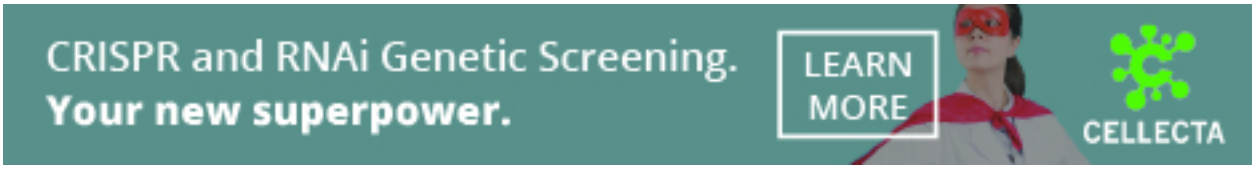
Untangling the effects of cellular composition on coexpression analysis

Marjan Farahbod and Paul Pavlidis

Genome Res. 2020 30: 849-859 originally published online June 24, 2020
Access the most recent version at doi:10.1101/gr.256735.119



Untangling the effects of cellular composition on coexpression analysis

Marjan Farahbod[1,2,3] and Paul Pavlidis[1,2]

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; [2]Department of Psychiatry, University of British Columbia, Vancouver, British Columbia V6T 2A1, Canada; [3]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia V5T 4S6, Canada

Coexpression analysis is widely used for inferring regulatory networks, predicting gene function, and interpretation of transcriptome profiling studies, based on methods such as clustering. The majority of such studies use data collected from bulk tissue, where the effects of cellular composition present a potential confound. However, the impact of composition on coexpression analysis has not been studied in detail. Here, we examine this issue for the case of human RNA analysis. Focusing on brain tissue, we found that, for most genes, differences in expression levels across cell types account for a large fraction of the variance of their measured RNA levels (median $R^2 = 0.68$). We then show that genes that have similar expression patterns across cell types will have correlated RNA levels in bulk tissue, due to the effect of variation in cellular composition. We demonstrate that much of the coexpression and the formation of coexpression clusters can be attributed to this effect for both brain and blood transcriptomes. For brain, we further show how this composition-induced coexpression masks underlying intra-cell-type coexpression observed in single-cell data. An attempt to correct for composition yielded mixed results. Our conclusion is that the dominant coexpression signal in brain, blood, and, likely, other complex tissues can be attributed to cellular compositional effects, rather than intra-cell-type regulatory relationships. These results have implications for the relevance and interpretation of coexpression analysis.

[Supplemental material is available for this article.]

Coexpression analysis is among the most-used methods in transcriptome data interpretation. The biological underpinnings of coexpression are well-established. Within a cell, genes whose products work together (either directly or indirectly) must be expressed together. This implies some commonality of regulation. Indeed, it is observed that genes with similar functions tend to be coexpressed (Eisen et al. 1998; Lee et al. 2004; Langfelder et al. 2011; Gaiteri et al. 2013). Based on these observations, clustering of genes based on coexpression has become a common approach, generally with the aim of predicting gene function and regulation (de la Fuente 2010; Amar et al. 2013; Rotival and Petretto 2013; Li et al. 2016; Saha et al. 2017). However, the effectiveness of these approaches has been questioned (Marbach et al. 2012; Larsen et al. 2019), suggesting there may be a disconnect between the actual biological meaning of coexpression and these common interpretations. Therefore, in this paper, we examine assumptions that underlie such applications of coexpression to "bulk" samples of tissues containing multiple cell types. In particular, we explore the role played by variation in cellular composition in the expression variation and coexpression.

In bulk brain tissue transcriptome data sets, gene expression clusters (sets of genes which are observed to be coexpressed) are often enriched for cell-type markers (Oldham et al. 2008). Recently it has been proposed that variation in cell-type composition between individual samples explains a substantial degree of variation in gene expression in human brain (Kelley et al. 2018). Furthermore, cell-type "deconvolution" methods rely on the idea that cell-type markers can be used to infer cellular composition, as their variation among samples is likely to represent variation in the content of the cell type they are expressed in (Newman et al. 2015; Patrick et al. 2019; Zaitsev et al. 2019). Inferred cellular composition is also used for adjusting statistical models, as in some expression quantitative trait locus (eQTL) analyses (Westra et al. 2015; Ng et al. 2017). Thus, there is at least implicit awareness that cellular composition is a factor in transcriptome data (Gaiteri et al. 2013; Crow et al. 2016). McCall et al. (2016) discuss cell-type heterogeneity in lung bulk tissue as a major source of variation with implications on gene correlation. However, to our knowledge, the connection between these observations and the interpretation of coexpression network analysis has not been described in detail and requires more attention.

In this study, we document the effect of cellular composition variability among samples in bulk tissue on the observed variance among the genes and their coexpression. We also demonstrate its downstream effect on network-based functional analyses. To conduct this study, we use a combination of bulk tissue and single-cell data analysis, supplemented by simulations. We focus on human brain tissue but replicate our key findings in blood transcriptomes as well, suggesting that this is a general phenomenon.

Results

Variance of gene expression is highly affected by variation of cellular composition

Our work builds on two empirically founded concepts. The first is that, for a given tissue, many genes are expressed at different levels

Corresponding author: paul@msl.ubc.ca
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.256735.119.

in different cell types. The second is that tissue samples vary in their precise cellular composition. The latter occurs due to technical (e.g., sampling variability) and biological effects (von Bartheld et al. 2016). The connection between the two in the context of bulk-tissue transcriptomics can be formalized in the following simple model, schematized in Figure 1 (for mathematical details, see the Supplemental Methods S1). For each gene, we define a cell type (CT) expression profile, which is a vector of expression levels of the gene in each of *k* cell types. In the model, the CT profile is treated as a fixed intrinsic feature of the gene. Second, each bulk tissue sample has a specific cellular composition for those same *k* cell types. This forms a cellular composition vector of length *k* for each sample, where each element represents the proportion of a cell type in the sample. The observed expression level of a gene in the sample can be modeled as a weighted sum of the values in the CT profile, where the weights are given by the cellular composition vector of the sample. In the example shown in Figure 1, Gene 1 is only expressed in cell type B and therefore its relative expression in the data precisely tracks the proportion of cell type B present in the samples. This special case is used in many approaches to "cell-type deconvolution," where Gene 1 is considered a "marker gene" for cell type B. In contrast, Gene 5 is expressed equally in all cell types, so it is completely insensitive to differences in cellular composition and its expression level is the same in all the samples. The expression pattern becomes more complicated for a case like Gene 4, which is expressed at different levels in each of the two cell types, but because it is expressed at higher levels in cell type A, its pattern in the bulk tissue is positively correlated with the proportion of cell type A. Furthermore, genes that have correlated CT profiles will also be correlated in the bulk tissue (illustrated in Fig. 1 by Genes 1 and 2, and Genes 3 and 4).

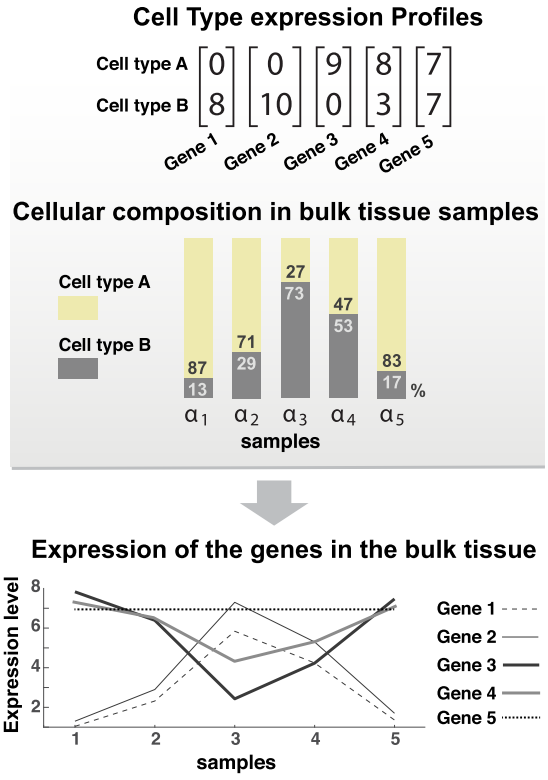


Figure 1. Schematic of cellular composition effects on gene expression variance in bulk tissue. *Top*: Cell type (CT) profiles for five genes in a hypothetical tissue with two cell types. Genes 1 and 2 are marker genes for cell type B. Gene 3 is a marker gene for cell type A. Gene 4 is expressed in both cell types but at different levels, whereas Gene 5 is expressed at equal levels. *Middle*: Hypothetical cellular compositions of five bulk tissue samples. Each sample α_i has the same amount of biological material but different proportions of each cell type. *Bottom*: The expected observed expression levels. Genes 1 and 2 are positively correlated and negatively correlated with Genes 3 and 4. Gene 5 is expressed at the same level in all the bulk tissue samples as it is equally expressed in all cell types.

In general, the model predicts that the more variable the elements in a gene's CT profile, the more its measured expression in bulk tissue will be affected by variability in cellular composition of the samples (see Supplemental Methods S1 for simulation results demonstrating this). It is important to note that this model ignores all other potential sources of variability, including noise or technical artifacts, interactions between genes or cells that can influence expression, and within-cell dynamic transcriptional regulation (the usual interest in doing coexpression analysis). Our goal is to explore how well compositional effects explain the observed variance and correlation of genes in bulk tissue data.

As an initial assessment of whether this model is broadly explanatory, we estimated CT profiles for human cortex from single-nucleus RNA-seq data (sNuc-seq data) (see Methods), yielding expression levels for 16,789 genes in each of 75 different cell types, including all of the major classes of cells expected to be present in bulk cortex. We compared these data to a bulk cortex transcriptome data set from GTEx (GTExBulk) (see Methods). As predicted by the model, the variance of a gene's expression in GTExBulk is correlated with the variance of its CT expression profiles (Spearman's rho = 0.18) (Supplemental Fig. S1). Given the many potential sources of error, including noise in the CT profiles as well as the GTExBulk data, the agreement with the naïve model indicates a strong contribution from compositional effects.

We next applied an approach related to many deconvolution methods to estimate the amount of variance attributable to cellular composition effects for each gene. As demonstrated in Figure 1, expression levels of cell-type marker genes in bulk tissue will reflect the variation of cellular composition among the samples. Therefore, in a given transcriptomic data set, the cellular composition-induced variance of the genes could be modeled by the variation of marker genes. Here, we used principal component regression (PCR) using the expression of marker genes to predict the variation of the nonmarker genes in the GTExBulk data set. In this method, component scores from a principal components analysis (PCA) of marker genes are used as explanatory variables to model the expression level of nonmarker genes (see Methods). The amount of variance explained by the model for each gene (R^2) is an estimate of the degree to which the gene's expression pattern is due to variability in cellular composition. In our first analysis, we used sets of genes that we identified as markers based on their average expression levels in different cell types in a high-quality sNuc-seq data set for five major brain cell types (Pyramidal, Microglia, Astrocyte, Oligodendrocyte, and Endothelial) (see Methods). The resulting gene-level values of R^2 range up to 0.97 (90th percentile is 0.85) with a median of 0.68 (see Supplemental Fig. S2 for distribution). In contrast, when we fit the model to sNuc-seq data, where we expect no effect of cellular composition (barring contamination of individual nuclei), the mean R^2 is 0.041, with only 135 genes having values greater than 0.25 (tested on cell type "Exc L2Exc L2-3 *LINC00507 FREM3*").

As predicted by our model, R^2 values are correlated with the variance of the CT expression profiles (rho = 0.28). To check the

robustness of these findings, we tested another set of (largely nonoverlapping) marker genes from Mancarci et al. (2017) with similar results (rho = 0.3) (see Supplemental Methods S2). The R^2 values from the two sets of marker genes are highly correlated (rho = 0.96) (see Supplemental Fig. S3). We also tested randomly selected sets of nonmarker genes instead of markers and found that R^2 values are significantly higher for the marker genes when PCs are obtained from marker genes compared to the random selection of genes with similar average expression levels ($P < 0.01$ for average of R^2 of marker genes for 100 trials) (see Supplemental Fig. S4). Likewise, the two marker sets also generated higher R^2 values for each other than the random gene sets despite their small overlap.

Motivated by reports that coexpression clusters are often associated with tissue-relevant gene functions, we next examined the relationship between gene function and cellular expression patterns. We observed that genes associated with brain-related functional terms (see Methods) tend to have higher R^2 values, consistent with expected cell-type-specific expression patterns in the brain (see Fig. 2A). That is, genes with a brain-related function tend to have more varying CT profiles—they are enriched in particular cell types—which leads to high variation in bulk tissue. For example, genes involved in synaptic transmission are expressed in neurons, whereas genes involved in myelination are expressed in oligodendrocytes. Examples are genes annotated with "Regulation of synaptic plasticity" (GO:0048167, mean $R^2 = 0.76$) and genes annotated with "Axon ensheathment" (GO:0008366, mean R^2: 0.68) (see Supplemental File S1). In contrast, terms for housekeeping functions tend to be associated with genes with lower R^2 values (examples: "Histone demethylation," GO:0016575, mean R^2 value = 0.61; "Spliceosomal snRNP assembly," GO:0000387, mean R^2 value = 0.53) (see Supplemental File S2). In a closer examination, we also see that genes associated with the brain-specific term "Regulation of synaptic plasticity" have significantly higher variance in the GTExBulk data set compared to genes associated with the housekeeping term "Histone demethylation" ($P = 0.005$, t-test). In contrast, in the sNuc-seq cell population (a data set expected to not have cellular composition effects), they have significantly lower variance ($P = 4.0 \times 10^{-4}$) (see Fig. 2B). In summary, these results demonstrate that some of the observed variance of genes can be attributed to cell-type composition variation, and this is especially true for genes with tissue-specific functions due to their tendency to also have cell-type-specific expression patterns.

Much bulk tissue coexpression is explained by cellular composition variation among samples

In the previous section, we demonstrated that variation in gene expression can be partly accounted for by variation in cellular composition. As illustrated in Figure 1, genes which have similar patterns of expression across cell types (as evidenced by correlated CT profiles) are also expected to have correlated expression in bulk tissue. It is noteworthy that this phenomenon will be observed for any gene that has variability in expression across cell types, not just highly cell-type-specific marker genes. For any two genes, in the absence of other factors, as the correlation between their CT expression profiles approaches one (or minus one), their correlation in bulk tissue is expected to approach one (or minus one) (see Supplemental Methods S1; Figs. 1, 3C). We call this "cellular composition-induced coexpression," to be distinguished from coexpression due to "within-cell" co-regulation. We hypothesized that it is a major source of observed coexpression in bulk tissue.

We first performed clustering of the GTExBulk gene expression profiles, yielding 69 clusters (minimum 20 genes each) (see Methods). As expected, some clusters are enriched with markers for one cell type (Fig. 3A; see Methods). Although many of the other genes in these clusters are not markers, they tend to be "quasi-markers"; they are enriched in expression in a cell type

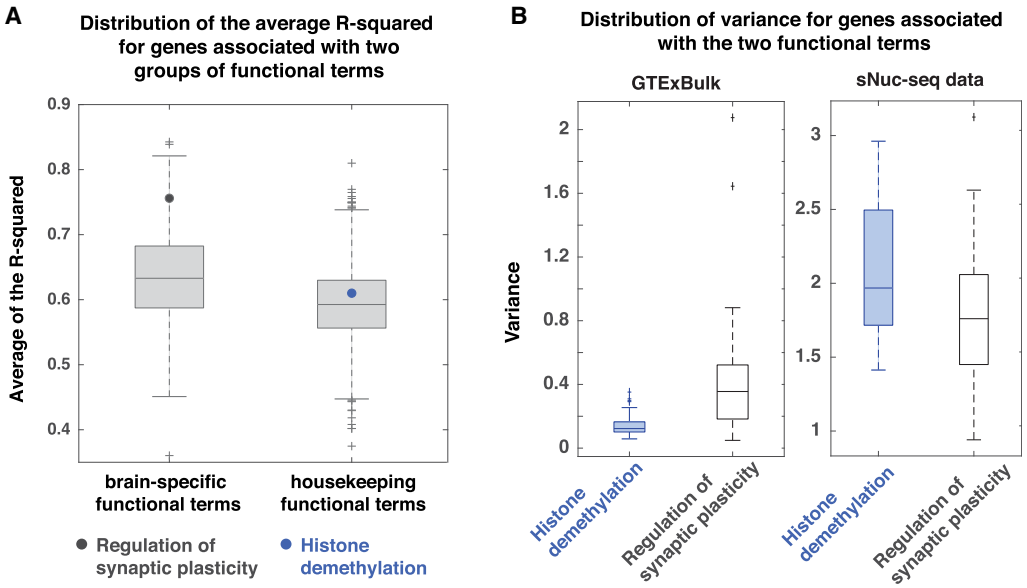


Figure 2. Much of the observed variance of brain-specific genes in bulk tissue is explained by cellular composition effects. (*A*) Groups of genes associated with brain-specific functional terms tend to have higher cellular composition-associated variance (indicated by higher average R^2 values) compared to groups of genes associated with housekeeping terms. The average R^2 values for two example terms are highlighted with black and blue dots. (*B*) Distribution of gene-level variance in the GTExBulk and sNuc-seq cell population (Exc L2-3 *LINC00507 FREM3*) for two groups of genes. Genes associated with brain-specific term "Regulation of synaptic plasticity" have higher variance than genes associated with the housekeeping term "Histone demethylation" in the GTExBulk data set, whereas they have slightly but significantly lower variance in the sNuc-seq cell population.

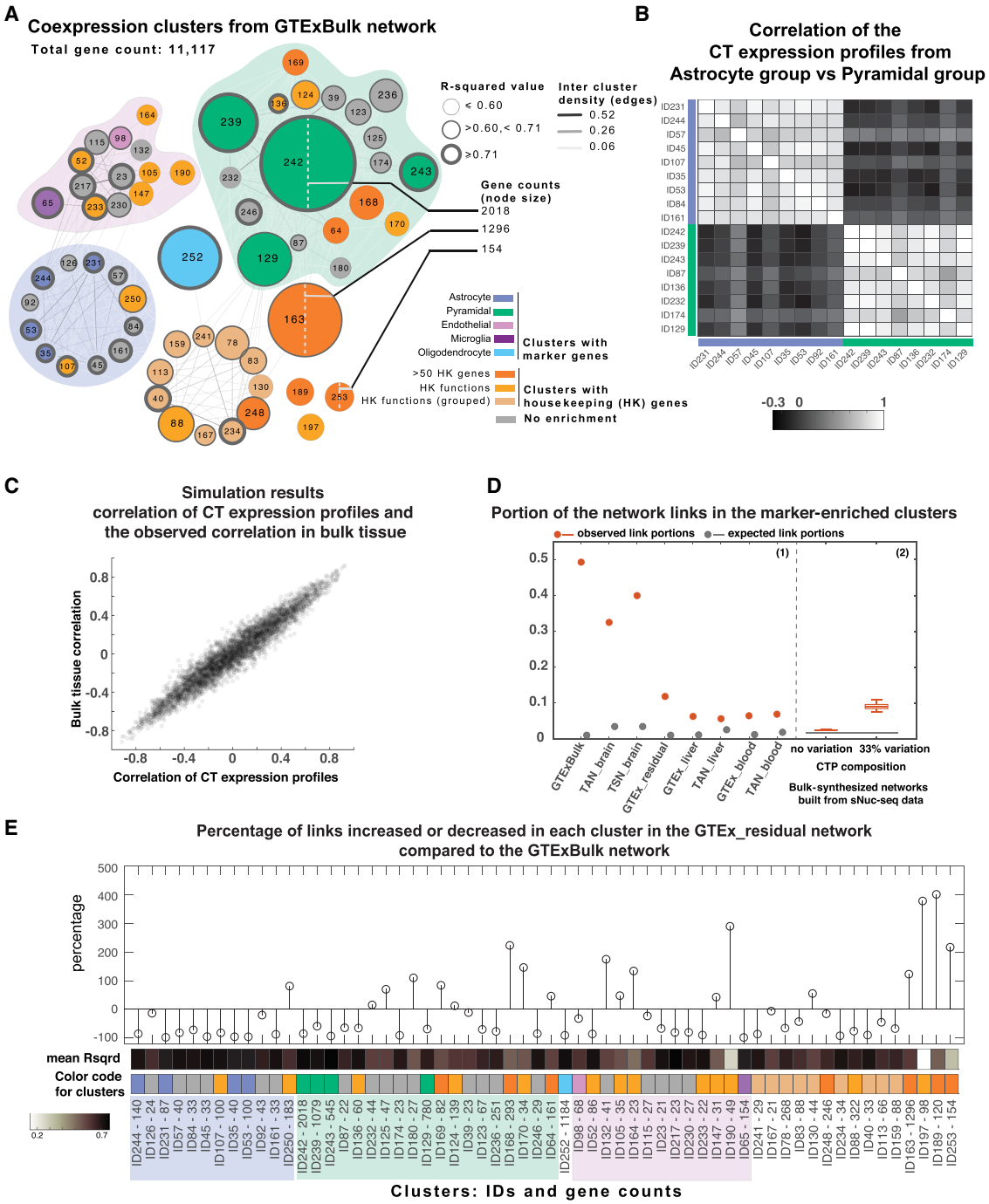


Figure 3. Much coexpression is explained by cellular composition effects. (*A*) Coexpression clusters from GTExBulk data set network. Clusters are labeled with their IDs. Color indicates if the cluster was enriched with markers of a specific cell type, is enriched with housekeeping functional terms, or has a high count of housekeeping genes. Thickness of the border reflects the mean R^2 value for genes in the cluster. (*B*) Coexpression of mean CT expression profiles for a group of clusters affiliated with Pyramidal cell expression patterns and a group of clusters affiliated with Astrocyte cell expression patterns (affiliation is indicated with presence of markers, high R^2, or high inter-cluster links). (*C*) Results from simulated bulk tissue data. Each dot represents data from a pair of genes. Plot shows data for 1000 gene pairs, sampled from a bulk tissue data set with 100 samples and 10 hypothetical cell types. As demonstrated, for a given pair of genes, their Pearson's correlation in the bulk tissue is highly correlated with the Pearson's correlation between their CT expression profiles. Also, the higher the correlation between their CT expression profiles, the more likely their correlation in the bulk tissue is the same as the correlation of their CT expression profiles. (*D*) Proportion of coexpression involving the set of genes from the clusters enriched with marker genes. Panel 1: The two brain networks (GTExBulk and the TAN-brain) and the brain-specific network (TSN-brain) have between ~30% and 50% intra-cluster links in clusters enriched with marker genes. Panel 2: Portion of links in the same set of clusters in two groups of synthesized bulk tissue networks, modeling the effect of cellular composition variation. (*E*) Percentage of links increased or decreased for GTExBulk clusters in the residual network (GTEx_residual). Cluster color code as in *A*. Clusters are ordered based on their grouping in *A*. The range for average R^2 is shown in the color bar. Many housekeeping clusters (orange) with low average R^2 values yield more links in the residual network.

(as for Genes 1, 2 and 3, 4 in the simplified model) (Fig. 1; see Supplemental Fig. S5 for expression levels of genes from clusters associated with different cell types, compared to marker genes). Furthermore, clusters that are enriched for markers for the same broad cell types and most of their neighboring clusters have correlated average CT expression profiles ($0.42 < \text{rho} < 0.993$, all $P < 5.0 \times 10^{-4}$) (Fig. 3A,B). In addition, the average R^2 values from the regression model are generally high for the marker-enriched clusters, consistent with composition-induced variance in expression (9/10 marker-enriched clusters have an average R^2 greater than the median for all clusters [median = 0.65]) (Fig. 3A; see Supplemental File S3 for information on each cluster).

The results so far make it apparent that some of the observed coexpression in the bulk brain tissue is explainable by cellular composition variation. Since cell-type-specific patterns of expression are likely to be relatively fixed and therefore reproducible, composition-induced coexpression is also likely to be reproducible and therefore contributing to the reported reproducibility of coexpression clusters among different bulk brain data sets. We examined this by comparing brain bulk tissue coexpression networks with each other and also with coexpression networks from other tissues. In the GTExBulk coexpression network, the intra-cluster coexpression links in the clusters enriched with brain marker genes contain 49% of the total links. This is up to 40 times higher than the null expected value for the count of genes in the clusters for the given density of this network (see Fig. 3D). The same set of genes has a high count of links in multi-data set brain coexpression networks (TAN-brain and TSN-brain) (see Methods) we previously described in Farahbod and Pavlidis (2019). This confirms that much of the reproducibility among bulk brain networks can be explained by cellular composition-induced coexpression. Also, although most (60%–80%) of the genes in these clusters are also expressed in blood and liver, the high degree of observed coexpression among these particular genes is a phenomenon specific to the brain. We also see a large increase of links between the genes in marker-enriched clusters in our simulated bulk tissue data upon the introduction of cellular composition variation (Fig. 3D; see Methods for details).

Apart from the marker-enriched clusters, many clusters in the GTEx-derived network are enriched with housekeeping genes and/or functions (see Fig. 3A; see Supplemental File S4 for enrichment of functional terms in clusters). Most of these clusters have low mean R^2 values (18/28 have a mean R^2 less than the median of all clusters [0.65]) (see Supplemental File S3), suggesting that their genes have small variability in their CT expression profiles, and their coexpression is less likely to be affected by the cellular composition variation (like Gene 5 in Fig. 1). We hypothesized that some of the coexpression signal among genes from these clusters could have remained obscured due to the prevalence of high correlation values induced by cellular composition variation among other genes. To investigate this, we compared counts of links in different clusters in GTExBulk with counts of links in a coexpression network built from the residuals of the PCR fits in the data set (GTEx_residual network). This could be considered as a form of correction for the cellular compositional effect in the data set. We observed large drops in the count of links in the marker-enriched clusters and an increase in the count of links in clusters with low R^2 values (Fig. 3E). The magnitude of these changes highlights how cellular composition-induced coexpression can mask underlying coexpression within cell types. We discuss the use of the GTEx_residual network as a "corrected network" in the next section.

Cellular composition effects can mask underlying intra-cell-type coexpression

We have shown that a major coexpression signal in bulk tissue comes from cellular composition effects. In our view, this presents a shift from the usual interpretation and raises the question of whether there is substantial coexpression attributable to other sources. This is especially relevant to attempts to infer coregulation. Specifically, the question remains as to whether coregulatory relationships in the sense typically sought are "visible" in bulk tissue data in the background of cellular composition effects. We do not attempt to fully address this question here but instead concern ourselves with a simpler one: In the common modes of coexpression analysis of bulk brain tissue, are coexpression patterns present within a cell type detectable? Composition-induced coexpression could, in principle, mask or amplify the bulk-tissue visibility of intra-cell-type coexpression. In this section, we examine the difference between robust intra-cell-type coexpression as measured in the sNuc-seq data and the observed coexpression in GTExBulk and show that much of the difference can be attributed to the cellular composition effect. We also examine the GTEx_residual network as a form of "corrected network" in retrieving intra-cell-type coexpression.

As a preliminary step, we examined the general agreement of the coexpression networks built from different sNuc-seq populations and the GTExBulk data set and found that the agreement of network links is up to two times (and in few cases, 3–5 times) higher for most of sNuc-seq populations than that expected by chance (see Supplemental Fig. S6). For reference, for our two bulk brain networks, TAN-brain and GTExBulk, the agreement of the networks is 14 times higher than that expected by chance. Conversely, some of the observed coexpression clusters in the GTExBulk are also reproducible in the larger sNuc-seq cell populations (see Supplemental Fig. S7), including those of many of the housekeeping and some of the brain-specific clusters. This shows that there is some level of agreement between coexpression observed in sNuc-seq and bulk data, in agreement with prior work (Crow et al. 2016).

We then hypothesized that some of the differences between the observed coexpression in sNuc-seq data and GTExBulk could be explained by the cellular composition effect in a way that is shown schematically in Figure 4A. To test this, we compared bulk tissue coexpression with robust intra-cell-type coexpression patterns and found that, for the most part, differences between the two are explained by the effect of cellular composition variation. To identify robust intra-cell-type coexpression patterns, we combined 64 sNuc-seq coexpression networks built from neuronal cell types (both excitatory and inhibitory) and obtained a consensus (sum) "intra-cell-type" network (see Methods and Supplemental File S5 for list of links). We focused on the highest-confidence set of links that were present in 10 or more networks, leading to a set of 464 genes with 7678 highly robust links. Of these links, 32% have correlation values above the 95th percentile in the GTExBulk network, but only 63% have correlation values above the median. This subnetwork forms two very distinct clusters (Fig. 4B). The gray cluster is enriched with multiple functional terms associated with neuronal processes (see the complete list of functions in Supplemental File S6), and the black cluster is enriched with a few housekeeping functions. We then looked at the coexpression of these genes in the GTExBulk network (Fig. 4C). Although the genes in the two clusters have partly distinguishable coexpression patterns in GTExBulk, a large number of

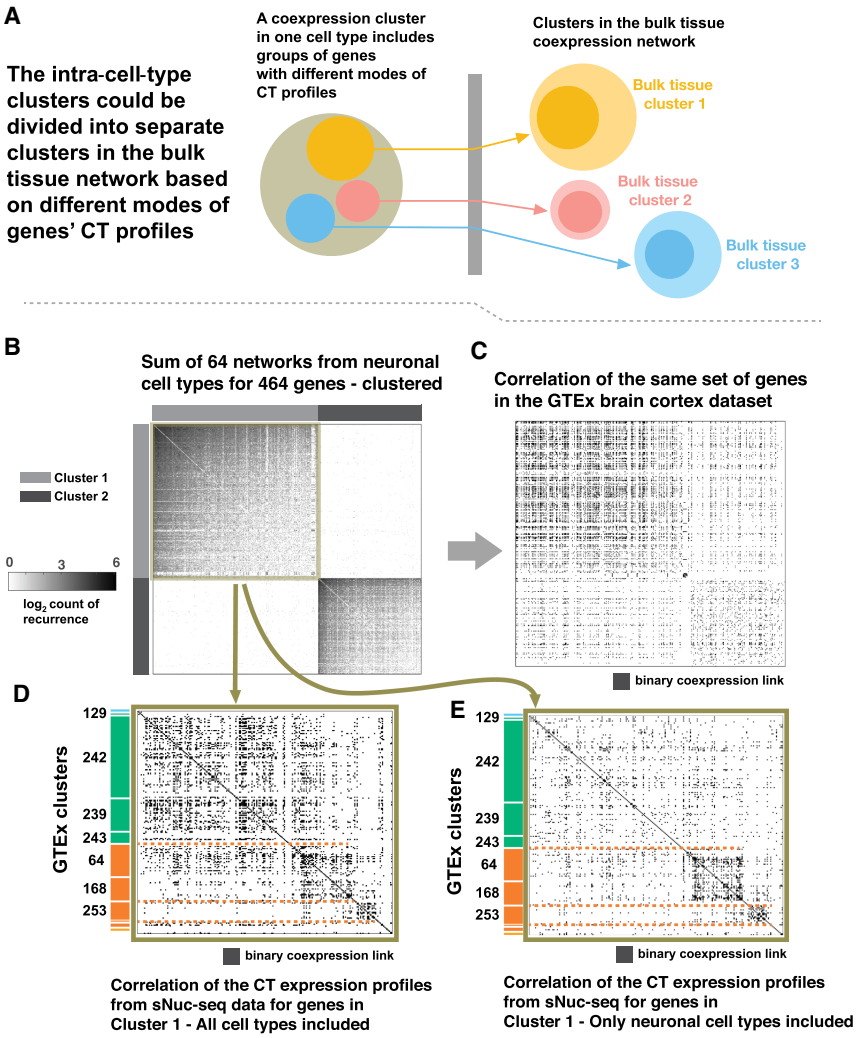


Figure 4. Modes of CT expression profiles shape gene clusters in the bulk tissue. (*A*) Schematic showing a coexpression cluster in a specific cell type could be divided into multiple clusters in the bulk tissue data set, as its genes might have different CT expression profiles. Each circle represents a group of genes. Blue, yellow, and coral represent different "modes" of CT expression profiles, similar to the mean CT expression profiles for the bulk tissue clusters in Figure 3. (*B*) The heat map shows part of the sum network from 64 neuronal sNuc-seq data sets where two coexpression clusters are identified. Clusters 1 and 2 (color bar gray and black) are well distinguished from each other. (*C*) Heat map showing the same set of genes with the same order in the GTExBulk coexpression network. Although the two clusters are somewhat distinguished, a great amount of inter-cluster links is present. (*D*) The heat map shows the network for the genes in Cluster 1, from the coexpression network built from the correlation of the CT expression profiles obtained from the 75 sNuc-seq data sets. Genes are ordered based on their belonging to different GTExBulk clusters, identified by the color bar and cluster IDs from Figure 3A. Three subclusters are mildly distinguished, separating two groups of housekeeping clusters from the Pyramidal clusters (orange vs. green). (*E*) Same plot as *D*, but the CT expression profiles are obtained from the 64 neuronal cell types only. The distinguished clusters demonstrate the group of genes with different expression levels in the neuronal cell types.

inter-cluster links are present; that is, the clusters are not as clearly separated. Indeed, the 464 genes appear in multiple clusters in the GTExBulk network (Fig. 4D). We hypothesized that this is due to similarity of CT expression profiles among some of the genes, causing composition-induced coexpression that "blurs" the underlying cell-type-specific coexpression pattern. In support of this hypothesis, correlations of the CT profiles are high for some of the genes (Fig. 4C,D). In particular, this can explain the differentiation between the two clusters ID253, ID168 and excitatory cell clusters (indicated by the green color bar in Fig. 4D; these are the clusters enriched with Pyramidal markers; see Fig. 3 for reference). The differentiation is even clearer when CT expression profiles are obtained from neuronal cells only (Fig. 4E), indicating different expression patterns among neuronal cell types for genes in clusters ID253, ID168, ID64 and excitatory clusters (Fig. 4E). Our conclusion is that the intra-cell-type coexpression patterns observed in single-cell data can be distorted and/or masked in bulk tissue by the effects of cellular composition.

In the previous analysis, we showed that cellular composition effects can mask intra-cell-type coexpression especially when there is a conflict between the correlation of the CT expression profiles and the intra-cell-type coexpression, resulting in loss of the intra-cell-type pattern. In general, there are various scenarios that could occur, and intra-cell-type coexpression patterns might happen to be observed in bulk tissue to varying degrees and for varying reasons. Here, we demonstrate this complexity with two genes, *CALM3* and *NRGN* (Fig. 5). They are robustly correlated in the sNuc-seq excitatory neurons (a link is present in 11 out of 23 of the networks for excitatory neurons), but there is no correlation between them in inhibitory neurons, since *NRGN* is not expressed in inhibitory neurons (Fig. 5A,B). Accordingly, they have relatively highly correlated CT expression profiles (rho = 0.46), driven by their high expression in excitatory neurons and close to zero expression in the nonneuronal cell types but moderated by their disjoint expression in inhibitory neurons. This suggests that they might be coexpressed in bulk tissue but for a reason different from that driving their coexpression within excitatory cells. As it happens, their correlation in bulk tissue is relatively high (97th percentile) (Fig. 5C) but not nearly high enough to pass our original 99.5 percentile filter for link selection. Their correlation ranks drop to the 82.3 percentile in the GTEx_residual network. We conclude that the observed coexpression of *CALM3* and *NRGN* in the bulk tissue is primarily caused by correlation of their cell-type expression profiles, rather than a reflection of their coexpression in excitatory cells. Also, although their coexpression in the bulk tissue resembles their coexpression in excitatory cells, it is in disagreement with their lack of coexpression in other nonneuronal cell types. In general, there is no simple relationship between coexpression within a cell type and coexpression in a tissue in which that cell type is one of several present.

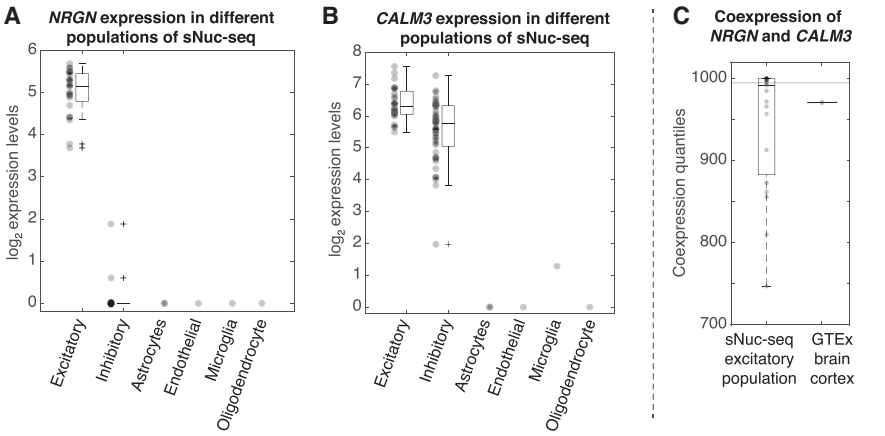


Figure 5. Coexpression of *NRGN* and *CALM3* in Excitatory cell types. (*A*) From sNuc-seq data: *NRGN* is only expressed in the Excitatory cell types. (*B*) From sNuc-seq data: *CALM3* is expressed in both Inhibitory and Excitatory cell types. (*C*) The two genes are highly correlated in the Excitatory cell types—based on coexpression networks built from sNuc-seq data. They are correlated in the GTExBulk data set but do not meet the threshold for the network (threshold is marked by the red line; it is the 99.5 percentile).

Given that the effects of cellular composition on coexpression can be viewed as a confound, it is natural to consider whether the data can be corrected. In the previous section, we observed that many of the clusters from GTExBulk had a much higher count of links in the residual network. In our framework, a natural choice for such a correction are the residuals from our PCR model fits used to obtain the R^2 estimates. We observe that most of the GTExBulk clusters are significantly reproduced in the GTEx_residual network (see Supplemental Fig. S7), and many of the brain-specific and housekeeping terms are enriched in the GTEx_residual network (see Supplemental File S7). However, there is no overall significant improvement in agreement of the links in sNuc-seq populations with the GTEx_residual compared to the GTExBulk network (Supplemental Fig. S6), indicating that correction for composition may not be a panacea. However, there is improvement of the precision in recovering sNuc-seq coexpression links for some of the GTExBulk-driven clusters. Results for the largest sNuc-seq population are presented in Supplemental Figure S8. As an example, we can consider the neuronal clusters with IDs 239, 242, and 243. These clusters are enriched for pyramidal cell markers, and their genes have high mean R^2 values of 0.72, 0.74, and 0.75, respectively, suggesting that correction will have a large effect. In GTEx_residual, the link density among the genes in these clusters decreases to approximately 1/2, 1/6, and 1/14 of the density in the GTExBulk network, whereas the link overlap with the sNuc-seq cell populations increases 24%–175%. We also see that, for the most robust sNuc-seq clusters (from Fig. 4B), which were substantially degraded in GTExBulk (Fig. 4C), their separation is largely restored in the GTEx_residual network (Fig. 6A–C). However, the restoration is mostly due to the removal of inter-cluster links, rather than an increase in the intra-cluster links. These findings suggest that correction for cellular composition effects can be beneficial, but confidence in the results is uncertain without matched single-cell data.

Cellular compositional effect in whole-blood transcriptomic data

The phenomenon we describe in brain might be expected to occur in any complex tissue, to the extent that there are differences in cell-type expression patterns and variations in cellular composition. To test whether our results generalize, we conducted a similar analysis of blood transcriptomes. For "markers," we used a set of genes differentially expressed among blood cell types provided by Newman et al. (2015). For the blood CT profiles, we used a cell-type transcriptomic data set from Abbas et al. (2005) (see Methods for details). We estimated compositional effects in three whole-blood data sets (GTEx as well as two microarray data sets we called DS1 and DS2). Fitting models using PCR as described above, the R^2 values (estimates of compositional effects at the gene level) had a median of 0.78, 0.48, and 0.43 in the three data sets, respectively (see Supplemental Fig. S9 for distributions). The per-gene

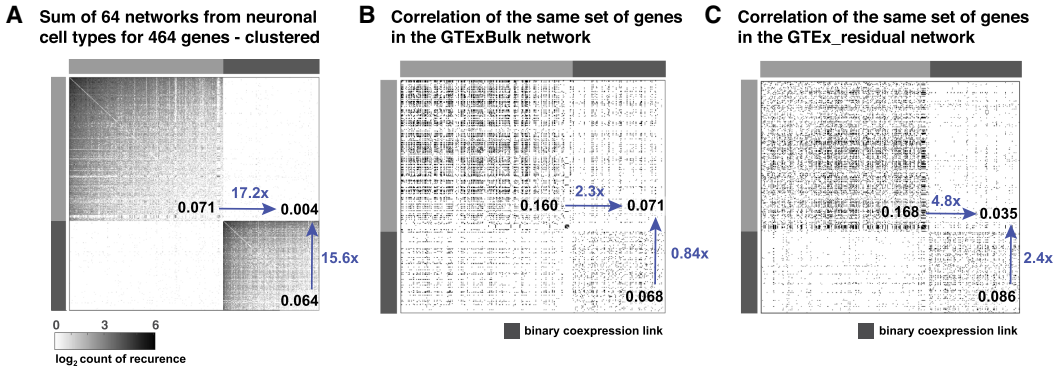


Figure 6. Reproducibility of robust correlation clusters from sNuc-seq networks in GTExBulk and the residual network. Inter- and intra-cluster density values are overlaid in black, and fold-changes in blue. The null density is 0.005 for all the networks. (*A*) Sum network as in Figure 4B. For visualization, the data are scaled to be comparable to *B* and *C*. Intra-cluster density is more than 15 times the inter-cluster density for both black and gray clusters. (*B*) As in Figure 4C. Intra-cluster density is less than the inter-cluster density for the black cluster; for the gray one it is less than 2. (*C*) Intra-cluster density is more than 10 times higher than the inter-cluster density for both black and gray clusters. Notice that the differences in the density (from GTExBulk to GTEx_residual network) is mostly explained by the relative reduction of inter-cluster links rather than an increase in the intra-cluster links.

R^2 values were correlated with each other across all three data sets ($0.30 < \text{rho} < 0.37$ for all three comparisons). Similar to the brain, we found that variance of blood CT expression profiles is correlated with the gene variance in bulk tissue (rho for DS1 = 0.18, DS2 = 0.35, and GTEx-whole-blood = 0.3) and with the R^2 values (rho for DS1 = 0.19, DS2 = 0.34, GTEx-whole-blood = 0.09). As for brain, the effect of cellular composition appears widespread in blood, contributing to formation of most of the coexpression clusters (Supplemental Fig. S10) and is also associated with relevant patterns of functional enrichment (Supplemental File S8).

Discussion

The term "coexpression" refers to two tightly linked concepts, one defined in the realm of molecular biology as the coordinated transcription of genes by regulatory mechanisms occurring within a cell (coregulation). The other is an observation of correlation of RNA levels in transcriptomic data; for clarity, we can refer to the latter as "observed coexpression." Coregulation has formed a foundation for understanding genome function for decades: that genes with collaborating products need to be expressed at the same time and are thus coregulated. Meanwhile, coexpression detected in high-throughput data sets has proven to be a reproducible signal with biological relevance: Genes which are coexpressed have a higher probability of having related function than those which are not coexpressed. Based on this, it is often assumed that observed coexpression is due to coregulation, and therefore the latter could be reverse-engineered from the former—despite limited success in benchmarking (Marbach et al. 2012). Our work sheds light on the challenge of inferring regulation by providing a deeper understanding of the causes of observed coexpression in bulk transcriptomic data.

We have demonstrated that, for a given gene, the variance of its expression level in bulk tissue is directly related to its variability across cell types. We then showed that this is strongly related to coexpression of genes with each other, such that the dominant signal in bulk tissue is simply due to variation in cellular composition across samples. Because many gene functions are highly associated with specific cell types, our results provide one reason why clusters enriched for functions are observed in expression data. A further implication is that the utility of bulk tissue coexpression to infer transcriptional regulatory networks beyond uncovering cell-type-specific expression patterns is greatly complicated. Although our study was mostly focused on expression in the human nervous system, the results of the blood analysis indicates similar effects, indicating the phenomena we document are likely to affect coexpression analyses in other tissues as well.

Coexpression is highly reproducible—that is, there are strong patterns of coexpression that are observed in many independent data sets for a given tissue (Oldham et al. 2008; Farahbod and Pavlidis 2019). Our results suggest that this is mostly due to the reproducibility of cell-type expression patterns, and the pervasive presence of variability in cellular composition between samples (Figs. 1, 3). We have shown that, when coexpression is observed across cell types or tissues, the dominant patterns are due to cell-type- or tissue-specificity of expression, and coexpression is merely a proxy for differential expression across cell types or tissues. Although genes which are expressed specifically in one cell type (for example) can be thought of as having a "shared function," that function is broad, only reflecting the function of that cell type. There is little expectation that function at the level of individual molecular interactions or pathways would be captured: The distinctness of a cell type cannot be fully described by the activity of a single pathway. Likewise, even for these genes, their coregulation may reflect the broad epigenetic state of the cell type (Yoshida et al. 2019), and finer-grained details of coregulation are unlikely to be easily captured.

We have also shown that cellular composition-induced coexpression can mask apparently robust cell-type-specific coexpression patterns (Fig. 4). Despite this, a remaining question is whether correction for cellular composition would enable more efficient extraction of coregulation. For this to be the case, underlying patterns due to coregulation would have to be present in the data and sufficiently separable from cellular proportion effects. For this to be effective, the regulation should ideally not be cell-type-specific (otherwise, the signal would be that much weaker) (Fig. 6), so the genes involved would have to be expressed in most cells. Since genes which are not cell-type-specific tend to have housekeeping functions, it stands to reason that the most apparent coregulatory relationships would be those among housekeeping genes. We note that schemes for correcting bulk tissue data for cell-type proportions (either directly or indirectly) are often used in expression QTL studies and have been shown to increase the number of *cis*-eQTLs that can be recovered (Ng et al. 2017). This suggests that correcting for cell-type proportions and recovering underlying biological signals is possible, but eQTL studies require large sample sizes (generally at least 100 but often far more, especially for *trans*-eQTLs). We expect that identification of coregulatory relationships from bulk tissue data will similarly require very large sample sizes and still be most effective at extracting regulation of housekeeping genes rather than cell-type-specific genes. Given these constraints, it would seem preferable to use coexpression data from a single cell type to extract regulatory relations. However, limitations of the most commonly used single-cell transcriptome methods suggest extracting high-quality regulation information is a challenge (Crow and Gillis 2018). Furthermore, the most commonly used computational method for doing so is designed to simultaneously identify cell types along with building a regulation network, so that the strongest patterns observed are likely dominated by differential expression across cell types, not coexpression within cell types (Aibar et al. 2017).

Our study does have some limitations. First, our analysis of cell-type-level coexpression is based on a single (albeit large and unusually deeply sequenced) data set that used different samples than the bulk tissue. Thus, we cannot rule out that the failure to recover some sNuc-seq coexpression patterns in bulk tissue might reflect data-specific effects. This might be resolved in the future with additional data sets. Second, we only considered the phenomenon in two tissues, and determining whether the inferred effects in other tissues are weaker or stronger than those we observe for brain and blood should be a topic of future research. Finally, the actual cellular composition of the bulk tissue samples we used is not known. Although the approach of using cell-type markers to infer composition has been validated many times (Newman et al. 2015; Mancarci et al. 2017; Patrick et al. 2019), we do not claim it is a perfect substitute for accurate direct counts. It remains formally possible that some of the variation we attribute to cellular composition is instead due to complex patterns of gene regulation that mimic compositional effects, but we feel the most parsimonious interpretation of the data is that cellular composition is a major contributor. It is also worth noting that imperfect cell-type-effect measurement could just as easily cause us to underestimate the impact of composition, as the residual would still contain compositional effects.

Beyond the implications for the goal of inferring regulation, our results have implications for any use of expression data-based gene clustering or module identification in which the patterns are driven by cellular composition effects. First, the representation of the data as a network is potentially misleading, because it is tempting to interpret a network as representing physical relationships. In particular, the idea that "hubs" in coexpression models are especially interesting is highly questionable if that pattern is simply a reflection of the cellular distribution of those transcripts. Second, if cellular composition is of interest, it would be reasonable to analyze composition more directly by inspecting the expression of known markers rather than by using indirect means via clustering and enrichment analysis. This parallels the situation for analysis of differential expression, where changes in measured expression levels can be due to changes in composition (Mancarci et al. 2017; Toker et al. 2018). On the other hand, machine learning applications of coexpression to tasks such as gene function prediction are not directly affected by our findings, as success in prediction does not necessarily depend on the biological meaning of the features used.

Methods

Data

We have three main sources: (1) a single-nucleus data set from Allen Brain Atlas from Middle Temporal Gyrus (available from: Allen Institute for Brain Science. Cell Diversity in the Human Cortex. [https://portal.brain-map.org/atlases-and-data/rnaseq#transcriptomics]); (2) GTEx RNA-seq expression data set from brain-cortex (Lonsdale et al. 2013); and (3) a set of coexpression networks: binary coexpression networks were built from GTEx RNA-seq blood and liver and a set of Tissue Aggregated Networks (TANs) from blood, brain, and liver, from our previous study (Farahbod and Pavlidis 2019). The TAN networks are built by aggregating several networks from each tissue, built from data sets on an Affymetrix platform. We also use the TSN-brain network from the same study. TSN-brain is a subset of the TAN-brain network, where the links are identified as specific to the brain among the five tissues using a total of 53 data sets. Supplemental Table S1 provides counts of genes and links in each of the networks. TSN and TAN networks, as well as all the scripts related to the manuscript, are available at GitHub (https://github.com/PavlidisLab/CellularComposition) and as Supplemental Code.

Single-nucleus data

The single-nucleus (sNuc-seq) data set has records from 15,928 nuclei for a total of 50,281 genes, grouped into 75 cell types (see Supplemental Fig. S11 for counts of samples for each cell type). We used the read counts from exons only and did not use the intronic reads. We used the labels for the cell types based on the clustering provided by the Allen Institute. We removed nuclei which had data for less than 2000 genes and nuclei for which the total read count was more than three times or less than 1/3 the median. Genes were filtered for *RBFOX3* negative and *RBFOX3* positive samples separately. We selected genes expressed in at least 2% of the nuclei or expressed at the highest quartile in the nuclei in which it is expressed. The final data set has data for 16,789 genes and 15,646 nuclei. Supplemental Figure S11 shows the count of cells in each group of the 75 cell types. To construct cell type vectors for each gene, we obtained the mean expression level of that gene in each of the 75 cell types. Each of the 16,789 genes yields a CT expression profile vector of 75 elements.

We built coexpression networks for 69 of the 75 cell types in the sNuc-seq data set (six cell-types had too few cells). For each cell type, correlations were computed for each pair of genes using only nuclei in which expression was greater than zero to reduce the impact of zeros due to data being left-censored. Gene pairs with less than 20 usable nuclei were removed. Because of differences in sample size for the correlations (causing different null distributions for the correlation), we omitted the correlation threshold filtering step used for the other data sets and therefore filtered the one-sided *P*-values of the correlations to identify the 0.5% of the gene pairs with the smallest *P*-values. Supplemental Table S2 has the link count and gene count for the 69 networks.

To construct combined networks, we summed the 64 binary coexpression networks built from inhibitory and excitatory neurons. Robust coexpression links were identified as those present in 10 or more of the networks, between genes with more than two such links. A total of 490 genes passed this criterion and were clustered using topological overlap and hierarchical clustering. Sixteen mitochondrial genes and 10 unclustered genes were removed (the presence of mitochondrial genes is likely due to variable mitochondrial contamination of the nuclei). The remaining gray and black clusters have 286 and 178 genes, respectively.

GTEx data sets and networks

We used GTEx data set version 6. The read counts per million reads (CPM) values from each of the three GTEx data sets—brain-cortex (sample count: 114), liver (sample count: 119), and blood (sample count: 393)—were filtered to include the genes with CPM > 0 in >20% of the samples. Expression values were $\log_2$ transformed, and binary coexpression networks were built using the Pearson's correlation, filtered to include the 0.5% of the links with the highest correlation values in each of the three networks. The counts of links and genes included in each network are provided in Supplemental Table S1. To cluster the GTExBulk network, we applied hierarchical clustering to the Topological Overlap (TOP) (Zhang and Horvath 2005). An initial set of 253 clusters were identified for 12,416 of genes, of which 69 clusters had at least 20 genes and were retained for further study. Clustering labels are in Supplemental File S9. Figure 3, A and E, provides information for 60 of the clusters. The remaining 10 clusters are included in Supplemental File S3.

Housekeeping genes

Housekeeping genes were obtained from Eisenberg and Levanon (2013). They identified 3804 genes as housekeeping based on their uniform expression among 16 human tissues from Human BodyMap data (NCBI Gene Expression Omnibus [GEO] accession number GSE30611).

Identification of marker genes

Marker genes were selected based on two sources. From sNuc-seq data, for each of the five major cell types, Astrocyte, Oligodendrocyte, Microglia, Endothelial, and Pyramidal (labeled as excitatory cell types), we identified genes with mean count per million greater than or equal to twofold-change in all other cell types as marker genes. Supplemental File S10 has list of markers identified in this manner for each of the five cell types.

As the second source, we used 1208 marker genes for 18 mouse cerebral cortex cell types identified by Mancarci et al. (2017). We mapped the marker genes to their human orthologs using the Ensembl database (Zerbino et al. 2018). Mancarci et al. further refined these markers based on their coexpression in bulk human tissue. To remain consistent with their method, for each

cell type, we only considered the subset of its marker genes which were highly correlated with each other, using the hierarchical clustering with topological overlap on the marker genes for each cell type. Genes in the cluster with the highest count of links were selected as the markers of the cell type. Our final list includes 256 markers for five major cell types (Supplemental File S11). Supplemental Figure S3C shows overlap of the marker gene sets from the two sources.

Enrichment of clusters with marker genes

Enrichment was determined using a hypergeometric test at FDR 0.05. Counts of markers in each cluster for all five cell types are provided in Supplemental File S3. Some clusters have markers from multiple cell types, but in all but one cluster (ID252), enrichment was significant for markers of one cell type only. For cluster ID252—where the enrichment is significant for two cell types, Endothelial and Oligodendrocyte—there is a huge difference in the count of markers (122/141 markers from Oligodendrocyte vs. 13/132 from Endothelial).

Modeling expression level of genes in the GTExBulk data set

We used linear models to estimate the expression level of genes based on the variation of the marker genes in each of the samples, using the first seven principal components of the whole set of marker genes (the sNuc-seq and Mancarci et al. [2017] marker sets separately) in the bulk tissue data set. Therefore, the expression level of gene A in sample j of the GTExBulk data set is modeled as

$$Exp_j(A) = \mu_A + \beta_1 s_1 + \beta_2 s_2 + \cdots + \beta_7 s_7 + \varepsilon_{A,j},$$

where μ_A is the average expression level of gene A, s_i's are the principal component scores for sample j, β_i's are the parameters of the model and $\varepsilon_{A,j}$ is residual error. See Supplemental Figure S12A for correlation of the mean expression level of different sets of marker genes with the principal component scores and Supplemental Figure S12B for percentage of variation explained by each of the principal components. The adjusted R^2 (for simplicity, called R^2) from the model is used for all downstream analyses.

Enrichment of functional terms

For a given network, each functional term is marked as enriched if the density of the links between the genes associated with the term is significantly higher than the density of the network, where the hypergeometric distribution is used as the null. The FDR was controlled at 0.1 using the method of Benjamini and Hochberg (1995). Brain-specific functional terms were identified as the terms enriched in either of the TAN-brain networks for GTExBulk network but not enriched in TAN-liver and TAN-blood networks. Likewise, housekeeping terms were identified as terms enriched in either of the TAN-brain or GTExBulk networks, as well as the TAN-blood and TAN-liver networks.

Synthesized bulk data sets

Each of synthesized samples was built using sNuc-seq samples (nuclei) from the Allen data described above, as follows. First, nuclei were grouped into five major cell types based on their provided labels. Then, for each synthetic sample, nuclei were randomly sampled with the following baseline proportions: Pyramidal (20%), Inhibitory (20%), Oligodendrocyte (43%), Astrocyte (12%), and Microglia (5%), based on the estimates of von Bartheld et al. (2016), and their gene expression values were added and divided by the total count of nuclei, yielding a final synthetic sample. This was repeated to create multiple synthetic data sets with 100 samples each. To add composition variability, the baseline proportions were randomly varied by drawing each proportion from a normal distribution with variance 33% of its mean.

Intra- and inter-cluster densities

For a given gene cluster in the network, the intra-cluster density refers to the count of coexpression links between the genes in the cluster divided by the total potential count of links in that cluster—that is, all the possible gene pairs in the cluster. For a cluster with n genes, the intra-cluster density is

$$\frac{\text{count of links between the genes}}{n^*(n-1)/2}.$$

For two clusters c_1 and c_2 with n_1 and n_2 gene counts, their inter-cluster density is calculated as the count of links between the genes in c_1 and the genes in c_2 divided by total potential count of links between genes in c_1 and c_2; that is,

$$\frac{\text{count of links between genes in } c_1 \text{ and } c_2}{n_1{}^* n_2}.$$

In all cluster analyses regarding the count of links in a cluster, the density of the network was used as the estimated background (null) expected count of links between the genes in clusters.

Data sets used for blood analysis

We used a set of genes marked as differentially expressed among major blood cell types from Newman et al. (2015). The list contained 372 genes with expression patterns identifying multiple blood cell types including: B cells, T cells, Monocytes, Neutrophils, and Natural Killers (NK) in different states (see Supplemental File S12 for list of genes for each cell type). We obtained cell-type expression profiles from GSE22886 (Abbas et al. 2005).

We used three bulk tissue data sets: (1) GTEx-whole-blood (GTEx-blood) data set from human (version 6, count of samples 393); (2) normal samples from microarray data set GSE27562 from LaBreche et al. (2011) (DS1, platform: Affymetrix Human Genome U133 Plus 2.0 Array, sample count: 31); and (3) microarray data set GSE16028 from Karlovich et al. (2009) (DS2, platform: Affymetrix Human Genome U133 Plus 2.0 Array, sample count:105). Both Affymetrix data sets were preprocessed as described in Farahbod and Pavlidis (2019). R^2 values were obtained from all three data sets using a PCR method with seven first component scores. Similar to the brain, we performed network clustering on GTEx-whole-blood network using WGCNA. Supplemental File S13 contains cluster labels and R^2 values for the genes. Functional enrichment analysis for clusters was performed similar to the brain.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the Allen Institute for Brain Science and the GTEx Consortium for making their data available, without which this study could not have been conducted. We thank Shreejoy Tripathy, Ogan Mancarci, and members of the Pavlidis lab for advice and discussion. This research was supported by the National Institute of Mental Health (R01 MH111099) and the

Natural Sciences and Engineering Research Council of Canada, Government of Canada (RGPIN-2016-05991).

Author contributions: P.P. and M.F. conceived the study. M.F. conducted research. P.P. provided oversight. M.F. and P.P. wrote the manuscript.

References

Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, Fong S, van Lookeren Campagne M, Godowski P, Williams PM, et al. 2005. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun* **6:** 319–331. doi:10.1038/sj.gene.6364173

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14:** 1083–1086. doi:10.1038/nmeth.4463

Amar D, Safer H, Shamir R. 2013. Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol* **9:** e1002955. doi:10.1371/journal.pcbi.1002955

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Crow M, Gillis J. 2018. Co-expression in single-cell analysis: saving grace or original sin? *Trends Genet* **34:** 823–831. doi:10.1016/j.tig.2018.07.007

Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. 2016. Exploiting single-cell expression to characterize co-expression replicability. *Genome Biol* **17:** 101. doi:10.1186/s13059-016-0964-6

de la Fuente A. 2010. From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases. *Trends Genet* **26:** 326–333. doi:10.1016/j.tig.2010.05.001

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* **95:** 14863–14868. doi:10.1073/pnas.95.25.14863

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29:** 569–574. doi:10.1016/j.tig.2013.05.010

Farahbod M, Pavlidis P. 2019. Differential coexpression in human tissues and the confounding effect of mean expression levels. *Bioinformatics* **35:** 55–61. doi:10.1093/bioinformatics/bty538

Gaiteri C, Ding Y, French B, Tseng GC, Sibille E. 2013. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes Brain Behav* **13:** 13–24. doi:10.1111/gbb.12106

Karlovich C, Duchateau-Nguyen G, Johnson A, McLoughlin P, Navarro M, Fleurbaey C, Steiner L, Tessier M, Nguyen T, Wilhelm-Seiler M, et al. 2009. A longitudinal study of gene expression in healthy individuals. *BMC Med Genomics* **2:** 33. doi:10.1186/1755-8794-2-33

Kelley KW, Nakao-Inoue H, Molofsky AV, Oldham MC. 2018. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat Neurosci* **21:** 1171–1184. doi:10.1038/s41593-018-0216-z

LaBreche HG, Nevins JR, Huang E. 2011. Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. *BMC Med Genomics* **4:** 61. doi:10.1186/1755-8794-4-61

Langfelder P, Luo R, Oldham MC, Horvath S. 2011. Is my network module preserved and reproducible? *PLoS Comput Biol* **7:** e1001057. doi:10.1371/journal.pcbi.1001057

Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J. 2019. *E. coli* gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res* **47:** 85–92. doi:10.1093/nar/gky1176

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome Res* **14:** 1085–1094. doi:10.1101/gr.1910904

Li X, Zheng Y, Hu H, Li X. 2016. Integrative analyses shed new light on human ribosomal protein gene regulation. *Sci Rep* **6:** 28619. doi:10.1038/srep28619

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45:** 580–585. doi:10.1038/ng.2653

Mancarci BO, Toker L, Tripathy SJ, Li B, Rocco B, Sibille E, Pavlidis P. 2017. Cross-laboratory analysis of brain cell type transcriptomes with applications to interpretation of bulk tissue data. *eNeuro* **4:** ENEURO.0212-17.2017. doi:10.1523/ENEURO.0212-17.2017

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, the DREAM5 Consortium, Kellis M, Collins JJ, et al. 2012. Wisdom of crowds for robust gene network inference. *Nat Methods* **9:** 796–804. doi:10.1038/nmeth.2016

McCall MN, Illei PB, Halushka MK. 2016. Complex sources of variation in tissue expression data: analysis of the GTEx lung transcriptome. *Am J Hum Genet* **99:** 624–635. doi:10.1016/j.ajhg.2016.07.007

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12:** 453–457. doi:10.1038/nmeth.3337

Ng B, White CC, Klein H-U, Sieberts SK, McCabe C, Patrick E, Xu J, Yu L, Gaiteri C, Bennett DA, et al. 2017. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci* **20:** 1418–1426. doi:10.1038/nn.4632

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. 2008. Functional organization of the transcriptome in human brain. *Nat Neurosci* **11:** 1271–1282. doi:10.1038/nn.2207

Patrick E, Taga M, Ergun A, Ng B, Casazza W, Cimpean M, Yung C, Schneider JA, Bennett DA, Gaiteri C, et al. 2019. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. bioRxiv doi:10.1101/566307

Rotival M, Petretto E. 2013. Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits. *Brief Funct Genomics* **13:** 66–78. doi:10.1093/bfgp/elt030

Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, Engelhardt BE, Battle A. 2017. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res* **27:** 1843–1858. doi:10.1101/gr.216721.116

Toker L, Mancarci BO, Tripathy S, Pavlidis P. 2018. Transcriptomic evidence for alterations in astrocytes and parvalbumin interneurons in subjects with bipolar disorder and schizophrenia. *Biol Psychiatry* **84:** 787–796. doi:10.1016/j.biopsych.2018.07.010

von Bartheld CS, Bahney J, 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. *J Comp Neurol* **524:** 3865–3895. doi:10.1002/cne.24040

Westra H-J, Arends D, Esko T, Peters MJ, Schurmann C, Schramm K, Kettunen J, Yaghootkar H, Fairfax BP, Andiappan AK, et al. 2015. Cell specific eQTL analysis without sorting cells. *PLoS Genet* **11:** e1005223. doi:10.1371/journal.pgen.1005223

Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, Desland F, Chudnovskiy A, Mortha A, Dominguez C, et al. 2019. The *cis*-regulatory atlas of the mouse immune system. *Cell* **176:** 897–912.e20. doi:10.1016/j.cell.2018.12.036

Zaitsev K, Bambouskova M, Swain A, Artyomov MN. 2019. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat Commun* **10:** 2209. doi:10.1038/s41467-019-09990-5

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761. doi:10.1093/nar/gkx1098

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4:** Article17. doi:10.2202/1544-6115.1128

Received September 3, 2019; accepted in revised form June 18, 2020.