Introduction

Gene expression analysis of complex tissues [e.g., the central nervous system (CNS)] from patients and controls is a common method to screen for potentially pathological molecular mechanisms of psychiatric disease. However, the substantial cellular heterogeneity of these tissues poses several difficulties in interpreting human gene expression data. First, it restricts the detection of gene expression changes. For example, biologically important changes in gene expression in one cell type may be masked by compensatory changes in another cell type. Also, even massive changes in gene expression may not be detected if the gene is only expressed in a rare cell type. Second, even if the changes in gene expression are detected in a heterogeneous tissue, they are difficult to interpret. Changes in gene expression could be due to the arrival or loss of certain cell types, an expression change in all cells or in a fraction of cells, or combinations of all of these. These challenges highlight the importance of understanding changes in cellular composition of a tissue. Recent work profiling whole human brains across regions and time have suggested that many, and perhaps most gene expression differences seen in these studies is driven by differences in cellular composition across the samples. If the human brain expression data can guide us to consistent cellular alterations, even in the context of distinct genetic or environmental causes of a disorder across different individuals, then treatments could be tailored to addressing the common cellular deficits.

Emerging methods are attempting to incorporate cellular information into analytical models for human brain data. However, these methods either tend to evade the complexity of cellular composition by using artificially predesigned cellular mixtures in tissues, or they rely on previous knowledge from low throughput histology-based approaches, which have identified only three or four cell type-specific genes for each cell type at best, and would preclude analysis of more poorly described cell types. Consequently, to predict the cellular composition in a new
experimental data set, these small numbers of known markers limit statistical power, and render the analysis more vulnerable to any measurement noise or technical artifacts present in a given marker, potentially resulting in poor sensitivity and specificity.

To improve our capability of interpreting human brain transcriptome data and predicting the cellular composition, there is a need for methodologies that robustly and systematically integrate prior information to prioritize, interpret, and visualize relationships among cell type-specific genes, and identify changes in cellular composition. To address this need, we propose a novel cell type-specific analysis method. We first identify genes that are significantly enriched in each cell type from our previous cell-specific transcriptome profiling experiments in mouse brain. Each cell type is assigned with a number of cell-specific genes as “prior” potential gene markers using our previously developed method validated for this purpose. We then perform co-expression network analysis on human brain transcriptome data of the prior gene markers of each cell type to uncover the genes with expressions highly correlated with each other. As changes in cellularity should result in correlated changes in cell-specific genes, the highly-correlated genes are then selected as the effective markers of the cell type. We believe that analysis based on these systematically selected gene markers provides an accurate prediction of cellular composition and alterations with better statistical power.

In this work, we apply our method to two problems, one where changes in cellularity are known, and one where they are not known. Analysis 1: we employ temporal transcriptome of human brain (specifically, from cortex and cerebellum) and successfully detect changes in cellularity over time that roughly correspond to known epochs of human brain development. Analysis 2: we apply our method to transcriptomes from brain samples (specifically, cortex) from individuals with the neurodevelopmental disorder of autism. The results indicate an increased signal from astrocytes and a reduced signal from neuronal cells, suggesting a corresponding change in cellularity may contribute to the disorder.

**Results and Discussion**

**Analysis 1: Determining temporal cellular changes in human brain**

Here we apply our method to predict temporal cellular changes from transcriptome data of developing and adult post-mortem human brains. The dendrograms in Figure 1A and Figure 2A show the gene co-expression analyses (Step 5) on prior gene markers (obtained from Steps 1–4) of astrocytes in cortex and Purkinje cells in cerebellum, and the identification of their effective gene markers. Of all four cell types in cortex and of all eight cell types in cerebellum, the marker lists are given in Table S1 and co-expression analyses are in Figure S1 and Figure S2, respectively. These plots also present an evaluation of the effective gene markers of all cell types by combining them into a single set and repeating the co-expression analysis. In cortex, the effective gene markers corresponding to the four cell types (astrocytes, mature oligodendrocytes, neurons, and progenitor cells) are completely separately classified, indicating distinct cellular behaviors. On the contrary, in cerebellum, while some cell types’ gene markers are nearly perfectly clustered, such as mature oligodendrocytes, progenitor cells, granule cells, and Purkinje cells, some are not as distinct. The largest sub-clusters for each of the remaining cell types are: astrocytes and Bergman glia (6, out of 13 markers), inner golgi neurons (4/6), stellate and basket cells (3/5). One possible reason the markers are intermingled might be that those cell types indeed share similar temporal patterns in cerebellum development.

Figure 1B and Figure 2B show the temporal changes of cell types across development in cortex and cerebellum, respectively. From these plots, at embryonic and early to middle fetal periods, the eigengene of progenitor cells is dominant, while the eigengenes of the other cell types suggest that they have not yet appeared. Consistent with the progenitor cells then differentiating to form specific cell types in succession, the progenitor eigengene expression reduces as the others increase. In addition to this broad shift from progenitor to mature cell types, these figures indicate that we also successfully detect the known epochs of the birth and maturation of individual cell types. For instance, in cortex, neurons are known to be born the first, followed by the astrocytes, and the oligodendrocytes are the last (Fig. 1B), and this is largely consistent across cortical regions (Fig. 3B). In cerebellum, our data indicate that the Purkinje cells are born before the granule cells (Fig. 2B). Then, the birth and maturation of granule cells in turn dilute the signal of Purkinje cells. The observations in the enlarged plots of changes of cell types in cortex and cerebellum around birth date (Fig. 1C and Fig. 2B, respectively) are generally consistent with the ontogeny we have summarized from literature (Fig. 1D and Fig. 2D, respectively), indicating that broadly our method accurately detects known changes in cellular composition of the tissue.

To examine robustness of the approach to marker selection, we performed a bootstrap analysis (Fig. 3A), sampling with replacement (n = 30) from the markers and examining the variance of the eigengenes at each time point. This analysis demonstrated robustness of the method relative to marker choice. As a test of data variation across regions, Figure 3B shows temporal cellular changes in human cortex development, represented by the average of the eigengenes computed for the four cortical regions available at all time points: frontal lobe, parietal lobe, occipital lobe, and temporal lobe. We observe that the cellular patterns are largely consistent across these cortical regions.

Though the order of arrival of each cell type in cerebellum is correct, the time scale we are detecting for these cell types does appear to be a bit compressed relative to the known ontogeny. We suspect that we are detecting the cell types as they arrive at an approximation of their mature gene expression profile, rather than the moment they are born, as our eigengenes are built using prior information from the adult state of these cells. This would be consistent with classical morphological studies of the development of the cerebellum: while Purkinje cells are born substantially earlier than their counterparts, their final morphological maturation depends on instructive cues from the later born cells.

Likewise, in the cortex we detect an unexpected astrocyte-like signal slightly preceding neurons. This is likely to derive from a transient population of astrogial-like cells (the radial glia) that...
are present during the period of neurogenesis, and which later give rise to the more mature astroglial phenotype. Overall, however, we expect that the success of our approach in roughly recapitulating the cellular maturation of the nervous system suggests it might be equally informative as a method to detect changes in cellularity that occur in gene expression studies of psychiatric disorder, in which the relevant changes are not known.

Analysis 2: Predicting cellular alterations in human neuro-developmental disorder of autism

Here we apply our method to predict cellular alterations of cerebral cortex in autistic patients from transcriptome data. Similar to Analysis 1, the dendrograms in Figure 4A show the gene co-expression analysis (Step 5) on prior gene markers (obtained from Steps 1–4) of astrocytes in cortex, and the identification
of their effective gene markers. Of all four cell types in cortex, the marker lists are given in Table S2 and co-expression analyses are in Figure S3. We observe that the effective gene markers corresponding to the four cell types are almost completely classified, except that three gene markers of progenitor cells are clustered with the markers of astrocytes (the largest sub-cluster of

Figure 2. Temporal cellular changes in human cerebellum development. (A) Identification of effective gene markers: co-expression analysis on prior gene markers of Purkinje cells as an example (top), where the effective markers are selected as the largest clustered nodes under the red-line cut-off. Effectiveness testing: co-expression analysis on effective gene markers of all cell types (bottom). Co-expression analyses of all eight cell types in cerebellum are in Supplementary Figure S2. (B) Cellular changes across development, represented by the changes of the eigengenes’ expressions. The emergence of the cell types is indicated by the date when the expressions of eigengenes start to go up sharply. (C) Enlarged plot of the changes around birth date. (D) Ontogeny of the cell types summarized from literature, where the oval-shape indicates time of birth and period of maturation of the cell types. The cell types in the dendrogram are color-coded with the legends in (B).
progenitor cells contains six out of nine markers). This perhaps reflects the fact that a subset of astrocytes can serve as adult progenitor cells in certain brain regions. 

Figure 4B presents average fold changes (log2 transformed) in the autism group relative to the controls for the expressions of the effective gene markers of the four cell types in cortex, as well as the corresponding P values from the t tests on the fold-changes. This plot shows a decreased expression of neuronal genes (P value = 7.38e-3) and an increased expression of astrocytic genes (P value = 6.41e-4) in cortices from autistic patients. The results suggest that one of the cellular mechanisms of autism may be a relative deficit in the function or presence of neurons, or an enrichment of astrocytes (perhaps to support or remove dysfunctional neuronal cells).

Here the data tested are mixed samples in temporal lobe of cortex (13 autistic samples and 13 controls) and frontal lobe of cortex (16 autistic samples and 16 controls). We also apply our method independently to the samples of each region. The lists of effective gene markers for the temporal lobe and the frontal lobe (also given in Table S2) both substantially overlap the markers of the mixed samples. The overall changes observed in individual regions (Fig. S4; Fig. S5) are consistent as those in Figure 4, confirming the robustness of the analysis. We note that here we have used a single category of “neurons” in cortex; future investigations can apply the method to investigate particular sub-types of cortical neurons.

Comparison of the cell type-specific markers derived from our method with markers from literature

To further evaluate the performance of our method, we also perform Analyses 1 and 2 for the four cell types in cortex based on the three or four gene markers collected in reference 12, which had been identified by classic tissue-based approaches. These markers are denoted as the “markers from literature” and are given in Table S3. We notice that six of the 13 markers from literature are also found in our method, a significant overlap, given the size of the genome (P value = 2.096e-10 [Fisher’s Exact test]).

Figure 5A compares the eigengenes computed from our effective gene markers and the markers from literature for Analysis 1. The expression patterns of the eigengenes from our markers coincide with those from the markers from literature, which once again verifies the effectiveness of the markers collected from our method. This concordance suggests that there are a number of different genes that can function effectively as markers in this context. By contrast, Figure 5B presents the average fold changes in the autism group relative to the controls for the expressions of the gene markers from literature and P values from the t test for Analysis 2. Though the trends of the fold changes from the literature markers are similar to ours, the P values from the t test are not significant. This suggests that the larger number of markers identified by our method provides superior statistical power for this type of analysis (Fig. 4B).

The results demonstrate several advantages of our method. It is efficient and robust to select cell type-specific gene markers in predicting cellular composition and alterations, for any given transcriptome data set. As verified, this data-driven method achieves high sensitivity and specificity. It enables us to work for cell types even when few or no markers are available in the literature. Moreover, our method provides a larger number of markers for further analysis, which improves statistical power.

Materials and Methods

Cell-specific transcriptome profiling experiments in mouse brain

We systematically examined and identified genes whose expressions were significantly enriched in each cell type relative to all others, from cell-specific transcriptome profiling experiments (conducted on a single platform for dozens of targeted cell
types) in mouse brain.\textsuperscript{4,5,14} Genes relatively specifically employed by a cell type are expected to contribute to the unique functions of that cell. These genes may both serve as proxies (markers) for detecting the cell type, and in addition may become useful targets for development of pharmacological tools for cell-specific manipulations in the future. The specificity index (SI) is a statistic that describes the relative uniqueness (specificity) of the expression of a certain gene in a certain cell type, when compared with all others. It is described in detail in reference \textsuperscript{4}. Essentially, for a given gene, it is the average rank, in terms of “fold-change,” across each pairwise comparison of the targeted cell type to each of the other available cell types (i.e., an SI of 1 would mean it was the most enriched gene in every pairwise comparison). Permutation tests are used to shuffle the gene expression values from the cell type of interest to calculate an empirical distribution of possible SI values. Then \textit{P} values are ascribed to the genes based on the comparison between their experimental SI values and the ones that are calculated from the permuted data. We denote these \textit{P} values as pSI. Previous validation with in situ hybridization databases indicated that genes with pSI < 1e-4 are likely to be nearly uniquely expressed in a particular cell type, while those with more modest \textit{P} values (pSI < 1e-2) will merely show enrichment.\textsuperscript{4} Lists of genes with pSI values below thresholds (from 0.05 to 1e-4) for each cell type in mouse brain are available online at “http://java.bactrap.org/bactrap/cells.jsp”.

In this work, the genes are denoted by gene symbols using Entrez CDF files, version 14.\textsuperscript{14} We focus on four cell types in cortex, which are astrocytes, mature oligodendrocytes, neurons, and progenitor cells; as well as eight cell types in cerebellum, which are astrocytes and Bergman glia, mature oligodendrocytes, inner golgi neurons, stellate and basket cells, granule cells, progenitor cells, and Purkinje cells.\textsuperscript{14} We note that the cell types of the same names in cortex and cerebellum were assayed independently in each region. Meanwhile, the closely related astrocytes and Bergman glia of the cerebellum are profiled in a single experiment.
Figure 5. Comparison between the effective gene markers from our method and the gene markers from literature. (A) Analysis 1: temporal changes of the cell types in human cortex development, represented by the computed eigengenes. (B) Analysis 2: Average fold changes (log₂ transformed) in the autism group relative to the controls for the expressions of the gene markers from literature and P values from the t tests.
Temporal transcriptome of human brain (Analysis 1)

The transcriptome data of developing and adult post-mortem human brains was downloaded from GEO: GSE25219. The data set was generated from 1,340 tissue samples collected from 57 human brains in the cortex and cerebellum regions covering 15 time periods (Table 1 in ref. 6) from embryonic development to late adulthood. The IDs of the transcript cluster probes were mapped to RefSeq accessions and then to gene symbols. The resulting data set includes 17,565 protein-coding genes, with expressions in each region and each time period.

Transcriptome of human brain with autism (Analysis 2)

The transcriptome data from the gene expression study of post-mortem human brains was downloaded from GEO: GSE28521. The IDs of the transcript cluster probes were mapped to gene symbols. For those genes with multiple probe sets, probe sets were averaged, resulting in a total of 17,942 genes with expressions in 58 cortex samples (29 autism and 29 controls). In autism or control samples, 13 samples are from temporal lobe of cortex and 16 samples are from frontal lobe of cortex.

Cell type-specific analysis method

Our method consists of the following steps, which are also summarized in Figure 6.

**Step 1:** Filter human brain transcriptome data based on common genes between human and mouse.

**Step 2:** Out of common genes, select genes with highest 20% coefficient of variation (CV) of expression levels across samples.

**Step 3:** Identify genes that are significantly enriched in each cell type relative to all others (cell type specific genes), from cell specific transcriptome profiling experiments in mouse brain.

**Step 4:** For each cell type, overlap the cell type specific genes with the top CV genes. The overlapped genes are prior gene markers.

**Step 5:** For each cell type, perform gene co-expression network analysis on the expressions of the prior gene markers across samples. The hierarchical clustering dendrogram is formed and the effective gene markers are found.

**Step 6:** Test the effectiveness of the gene markers of each cell type by combining them all into a single set and re-performing the clustering.

**Step 7:** Statistical analysis based on the effective gene markers of each cell type.

---

**Temporal transcriptome of human brain (Analysis 1)**

The transcriptome data of developing and adult post-mortem human brains was downloaded from GEO: GSE25219. The data set was generated from 1,340 tissue samples collected from 57 human brains in the cortex and cerebellum regions covering 15 time periods (Table 1 in ref. 6) from embryonic development to late adulthood. The IDs of the transcript cluster probes were mapped to RefSeq accessions and then to gene symbols. The resulting data set includes 17,565 protein-coding genes, with expressions in each region and each time period.

**Transcriptome of human brain with autism (Analysis 2)**

The transcriptome data from the gene expression study of post-mortem human brains was downloaded from GEO: GSE28521. The IDs of the transcript cluster probes were mapped to gene symbols. For those genes with multiple probe sets, probe sets were averaged, resulting in a total of 17,942 genes with expressions in 58 cortex samples (29 autism and 29 controls). In autism or control samples, 13 samples are from temporal lobe of cortex and 16 samples are from frontal lobe of cortex.

**Cell type-specific analysis method**

Our method consists of the following steps, which are also summarized in Figure 6.

**Step 1:** Because cell-specific transcriptome profiles are only available from the mouse brain for most cell types, we start from the assumption that most gene markers’ expressions will be conserved across species. Previous genomic analysis of expression across tissues and species suggest this is a reasonable assumption. We filter the human brain transcriptome data to the subset with clear homologs and data available in both human and mouse using gene symbol. We find 8,380 genes in common for the data set for Analysis 1, and 8,846 genes for Analysis 2.

**Step 2:** Out of the “common genes”, we select the genes having the highest 20% coefficient of variation (CV) of expression levels across samples, as invariant genes will be uninformative as markers. For Analysis 1, the samples are tissue samples of cortex and cerebellum regions covering 15 developing time periods; while for Analysis 2 are cortex tissue samples of 29 autism and 29 controls. CV is a normalized measure of variability: the ratio of the standard deviation of a gene’s expressions in the samples to the mean of the expressions. The list of top CV genes is utilized in Step 4.

**Step 3:** As mentioned, we identify cell type-specific genes from cell-specific transcriptome profiling experiments in mouse brain. Here, we select genes with pSI values below a threshold of 5e-3 for each cell type for further analysis. The use of 5e-3 ensures sufficient cell type-specificity and moderate number of genes.

**Step 4:** For each cell type, we overlap its cell type-specific genes obtained in Step 3 with the top CV genes obtained in Step 2. We denote these overlapped genes as prior gene markers for their corresponding cell type.

**Step 5:** To further uncover the most representative and correlated gene markers and remove uncorrelated noisy genes, we perform gene co-expression analysis for individual cell types, using the weighted gene co-expression network analysis (WGCNA) package in R. In each cell type, the Pearson correlation matrix for the expressions of its gene markers is computed first, and a dissimilarity matrix using the Euclidean distance is calculated based on the correlation matrix. Then average linkage hierarchical clustering is performed on the dissimilarity matrix and the clustering tree (dendrogram) is formed (Fig. 1A; Fig. 2A; Fig. 3A). In the dendrogram, the nodes in bottom row represent the genes, and the nodes in other rows represent the clusters to which the genes belong, with the branches connecting the nodes representing the distances (dissimilarities). The distance between merged clusters is monotone decreasing with the level of the splitting: the height of each node is the intergroup dissimilarity (two
genes with exactly the same expression pattern across samples will have heights of zero).

In the dendrogram of each cell type, by setting a static (fixed) height of cut-off as 1 and finding the largest cluster (module) of genes, we identify the most correlated set of genes, which we regard as effective cell type-specific gene markers. If more than one cluster has the largest number of genes, we accept them all as the effective gene markers.

**Step 6:** To confirm the consistency of the gene markers, we combine the obtained modules in Step 5 of all cell types into a single set and redo the gene co-expression analysis (Fig. 1A; Fig. 2A; Fig. 3A, where the different cell types are coded with different colors). Intuitively, distinct clustering of the gene markers suggests that they are effectively representing the behaviors of distinct cell types. We can quantify this by examining the number of markers in the cell type’s major cluster compared with its total number of potential markers.

**Step 7:** Based on the effective gene markers of each cell type, various statistical analyses can be performed to identify potential changes in cellularity from the human brain transcriptome data.

For Analysis 1, we aim to determine if our method will detect known changes in cellularity across human brain development, by analyzing the temporal transcriptome data. Thus for each cell type we compute an eigengene, which summarizes the temporal expressions of its effective gene markers and is a representative of the cell type’s temporal behavior. Explicitly, in each cell type, the temporal expression profile of each gene marker is scaled so that it has mean 0 and variance 1. Combining the standardized expression profiles of all the gene markers into one matrix, the temporal expression profile of the eigengene is defined as the first principal component of the matrix. Compared with individual gene marker, the eigengene, as a scaled weighted average of gene markers, is less likely to be affected by noise and thus has a better interpretability of the corresponding cell type.

For Analysis 2, we aim to predict relevant cellular alterations that may contribute to human neurodevelopmental disorder of autism, by analyzing the human brain transcript data of autism samples and control samples. Thus for each cell type we compute the expression fold changes (autism/control) of the effective gene markers, and then compute the mean and standard deviation of the fold changes which represents the cellular difference between autism and control. To test the significance of the fold changes’ mean, we perform a paired *t* test of the fold changes of the gene markers compared with the fold changes of genes randomly selected from the “common genes” obtained in Step 1. Each randomly selected “common gene” list is of the same size as that of the gene marker list, and the random lists are selected 1,000 times for 1,000 *t* tests. The *P* values from these permuted *t* tests are averaged in their log scales, and the average is converted back to its normal scale, which we denote as the final *P* value.

**Conclusion**

Our cell type-specific analysis method provides a novel way to systematically integrate high throughput cellular transcriptome profiling experiments in mouse brain and gene co-expression networks to prioritize and select gene markers of individual cell types. The applications of our method to different human brain transcriptome data sets confirm its effectiveness and as a good foundation for further statistical analysis on these data sets to detect cellular composition and alterations in human tissues. Particularly, in this work, this method successfully detected temporal cellular changes that roughly correspond to known epochs of human brain development. We also applied this method to human autism cortical transcriptome data and identified potentially relevant cellular alterations that may contribute to the disorder. Validation of these predictions awaits future anatomical studies in post-mortem tissue from individuals with autism, when available. However, we believe that our method is generally applicable to human transcriptome data of other disorders of the nervous system where the mechanisms are unknown. Furthermore, where appropriate cell-specific expression profiles exist, we believe this method could be adapted to other tissues as well. We hope that the identification of cellular disruptions will provide insights to the mechanisms and help identify cell type based treatments of these disorders.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**Acknowledgment**

This work was supported by the National Science Foundation Grant CIF: IHCS-0963742 and the NINDS (4R00NS067239-03 to JDD). We would like to thank Natalia Rivera and Sukrit Singh for their assistance, and Matthew Tso for his helpful comments.

**Supplementary Materials**

Supplemental materials may be found here: www.landesbioscience.com/journals/systemsbiomedicine/article/25630

---

**References**


