Using multiple measurements of tissue to estimate 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

Abstract

Quantification of gene expression in cells or tissue can inform on etiology of disease. Complementing these approaches, we propose to estimate subject- and cell-type-specific (CTS) gene expression from tissue using an empirical Bayes method that borrows information across multiple measurements of the same tissue per subject. Analyzing multiple brain regions from the Genotype-Tissue Expression project (GTEx) reveals a subset of expression quantitative trait loci specific to neurons, others specific to astrocytes, and others active across all cell types. In another example, CTS expression of the BrainSpan atlas, which profiles expression patterns of the developing human brain, demonstrates potential insights into processes associated with neurodevelopmental disorders. Our analyses reveal clear CTS co-expression networks that, when combined with genetic findings in autism spectrum disorder (ASD), identify a cluster of co-expressed ASD-associated genes and implicate immature neurons in ASD risk.

Introduction

Altered gene expression is one mechanism by which genetic variation confers risk for complex disease. Thus, many studies have quantified bulk gene expression from tissue, thereby assessing expression averaged over the individual cells comprising the tissue. Recently, using single-cell RNA sequencing (scRNA-seq)[1,2,3], studies have quantified gene expression at the level of cells and cell types; such data could be especially informative for brain tissue, which harbors myriad cell types whose functions are not fully resolved. Drawbacks to scRNA-seq data include its noisy nature and the challenge of characterizing such cells from many subjects, which limits its potential for genetic analyses. Alternatively, there are established resources, such as BrainSpan[4,5] and GTEx[6], among others, that have collected bulk transcriptome data from many subjects and multiple brain regions. Here we present a method, MIND for Multi-measure INdividual Deconvolution (**Fig. 1**), to exploit such resources to learn about subject-level and CTS gene expression. For each subject and gene, MIND's CTS estimate represents the average expression of the gene for fundamental cell types, such as neurons, astrocytes and oligodendrocytes in brain.

Two ideas are key for obtaining CTS gene expression from tissue. First, because a tissue sample's bulk transcriptome is a convolution of gene expression from cells belonging to various cell types, deconvolution methods[7,8,9,10] can estimate the fraction of each cell type within this tissue. Most methods deconvolve a single tissue sample per subject and require prior information, specifically sets of genes that are expressed in certain cell types (marker genes), the collection of which we call the signature matrix. The second key idea is that multiple transcriptomes from the same subject, but different brain regions, share common cell types. MIND uses empirical Bayes techniques to exploit this commonality, together with the estimated cell type fractions, to estimate CTS gene expression. Using MIND, we analyze data from GTEx and BrainSpan to obtain CTS gene expression, from which we determine expression quantitative trait loci (eQTLs) and co-expression networks, as well as to further our understanding of the etiology of ASD.

Results

GTEx and BrainSpan data

GTEx[6] 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[6]. Samples of brain tissue from different brain regions share common cell types and thus can be deconvolved together. To ensure more reliable estimates, we remove subjects with less than nine collected brain tissue samples, resulting in data from 105 subjects for analysis. Among these subjects, 95 also have genotype data that can be used in the eQTL analysis for each cell type. To derive a signature matrix for all GTEx-related analyses, we use the NeuroExpresso database[11], which holds gene expression data for purified-cell samples from multiple mouse brains and regions. We restrict our analysis to fundamental cell types, namely astrocytes, oligodendrocytes, microglia, and GABAergic and pyramidal neurons. We apply MIND to log-transformed expression data, first calculating cell type fractions for each brain region and then estimating subject-level and CTS gene expression (**Fig. 1**).

BrainSpan quantified gene expression from multiple brain regions and subjects from 8 post-conceptional weeks to 40 years of age. These data are ideal for analysis of spatio-temporal patterns of transcription of human brain[4]. 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. Because BrainSpan represents a developmental series and most of its samples are fetal in origin, we tailor the signature matrix accordingly. We leverage scRNA-seq data from human adult and fetal cortical samples[1], specifically 466 cells that were clustered into astrocytes, oligodendrocytes, OPC (oligodendrocyte progenitor cells), microglia, endothelial cells, and immature and mature neurons.

Validating model assumptions

MIND models cell type fraction as subject- and region-specific. It is natural to assume CTS expression is subject-specific, which allows for differences among subjects due to age, phenotype, genotype and other measured variables and thereby permits downstream analyses not formerly possible (**Fig. 1f**). MIND also assumes CTS expression is similar across brain regions of the

same subject, thereby avoiding overfitting the data. For this assumption to hold, cells from the same cell type, but from different brain regions, should show similar patterns of gene expression; whereas cells of different cell types from the same region should show distinct expression profiles. This is the observed pattern in the NeuroExpresso database of purified brain cells from multiple brain regions (**Fig. 2a**). Fitting a mixed-effects model for each gene and decomposing the variance into that explained by cell types versus brain regions, as well as studies and error, cell types account for a larger amount of the variance than region, (25% versus 12%), while the largest variance comes from study (39%). Next, examination of the correlation of gene expression over regions for the GTEx data shows that bulk gene expression is highly correlated over all regions, with cerebellum and spinal cord showing slightly lower correlation (**Fig. 2b**). Reversing the role of region and subject in MIND, to estimate CTS expression for every region, shows that the imputed expression is quite similar across regions as illustrated by marker genes (**Fig. 2c**), with the strongest deviation observed for cerebellum. Fitting a mixed-effects model for each gene and decomposing the variance into that explained by cell types and brain regions, the variance explained by cell types (25%) is substantially larger than that for regions (5%). These results lead to the expectation that gene co-expression patterns for brain should be correlated with cell type fractions. Indeed, leading principal components of GTEx bulk transcriptome data are strongly correlated with cell type fractions (**Supplementary Fig. 1**).

It is reasonable to ask if MIND requires repeated measures of gene expression in the same or similar tissue. Using the deconvolved GTEx data, we calculate the standard deviation and mean of the imputed CTS expression per subject and assess their relationship with the number of measures. For subjects with fewer measures, the deconvolved CTS expression has less variability and lower mean, on average (**Fig. 2d** and **Supplementary Fig. 2**), implying that the model typically imputes similar expression for each cell type when the number of measurements is small and it lacks strong information to the contrary. Thus, while the model is identifiable when there is a single measure, the results are not very informative. The number of measurements provides an indicator of the reliability of the deconvolved expression. Because marker genes tend to show the greatest expression in the cell type they mark, the accuracy of CTS expression can be evaluated using known marker genes. Of the 189 marker genes found in GTEx brain tissue, 71% show greatest expression in the cell type they mark when the number of measures per subject is high and it falls off rapidly when the number of measures per subject approaches one **(Fig. 2d**; **Supplementary Fig. 3**).

Validating model estimates

We evaluate the performance of MIND for various scenarios, including pure simulation, simulation based on real CTS expression and analysis of the GTEx brain tissue data. Importantly, GTEx[3] produced scRNA-seq data from the prefrontal cortex (3 subjects) and hippocampus (4 subjects). For these same samples, bulk transcriptomes were also characterized[6]. From the scRNA-seq data, we can calculate CTS expression by averaging over cells of each cell type for each subject. Then, existence of both bulk and scRNA-seq data enables a direct comparison of MIND's performance and reveals highly concordant estimates for most cell types and donors (**Fig. 3a**). If MIND's estimates are accurate, bulk gene expression should be a convolution of its estimated CTS gene expression and the estimated cell type fraction for the tissue sample. Using MIND's estimates to predict region level expression for each subject shows excellent correspondence between predicted versus measured bulk gene expression (**Fig. 3b**).

Next, we conduct a simulation study by generating bulk gene expression data based on parameters estimated from GTEx. Specifically, with the estimated CTS 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 relative to the variance of the measured CTS expression. We use MIND and an approach based on least-squares to analyze the simulated bulk data, treating cell type fractions as known. Note that least-squares treats multiple measures as independent samples. To assess performance, we calculate the correlation between the deconvolved and measured gene expression for each cell type. MIND provides consistently high correlation for all cell types and is robust to increasing noise (**Fig. 3c**). This conclusion still holds when we simulate bulk data with region-specific CTS expression (**Supplementary Fig. 4a**). Moreover, the performance of MIND improves with the number of measures (regions) (**Fig. 3d**). When there are three or more measures, the correlation between the estimated and true CTS expression for four cell types can reach 0.8, assuming the error variance equals the variance of the measured CTS expression, and this reliability is also confirmed by our earlier results regarding marker gene expression (**Fig. 2d**). Moreover, MIND yields approximately unbiased estimates of all parameters (**Supplementary Table 1**) when the number of measures is large. Overall, the least-squares approach does not perform as well as MIND (**Supplementary Fig. 4**), highlighting the advantages of considering correlations between measures and assuming random CTS expression in MIND, an assumption that is particularly valuable when the number of measures is small, which is usually the case in practice.

Analysis of the GTEx brain tissue

In our early analyses of cell type fractions of the GTEx tissue, the estimated fractions for microglia were always close to zero and thus we dropped microglia from our analyses. To build the signature matrix and then estimate cell type fractions, we used CIBERSORT[9] (see Methods and **Supplementary Table 2** for approach and discussion.) Results for cell type fractions (**Fig. 4a**) were 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[12]; 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[13,14]. 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.[15] and results from Zeisel et al.[2]. We find that the estimated cell size is highly positively correlated with level of gene expression (**Supplementary Fig. 5**), and neurons tend to have a larger cell size than non-neurons, which agrees with previous findings[16]. 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.

We next examine the estimated CTS expression values, by subject, to determine if the estimates conform to expected patterns. It is reasonable to predict that RNA showing specificity for certain brain regions would also show specificity to a cell type prominent in that

region. This is indeed the case. For example, consider *ZP2* and *LINC00507*, the former is highly expressed in cerebellum, and the latter in cortical brain tissue (**Fig. 4b**). By contrasting the region-level expression for these genes with their estimated CTS expression (**Fig. 4c**), we find that *ZP2* is expressed largely in GABAergic neurons in the cerebellum, which contains large GABAergic Purkinje cells and other types of GABAergic neurons, while *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[17], we would also expect cell type to be a strong predictor of gene co-expression. Moreover, because GTEx subjects were all adults at death, but not elderly, recent findings[17] suggest that age would not be a strong predictor of gene co-expression. Thus, we asked if the estimated CTS expression clusters by cell type or by age of the subject using estimates from 98 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. 4d**), consistent with MIND's modeling assumptions.

Nonetheless, CTS expression by age reveals interesting patterns that are not always apparent at the tissue level. For example, for *GRIN3A*, expression is nearly constant across age in tissue, but GABAergic and pyramidal neurons show opposite trends in expression by age (**Fig. 4e**). Overall, 18% of genes show age trends at the region level or cell-type level, with the false discovery rate (FDR)[18] 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.

Because MIND yields subject-level and CTS gene expression, we can identify eQTLs for each cell type. To do so, CTS gene expression data were analyzed using MatrixEQTL[19], with FDR controlled at 0.05 for each cell type. We then compared the MIND-identified eQTLs with region-specific eQTLs identified by the GTEx project[6]. Notably, the rate at which eQTLs are both region-specific and CTS increases as the cell type becomes more prominent in the region (**Fig. 5a**; **Supplementary Fig. 6a**). Moreover, when an eQTL was jointly identified in more brain cell types, it was more likely to be detected across a variety of tissues and especially across brain regions[20] (**Fig. 5b** and **Supplementary Fig. 6b**). We find that the absolute effects of eQTLs increase with the number of cell types in which they are identified (**Supplementary Fig. 6c**; correlation test p-value $= 2.2 \times 10^{-16}$). Finally, 52% of eQTLs that were identified in one or more brain cell types were not identified from any GTEx brain region, which suggests MIND's results can identify novel eQTLs. Moreover, some eQTLs were shared by all four cell types, while others are specific to certain cell types, especially for neuronal cells (**Fig. 5c**), which implies that eQTL analysis based on MIND's results can shed light on gene expression regulation within cell types. Interestingly, those genes that have eQTLs in fewer cell types are more likely to be marker genes (Chi-squared-test of independence, p-value $= 5.9 \times 10^{-4}$).

Analysis of the BrainSpan data yields insights into autism

We observed that five cell types had non-negligible estimated cell type fractions in regional BrainSpan tissue (astrocytes, OPC, oligodendrocytes, immature neurons, and mature neurons). Consistent with expectation, the fraction of immature neurons decreased and that of mature neurons increased with age (**Supplementary Fig. 7**); likewise, oligodendrocytes replaced OPC, consistent with the myelination process. As the brain develops, the overall neuronal fraction (immature neuron plus mature neuron) decreased relative to other cell types, again

consistent with what is known about brain maturation[21].

Because the BrainSpan resource represents a dynamic period of development, results from the MIND algorithm provide a developmental expression profile for each cell type, which should prove useful for the study of typical and atypical neurodevelopment. For example, the neurodevelopment of subjects diagnosed with ASD probably diverges from typical development during the fetal period[5,22]. These profiles also permit construction of a co-expression network for each cell type. To do so, we calculated the correlation of expression for each pair of genes over subjects. To make the analysis relevant to ASD, we next evaluated a set of 65 genes previously implicated in risk for ASD on the basis of analysis of rare variation by the Autism Sequencing Consortium[23]. We find that the correlations between ASD genes are higher than those between non-ASD and ASD genes only in immature neurons (**Fig. 6a**).

On the basis of the CTS correlations, we regarded genes as connected in an adjacency matrix if the absolute correlation passed a threshold, here taken to be 0.9. We counted the number of connections for each gene and tested if there was a difference between the 65 ASD genes and random sets of 65 non-ASD genes matched on size of ASD genes (**Fig. 6b**). ASD genes were more connected than non-ASD genes in immature neurons (p-value $= 3.0 \times 10^{-4}$), while other cell types showed no more connections than expected by chance (all p-values > 0.05). When we performed the same CTS network analysis using a scRNA-seq dataset[1], however, we did not observe similar findings (**Supplementary Fig. 8**); apparently these scRNA-seq data were too noisy to calculate accurate correlations or, because the scRNA-seq data were derived from only a few subjects, cells of the same type lacked sufficient variability to reveal correlation patterns. However, when we examined gene expression in these scRNA-seq data, not co-expression, immature neurons were the most enriched cell type for ASD genes (odds ratio $= 7.9$; Fisher's exact p-value $= 2.8 \times 10^{-8}$; **Supplementary Fig. 9**).

Fifteen of the 65 putative ASD genes were connected in the immature neuron network (**Fig. 6c**), all were positively and highly correlated and, remarkably, all of these genes played a regulatory role according to Gene Ontology annotation for biological processes. Sixteen genes were highly correlated to more than six ASD genes in this network (**Fig. 6d**), although they lacked genetic evidence for ASD association[23]. We refer to them as ASD-correlated genes. The products of these ASD-correlated genes also tend to play regulatory or developmental roles, including acetyltransferase activity (*EPC1*[24], *KAT6A*, *KAT6B*[25]), transcriptional regulation in some form (*AFF4*, *CNOT2*, *GATAD2B*, *PCF11*, *SUPT20H*, *TUG1*[26,27,28,29,30,31,32]) and DNA replication (*HNRNPUL1*[33]). Intriguingly, the encoded protein of *FUBP* could be a key regulator of cell differentiation[34,35], specifically the transition from progenitor cells to neurons, and it is possible that all 31 genes (**Fig. 6c,d**) play a part in this transition. Of the 16 ASD-correlated genes, 13 have pLI = 1 (the probability of being Loss of Function intolerant)[36]; 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 three genes in the 12q15 deletion region[37]) and *UBXN7* (3q29 microdeletion/microduplication region).

Discussion

We develop an algorithm, MIND, to obtain gene expression by cell type and subject, even though gene expression is measured from tissue. There are notable advantages to the MIND algorithm. Because its estimates are CTS for each subject, they represent the cell-specific features inherent in the database, such as the change in CTS gene expression over development for BrainSpan or eQTLs from CTS expression for GTEx. While we have concentrated our analyses on brain tissue, MIND is not specific to brain, any tissue could be appropriate, given these two conditions: there are a group of subjects for which transcriptomes have been assessed repeatedly; and the repeatedly sampled tissue, per subject, has cell types in common. For example, several other GTEx tissues meet these requirements, including artery and esophagus[6]. Other experimental settings fit these requirements too, such as organoids[38,39]. It is also possible that one could substitute repeated measures per subject with repeated measures of genetically similar subjects, such as sibships for model organisms. Importantly, the number of repeated measures needed to obtain accurate estimates of CTS gene expression is not large, it appears three is sufficient (**Fig. 2d**; **Fig. 3d**).

There are also limitations to the current version of MIND, which relies on reference samples to identify genes whose expression are largely specific to cell type, so-called marker genes. Identifying which reference samples are appropriate can be challenging. A different challenge is presented when there are a large number of cell types in the tissue. Reliably estimating expression by cell type and subject will require a large number of repeated measures per subject, something most resources do not have at this time. For this reason, we limit our analyses to major cell types. Furthermore, MIND is limited to estimating the average gene expression across cells of the same type within a subject, ignoring the diversity of expression within single cells.

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 specific cell types. When we tried to construct such networks, however, they show very little coherent structure. By contrast, results from MIND yield coherent and interpretable networks, which show relevance to risk for ASD, potentially highlighting new genes in risk, their functional impact, and periods during which neurodevelopment begins to diverge from typical patterns.

Methods

The MIND algorithm

For a single measure (t) from subject i, let X_{ijt} be the observed expression of gene j. When the tissue consists of K cell types, typically the goal of gene expression deconvolution is to find $\boldsymbol{W}_{it}$, the K cell type fractions for subject i in measure t, such that

$$\underset{(1\times1)}{X_{ijt}} = \underset{(1\times K)}{\boldsymbol{W}_{it}} \; \underset{(K\times1)}{\boldsymbol{A}_{j}} + \underset{(1\times1)}{e_{ijt}}\,, \tag{1}$$

where $\boldsymbol{A}_j$ is the cell type gene expression and e_{ijt} is the error term (csSAM[8] is an exception to this rule.) When reference samples are available, such as purified cells or scRNA-seq data, the signature matrix can be estimated for the marker genes by differential expression analysis of cell types from the reference samples. Plugging in $\boldsymbol{A_j}$, deconvolution becomes a standard regression problem[7,9] and $\boldsymbol{W}_{it}$ can be estimated directly.

We extend the single-measure deconvolution in Eq. 1 by borrowing information across multiple measurements, $t = 1, \ldots, T_i$ from the same tissue for subject i to estimate subject-specific and CTS gene expression (T_i can vary by subject.) **Step 1** of the MIND algorithm is to estimate cell type fractions for subject i and measure t, $\boldsymbol{W}_{it}$, for $t = 1, \ldots, T_i$. Combining estimated information across measures yields $\boldsymbol{W}_i$, a $T_i \times K$ matrix, of cell type fractions. **Step 2**, treating $\boldsymbol{W}_i$ as known, we reverse the problem from single-measure deconvolution, estimating instead CTS gene expression. For gene j in subject i, the observed gene expression $\boldsymbol{X}_{ij}$ is a $T_i \times 1$ vector that represents T_i quantified measurements (**Fig. 1e**), rather than a scalar as in Eq. 1. We model $\boldsymbol{X}_{ij}$ as a product of cell type fraction ($\boldsymbol{W}_i$) and CTS expression ($\boldsymbol{A}_{ij}$),

$$\underset{(T_i \times 1)}{\boldsymbol{X}_{ij}} = \underset{(T_i \times K)(K \times 1)}{\boldsymbol{W}_i \quad \boldsymbol{A}_{ij}} + \underset{(T_i \times 1)}{\boldsymbol{e}_{ij}}, \tag{2}$$

where $\boldsymbol{e}_{ij}$ is the error term that captures the unexplained random noise and $\boldsymbol{e}_{ij} \sim N\left(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_{T_i}\right)$.

In summary, with $\boldsymbol{W}_i$ pre-estimated using an existing deconvolution method (e.g., CIBERSORT), our goal is to estimate CTS expression $\boldsymbol{A}_{ij}$. To ensure robustness, we assume that $(\boldsymbol{A}_{ij})$ is randomly distributed as $\boldsymbol{A}_{ij} \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_c\right)$, where $\boldsymbol{\Sigma}_c$ is a $K \times K$ covariance matrix for K cell types. Estimation is performed across all subjects and genes simultaneously. In contrast to single-measure deconvolution, we assume

1) cell type fraction ($\boldsymbol{W}_i$) is subject- and measure-specific;

2) CTS expression ($\boldsymbol{A}_{ij}$) is subject-specific but constant across measures.

We estimate the parameters through maximum likelihood via a computationally efficient EM (Expectation-Maximization) algorithm (see **Supplementary Note**). CTS expression ($\boldsymbol{A}_{ij}$) 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 (e.g., brain regions) per subject, whereas all genes in the genome can be efficiently deconvolved together. Tissue expression can be centered beforehand to meet the prior distribution of CTS expression and ensure more precise estimates. When comparing the deconvolved and measured expression, however, we keep the tissue expression as uncentered to impose a fair comparison. Centering is an option that can be chosen by users of MIND software and can be done for each tissue sample or over all samples. If centered, the subtracted mean of expression can be added back after deconvolution.

MIND ignores gene-gene correlation in the prior distribution of CTS expression to achieve efficient computation, deconvolving for the whole genome in several minutes. Gene-gene correlation can be recovered from the CTS gene expression estimates. To compute correlations, CTS estimates are weighted by the average cell type fraction per subject.

Software availability

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

Validating model assumptions

To partition variation in gene expression by cell type and brain region, we analyzed NeuroExpresso normalized data on expression of 11,546 genes[11]. To evaluate the correlation of gene expression across brain regions, we analyzed the read count data for all genes detected in

brain by GTEx[6]. Expression was transformed as $\log_2(X+1)$ prior to analysis. Unless otherwise noted, all expression count data analyzed herein were log-transformed in this way. (See **Supplementary Note** for discussion and analysis of log-transformation of the data.) GTEx brain data were further quantile normalized. To estimate principal components (PCs) from gene co-expression, we first quantile normalized and scaled the expression data. All genes were used to compute the co-expression matrix 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 performance of MIND is a function of the number of measures of gene expression, but its exact nature was unknown. We addressed this question in two ways: (1) by evaluating the variability of gene expression as a function of the number of brain regions measured; and (2) how marker gene expression behaved in cell types they are reported to mark in mouse brain. The key idea of (1) is that MIND will tend to shrink CTS expression toward a common mean and thereby estimates will be less variable when there is little information about CTS expression. For (2), we expect marker genes will tend to be expressed at highest levels in the cell type they mark; however, when there is little information about CTS expression, this is not the expected pattern due to shrinkage to the mean. As described previously, CIBERSORT was used to select marker genes from NeuroExpresso normalized data on expression of 11,546 genes; 192 genes were selected, of which 189 are found to be expressed in GTEx brain tissue samples.

Validating model estimates

Habib et al.[3] quantified single-nucleus RNA-seq data from seven brain tissue samples from five GTEx donors. Because the authors classified the cells into cell types, we could average their read count data for cells of each type to obtain CTS expression on a scale similar to that produced by MIND. One of the five subjects that only has hundreds of cells and thus cannot provide accurate CTS expression was excluded from our analyses. For a fair comparison, we converted the read counts to count per million (CPM) and then compared the directly measured subject-specific and CTS expression to MIND's estimated quantities from bulk transcriptomes (in CPM) from the same subjects. In **Fig. 3b**, we showed MIND's predicted bulk transcriptome data for two brain regions, frontal cortex and cerebellum: we chose the former because it is the most studied region of brain; we chose the latter because it deviates greatly from other regions; and we noted that all brain regions showed similar patterns. **Figs. 3a** and **3b** showed results using the R function smoothScatter, which was implemented using 128 bins for the density estimation and default settings.

To evaluate MIND via simulations, in **Supplementary Table 1**, we generated artificial gene expression from the multi-measure deconvolution model Eq. 2. We systematically varied the values of the true variance parameters, σ_e^2 and $\mathbf{\Sigma}_c$, which denote the error variance and the covariance of CTS expression. Here we let $\mathbf{\Sigma}_c$ have equal variance σ_c^2 and equal covariance $\sigma_c^{kk'}$ across cell types, where k and k' denote cell types. The cell type fraction was estimated from the GTEx brain data and we focused on the 105 subjects with at least nine measures and allowed some brain regions to be unmeasured. The number of cell types was set at four. We simulated 100 replicated datasets with 100 genes and 9-13 measurements of the same tissue. We produced data for 100 genes to reduce simulation time; these genes were randomly generated, they were not necessarily cell type marker genes.

Using single-cell measurements from 4 GTEx subjects as a guide, we simulated bulk expression for 4 subjects from Eq. 2. The measured CTS expression was taken to be the

average expression values across cells derived from the Habib et al.[3] single-nucleus RNA-seq data as described above, for four subjects, four cell types and 31,496 genes. The cell type fractions ($\boldsymbol{W}_i$) were derived from those estimated in GTEx for these subjects. In **Fig. 3c**, we varied the error variance σ_e^2 via the noise level defined as σ_e^2/σ_c^2, where σ_c^2 is the variance calculated from the measured CTS expression[3]. For this display, the number of measures is 13 as in the GTEx brain data. In **Fig. 3d**, we fixed $\sigma_e^2 = \sigma_c^2$ and varied the number of measures from 1 to 13. For **Supplementary Fig. 4a**, on the basis of the simulation in **Fig. 3c**, we added region-specific variation to $\boldsymbol{A}_{ij}$ (CTS expression per subject). The variation was simulated from a normal distribution with zero mean and variance the same as the error variance (σ_e^2), which increased up to the variance of the measured CTS expression (σ_c^2).

To assess whether MIND produces approximately unbiased parameter estimates, we calculated the average of the variance parameter estimates from the 100 replications (**Supplementary Table 1**). To evaluate whether MIND can recover the true CTS expression, we computed the correlation between MIND's and true CTS expression.

Analysis of the GTEx brain tissue

We have described the processing of the GTEx gene expression data from brain regions previously. For the Remark about cell size and level of gene expression, our analyses made use of the scRNA-seq data in Zeisel et al.[2], which also contained spike-in information. We leveraged the spike-in information to estimate cell size[15] for neurons and non-neurons (**Supplementary Fig. 5**) and thus interpreted the impact of size versus cell type composition in the deconvolution of bulk transcriptomes. To identify genes that show the greatest variability across regions of the brain, we selected the top 10 genes that have the most significant difference in expression between each region and other regions. Pooling these genes from 13 regions, we obtained 98 unique genes. As described previously, eQTLs from CTS expression were estimated using MatrixEQTL. To compare MIND's results to eQTLs from GTEx data, we downloaded eQTLs from GTEx portal, https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz. To get eQTLs specific to region, we removed any eQTLs shared by two or more regions.

Analysis of the BrainSpan data yields insights into autism

Because we used BrainSpan's exon microarray data with normalized expression values, no transformation of the data was performed. For the signature matrix to estimate cell type fractions for regions and subjects, we used the scRNA-seq data in Darmanis et al.[1], which includes both fetal and adult cells. We also used this dataset for the enrichment analysis of ASD genes (**Supplementary Fig. 9**); we defined a gene as "expressed" in a cell type if at least 15% of the cells of that type contain at least one RNA-seq read attributed to that gene. We restricted the enrichment analysis to the 11,215 genes that were 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 tabulated whether the gene is expressed and whether it is associated with ASD risk. When comparing the number of connections for ASD and non-ASD genes, we calculated the tail probability for the average number of connections for ASD genes in the reference distribution of average number of connections. To construct the reference distribution, we first matched each ASD gene with the top 100 genes with the closest gene size. We then randomly sampled one gene from those matched genes for each ASD gene and constituted a gene set. We

calculated the average number of connections for this gene set and repeated this process 10,000 times. Gene Ontology[40] was performed using Enrichr[41]. pLI was obtained from the EXAC[36] browser, http://exac.broadinstitute.org.

Acknowledgments

We are grateful for the insightful comments from Michael Breen, Joseph Buxbaum, Lin Chen, Serkan Erdin, Dadi Gao, Lambertus Klei, Maria Jalbrzikowski, Silvia De Rubeis, Stephan Sanders, Michael Talkowski, 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 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

1. 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).

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. Habib, N. *et al.* Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods* **14**, 955–958 (2017).

4. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483 (2011).

5. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).

6. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).

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

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

9. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015).

10. Wang, X., Park, J., Susztak, K., Zhang, N. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *bioRxiv* 354944 (2018).

11. 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).

12. Bekkers, J. M. Pyramidal neurons. *Current Biology* **21**, R975 (2011).

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

14. Azevedo, F. A. *et al.* 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 (2009).

15. Jia, C. *et al.* Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic acids research* **45**, 10978–10988 (2017).

16. Wang, J. *et al.* Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences* **115**, E6437–E6446 (2018).

17. Soreq, L. *et al.* Major shifts in glial regional identity are a transcriptional hallmark of human brain aging. *Cell reports* **18**, 557–570 (2017).

18. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

19. Shabalin, A. A. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

20. McKenzie, M., Henders, A. K., Caracella, A., Wray, N. R. & Powell, J. E. Overlap of expression quantitative trait loci (eqtl) in human brain and blood. *BMC medical genomics* **7**, 31 (2014).

21. Tau, G. Z. & Peterson, B. S. Normal development of brain circuits. *Neuropsychopharmacology* **35**, 147 (2010).

22. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).

23. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).

24. Searle, N. E. & Pillus, L. Critical genomic regulation mediated by Enhancer of Polycomb. *Current Genetics* **64**, 147–154 (2018).

25. Huang, F., Abmayr, S. M. & Workman, J. L. Regulation of KAT6 Acetyltransferases and Their Roles in Cell Cycle Progression, Stem Cell Maintenance, and Human Disease. *Molecular and Cellular Biology* **36**, 1900–1907 (2016).

26. Luo, Z. *et al.* The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output. *Molecular and Cellular Biology* **32**, 2608–2617 (2012).

27. Rambout, X. *et al.* The transcription factor ERG recruits CCR4-NOT to control mRNA decay and mitotic progression. *Nature Structural & Molecular Biology* **23**, 663–672 (2016).

28. Bornelov, S. *et al.* The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression. *Molecular Cell* **71**, 56–72 (2018).

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

30. Watanabe, K. & Kokubo, T. 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 (2017).

31. Baptista, T. *et al.* SAGA Is a General Cofactor for RNA Polymerase II Transcription. *Molecular Cell* **68**, 130–143 (2017).

32. Guo, Y. *et al.* Interplay between FMRP and lncRNA TUG1 regulates axonal development through mediating SnoN-Ccd1 pathway. *Human Molecular Genetics* **27**, 475–485 (2018).

33. Ideue, T. *et al.* U7 small nuclear ribonucleoprotein represses histone gene transcription in cell cycle-arrested cells. *Proceedings of the National Academy of Sciences* **109**, 5693–5698 (2012).

34. Quinn, L. M. FUBP/KH domain proteins in transcription: Back to the future. *Transcription* **8**, 185–192 (2017).

35. Zhou, W. *et al.* 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 (2016).

36. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

37. Alesi, V. *et al.* Reassessment of the 12q15 deletion syndrome critical region. *Eur J Med Genet* **60**, 220–223 (2017).

38. Mariani, J. *et al.* Foxg1-dependent dysregulation of gaba/glutamate neuron differentiation in autism spectrum disorders. *Cell* **162**, 375–390 (2015).

39. Takasato, M. *et al.* Kidney organoids from human ips cells contain multiple lineages and model human nephrogenesis. *Nature* **526**, 564 (2015).

40. Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research* **45**, D331–D338 (2016).

41. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016).

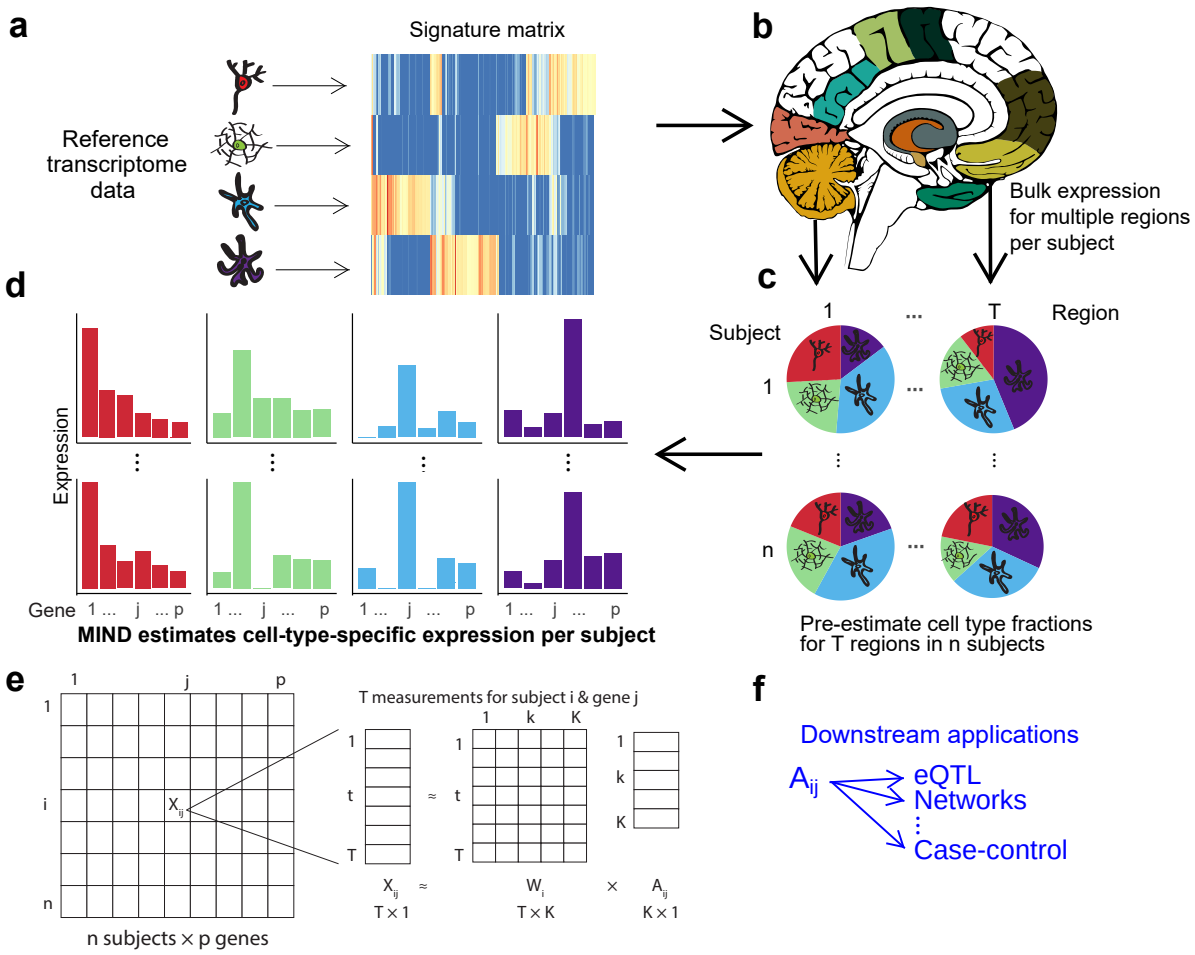

Figure 1. Flow diagram for the MIND algorithm. (**a**) For a set of relevant cell types, select cell type marker genes and build a signature matrix using reference samples. (**b**) Multiple transcriptomes are measured from each subject; here, one transcriptome for each of multiple regions. (**c**) Using an existing deconvolution method, e.g., CIBERSORT, estimate the cell type fractions for each brain region and subject. Here we depict $K = 4$ cell types for which their fractions will be estimated per brain region. (**d**) With results from (**b**) and (**c**), MIND estimates cell-type-specific (CTS) expression for each of p genes for each subject and cell type. Colors map to the cell types in (**c**) and (**d**) and we depict two of n subjects, 1 and n. (**e**) Matrix representation of key data elements of the MIND algorithm: for each of T brain regions for subject i, expression of p genes from the transcriptome is measured, $\boldsymbol{X}_{ij}$; and the key outputs are the subject level CTS gene expression ($\boldsymbol{A}_i$) and the subject and measurement level cell type fractions ($\boldsymbol{W}_i$). (**f**) Examples of downstream applications for MIND.

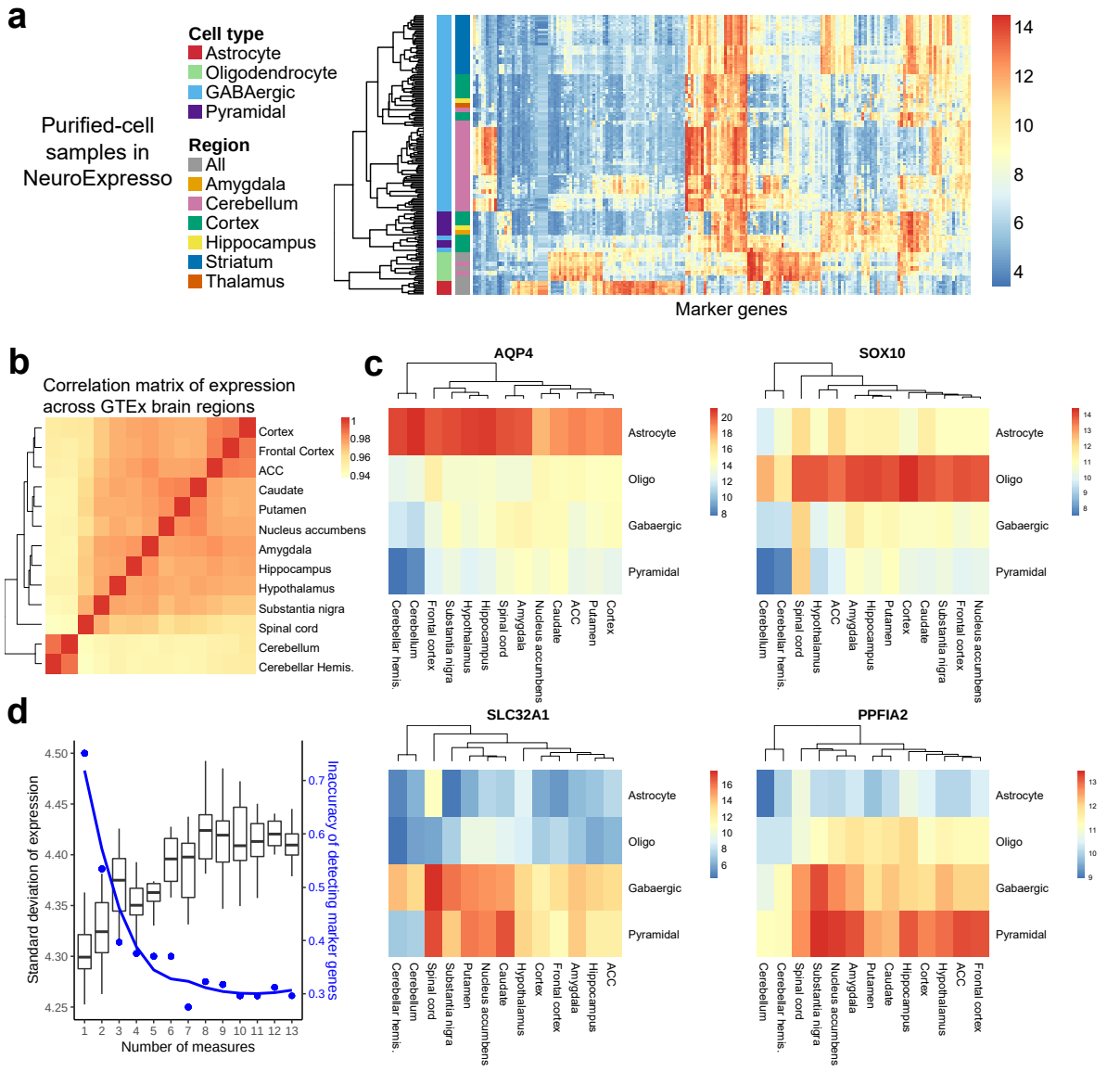

Figure 2. Validation of the assumptions of MIND. (**a**) Heatmap of expression of cell type marker genes in the NeuroExpresso database of purified-cell samples. Columns denote 192 marker genes selected by CIBERSORT from NeuroExpresso. Rows represent 185 purified-cell samples that we use to estimate fractions of four cell types in GTEx. Purified-cell samples are clustered, then annotated by cell type and brain region (labels on left, scale of expression on right). (**b**) Correlation matrix of gene expression (heatmap) for brain regions from GTEx samples. (ACC: anterior cingulate cortex; hemis.: hemisphere.) (**c**) Heatmaps of region-specific and CTS expression of marker genes estimated by reversing the role of subject and measure in MIND. The four marker genes correspond to astrocyte, oligodendrocyte (oligo), GABAergic, and pyramidal neurons, respectively. (**d**) Left scale: variance of expression across all genes, per subject, as a function of the number of measures in GTEx brain data. Right scale: fraction of marker genes showing greatest expression in a different cell type than the cell type they mark. Marker genes are selected by CIBERSORT using the reference data of NeuroExpresso.

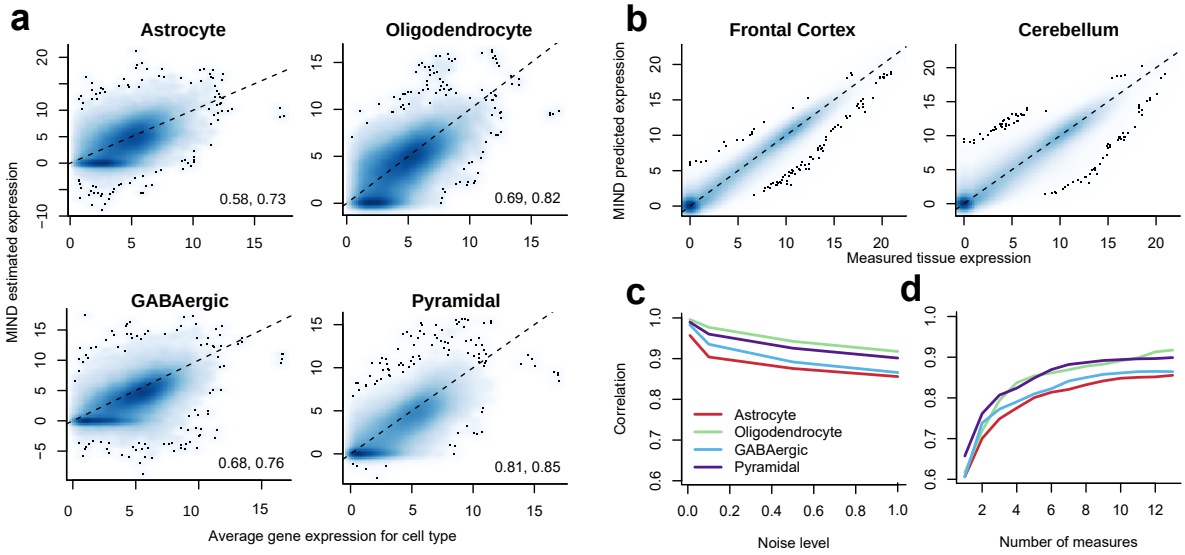

Figure 3. Validation of the estimates of MIND. (**a**) Direct quantification of average gene expression from single cells (observed)[3] from GTEx brain samples of the same subjects as the CTS expression estimated by MIND. Shown are scatter plots represented as a smoothed two-dimensional color density. For each panel, two summary statistics are given, correlation for all genes with positive observed expression (left) and for all genes (right). On average, there are 17,223 out of 31,496 genes that have positive observed CTS expression. Smooth line at $y = x$. (**b**) Smoothed scatter plots of the observed GTEx brain tissue expression and MIND predicted expression for frontal cortex and cerebellum. Smooth line at $y = x$. (**c-d**) Correlation between the true and MIND estimated expression for each cell type in simulation. We simulated cell mixture data following Eq. 2 using the measured CTS expression[3] and the estimated cell type fractions from GTEx data, with increasing noise levels (the error variance relative to the variance of CTS expression, **c**) and number of measures (**d**). For (**c**), the number of measures is 13 as in the GTEx brain data; for (**d**), the noise level is set as 1, which means that the error variance equals the variance of CTS expression.

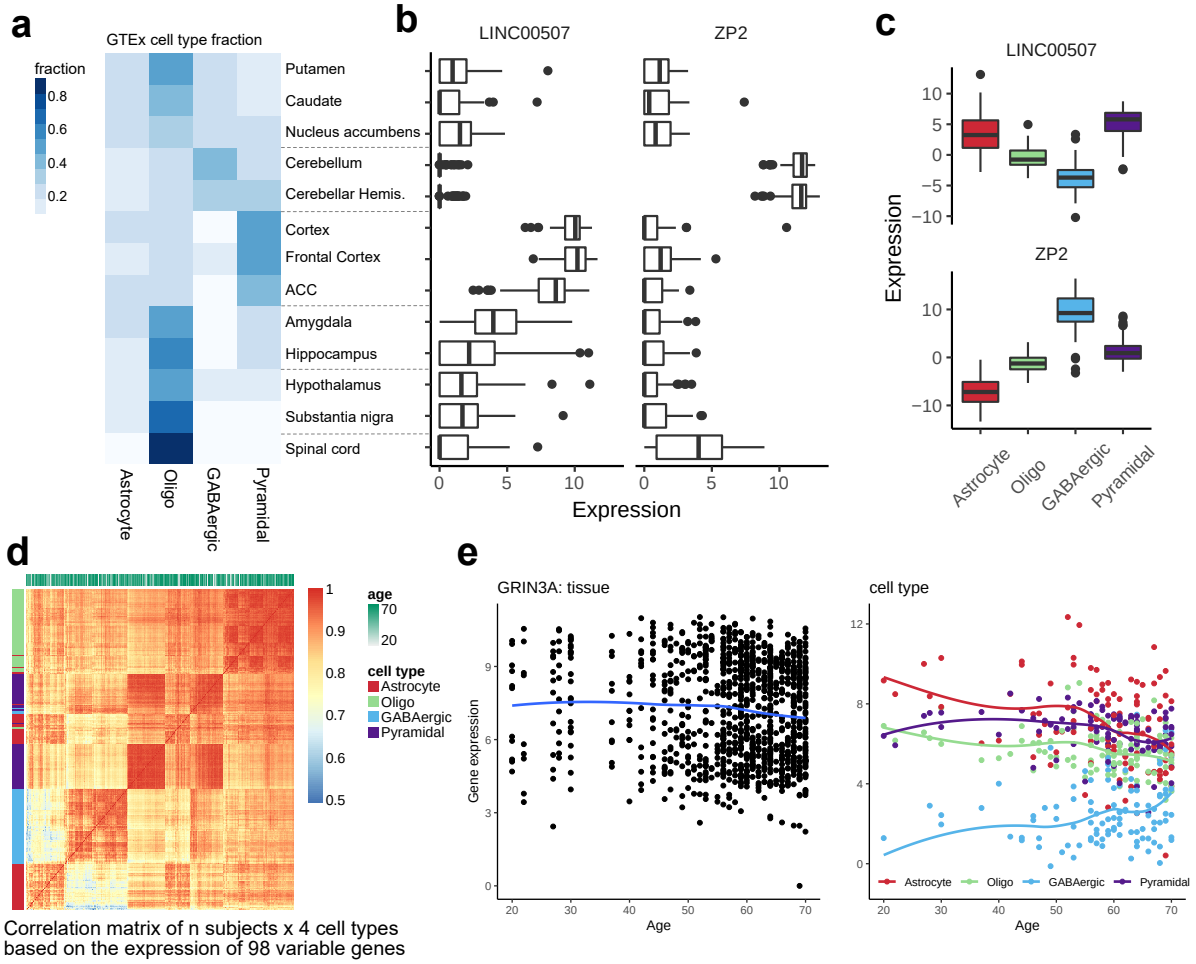

Figure 4. Analyses of CTS gene expression of the GTEx brain data. (**a**) Estimated cell type fractions in each GTEx brain region, averaged over subjects. Putamen, caudate, and nucleus accumbens are the three basal ganglia structures. (**b-c**) For two transcripts selected for differential expression in cortex versus cerebellum, (**b**) boxplots of tissue-level expression across brain regions and (**c**) CTS expression estimated by MIND from tissue-level expression across brain regions. (**d**) The heatmap and clustering of estimated CTS expression from MIND by cell type and age. Here we visualize a $4n \times 4n$ correlation matrix for the 4 cell types and $n = 105$ subjects, based on the expression of 98 genes that have the largest variability across brain regions. (**e**) Age trends for expression of gene *GRIN3A* in tissue and its estimated CTS expression from MIND.

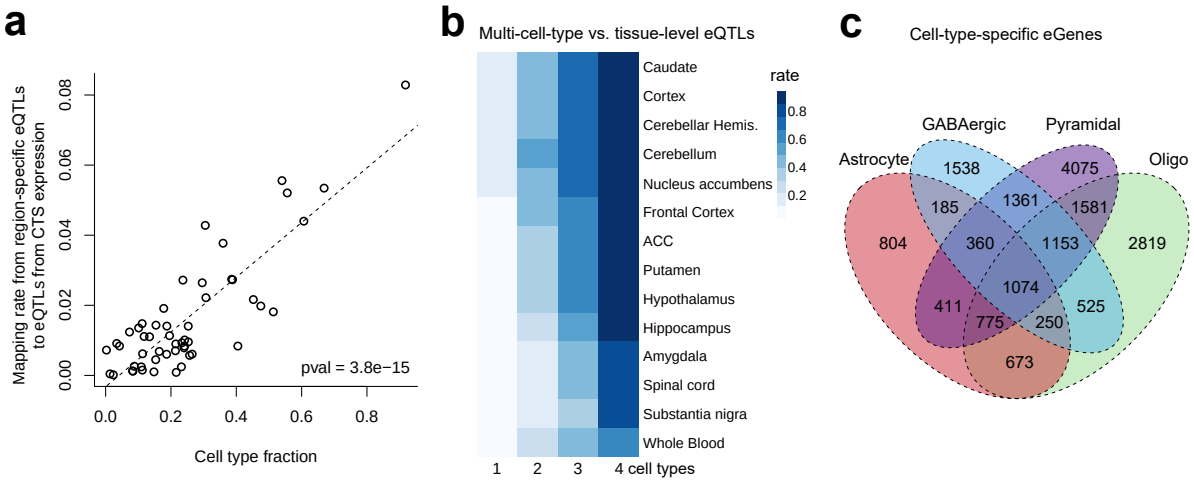

Figure 5. Expression quantitative trait loci (eQTL) discovered from tissue-level or CTS gene expression. (**a**) Scatter plot of eQTL mapping rate versus the estimated cell type fraction. The rate is for mapping region-specific eQTLs identified by the GTEx consortium to eQTLs from CTS expression estimated by MIND. Each point denotes a brain region and cell type. The dashed line depicts the fitted linear regression model and the p-value (pval) is for the test of the regression slope. (**b**) Rate of correspondence between eQTLs appearing in one to more cell types and those in each 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. We show brain regions and whole blood here. For results from all GTEx tissues, see **Supplementary Fig. 6b**. (**c**) Overlap among eGenes (genes with eQTLs) for each cell type.

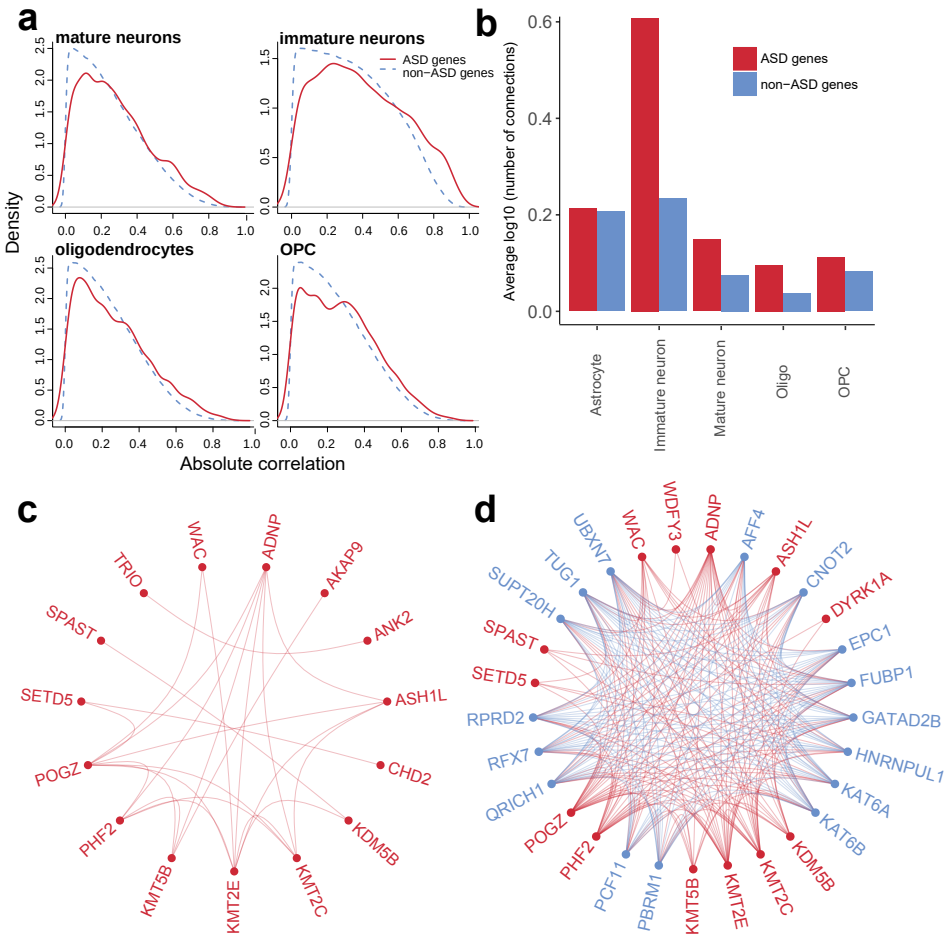

Figure 6. Analyses of MIND-estimated CTS networks from BrainSpan and their relationship to genes implicated in risk for ASD[23]. (**a**) The density distribution of absolute weighted correlations for pairs of ASD genes (red solid line) and pairs involving one ASD and one non-ASD gene (blue dashed line) for each of four cell types. The weights are the average cell type fractions per subject. (**b**) The average number of connected genes for ASD genes and non-ASD genes in different cell types (in log10 scale) based on CTS networks. A connection between genes is indicated if the absolute weighted pairwise correlation of expression is greater than 0.9. (**c**) Co-expression network of 15 out of 65 ASD genes in the immature neuron. (**d**) For the network in immature neurons, 16 genes are connected to more than six ASD risk genes (red) and we call them ASD-correlated genes (blue). These ASD-correlated genes were not detected as risk genes by Sanders et al.[23]. Here we show only the 13 ASD risk genes that are connected to those 16 ASD-correlated genes. The interactive version of this figure is available at http://rpubs.com/randel/ASDnetwork.