Using multiple measurements of tissue to estimate individual- and cell-type-specific gene expression via deconvolution

Supplemental Material

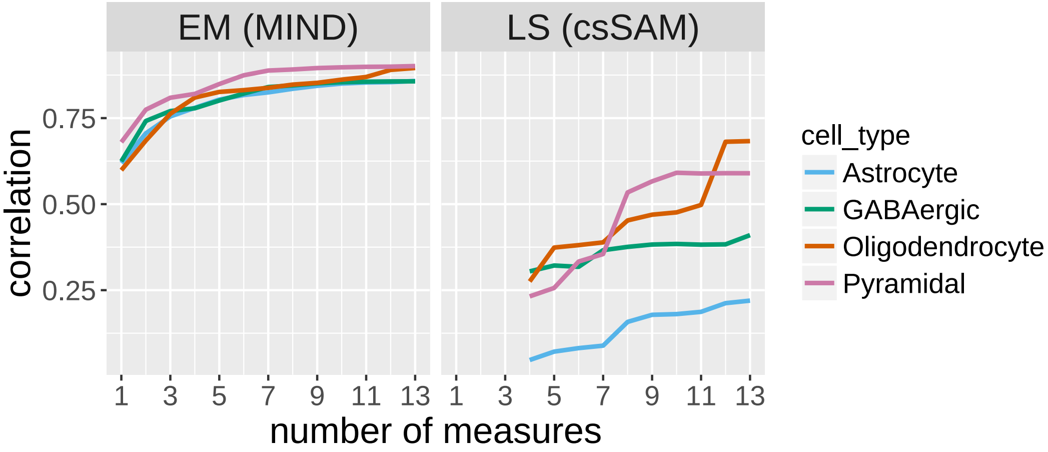

Figure S1. The correlation between the measured and deconvolved expression for each cell type as a function of the number of measures. We simulate cell mixture data using the measured cell-type-specific expression and the estimated cell type fractions from the GTEx data. We compare our proposed EM component of the MIND algorithm with the least squares (LS) based method of csSAM (Shen-Orr et al., 2010).

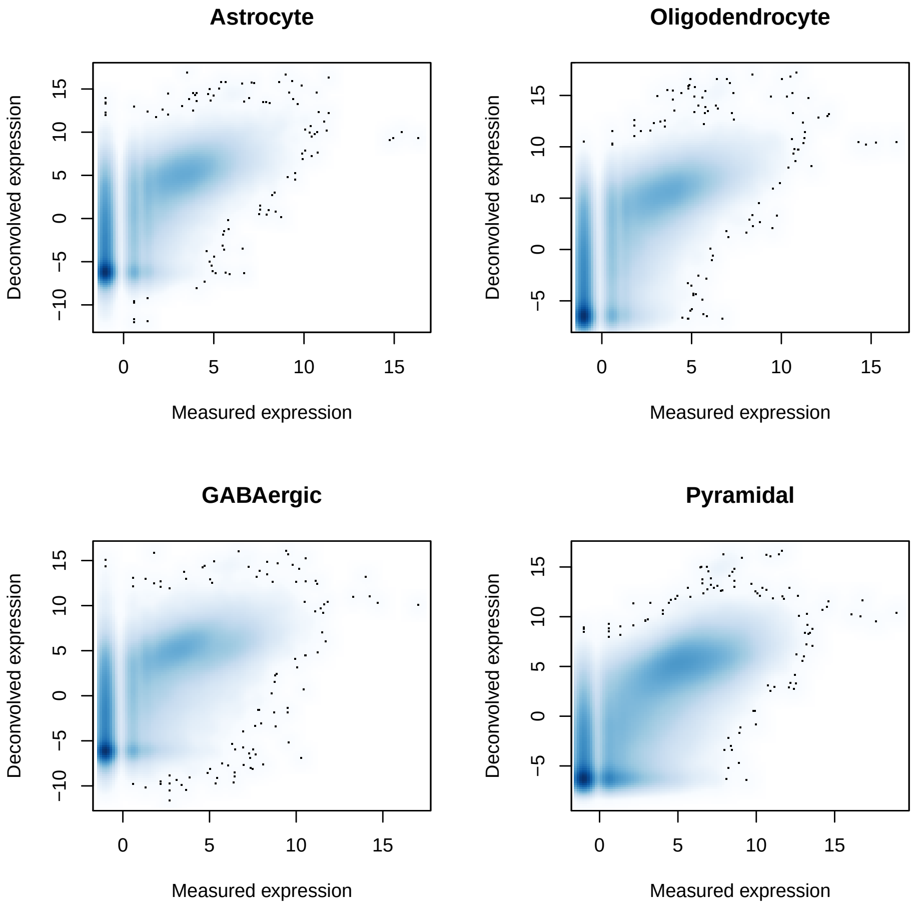

Figure S2. The comparison of the measured and deconvolved cell-type-specific expression. The measured cell-type-specific expression is calculated as the sum of gene expression across all cells of one cell type for each donor in (Habib et al., 2017). Note that the deconvolved and measured cell-type-specific expression differ slightly in scale because of data normalization. The scRNA-seq data have dropout (zero values) in the majority of gene expression.

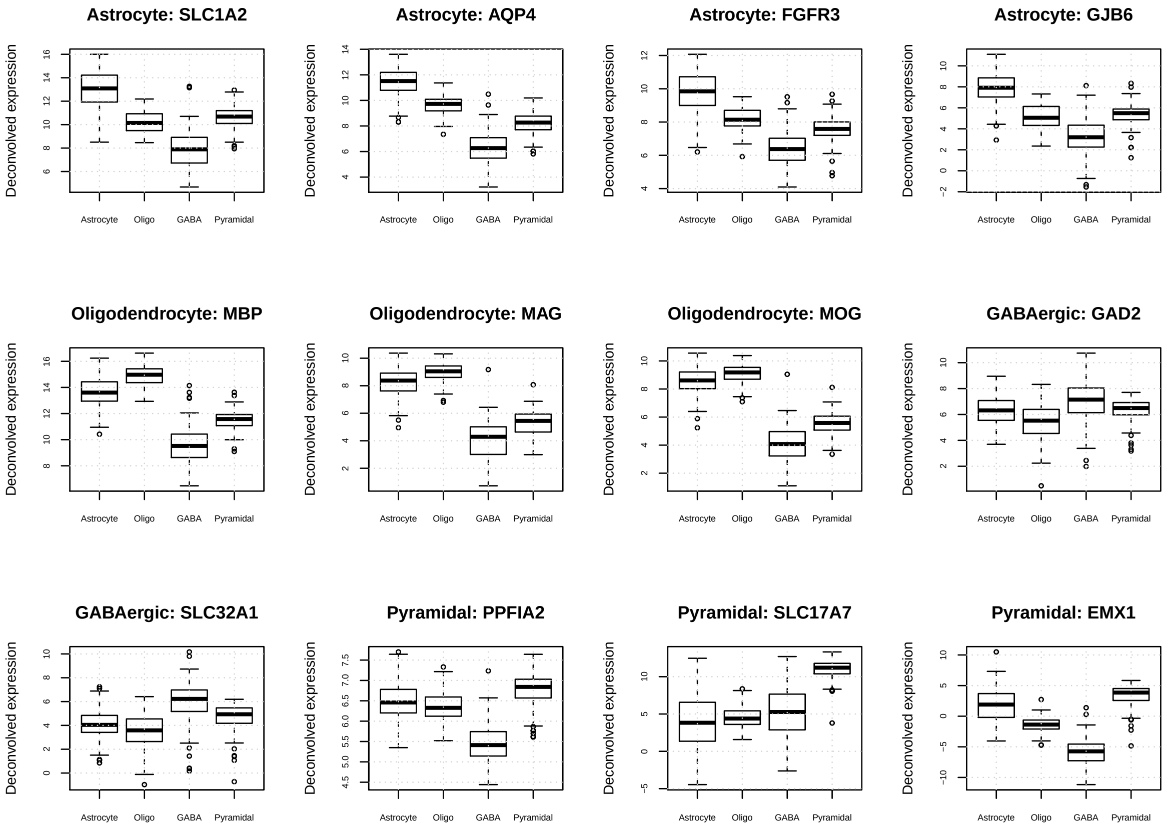

Figure S3. The deconvolved expression distinguishes cell types according to marker genes. The boxplots visualize the distribution of cell-type-specific expression for GTEx subjects with at least nine tissues. The subtitles show the marker gene and its corresponding cell type. For each marker gene, its corresponding cell type matches with the cell type that has the maximum average deconvolved expression.

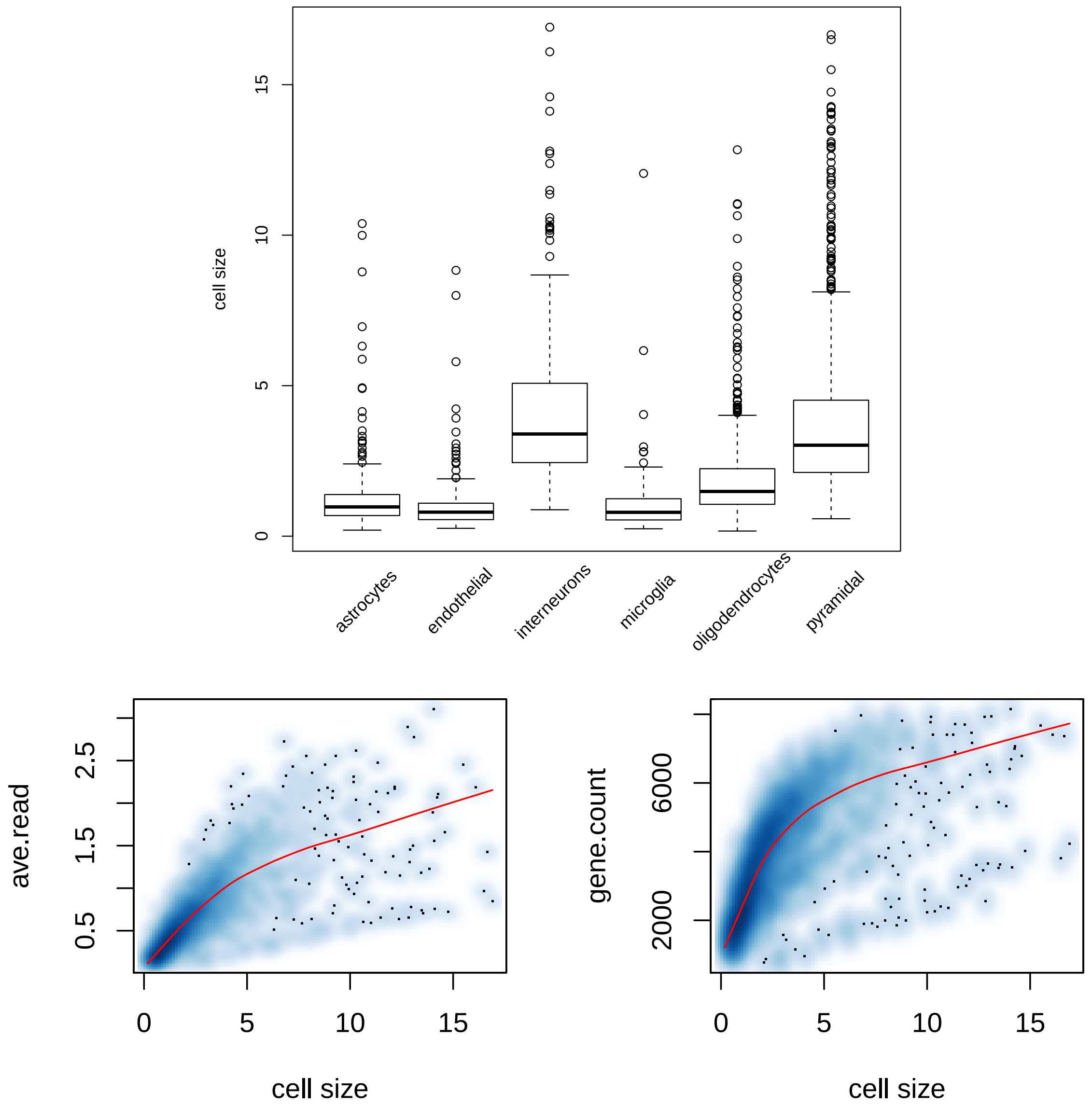

Figure S4. The estimated cell size in the scRNA-seq data of Zeisel et al. (2015). Top: neurons (interneurons and pyramidal neurons) have larger cell sizes as compared to non-neurons. Bottom: the average read (left) and gene count with nonzero read (right) vs. cell size. Both have a correlation of 0.6-0.7. The red line is the smooth curve.

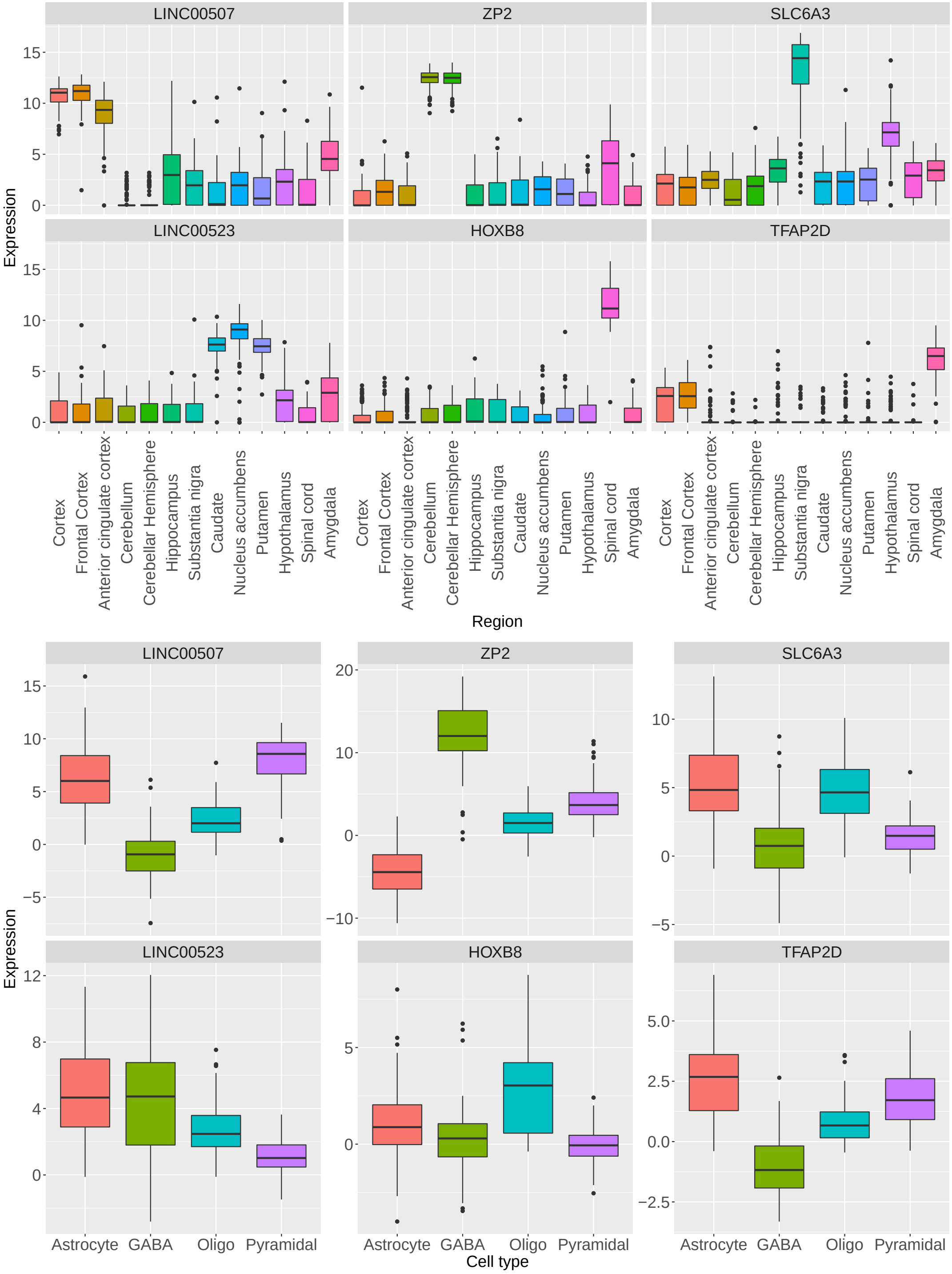

Figure S5. The mapping of variable expression across brain regions onto cell-type-specific expression. The top panel shows the boxplots of tissue-level expression over individuals for six genes/RNA markers and each brain region, and the bottom panel visualizes the cell-type-specific expression for each cell type.

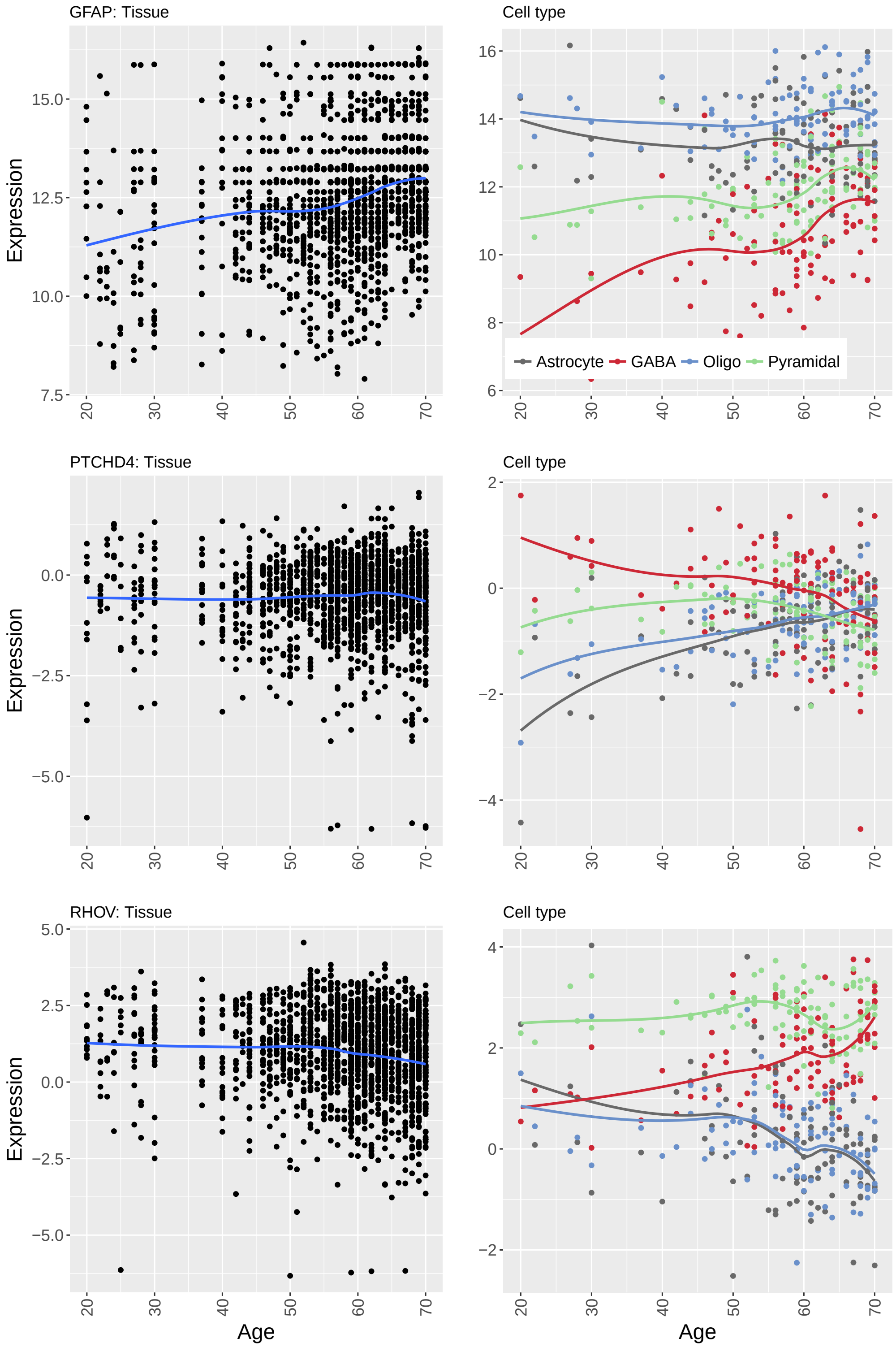

Figure S6. The age trends for tissue-level expression and cell-type-specific expression. Each row is a gene. The first column shows the tissue-level data and the second column shows the cell-type-specific expression. The tissue-level data have been centered per tissue sample.

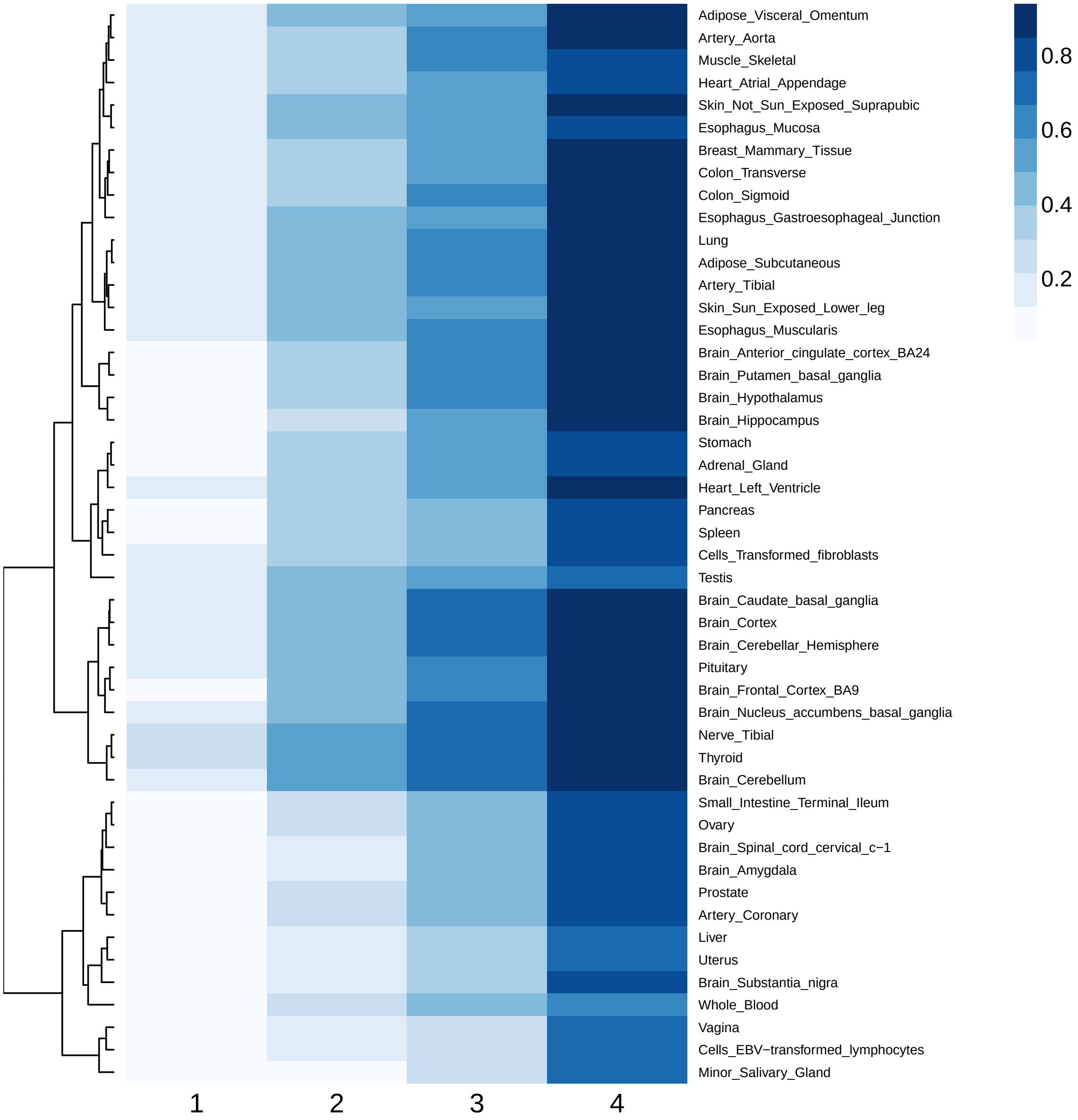

Figure S7. The overlap between eQTLs appearing in multiple cell types and those in each GTEx tissue type. For eQTLs that appear in one, two, three, and four cell types, respectively, we calculate their probability of being identified in each tissue type.

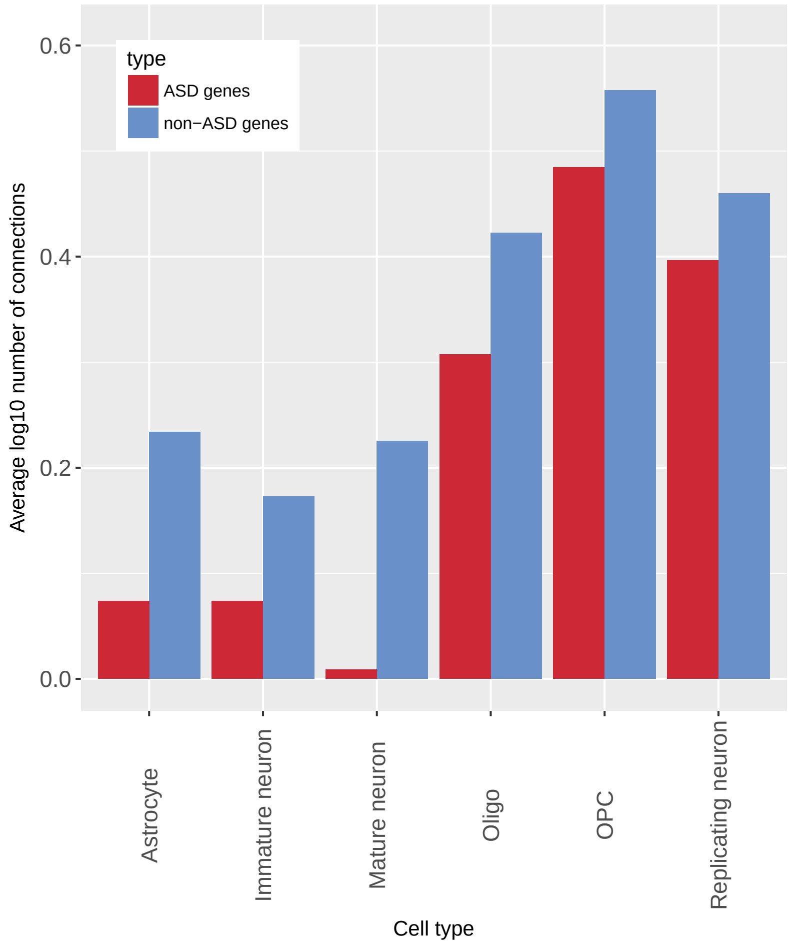

Figure S8. The average number of connected genes for ASD genes and non-ASD genes in different cell types (in log10 scale) based on scRNA-seq data from Darmanis et al. (2015). Endothelial and microglia cells are excluded since the numbers of cells are small (≤20).

Table S1. 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 tissues. The expression fraction of the marker gene within each tissue sample is calculated as the ratio of the expression of marker gene over the sum of the expression of all genes (Zhu et al., 2018). This is to compare the performance of different deconvolution schemes. The scheme with the highest correlation for each marker gene is in boldface. The low correlation for GABAergic neuron may be caused by its diversity. Scheme 4 can be used to estimate cell type fractions when treating neuron as a whole. We use Scheme 3 to estimate cell type fractions in GTEx since it has better performance for the two neuronal subtypes (GABAergic and pyramidal).

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

Scheme 1: uses 269 NeuroExpresso samples with 11 types of neurotransmitter and 2 single-cell clusters of endothelial cell and deconvolves GTEx brain tissues into 12 cell types.

Scheme 2: uses 212 NeuroExpresso samples and deconvolves GTEx brain tissues into 7 cell types, including three glial cells and four neuronal cells (six oligodendrocyte samples in NeuroExpresso that may be contaminated are excluded).

Scheme 3: excludes cholinergic and glutamatergic neurons on the basis of Scheme 2, and thus uses 188 NeuroExpresso samples and deconvolves GTEx brain tissues into 5 cell types.

Scheme 4: uses 212 NeuroExpresso samples and deconvolves GTEx brain tissues into 4 cell types, including three glial cells and neuron. Since it has no specific fractions for GABAergic and pyramidal neurons, the correlations for the two cell types are italicized.

S1 Appendix. Derivation of the algorithm.

An EM algorithm framework for the full model

The parameters in the deconvolution model (3) are the two covariance matrices for random effects, Σ_g and Σ_c, and the error variance σ_e^2. The dimension of Σ_c is relatively low and thus it can be estimated directly via an EM algorithm.

The complete data log-likelihood is given by

$$\ell\left(\Sigma_g, \Sigma_c, \sigma_e^2\right) = const - \frac{p}{2}\sum_{i=1}^{n} t_i \log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - W_i\boldsymbol{\alpha}_i\right)'\left(\boldsymbol{x}_i - W_i\boldsymbol{\alpha}_i\right)$$
$$-\frac{1}{2}nk\log|\Sigma_g| - \frac{1}{2}np\log|\Sigma_c| - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\alpha}_i'\left(\Sigma_g^{-1}\otimes\Sigma_c^{-1}\right)\boldsymbol{\alpha}_i.$$

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

$$E\left(\ell\left(\Sigma_g, \Sigma_c, \sigma_e^2\right)|\boldsymbol{x}, \boldsymbol{\gamma}^{(t)}\right) = const - \frac{p}{2}\sum_{i=1}^{n} t_i\log(\sigma_e^2) - \frac{1}{2}nk\log|\Sigma_g| - \frac{1}{2}np\log|\Sigma_c|$$
$$-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left[E\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right)' E\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \gamma^{(t)}\right)\right)\right]$$
$$-\frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{\mu}_i^{(t)'}\left(\Sigma_g^{-1}\otimes\Sigma_c^{-1}\right)\boldsymbol{\mu}_i^{(t)} + \mathrm{tr}\left(\left(\Sigma_g^{-1}\otimes\Sigma_c^{-1}\right)\Sigma_i^{(t)}\right)\right], \tag{1}$$

where

$$\boldsymbol{\mu}_i^{(t)} = E\left(\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) = \Sigma_\alpha^{(t)}W_i'\left(W_i\Sigma_\alpha^{(t)}W_i' + \sigma_e^{2(t)}I_{pt_i}\right)^{-1}\boldsymbol{x}_i = \Sigma_i^{(t)}W_i'\boldsymbol{x}_i/\sigma_e^{2(t)}$$
$$= \Sigma_i^{(t)}\mathrm{vec}(W_i^{*'}X_i)/\sigma_e^{2(t)}$$

is the empirical Bayes estimate of $\boldsymbol{\alpha}_i$ and its covariance matrix is

$$\Sigma_i^{(t)} = \mathrm{var}\left(\boldsymbol{\alpha}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) = \Sigma_\alpha^{(t)} - \Sigma_\alpha^{(t)}W_i'\left(W_i\Sigma_\alpha^{(t)}W_i' + \sigma_e^{2(t)}I_{pt_i}\right)^{-1}W_i\Sigma_\alpha^{(t)}$$
$$= \left(W_i'W_i/\sigma_e^{2(t)} + \left(\Sigma_\alpha^{(t)}\right)^{-1}\right)^{-1} = \left[\left(I_p\otimes W_i^{*'}W_i^*\right)/\sigma_e^{2(t)} + \left(\Sigma_\alpha^{(t)}\right)^{-1}\right]^{-1}.$$

For error term $\boldsymbol{e}_i$, $E\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) = \sigma_e^{2(t)}\left(R_i^{(t)}\right)^{-1}\boldsymbol{x}_i, \mathrm{var}\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) =$
$\sigma_e^{2(t)}I_{pt_i} - \sigma_e^{4(t)}\left(R_i^{(t)}\right)^{-1}, R_i^{(t)} = W_i\Sigma_\alpha^{(t)}W_i' + \sigma_e^{2(t)}I_{pt_i}$

In the M-step, to obtain closed-form solutions for variance parameters when Kronecker products involved, we rewrite $\Sigma_i^{(t)}$ as a Kronecker product singular value decomposition (KPSVD) (Van Loan, 2000), $\Sigma_i^{(t)} = \sum_l \delta_{il}\mathbf{G}_{il}\otimes\mathbf{H}_{il}$, where $\mathbf{G}_{il}$ is a $p\times p$ matrix and $\mathbf{H}_{il}$ is a $k\times k$ matrix. Based on the properties of the Kronecker product, we derive the closed-form estimates:

$$\hat{\Sigma}_c^{(t+1)} = \frac{1}{np}\sum_{i=1}^{n}\left[A_i\Sigma_g^{-1}A_i' + \sum_l \delta_{il}\mathrm{tr}\left(\mathbf{G}_{il}\Sigma_g^{-1}\right)\mathbf{H}_{il}'\right],$$

where A_i is a $k\times p$ matrix such that $\mathrm{vec}(A_i) = \boldsymbol{\mu}_i^{(t)}$. As noted in Glanz and Carvalho (2013), there is a non-identifiability issue in the Kronecker product. To resolve this issue, we scaled Σ_c with its first element and set the first element as one.

The error variance estimate is

$$\sigma_e^{2(t)} = \sum_{i=1}^{n} \left[E\left(\boldsymbol{e}_i^{'}|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) E\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_i|\boldsymbol{x}_i, \gamma^{(t)}\right)\right)\right] / \left(p \sum_{i=1}^{n} t_i\right).$$

The dimension of Σ_g is hundreds and we assume a sparse inverse covariance (precision) matrix. It can be estimated via an alternating direction method of multipliers (ADMM) algorithm.

An ADMM algorithm to estimate a sparse precision matrix for genes

The major challenge in the model is the estimation of the gene precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_g^{-1}$. To estimate $\boldsymbol{\Theta}$ in each M-step, we first rewrite the expected complete data log-likelihood function in Eq (1) as a function of the gene precision matrix $l\left(\boldsymbol{\Theta}\right)$. We then obtain the penalized MLEs for $\boldsymbol{\Theta}$ by maximizing the penalized log-likelihood, which is the likelihood function plus a graphical lasso penalty on $\boldsymbol{\Theta}$.

Following the idea in Danaher et al. (2014), we use the ADMM algorithm. The problem is equivalent to $\mathrm{minimize}_{\boldsymbol{\Theta}}\ -2l\left(\boldsymbol{\Theta}\right) + \lambda \left|\mathbf{T}\right|_1$ subject to $\mathbf{T} = \boldsymbol{\Theta}$. With an additional penalty (the augmentation) and a Lagrange multiplier matrix $\mathbf{U}$ scaled by a penalty parameter ρ, we write the scaled augmented Lagrangian as

$$L_\rho\left(\boldsymbol{\Theta}, \mathbf{T}, \mathbf{U}\right) = -2l\left(\boldsymbol{\Theta}\right) + \lambda n \left|\mathbf{T}\right|_1 + \frac{n\rho}{2}\left\|\boldsymbol{\Theta} - \mathbf{T} + \mathbf{U}\right\|_F^2 - \frac{n\rho}{2}\left\|\mathbf{U}\right\|_F^2,$$

and here we use $\rho = 1$. The idea is to minimize $L_\rho\left(\boldsymbol{\Theta}, \mathbf{T}, \mathbf{U}\right)$ with respect to $\boldsymbol{\Theta}$ and $\mathbf{T}$, respectively, and then update $\mathbf{U}$ at each iteration. Algorithm S1 provides the details for re-estimating the regularized $\boldsymbol{\Theta}$ within each iteration of the M-step. The tuning parameter (λ) can be selected by Akaike information criterion (Danaher et al., 2014).

Algorithm S1 The ADMM algorithm for re-estimating regularized $\boldsymbol{\Theta}$ within the M-step

```
1. Initialize with $\boldsymbol{\Theta} = \mathbf{I}, \mathbf{U} = \mathbf{T} = \mathbf{0}$.
2. Minimize the target function $-2l\left(\boldsymbol{\Theta}\right) + n\rho\left\|\boldsymbol{\Theta} - \mathbf{T} + \mathbf{U}\right\|_F^2/2$ with respect to $\boldsymbol{\Theta}$. Let $\boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}^{'}$ be the eigendecomposition of the derivative of the target function, $\sum_{i=1}^{n}\mathbf{S}_i/nk + \rho\left(\mathbf{U}^{(t)} - \mathbf{T}^{(t)}\right)/k$, where $\mathbf{S}_i = A_i^{'}\Sigma_c^{-1}A_i + \sum_l \delta_{il}\mathrm{tr}\left(\mathbf{H}_{il}\Sigma_c^{-1}\right)\mathbf{G}_{il}^{'}$. We have the estimate $\boldsymbol{\Theta}^{(t+1)}$ as $\boldsymbol{\Lambda}\tilde{\boldsymbol{\Omega}}\boldsymbol{\Lambda}^{'}$, where $\tilde{\boldsymbol{\Omega}}$ is a diagonal matrix with the $i$th diagonal element as $k\left(-\omega_{ii} + \sqrt{\omega_{ii}^2 + 4\rho/k}\right)/2\rho$, where $\omega_{ii}$ is the $i$th diagonal element of $\boldsymbol{\Omega}$.
3. Minimize $\lambda\left|\mathbf{T}\right|_1 + \rho\left\|\mathbf{T} - \boldsymbol{\Theta} - \mathbf{U}\right\|_F^2/2$ with respect to $\mathbf{T}$, where $\left|\mathbf{T}\right|_1 = \sum_{i \neq j}\left|\mathbf{T}_{ij}\right|$. Let $\mathbf{A} = \boldsymbol{\Theta} + \mathbf{U}$. We have $T_{ii}^{(t+1)} = A_{ii}^{(t)}, i = 1, \ldots, K$, for diagonal elements, and for $i \neq j$, $T_{ij}^{(t+1)} = \mathrm{sgn}\left(A_{ij}^{(t)}\right)\left(\left|A_{ij}^{(t)}\right| - \lambda/\rho\right)_+$.
4. Update $\mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} + \boldsymbol{\Theta}^{(t+1)} - \mathbf{T}^{(t+1)}$.
5. Iterate Step 2-4 until convergence.
```

An EM algorithm for the simplified deconvolution model

The complete data log-likelihood is given by

$$\ell\left(\Sigma_c, \sigma_e^2\right) = const - \frac{p}{2}\sum_{i=1}^{n} t_i \mathrm{log}(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\boldsymbol{x}_{ij} - W_i^*\boldsymbol{\alpha}_{ij}\right)^{'}\left(\boldsymbol{x}_{ij} - W_i^*\boldsymbol{\alpha}_{ij}\right)$$

$$-\frac{1}{2}np\mathrm{log}|\Sigma_c| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\boldsymbol{\alpha}_{ij}^{'}\Sigma_c^{-1}\boldsymbol{\alpha}_{ij}.$$

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

$$\begin{aligned}
E\left(\ell\left(\Sigma_c, \sigma_e^2\right)|\boldsymbol{x}, \boldsymbol{\gamma}^{(t)}\right) &= const - \frac{p}{2}\sum_{i=1}^{n} t_i \log(\sigma_e^2) - \frac{1}{2}np\log|\Sigma_c| \\
-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\sum_{j=1}^{p}&\left[E\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right)' E\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \gamma^{(t)}\right)\right)\right] \\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[\boldsymbol{\mu}_{ij}^{(t)\prime}\Sigma_c^{-1}\boldsymbol{\mu}_{ij}^{(t)} + \mathrm{tr}\left(\ \Sigma_c^{-1}\Sigma_{ij}^{(t)}\right)\right], \qquad (2)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{ij}^{(t)} = E\left(\boldsymbol{\alpha}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) &= \Sigma_c^{(t)} W_i^{*\prime}\left(W_i^{*}\Sigma_c^{(t)}W_i^{*\prime} + \sigma_e^{2(t)}I_{t_i}\right)^{-1}\boldsymbol{x}_{ij} \\
&= \Sigma_{ij}^{(t)} W_i^{*\prime}\boldsymbol{x}_{ij}/\sigma_e^{2(t)}
\end{aligned}$$

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

$$\begin{aligned}
\Sigma_{ij}^{(t)} = \mathrm{var}\left(\boldsymbol{\alpha}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) &= \Sigma_c^{(t)} - \Sigma_c^{(t)}W_i^{*\prime}\left(W_i^{*}\Sigma_c^{(t)}W_i^{*\prime} + \sigma_e^{2(t)}I_{t_i}\right)^{-1} W_i^{*}\Sigma_c^{(t)} \\
&= \left(W_i^{*\prime}W_i^{*}/\sigma_e^{2(t)} + \left(\Sigma_c^{(t)}\right)^{-1}\right)^{-1}.
\end{aligned}$$

For the error term,
$E\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \sigma_e^{2(t)}\left(R_{ij}^{(t)}\right)^{-1}\boldsymbol{x}_{ij}, \mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) = \sigma_e^{2(t)}I_{t_i} - \sigma_e^{4(t)}\left(R_{ij}^{(t)}\right)^{-1}$,
$R_{ij}^{(t)} = W_i^{*}\Sigma_c^{(t)}W_i^{*\prime} + \sigma_e^{2(t)}I_{t_i}$.

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

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

The error variance estimate is

$$\sigma_e^{2(t)} = \frac{1}{p\sum_{i=1}^{n} t_i}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[E\left(\boldsymbol{e}_{ij}^{\prime}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right)E\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathrm{var}\left(\boldsymbol{e}_{ij}|\boldsymbol{x}_{ij}, \gamma^{(t)}\right)\right)\right].$$

References

Danaher P, Wang P, and Witten DM. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**: 373–397.

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, and Quake SR. 2015. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112**: 7285–7290.

Glanz H and Carvalho L. 2013. An expectation-maximization algorithm for the matrix normal distribution. *arXiv preprint arXiv:1309.6609* .

Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, et al.. 2017. Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods* **14**: 955–958.

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, and Butte AJ. 2010. Cell type–specific gene expression differences in complex tissues. *Nature Methods* **7**: 287–289.

Van Loan CF. 2000. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics* **123**: 85–100.

Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al.. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science* **347**: 1138–1142.

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