, A. J. Gene expression deconvolution in linear space. *Nature methods* **9**, 9 (2012).

7. Shannon, C. P. *et al.* Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PloS one* **9**, e95224 (2014).

8. Zhu, L., Lei, J., Devlin, B., Roeder, K. *et al.* A unified statistical framework for single cell and bulk rna sequencing data. *The Annals of Applied Statistics* **12**, 609–632 (2018).