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

Jiebiao Wang[1], Bernie Devlin[2], Kathryn Roeder[1,3*]

1 Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
2 Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA.
3 Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
* Corresponding author: roeder@andrew.cmu.edu

Keywords: autism, brain, deconvolution, expression quantitative trait loci (eQTL), network analysis, single-cell RNA sequencing

Abstract

Quantification of gene expression in tissue, bulk gene expression, can be a critical step towards characterizing the etiology of complex diseases. Still, cell-level gene expression could be more informative, motivating recent studies that characterize expression of cells and cell types. While there are many strengths to single cell expression, the data tend to be noisy. Instead, we propose a method to glean more insights from bulk gene expression. Our objective is to borrow information across multiple measurements of the same tissue per individual, such as multiple regions of the brain, using an empirical Bayes approach to estimate individual- and cell-type-specific gene expression. Simulations and data analyses demonstrate the advantages of our novel method, which is used to analyze multiple measurements of brain tissue (i.e., brain regions) from the GTEx (Genotype-Tissue Expression) project and the BrainSpan atlas of the developing human brain. This deconvolved expression complements single-cell expression data and provides new insights into tissue-level expression. For example, from GTEx data we identify a subset of expression quantitative trait loci (eQTLs) that are specific to neurons, others specific to astrocytes, and yet others active across all cell types. From the individual- and cell-type-specific gene expression from BrainSpan data, we estimate gene co-expression networks in specific cell types, which are then interpreted in light of genetic findings in autism spectrum disorder (ASD). The co-expression network from immature neurons identifies a cluster of ASD risk genes and additional ASD-correlated genes, many of which play a regulatory role in cell development and differentiation.

Introduction

To characterize the etiology of complex phenotypes and diseases, many studies have quantified gene expression in tissue. For instance, such a characterization, often called bulk gene expression data, has been key to identifying expression quantitative trait loci (eQTLs), which in turn have

been linked to disease risk (Fromer et al. 2016; Zhu et al. 2016, 2018; Dobbyn et al. 2018). Hence many datasets from tissue have been generated. Bulk gene expression data, however, probably lack the resolution to understand disease etiology fully. Recently, studies of single-cell RNA sequencing (scRNA-seq) have quantified gene expression at the level of cells and cell types (Darmanis et al. 2015; Camp et al. 2015; Zeisel et al. 2015; Habib et al. 2017). This is potentially critical for brain tissue, which harbors myriad cell types whose functions are not fully resolved. Despite the many strengths to cell-specific expression, scRNA-seq data tend to be expensive, noisy (due to dropout or lack of sufficient variability), and difficult to generate, especially for brain tissue. By contrast, there are established resources, such as the BrainSpan atlas of the developing human brain (Kang et al. 2011; Willsey et al. 2013; Miller et al. 2014) and the GTEx (Genotype-Tissue Expression) project (GTEx Consortium 2015, 2017), among others, which have collected and/or are continuing to collect bulk transcriptome data from multiple brain regions. Here we present a method to exploit such resources to learn about individual- and cell-specific gene expression from the tissue-level transcriptomes.

Bulk transcriptome data are a convolution of gene expression from myriad cells and cell types. Many studies have proposed computational methods of deconvolution (Abbas et al. 2009; Shen-Orr et al. 2010; Gaujoux and Seoighe 2013; Newman et al. 2015; Li et al. 2016), which typically analyze gene expression from a series of subjects to estimate the fraction of each cell type within the tissue of each subject. Such analyses use information either from genes that are expressed specifically in certain cell types (marker genes) or the profile of expression of all genes within each cell type. We shall call either the signature matrix (Fig. 1A). These methods, typically designed for deconvolving a single measure of the same tissue across individuals, lend some cell-type-specific interpretation for bulk gene expression data. We shall call these deconvolution methods "single-measure deconvolution" hereafter.

Most single-measure deconvolution methods target estimation of cell fractions or some function of them, rather than obtain an estimate of cell-type-specific gene expression within the tissue. However, repeated measures of gene expression, such as assessment of gene expression over brain regions, together with purified-cells data (Mancarci et al. 2017) or scRNA-seq data, provide an unprecedented opportunity to study the cell-type-specific expression, even from bulk expression. Multiple measures from the same individual, but different brain regions, share common cell types. This commonality will allow us to extend the existing single-measure deconvolution methods to achieve more insights into gene expression within cell types. Indeed, resources such as GTEx and BrainSpan, together with transcriptome databases of purified cells (Mancarci et al. 2017) and scRNA-seq data (Darmanis et al. 2015; Camp et al. 2015; Zeisel et al. 2015; Habib et al. 2017), offer an ideal setting for deconvolution.

In this work, we extend single-measure deconvolution to multi-measure deconvolution. This enables the estimation of the individual- and cell-type-specific gene expression for a large number of individuals from existing gene expression resources. Such analyses can also lend insight into what determines cell level gene expression; for instance, whether it changes with development and how it is affected by genetic variation. We implement the Multi-measure INdividual Deconvolution (MIND) algorithm through two steps (Fig. 1B). First, we estimate the cell type composition of tissue from its expression and that of references samples of purified-cells or single-cell expression data. In the second step, we calculate the empirical Bayes estimates of average cell-type gene expression, per individual, via a computationally efficient algorithm. Simulations and data analyses demonstrate the advantages and robustness of MIND, which is used to analyze multiple brain regions from the GTEx project and the BrainSpan atlas of the

developing human brain. The deconvolved individual-level expression complements single-cell expression data, which are currently collected from a limited number of individuals, and provides new insights into gene expression data, for example, via cell-type-specific analysis of eQTLs and co-expression networks. Finally, we show how results from MIND analyses further our understanding of the etiology of autism spectrum disorder (ASD).

Results

Model overview

The numerous studies using gene expression deconvolution (Abbas et al. 2009; Shen-Orr et al. 2010; Gaujoux and Seoighe 2013; Newman et al. 2015; Li et al. 2016) essentially share a similar idea, deconvolving a single measure from each subject (Fig. 1A). Let X be the observed tissue gene expression matrix ($p \times n$) for a single measure per subject, for p genes in n individuals. When the tissue consists of k cell types, the goal of gene expression deconvolution is to find matrices A and W such that

$$\underset{(p\times n)}{X} = \underset{(p\times k)}{A}\,\underset{(k\times n)}{W} + \underset{(p\times n)}{E}\ , \tag{1}$$

where A is the signature matrix, W represents the mixing fractions of k cell types for each individual and the sum of each column is equal to one, and E is the error term. While there are exceptions (Shen-Orr et al. 2010), typically inference is focused on estimating and studying the cell type fractions (W), which are individual specific. When reference samples are available, such as purified cells or scRNA-seq data, the signature matrix (A) can be estimated by differential expression analysis of cell types from the reference samples. With known A, the deconvolution becomes a standard regression problem (Abbas et al. 2009).

We extend the single-measure deconvolution in Eq. 1 by borrowing information across multiple measures from the same tissue and individual to estimate individual- and cell-type-specific gene expression. The MIND algorithm unfolds as follows (Fig. 1B). First, we select cell type marker genes and build a signature matrix (A) using the reference samples (e.g., purified cells from the NeuroExpresso database or scRNA-seq data). Then, we use existing deconvolution methods to estimate the cell type fractions (W) for samples (e.g., GTEx brain tissue or the BrainSpan atlas). Finally, with the pre-estimated cell type composition, we deconvolve the expression of multiple measures from the same individuals. The output of MIND is the cell-type-specific gene expression for each individual ($\boldsymbol{\alpha}_{ij}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$).

Note that as compared to existing single-measure deconvolution models, we allow both cell-type-specific expression ($\boldsymbol{\alpha}_{ij}$) and cell type fraction ($W_i^*; i = 1, 2, \ldots, n$) to be individual-specific. This is useful if the individuals differ in a key feature, such as age. For gene j in individual i, the observed gene expression $\boldsymbol{x}_{ij}$ is a vector of $t_i \times 1$ that represents t_i quantified measurements (Fig. 1C), rather than a scalar as in Eq. 1. We model centered $\boldsymbol{x}_{ij}$ as a product of cell type fraction (W_i^*; $t_i \times k$) and cell-type-specific expression ($\boldsymbol{\alpha}_{ij}$; $k \times 1$),

$$\boldsymbol{x}_{ij} = W_i^* \boldsymbol{\alpha}_{ij} + \boldsymbol{e}_{ij}, \tag{2}$$

where $\boldsymbol{e}_{ij}$ is the error term that captures the unexplained random noise and $\boldsymbol{e}_{ij} \sim N\left(\mathbf{0}, \sigma_e^2 I_{t_i}\right)$. To achieve robust estimation, we assume that cell-type-specific expression ($\boldsymbol{\alpha}_{ij}$) is randomly distributed as $\boldsymbol{\alpha}_{ij} \sim N\left(\mathbf{0}, \Sigma_c\right)$. We estimate the parameters through maximum likelihood via a computationally efficient EM (Expectation-Maximization) algorithm (see Methods and

Supplemental Material). Cell-type-specific expression ($\boldsymbol{\alpha}_{ij}$) is estimated using an empirical Bayes procedure. To achieve reliable results, the number of cell types (k) to be estimated is limited by the number of measures per individual; see the next section for analysis and discussion. In contrast, all genes in the genome can be efficiently deconvolved together.

Model assessment and validation

In this section, we evaluate the performance of MIND in various scenarios, including pure simulation, simulation based on real cell-type-specific expression, and analysis of the GTEx brain tissue data. We find that MIND can produce approximately unbiased parameter estimates and accurately recover the cell-type-specific expression.

Simulations based on artificial data

We first generate artificial gene expression from Eq. 2 while systematically varying the values of the true variance parameters, σ_e^2, σ_c^2, and σ_c^{ij}, which denote the error variance, and the variance and covariance of cell-type-specific expression ($\boldsymbol{\alpha}_{ij}$), respectively. We simulate 100 replicated datasets with 100 genes, 13 measurements of the same tissue, and 100 individuals. The number of cell types is set at five. For each simulation setting, we calculate the average estimates for the variance parameters estimated by MIND and the correlation between the deconvolved and true cell-type-specific expression, based on the 100 replications. As shown in Table 1, MIND produces approximately unbiased estimates of all parameters. Moreover, the correlation between the deconvolved and true cell-type-specific expression is 0.95, on average. Thus MIND can recover cell-type-specific expression.

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 σ_c^{ij}, which denote the error variance, and the variance and covariance of cell-type-specific expression, respectively. We present the average estimates of variance parameters and the correlation between the deconvolved and true cell-type-specific expression. The correlation is calculated for each of five cell types. The results are based on 100 replications.

setting	true value			parameter estimate			correlation of expression for each cell type				
	σ_e^2	σ_c^2	σ_c^{ij}	$\hat{\sigma}_e^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_c^{ij}$	type 1	type 2	type 3	type 4	type 5
A	1	1	0.5	1.00	1.01	0.50	0.946	0.945	0.947	0.948	0.948
B	2	2	1.0	1.98	2.02	1.00	0.946	0.945	0.947	0.948	0.948
C	3	3	1.5	2.96	3.03	1.50	0.947	0.945	0.947	0.948	0.948
D	4	4	2.0	3.94	4.04	2.00	0.947	0.945	0.947	0.948	0.948
E	5	5	2.5	4.93	5.05	2.50	0.947	0.945	0.947	0.948	0.948

Simulations based on mixtures of single-cell expression

Next, we conduct a simulation study on the basis of measured cell-type-specific expression, generating bulk gene expression data from measured cell-type-specific gene expression data by convolution. To do so, we leverage the scRNA-seq data from Habib et al. (2017), which profiled nuclei from seven archived brain tissue samples of the GTEx project using massively parallel single-nucleus RNA-seq. The seven GTEx tissue samples in Habib et al. (2017) are

from prefrontal cortex and hippocampus. We calculate the cell-type-specific expression by averaging across cells of each cell type for each individual and estimate cell type fractions for the GTEx brain tissue (see discussion in the section of Analysis of the GTEx brain tissue). With the estimated cell-type-specific expression and fraction, we simulate bulk data by mixing cell expression as in Eq. 2 and sequentially add more random noise to the mixture of cell expression by increasing the error variance (σ_e^2) relative to the variance of the measured cell-type-specific expression (σ_c^2). We employ the EM component within MIND and the least-squares-based approach in csSAM (Shen-Orr et al. 2010) to analyze these bulk data, treating cell type fractions as known. Note that csSAM is designed for single-measure deconvolution, and we adapt it to multi-measure deconvolution by treating multiple measures as samples. To assess the results, we calculate the correlation between the deconvolved and measured gene expression for each cell type. As compared to the results from csSAM, MIND provides consistently higher correlation for all cell types and is more robust to increasing noise (Fig. 1D).

We further assess how many measures we need to achieve reliable deconvolved expression. Following the above simulation procedure, we decrease the number of measures from 13 (the maximum number of measures of brain tissue in GTEx) to 1, and calculate the correlation between the deconvolved and measured expression. The error variance level is set to be the same as the variance of the observed cell-type-specific expression in Habib et al. (2017). We find that MIND is robust to the number of measures: the correlation is above 0.6 even when there is only a single measure (Supplemental Fig. S1). As expected, the correlation increases as the number of measures increases. For a comparison, we repeat the analysis using the least-squares-based approach in csSAM. This approach is limited by the design of least squares, requiring the number of measures (t) to be greater than or equal to the number of cell types (k). csSAM's performance improves with the number of measures, but overall does not perform as well as MIND. This highlights the advantages of assuming random cell-type-specific expression in MIND, an assumption that is particularly valuable when the number of measures is small, which is usually the case in practice.

Comparison of the deconvolved and measured cell-type-specific expression

The existence of both bulk and scRNA-seq data from GTEx brain samples (GTEx Consortium 2017; Habib et al. 2017) enables us to perform a perfect comparison between the deconvolved and measured cell-type-specific expression. We compare the deconvolved and measured gene expression for each cell type and each individual by calculating the correlation between them. The measured cell-type-specific expression is the log-transformed average value across cells of the same type from each of the GTEx donors with thousands of cells (Habib et al. 2017). The deconvolved expression is estimated by MIND. The correlation is around 0.8 for most cell types and donors (Fig. 1E) and the deconvolved and measured expression estimates are highly concordant (Supplemental Fig. S2). Moreover, we observe that known marker genes have excess expression for their corresponding cell type (Supplemental Fig. S3).

Analysis of the GTEx brain tissue

Our goal is to provide cell-type-level interpretation for gene expression data from multiple measurements from the same tissue. The GTEx program (GTEx Consortium 2015, 2017) is an ongoing project that collects both gene expression data from multiple tissue types, including brain, and genotype data from blood for hundreds of post-mortem adult donors. Here we focus

on 1671 brain tissue samples from 254 donors and 13 brain regions in the GTEx V7 data (GTEx Consortium 2017). Samples of brain tissue from different brain regions still share common cell types and thus can be deconvolved together. Because some of the data from brain regions by GTEx donor are missing, we remove subjects with less than nine collected brain tissue samples to ensure more reliable estimates, resulting in data from 105 subjects for analysis. Among these subjects, 95 also have genotype data that can be used in the cell-type-specific eQTL analysis. We apply MIND to log-transformed expression data, first calculating cell type fractions for each brain region and then estimating the individual- and cell-type-specific gene expression.

Cell type composition in 13 brain regions

To estimate cell type composition by brain region, we require information on gene expression specific to cell types. Such a reference is available from the NeuroExpresso database (Mancarci et al. 2017), which holds gene expression data from purified-cell samples from multiple mouse brains and regions. While NeuroExpresso provides data from a larger array of cell types, we restrict our analysis to basic cell types, namely astrocyte, oligodendrocyte, microglia, and GABAergic and pyramidal neurons, because GTEx's limited regional repetition also limits the reliability of estimates for a wider diversity of cell types. In our analyses, the estimated fractions for microglia are always close to zero, and thus we drop microglia from our analyses. To build the signature matrix (A) and then estimate W, we use CIBERSORT (Newman et al. 2015); see Supplemental Table S1 for approach and discussion. Results for W (Fig. 2A) are consistent with previous findings and what is known about the brain: (i) related brain regions have similar cell type composition, for example, the three basal ganglia structures, two cerebellum samples, and three cortical samples; (ii) the abundance of pyramidal neurons in cortex, hippocampus, and amygdala also matches with previous findings (Bekkers 2011); and (iii) spinal cord (cervical c-1) is estimated to consist of 91% oligodendrocytes, which agrees with the prominence of white matter tracts present at c-1 and glial cells in white matter.

Remark: While our estimates of the abundance of pyramidal neurons, for example, match previous findings, such estimates can be inconsistent with those from neuroanatomical and other direct studies of cell representation (Pelvig et al. 2008; Azevedo et al. 2009; Bartheld et al. 2016). To better understand the estimated cell type fractions, we studied the relationship between cell size and gene expression in GTEx data using techniques in Jia et al. (2017) and results from Zeisel et al. (2015). We find that the estimated cell size is highly positively correlated with gene expression (Supplemental Fig. S4), and neurons tend to have a larger cell size than non-neurons, which agrees with previous findings (Wang et al. 2018). Thus, while most deconvolution studies present their results in terms of estimated fractions of cell types, we believe these methods, including MIND, estimate the fraction of RNA molecules from each cell type instead.

Cell-type-specific expression patterns

We next examine the estimated cell-type-specific expression values, by subject, to determine if the estimates conform to expected patterns. First, it is reasonable to predict that genes/RNA markers showing specificity for certain brain regions would also show specificity to a cell type prominent in that region. This is indeed the case (Supplemental Fig. S5). For example, consider *ZP2* and *LINC00507*, the former is highly expressed in cerebellar and the latter in cortical brain tissue (Fig. 2B). By contrasting the region level expression for these genes with their estimated cell-type-specific expression (Fig. 2C and Supplemental Fig. S5), we find that *ZP2* is

expressed largely in GABAergic neurons in the cerebellum, which contains large GABAergic Purkinje cells, among GABAergic neurons. Likewise, *LINC00507* tends to be expressed solely in pyramidal cells, which make up a substantial fraction of the neuronal cells of the cortex.

A priori, and based on recent findings (Soreq et al. 2017), we would also expect cell type to be a strong predictor of gene co-expression. By contrast, because GTEx subjects were all adults at death, but not elderly, recent findings (Soreq et al. 2017) suggest that age would not be a strong predictor of gene co-expression. Thus, we asked if the estimated cell-type-specific expression clusters by cell type or by age of the subject using estimates from 128 genes with the largest variability in expression across brain regions. Based on these genes, we compute the correlation matrix for the $4n$ subject-cell-type configurations (4 cell types and $n = 105$ subjects). Hierarchical clustering of the entries in the correlation matrix reveals that cell-type is a strong predictor of co-expression, while age is not (Fig. 2D).

Nonetheless, cell-type-specific expression by age reveals interesting patterns that are not always apparent at the tissue level. For example, for *GFAP*, expression increases with age and this increase is almost entirely attributable to increased expression of both neuronal cell types (Fig. 3A). *PTCHD4* and *RHOV* show similarly intriguing patterns by cell type (Supplemental Fig. S6): no age trend is present at the tissue level, but for *PTCHD4*, GABAergic expression is decreasing while astrocyte and oligodendrocyte expression are increasing and, for *RHOV*, the opposite pattern holds. Overall, 18% of genes show age trends at the region level or cell-type level, based on correlation test of gene expression and age for each brain region or cell type, with the false discovery rate (FDR) (Storey and Tibshirani 2003) controlled at 0.05: 7% show age trends in at least one brain region and at least one cell type; 7% show age trends in at least one brain region, but not in any cell type; and 4% show age trends in at least one cell type, but not in any brain region.

Cell-type-specific eQTL analysis based on deconvolved expression

Because MIND yields individual- and cell-type-specific gene expression, we can identify eQTLs for each cell type. To do so, the data are analyzed using MatrixEQTL (Shabalin 2012), with FDR controlled at 0.05 for each cell type. We then compare the MIND-identified eQTLs with single-tissue eQTLs identified by the GTEx project (GTEx Consortium 2017). As expected, for eQTLs identified by GTEx to be brain-region specific, there is a strong enrichment for eQTLs that are specific to a cell type prominent in that region (Fig. 3B). This pattern corresponds to the patterns of cell type composition for each brain region (Fig. 2A). Next, for eQTLs that appear in one to four cell types, we calculate their probability of being identified in each brain region and over tissue types. Based on the findings of McKenzie et al. (2014), we expect that when an eQTL is jointly identified in more brain cell types, it is more likely to be detected across a variety of tissues and especially across brain regions; indeed, this is what we find (Fig. 3C and Supplemental Fig. S7). Finally, 52% of eQTLs that are identified in one or more brain cell types are not identified by any brain region, which suggests MIND's results can identify novel eQTLs. Moreover, some eQTLs are shared by all four cell types, while others are specific to certain cell types, especially for the neuron cells (Fig. 3D), which implies that eQTL analysis based on MIND's analysis of bulk gene expression data can shed light on gene expression regulation of cells.

Analysis of the BrainSpan data yields insights into autism spectrum disorder

The BrainSpan atlas of the developing human brain quantified gene expression from multiple brain regions and subjects from 8 post-conceptional weeks (pcw) to 40 years of age (Miller et al. 2014). This rich dataset is ideal for analysis of spatio-temporal patterns of transcription of human brain (Kang et al. 2011). Here we make use of the exon microarray data with normalized expression values, which include 492 tissue samples from 26 brain regions and 35 subjects. Similar to the GTEx data, we restrict analysis to the 33 subjects with more than nine brain tissue samples.

To obtain the signature matrix (A), we leverage scRNA-seq data from human adult and fetal cortical samples obtained from Darmanis et al. (2015). There are 466 cells that were clustered into astrocyte, oligodendrocyte, OPC (oligodendrocyte progenitor cells), microglia, endothelial, mature neuron, replicating fetal neuron and quiescent fetal neuron (i.e., immature neuron). By preliminary analyses, we determined that three cell types have small estimated cell type fractions, namely microglia, endothelial, and replicating fetal neuron, and these were dropped from our analysis. Then W is estimated using CIBERSORT (Newman et al. 2015) for five cell types: astrocytes, OPC, oligodendrocytes, immature neurons, and mature neurons. Consistent with expectation, the fraction of immature neurons decreases and that of mature neurons increases with age (Fig. 4A) and likewise oligodendrocytes replace OPC, the latter consistent with the myelination process. As the brain develops, the overall neuronal fraction (immature neuron plus mature neuron) decreases relative to other cell types, again consistent with what is known about brain maturation (Tau and Peterson 2010).

With A and W in hand, MIND then yields gene expression estimates for each individual and cell type. Because the contributors have varying ages, results from the MIND algorithm represent a developmental expression profile for each cell type. These profiles should prove useful for the study of typical and atypical neurodevelopment. The neurodevelopment of individuals diagnosed with ASD, a disorder with onset prior to the age of three, is now thought to diverge from typical development during the fetal period (Willsey et al. 2013; Parikshak et al. 2013). Thus, MIND gene expression profiles could yield insights into the etiology of ASD.

These profiles allow one to construct a co-expression network for each cell type. To do so, we calculate the correlation of expression for each pair of genes, over individuals, weighted by the average cell type fraction for each individual; genes are regarded as connected in an adjacency matrix if the absolute correlation passes a threshold, here taken to be 0.9. To make the analysis relevant to ASD, we evaluate a set of 65 genes previously implicated in risk for ASD on the basis of analysis of rare variation (He et al. 2013) by the Autism Sequencing Consortium (Sanders et al. 2015). Fifteen of the 65 genes are connected in the immature neuron network (Fig. 4B). Remarkably all of these genes play a regulatory role, according to Gene Ontology annotation for biological processes (Kuleshov et al. 2016).

There are 16 genes not detected as ASD genes by Sanders et al. (2015) but highly correlated to more than six ASD genes in the network of immature neurons (Fig. 4C). We shall call them ASD-correlated genes. The products of these ASD-correlated genes also tend to play regulatory or developmental roles, including acetyltransferase activity (*EPC1* (Searle and Pillus 2018), *KAT6A*, *KAT6B* (Huang et al. 2016)), transcriptional regulation in some form (*AFF4*, *CNOT2*, *GATAD2B*, *PCF11*, *SUPT20H*, *TUG1* (Luo et al. 2012; Rambout et al. 2016; Bornelov et al. 2018; Volanakis et al. 2017; Watanabe and Kokubo 2017; Baptista et al. 2017; Guo et al. 2018)) and DNA replication (*HNRNPUL1* (Ideue et al. 2012)). Intriguingly, the encoded protein of *FUBP* could be a key regulator of cell differentiation (Quinn 2017; Zhou

et al. 2016), specifically the transition from progenitor cells to neurons, and it is possible that all the 15 connected ASD genes in Figure 4B and these 16 ASD-correlated genes play a part in this transition. Of the 16 ASD-correlated genes, 13 have pLI = 1 (the probability of being Loss of Function intolerant) (Lek et al. 2016); exceptions are *UBXN7* (pLI = 0.99), *SUPT20H* (pLI = 0) and *TUG1* (pLI undetermined, it is a long non-coding RNA). According to DECIPHER 9.23 (https://decipher.sanger.ac.uk/), four genes have been previously implicated in neurodevelopmental disorders (*QRICH1*, *KAT6A*, *KAT6B*, and *GATAD2B*), while two others lie in syndromic regions defined by structural variation associated with developmental disorders, specifically *CNOT2* (one of 3 genes in the 12q15 deletion, Alesi et al. (2017)) and *UBXN7* (3q29 microdeletion/microduplication).

On the basis of the cell-type-specific correlations, we then count the number of connections for each gene and test if there is a difference between the 65 ASD genes and the 17,217 other measured genes (Fig. 4D). ASD genes are more connected than non-ASD genes in immature neurons (Mann-Whitney U test p-value = 2.2×10^{-8}). When we perform the same cell-type-specific network analysis using the scRNA-seq data from Darmanis et al. (2015), however, we do not observe similar findings (Supplemental Fig. S8). Due to low call rate and dropout events, it is possible that scRNA-seq data are too noisy to calculate accurate correlations. Alternatively, because the scRNA-seq data are derived from only a few individuals, cells of the same type may lack sufficient variability to reveal the correlation pattern of the genes. The network derived from MIND results is built across a large number of individuals, spanning a variety of ages, facilitating calculation of the gene-gene correlations. This comparison of the estimated correlation structures, derived directly from scRNA-seq expression versus cell-type-specific expression imputed from bulk data, reveals a surprising benefit of the MIND algorithm.

While gene co-expression can be challenging from brain scRNA-seq data, simple presence/absence of expression is more straightforward. Therefore, we evaluate the enrichment of ASD genes in immature neurons using data from Darmanis et al. (2015). Because a large fraction of genes have no recorded expression for many cells, we say a gene is "expressed" in a cell type if it is expressed in at least 15% of the cells of that type. We restrict our analysis to the 11,215 genes that are expressed in one or more cell types. To determine if the genes expressed in a particular cell type are enriched for ASD risk genes, we tabulate whether the gene is expressed and whether it is associated with ASD risk. As shown in Figure 5, the immature neuron is most enriched for ASD genes (odds ratio = 7.9; Fisher's exact p-value = 2.8×10^{-8}).

Discussion

Most tissues are comprised of many cells, which are of various cell types. Thus, when gene expression is measured from tissue, the measurements are a convolution over cells and cell types, which could confound our understanding of determinants of differential gene expression. For example, consider a case-control study in which some genes show differential expression between case and control subjects. The differences could be due to disease status, to variation in cell composition of the tissue or both. For this reason, among others, direct measurement of gene expression in specific cells and cell types has become commonplace. Nonetheless, direct measurements are not without drawbacks and this approach ignores rich sources of data for gene expression from tissue. Instead, we develop an algorithm, MIND, to obtain gene expression by cell type and subject, even though gene expression is measured from tissue. The MIND deconvolution algorithm borrows information across multiple measurements of gene

expression from tissue from the same subject, such as regions of the human brain, to estimate gene expression from cell types. Most deconvolution algorithms solely estimate the fractional representation of cell types within a tissue, or some function thereof. For the human brain, for example, the methods typically seek to determine if the representation of cell types within a region is similar across subjects, which can be critical for interpreting case-control contrasts of gene expression (Fromer et al. 2016). In addition to MIND, the csSAM algorithm is an exception (Shen-Orr et al. 2010) to this rule. csSAM estimates average expression of cell types over individuals; by contrast, MIND estimates expression of the cell types for each individual.

In summary, variation in bulk gene expression data arises due to differences in the proportion of cells of each type and differences in gene expression per cell type. The vast majority of single-measure deconvolution methods use either cell-type-specific marker genes or cell-type-specific expression profiles to deconvolve cell-type proportion in bulk expression data. The MIND algorithm uses both cell-type proportion and variation in gene expression per cell type to achieve more informative results. As a result, MIND is unique in its estimate of cell-type-specific gene expression by individual.

In Results, we highlighted a few of the myriad uses for MIND's results. For instance, imagine the samples of subjects represent a developmental series, such as the BrainSpan atlas of the developing human brain. BrainSpan sampled multiple regions of the brain, per subject, which MIND exploits for its estimates. Because the estimates are cell- and individual-specific, they represent the cell-specific change in gene expression over development. Remarkably, one can also estimate cell-specific co-expression networks from the results, a feature we use to determine that immature neurons show the greatest preponderance of co-expressed genes affecting risk for ASD. By exploiting the repeated measurements of gene expression in brain from the GTEx sample, we identify eQTLs specific to certain cell types, as well identifying other eQTLs that are ubiquitous. Such eQTL results could open new avenues for understanding the etiology of brain disorders, for example, by reinterpreting GWAS (genome-wide association study) results in light of their eQTL effects within cells.

There are limitations to the current version of the MIND algorithm. The method relies on reference samples that provide genes whose expression is largely specific to cell type, so-called marker genes. Identifying which reference samples are appropriate can be challenging. There are resources to address this challenge, such as NeuroExpresso and scRNA-seq databases (Mancarci et al. 2017; Darmanis et al. 2015), and it will be beneficial to continue to develop and integrate such reference samples. A different challenge is presented when there are a large number of cell types in the tissue. Reliably estimating expression by cell type and individual will require a large number of repeated measures per subject, something most resources do not have at this time. One option is to impute the expression of unmeasured tissue (Wang et al. 2016) before the deconvolution, although this approach remains to be explored. Furthermore, as it features bulk gene expression, MIND is limited to estimating the average gene expression across cells of the same type within an individual, which ignores the diversity of expression within single cells. The true diversity of expression in single cells cannot be recaptured with this approach. Finally, in its current implementation, MIND does not account for the gene-gene correlation; this information could be informative and more computationally intensive versions of MIND will be explored in the future.

There are some surprising advantages to the MIND algorithm. For instance, one might imagine that scRNA-seq methods can be used to obtain many of the features captured by MIND, for example, gene co-expression networks for cell types. When we tried to construct

such networks from scRNA-seq results, however, they show very little coherent structure. This could be, in part, due to the limited number of subjects from which the cells were obtained. By contrast, results from MIND yield coherent and interpretable networks, which show relevance to risk for ASD. Perhaps the greater coherence arises from the larger number of subjects available for MIND's analysis of gene expression data. Intriguingly, it is also possible that MIND smooths out the noise inherent in scRNA-seq data, including noise such as that from dropout. It will be interesting to determine, molecularly, how the MIND algorithm is able to obtain coherent networks while scRNA-seq data are not and to document under what conditions we would expect these different approaches to converge to the same solution.

Methods

A new deconvolution method for multiple measurements of tissue expression

Given the availability of multiple (t) measures per individual (e.g., brain regions), the expression data expand from a matrix (X; $p \times n$; for p genes and n individuals) to a three-dimensional array ($p \times n \times t$). With a new dimension of measurement, we aim to estimate individual-level expression through deconvolution. Let $X_1, X_2, \ldots, X_t$ denote the expression matrix for measurement types $1, 2, \ldots, t$. To see the expanded opportunities for deconvolution with multi-measurements, we vectorize the expression matrix for each measurement type and construct the expression matrix as $X^* = (\mathrm{vec}\,(X_1)\,, \mathrm{vec}\,(X_2)\,, \ldots, \mathrm{vec}\,(X_t))$, where the dimension is $np \times t$. Within this setting, we can extend the single-measure deconvolution in Eq. 1 to multi-measure deconvolution

$$\underset{(pn\times t)}{X^*} = \underset{(pn\times k)}{A^*}\,\underset{(k\times t)}{W^*} + \underset{(pn\times t)}{E^*}\,, \tag{3}$$

where A^* is cell-type-specific gene expression for each individual and k cell types, W^* is the cell-type fraction for each measurement type, and E^* is the error term. If W^* is pre-estimated via single-measure deconvolution, then we can estimate A^* directly. This representation reveals why multi-measures facilitate estimation of the individual- and cell-type-specific expression (A^*), but there are also drawbacks to this direct extrapolation of traditional deconvolution. First, the number of measurement types (t) usually varies across individuals, which leads to an unbalanced sample. Second, the approach requires the cell type fraction to be fixed across individuals, which is an unreasonable assumption when the individuals are heterogeneous.

To allow for individual-specific fractions and unbalanced number of measurement types per individual, we rewrite the deconvolution model (Eq. 3) in a more flexible vector form. For the ith individual, we stack its multi-measure expression gene by gene and model it as

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{x}_{i1} \\ \boldsymbol{x}_{i2} \\ \vdots \\ \boldsymbol{x}_{ip} \end{pmatrix} &= \begin{pmatrix} W_i^* & & & \\ & W_i^* & & \\ & & \ddots & \\ & & & W_i^* \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{i1} \\ \boldsymbol{\alpha}_{i2} \\ \vdots \\ \boldsymbol{\alpha}_{ip} \end{pmatrix} + \boldsymbol{e}_i \\
\boldsymbol{x}_i &= (I_p \otimes W_i^*)\,\boldsymbol{\alpha}_i + \boldsymbol{e}_i = W_i \boldsymbol{\alpha}_i + \boldsymbol{e}_i,
\end{aligned} \tag{4}$$

where $\boldsymbol{x}_{ij}$ ($i = 1, \cdots, n$; $j = 1, \cdots, p$) is a $t_i \times 1$ vector representing the expression of the jth gene in the t_i quantified measurement of the ith individual, W_i^* is a $t_i \times k$ matrix denoting the pre-estimated cell type fractions for each measurement on the ith individual, $\boldsymbol{\alpha}_{ij}$ is a $k \times 1$ vector

representing cell-type-specific gene expression, and $\boldsymbol{e}_i$ is the error term with $\text{var}(\boldsymbol{e}_i) = \sigma_e^2 I_{pt_i}$. With centered expression ($\boldsymbol{x}_i$), we assume $\boldsymbol{\alpha}_i$ to be a random vector with mean $\mathbf{0}$ and covariance matrix

$$\text{var}(\boldsymbol{\alpha}_i) = \Sigma_\alpha = \Sigma_g \otimes \Sigma_c, \tag{5}$$

where Σ_g and Σ_c denote the covariance matrix for genes and cell types, respectively. The Kronecker product is introduced to simplify the covariance structure with an assumption that $\boldsymbol{\alpha}_i$ follows a matrix-variate distribution. The parameters can be estimated through maximum likelihood via an EM algorithm and the realization of $\boldsymbol{\alpha}_i$ can be obtained from empirical Bayes estimation.

There are many parameters in Eq. 4, mainly in the covariance matrix for all genes (Σ_g in Eq. 5). To improve the estimation efficiency, we need to simplify the form of the covariance matrix. As shown in the algorithm in the Supplemental Material, the main computational burden lies in the matrix computations such as matrix inverse. To reduce the computational burden, we can group genes into clusters of hundreds of genes first, e.g., using WGCNA (Langfelder and Horvath 2008), and then fit the deconvolution model for each cluster of genes. This largely reduces the number of parameters by assuming independence across gene clusters. We can further impose a sparse partial correlation structure among the genes within each cluster via the graphical lasso penalty (Friedman et al. 2008; Danaher et al. 2014). Details are provided in the Supplemental Material. However, even with these simplifications, the algorithm is still computationally challenging for tens of thousands of genes because of the estimation of the gene-gene correlation.

A simplified and computationally efficient multi-measure deconvolution model

We propose a simplified model that allows all the genes in the genome to be deconvolved simultaneously in a computationally efficient way. This method is utilized for simulations and data analysis presented in this paper. For the jth gene of the ith individual, the deconvolution model can be written as

$$\begin{aligned} \boldsymbol{x}_{ij} &= W_i^* \boldsymbol{\alpha}_{ij} + \boldsymbol{e}_{ij}, \\ \boldsymbol{\alpha}_{ij} &\sim N\left(\mathbf{0}, \Sigma_c\right), \\ \boldsymbol{e}_{ij} &\sim N\left(\mathbf{0}, \sigma_e^2 I_{t_i}\right). \end{aligned} \tag{6}$$

where $\boldsymbol{x}_{ij}$ ($i = 1, \cdots, n$; $j = 1, \cdots, p$) is a $t_i \times 1$ vector representing the centered expression of jth gene in the t_i quantified measurement of the ith individual, W_i^* is a $t_i \times k$ matrix denoting the pre-estimated cell type fractions for each measure in k cell types, $\boldsymbol{\alpha}_{ij}$ is a $k \times 1$ vector representing cell-type-specific gene expression, and $\boldsymbol{e}_{ij}$ is the error term. Note that as compared to single-measure deconvolution (Eq. 1), we allow both cell-type-specific expression and cell type fraction to be individual-specific.

The parameters, the covariance matrix for random effects (Σ_c) and the error variance (σ_e^2), can be estimated through maximum likelihood via EM algorithm. The realization of the random effects $\boldsymbol{\alpha}_{ij}$ can be obtained from empirical Bayes estimation. The computational details for each step are described in the Supplemental Material. The multi-measure deconvolution algorithm is described in Algorithm 1 and visualized as a flowchart (Fig. 1B).

Algorithm 1 The multi-measure deconvolution (MIND) algorithm

1. Build a marker gene signature matrix (A) following a standard method, such as CIBERSORT (Newman et al. 2015), and using gene expression reference samples of different cell types, for example, from the NeuroExpresso database (Mancarci et al. 2017).
2. Calculate cell type fractions for each individual in each measurement type (W) by deconvolving the expression of multiple measures (X) using a standard deconvolution method such as CIBERSORT.
3. Estimate individual-level cell-type-specific gene expression by multi-measure deconvolution of bulk measurements (Eq. 6), with the fraction matrix estimated in Step 2.
4. Output a three-dimensional array ($p \times n \times k$), i.e., the deconvolved expression levels for each gene of each individual in each cell type.

Usage of single cell RNA-seq data of brain

In this study, we use three public brain scRNA-seq datasets. The scRNA-seq data in Zeisel et al. (2015) provide spike-in information and we leverage it to estimate cell size (Jia et al. 2017) for neurons and non-neurons. This helps enhance the interpretation of the cell type composition in the deconvolution. Habib et al. (2017) quantified scRNA-seq data of seven brain tissue samples from five GTEx donors. We employ it to construct the cell-type-specific expression in the simulation and assess our deconvolved individual- and cell-type-specific expression. Lastly, the scRNA-seq data in Darmanis et al. (2015) include fetal cells and we utilize them as reference samples to deconvolve the BrainSpan data of the developing human brain.

Software availability

We implement the method discussed in this paper as an R package `MIND` to deconvolve the expression of multiple measurements of tissues. The package is publicly hosted on GitHub https://github.com/randel/MIND.

Acknowledgments

We are grateful for the insightful comments from Professors Lin Chen, Stephan Sanders, and Haiyuan Yu, who read a previous version of the manuscript. This work was supported, in part, by National Institute of Mental Health (NIMH) grants R37MH057881 and MH109900 and by Simons Foundation Autism Research Initiative (SFARI) grants SF402281 and SF367561. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods

development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v7.p2. This BrainSpan Atlas of the Developing Human Brain was supported by RC2MH089921, RC2MH090047 and RC2MH089929 from the National Institute of Mental Health.

References

Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, and Clark HF. 2009. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one* **4**: e6098.

Alesi V, Loddo S, Grispo M, Riccio S, Montella AC, Dallapiccola B, Ulgheri L, and Novelli A. 2017. Reassessment of the 12q15 deletion syndrome critical region. *Eur J Med Genet* **60**: 220–223.

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, Filho WJ, Lent R, and Herculano-Houzel S. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology* **513**: 532–541.

Baptista T, Grunberg S, Minoungou N, Koster MJE, Timmers HTM, Hahn S, Devys D, and Tora L. 2017. SAGA Is a General Cofactor for RNA Polymerase II Transcription. *Molecular Cell* **68**: 130–143.

Bartheld CS, Bahney J, and Herculano-Houzel S. 2016. The search for true numbers of neurons and glial cells in the human brain: a review of 150 years of cell counting. *Journal of Comparative Neurology* **524**: 3865–3895.

Bekkers JM. 2011. Pyramidal neurons. *Current Biology* **21**: R975.

Bornelov S, Reynolds N, Xenophontos M, Gharbi S, Johnstone E, Floyd R, Ralser M, Signolet J, Loos R, Dietmann S, et al.. 2018. The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression. *Molecular Cell* **71**: 56–72.

Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, Lewitus E, Sykes A, Hevers W, Lancaster M, et al.. 2015. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences* **112**: 15672–15677.

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.

Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, Hoffman GE, Perumal TM, Girdhar K, Jiang Y, et al.. 2018. Landscape of conditional eqtl in dorsolateral prefrontal cortex and co-localization with schizophrenia gwas. *The American Journal of Human Genetics* **102**: 1169–1184.

Friedman J, Hastie T, and Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**: 432–441.

Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, Shah HR, et al.. 2016. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience* **19**: 1442.

Gaujoux R and Seoighe C. 2013. Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**: 2211–2212.

GTEx Consortium. 2015. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science* **348**: 648–660.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204.

Guo Y, Chen X, Xing R, Wang M, Zhu X, and Guo W. 2018. Interplay between FMRP and lncRNA TUG1 regulates axonal development through mediating SnoN-Ccd1 pathway. *Human Molecular Genetics* **27**: 475–485.

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.

He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, et al.. 2013. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics* **9**: e1003671.

Huang F, Abmayr SM, and Workman JL. 2016. Regulation of KAT6 Acetyltransferases and Their Roles in Cell Cycle Progression, Stem Cell Maintenance, and Human Disease. *Molecular and Cellular Biology* **36**: 1900–1907.

Ideue T, Adachi S, Naganuma T, Tanigawa A, Natsume T, and Hirose T. 2012. U7 small nuclear ribonucleoprotein represses histone gene transcription in cell cycle-arrested cells. *Proceedings of the National Academy of Sciences* **109**: 5693–5698.

Jia C, Hu Y, Kelly D, Kim J, Li M, and Zhang NR. 2017. Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic acids research* **45**: 10978–10988.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al.. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478**: 483.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al.. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**: W90–W97.

Langfelder P and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* p. 559.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al.. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al.. 2016. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology* **17**: 174.

Luo Z, Lin C, Guest E, Garrett AS, Mohaghegh N, Swanson S, Marshall S, Florens L, Washburn MP, and Shilatifard A. 2012. The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output. *Molecular and Cellular Biology* **32**: 2608–2617.

Mancarci BO, Toker L, Tripathy SJ, Li B, Rocco B, Sibille E, and Pavlidis P. 2017. Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. *eneuro* pp. ENEURO.0212–17.2017.

McKenzie M, Henders AK, Caracella A, Wray NR, and Powell JE. 2014. Overlap of expression quantitative trait loci (eqtl) in human brain and blood. *BMC medical genomics* **7**: 31.

Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al.. 2014. Transcriptional landscape of the prenatal human brain. *Nature* **508**: 199.

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, and Alizadeh AA. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**: 453–457.

Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, and Geschwind DH. 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**: 1008–1021.

Pelvig D, Pakkenberg H, Stark A, and Pakkenberg B. 2008. Neocortical glial cell numbers in human brains. *Neurobiology of Aging* **29**: 1754–1762.

Quinn LM. 2017. FUBP/KH domain proteins in transcription: Back to the future. *Transcription* **8**: 185–192.

Rambout X, Detiffe C, Bruyr J, Mariavelle E, Cherkaoui M, Brohee S, Demoitie P, Lebrun M, Soin R, Lesage B, et al.. 2016. The transcription factor ERG recruits CCR4-NOT to control mRNA decay and mitotic progression. *Nature Structural & Molecular Biology* **23**: 663–672.

Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, et al.. 2015. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**: 1215–1233.

Searle NE and Pillus L. 2018. Critical genomic regulation mediated by Enhancer of Polycomb. *Current Genetics* **64**: 147–154.

Shabalin AA. 2012. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics* **28**: 1353–1358.

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.

Soreq L, Rose J, Soreq E, Hardy J, Trabzuni D, Cookson MR, Smith C, Ryten M, Patani R, Ule J, et al.. 2017. Major shifts in glial regional identity are a transcriptional hallmark of human brain aging. *Cell reports* **18**: 557–570.

Storey JD and Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Tau GZ and Peterson BS. 2010. Normal development of brain circuits. *Neuropsychopharmacology* **35**: 147.

Volanakis A, Kamieniarz-Gdula K, Schlackow M, and Proudfoot NJ. 2017. WNK1 kinase and the termination factor PCF11 connect nuclear mRNA export with transcription. *Genes Dev.* **31**: 2175–2185.

Wang J, Gamazon ER, Pierce BL, Stranger BE, Im HK, Gibbons RD, Cox NJ, Nicolae DL, and Chen LS. 2016. Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics* **98**: 697–708.

Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, Raj A, Li M, and Zhang NR. 2018. Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences* **115**: E6437–E6446.

Watanabe K and Kokubo T. 2017. SAGA mediates transcription from the TATA-like element independently of Taf1p/TFIID but dependent on core promoter structures in Saccharomyces cerevisiae. *PLoS ONE* **12**: e0188435.

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, Reilly SK, Lin L, Fertuzinhos S, Miller JA, et al.. 2013. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**: 997–1007.

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.

Zhou W, Chung YJ, Parrilla Castellar ER, Zheng Y, Chung HJ, Bandle R, Liu J, Tessarollo L, Batchelor E, Aplan PD, et al.. 2016. Far Upstream Element Binding Protein Plays a Crucial Role in Embryonic Development, Hematopoiesis, and Stabilizing Myc Expression Levels. *American Journal of Pathology* **186**: 701–715.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al.. 2016. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics* **48**: 481.

Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, et al.. 2018. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications* **9**: 224.

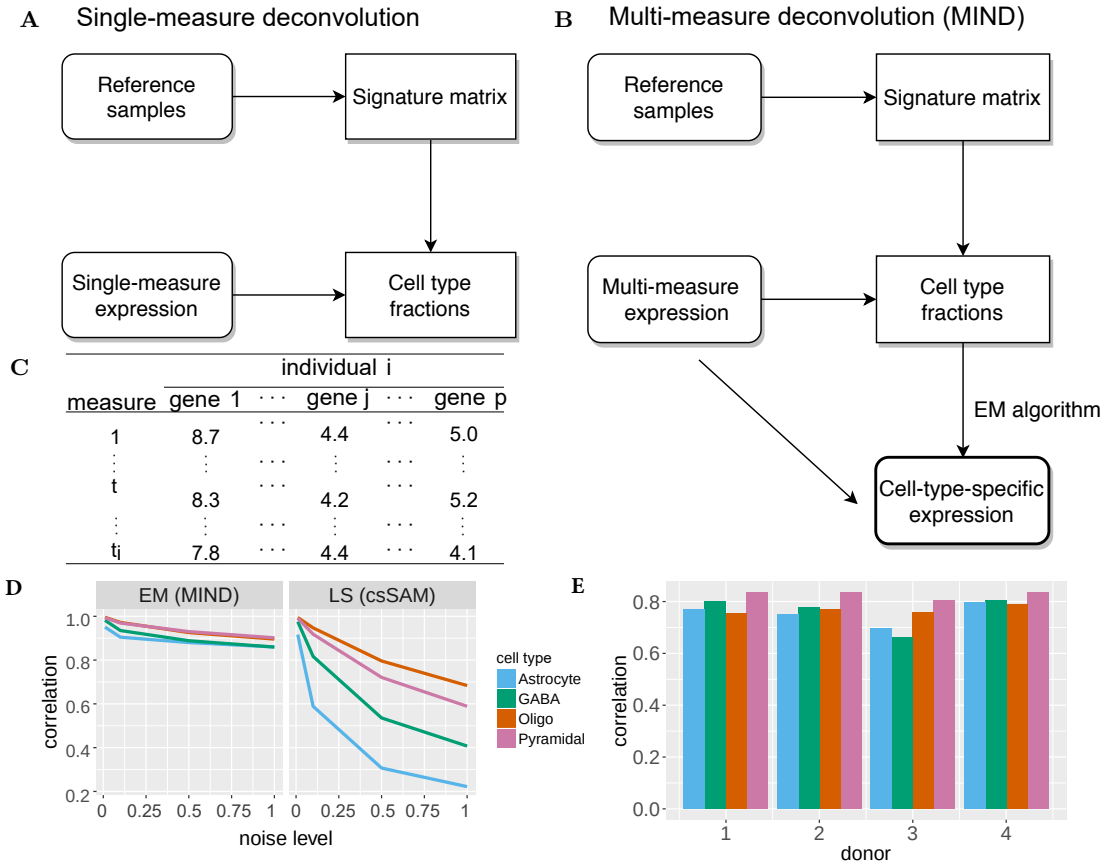

Figure 1. The flowchart and performance of the proposed multi-measure deconvolution algorithm. (A) Flowchart of traditional single-measure deconvolution algorithms. (B) Flowchart of the multi-measure deconvolution algorithm (MIND), which estimates cell-type-specific expression and fractions at the individual level. (C) The data structure of multi-measure expression for individual i, who has observed tissue expression of t_i measures for p genes. (D-E) The correlation between the measured and deconvolved expression for each cell type in simulation (D) and data analysis (E). (D) We simulate cell mixture data using the measured cell-type-specific expression and the estimated cell type fractions from the GTEx data, with increasing noise levels, and we compare our proposed EM component of the MIND algorithm with the least squares (LS) based method of csSAM (Shen-Orr et al. 2010). (E) In the analysis of the GTEx data, we calculate the correlation for each of the four GTEx donors with thousands of cells. The measured cell-type-specific expression is the average value across cells from the single-cell data of GTEx donors (Habib et al. 2017), in log scale.

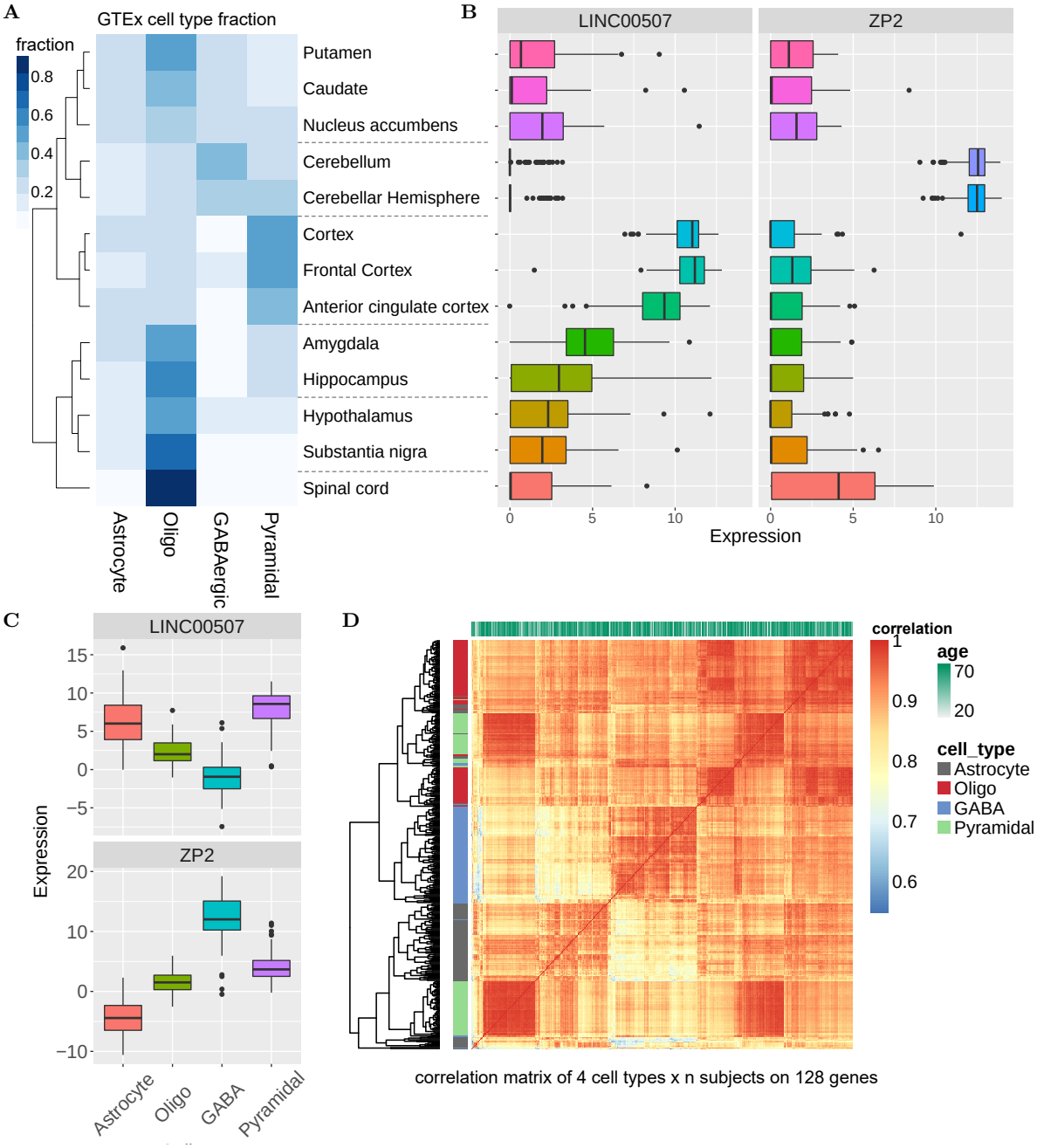

Figure 2. The analysis of the GTEx brain data (I). (A) The average estimated cell type fractions in each GTEx brain region. (B) The mapping of variable expression across brain regions onto cell-type-specific expression. Here we show the boxplots of tissue-level expression over individuals for the two genes/RNA markers of cortex and cerebellum. The full results for all brain regions are presented in Supplemental Figure S5. (C) The deconvolved expression for each cell type for the same two genes/RNA markers of cortex and cerebellum. (D) The clustering of cell-type-specific expression by cell type and age. Here we visualize a $4n \times 4n$ correlation matrix for 4 cell types and $n = 105$ subjects, based on the expression of 128 genes that have the largest variability across brain regions.

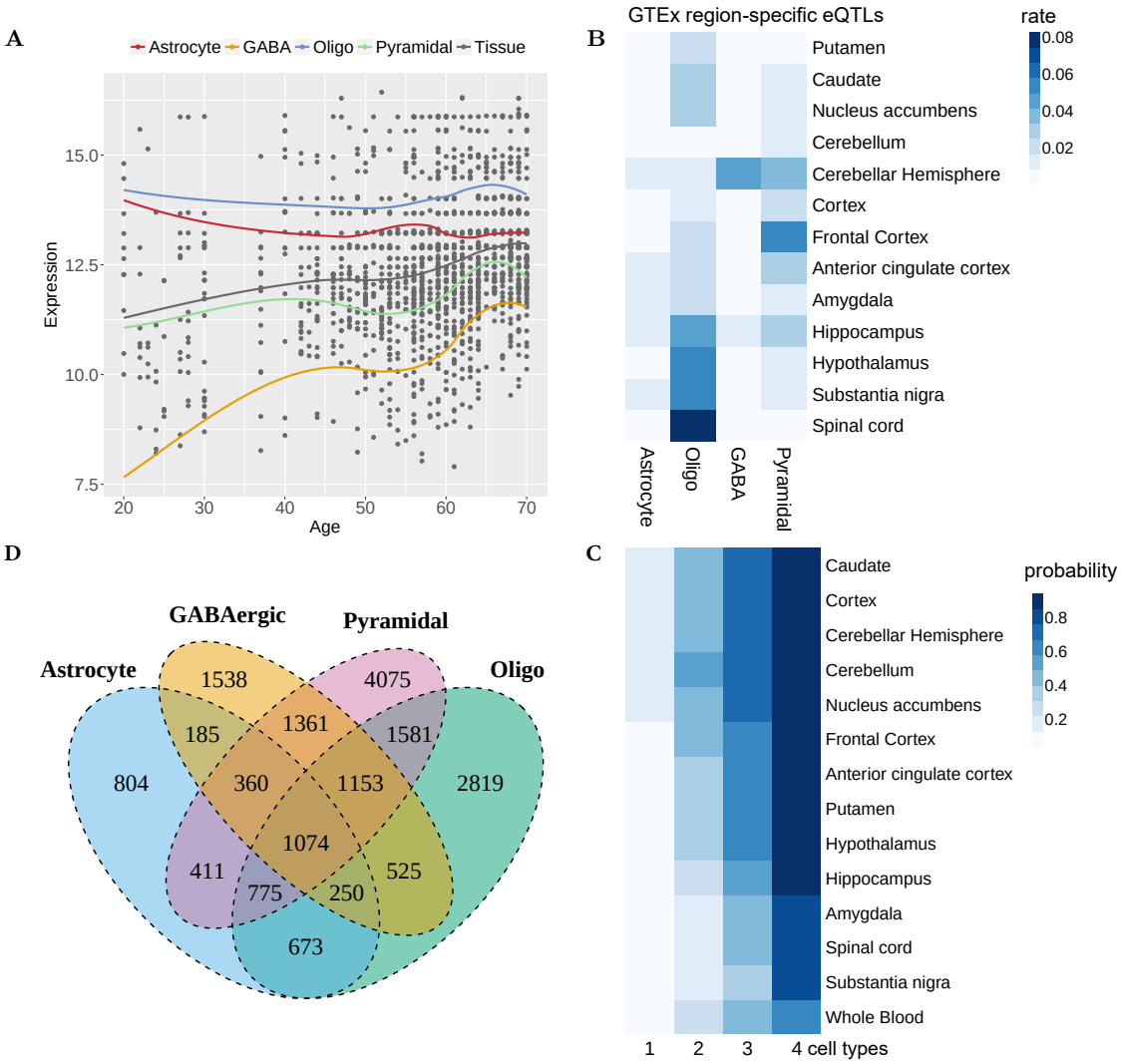

Figure 3. The analysis of the GTEx brain data (II). (A) The age trends for tissue-level expression and cell-type-specific expression for gene *GFAP*. (B) The rate of mapping region-specific eQTLs to each cell type. (C) The overlap between eQTLs appearing in multiple cell types and those in each tissue type. For eQTLs that appear in one, two, three, and four cell types, respectively, we calculate the probability of being identified in each tissue type. We show brain regions and whole blood here and all tissues in Supplemental Figure S7. (D) The overlap between cell-type-specific eQTLs.

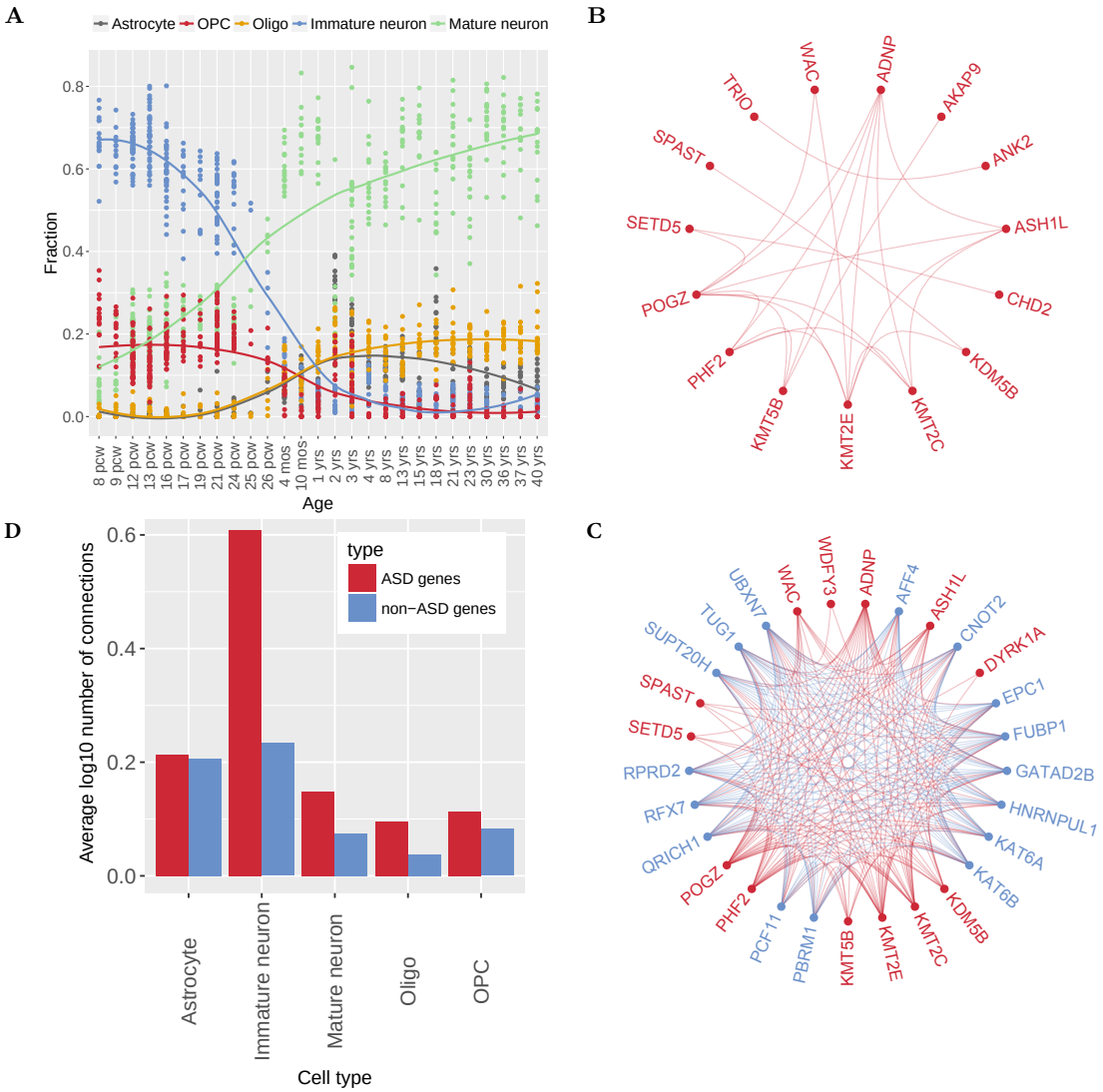

Figure 4. The analysis of the BrainSpan atlas of the developing human brain. (A) The cell type compositions across the lifespan in human brains of BrainSpan data. The curves denote the smooth lines of estimated fractions (represented by dots). Microglia, endothelial cells, and fetal replicating neurons have deconvolved cell fractions close to zero and thus are not shown. (B) The co-expression network of 15 out of 65 ASD genes in the immature neuron. A connection between genes is indicated if the absolute pairwise correlation of expression is greater than 0.9. Genes with no connections and the three genes not quantified in the BrainSpan data are not depicted here. (C) The network, in immature neuron, of 16 ASD-correlated genes that are not detected as ASD genes by Sanders et al. (2015) but are connected to more than six ASD genes. Here we only show the 13 ASD genes that are connected to those 16 ASD-correlated genes. The ASD-correlated genes are colored in blue and ASD-genes are in red. The interactive version of this figure is available at http://rpubs.com/randel/ASDnetwork. (D) The average number of connected genes for ASD genes and non-ASD genes in different cell types (in log10 scale) based on our deconvolved BrainSpan data using MIND.

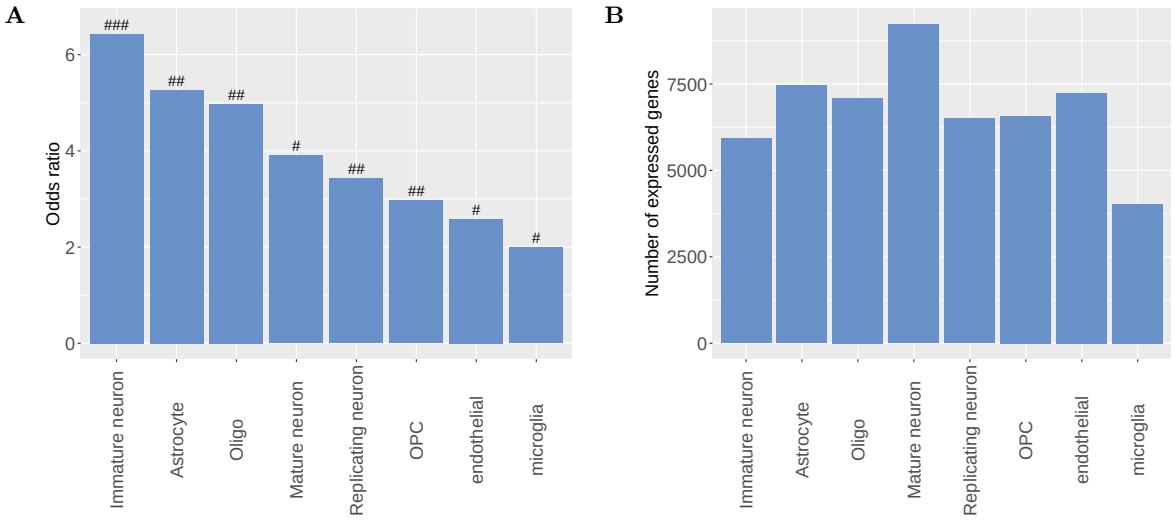

Figure 5. The enrichment analysis of ASD genes expressed in the scRNA-seq data from Darmanis et al. (2015). We focus on 11,215 genes that are expressed in at least 15% cells of one or more cell types. (A) The OR (odds ratio) assessing the association between being expressed and ASD genes. We test if 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.