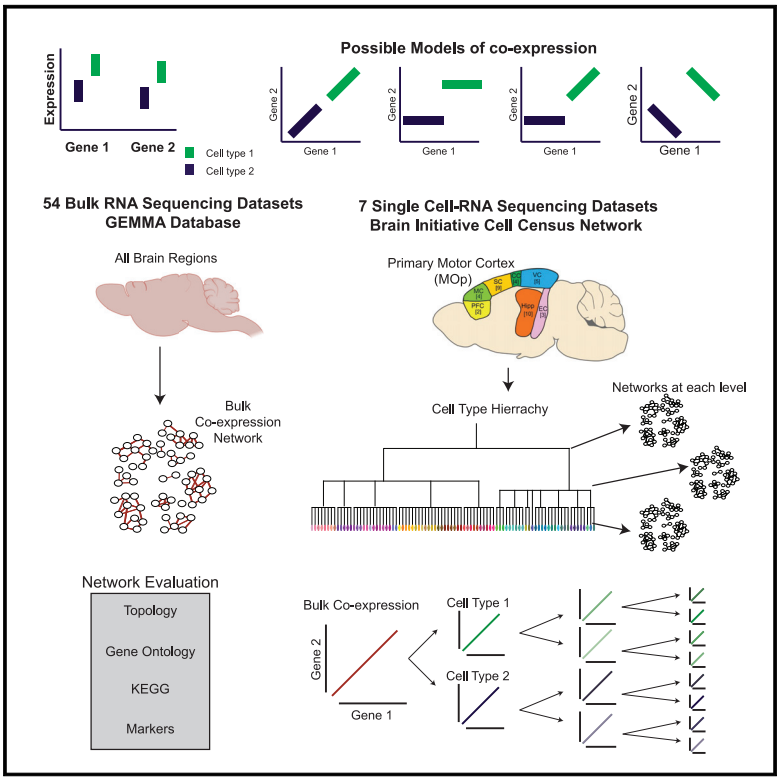
Report

Cell Systems

Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain

Graphical abstract



Highlights

- Co-expression describes how shared gene expression patterns reflect shared function
- Genes that define cell types are co-expressed across different scales of heterogeneity
- Cell-type-specific co-expression relationships are the exception, not the norm

Authors

Benjamin D. Harris, Megan Crow, Stephan Fischer, Jesse Gillis

Correspondence

jgillis@cshl.edu

In brief

Gene-gene relationships are commonly measured via the co-variation of gene expression across samples, also known as gene co-expression. The brain initiative cell census network (BICCN) single-cell RNA-sequencing (scRNA-seq) datasets provide an unparalleled opportunity to understand how gene-gene relationships shape cell identity. By evaluating the co-expression networks built at different levels of heterogeneity, we reveal the consistency of functional relationships across cell types in the brain.

Harris et al., 2021, Cell Systems *12*, 1–9
July 21, 2021
https://doi.org/10.1016/j.cels.2021.04.010

CellPress

Cell Systems

CellPress

Report

Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain

Benjamin D. Harris,[1,2] Megan Crow,[1] Stephan Fischer,[1] and Jesse Gillis[1,2,3,*]
[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Cold Spring Harbor School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[3]Lead contact
*Correspondence: jgillis@cshl.edu
https://doi.org/10.1016/j.cels.2021.04.010

SUMMARY

Gene-gene relationships are commonly measured via the co-variation of gene expression across samples, also known as gene co-expression. Because shared expression patterns are thought to reflect shared function, co-expression networks describe functional relationships between genes, including co-regulation. However, the heterogeneity of cell types in bulk RNA-seq samples creates connections in co-expression networks that potentially obscure co-regulatory modules. The brain initiative cell census network (BICCN) single-cell RNA sequencing (scRNA-seq) datasets provide an unparalleled opportunity to understand how gene-gene relationships shape cell identity. Comparison of the BICCN data (500,000 cells/nuclei across 7 BICCN datasets) with that of bulk RNA-seq networks (2,000 mouse brain samples across 52 studies) reveals a consistent topology reflecting a shared co-regulatory signal. Differential signals between broad cell classes persist in driving variation at finer levels, indicating that convergent regulatory processes affect cell phenotype at multiple scales.

INTRODUCTION

Co-expression networks characterize related genes on the basis of their shared expression profiles across samples. A shared profile suggests that their activity is driven by the same factors or that they are functionally related (Eisen et al., 1998). Networks built from bulk gene expression data have been widely observed to recapitulate known gene functions (Crow et al., 2016; Lee et al., 2004). As a result, co-expression analysis serves many applications in genomics. For example, co-expression has been used to infer transcription factor binding and causal regulation of downstream targets (Fiers et al., 2018; Kulkarni et al., 2018; Qiu et al., 2020; Song et al., 2016), characterize disease (Torkamani et al., 2010), and to predict which cells interact with each other based on ligand-receptor pairs (Cabello-Aguilar et al., 2020; Efremova et al., 2020).

Yet, because cell-type composition is a major factor driving expression variation in bulk expression data, a substantial fraction of co-expression in bulk data is likely to be driven by variation in cell-type abundance, even if only indirectly through changes in abundance across other conditions (e.g., disease) (Farahbod and Pavlidis, 2020; McCall et al., 2016; Zhang et al., 2021). Although some work has been done to use deconvolution to identify cell-type-specific co-expression from bulk data (Kelley et al., 2018), other analyses show that compositional differences confound a co-regulatory signal (Farahbod and Pavlidis, 2020; Zhang et al., 2021). Building networks from pure cell-type data, as from single-cell RNA-seq (scRNA-seq), has the potential to identify co-regulatory relationships between genes that may be hidden due to cell-type composition in bulk data (Trapnell, 2015). However, if single-cell co-expression data differ dramatically from bulk data, it could be considered as a surprise, given the longstanding utility of co-expression from bulk data (i.e., if bulk co-expression has been useful at capturing gene-gene relationships, how different should single cell be?). Characterizing the overlapping and distinct signals from single-cell and bulk data remains a major challenge (Crow and Gillis, 2018), and most previous research into single-cell co-expression has been limited to individual datasets or meta-analysis across unrelated biological conditions (Feregrino et al., 2019; Mohammadi et al., 2019; Skinnider et al., 2019; Smillie et al., 2019). Further analysis using more specific and powered data will help advance our understanding of both regulatory and compositional co-expression signals.

The seven mouse primary motor cortex scRNA-seq datasets from the brain initiative cell census network (BICCN), totaling over 500,000 cells/nuclei, provide a rich opportunity to comprehensively study cell-type-specific co-expression networks in scRNA-seq data (Yao et al., 2020). The BICCN data are particularly useful for studying composition and co-regulation in networks because of the diversity and specificity of cell types available. Specifically, cell types are annotated at multiple levels of resolution and are replicable across datasets, enabling meta-analysis of cell-type-specific co-regulatory modules.

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

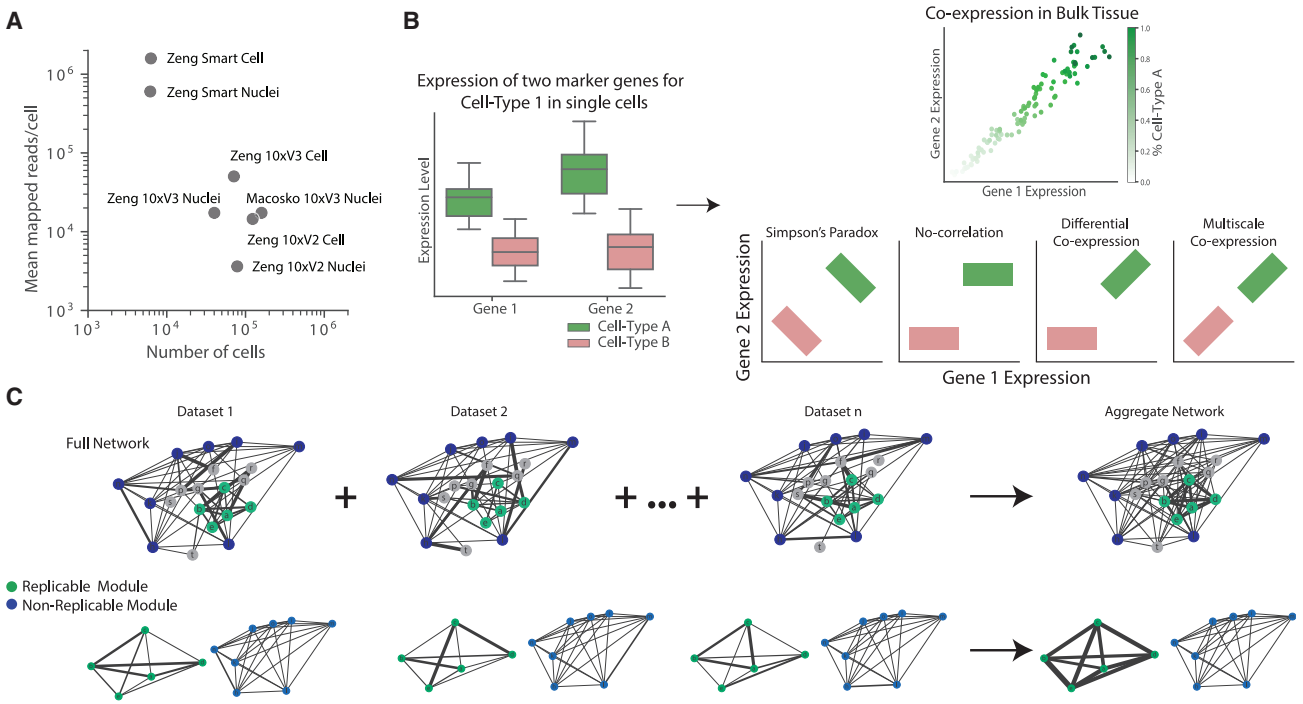


Figure 1. Generation of meta-analytic co-expression networks from scRNA-seq data

(A) The single-cell datasets were generated with multiple technologies, resulting in varying numbers of cells and varying read depths across datasets.

(B) In bulk RNA-seq data, marker genes must be co-expressed because of compositional differences in samples. Genes 1 and 2 represent hypothetical markers for cell-type A. In bulk, the co-expression of the genes is co-linear with the percent of cell-type A within each bulk sample. Co-expression of the genes within the two cell types could be any of the four models.

(C) Meta-analysis across dataset networks identifies robust co-expression relationships. Thicker edges represent stronger co-expression. Aggregate networks give strong weight to replicable co-expression.

We investigate co-expression by comparing networks built from heterogeneous data and pure cell types. We show that there is no dichotomy between cell-type composition and co-regulatory signals in co-expression. In other words, the same gene-gene relationships that differentiate cell types are evident on both finer and broader scales. We illustrate these conserved regulatory relationships using direct topological comparisons, reference functional annotations such as the gene ontology and KEGG, and most importantly, marker gene lists that define cell types. All of our analyses show overlapping connectivity in both compositional and cell-type-specific networks, revealing a consistent regulatory landscape that can be defined across all BICCN cell types. Finally, we show that finding cell-type-specific co-expression relationships will require substantially more data than is currently available.

RESULTS

The BICCN data consist of seven datasets produced using both SmartSeq2 and 10X genomics library preparation methods. There are datasets using both whole-cell and nuclei samples and both the V2 and V3 chemistries from 10x platform (Figure 1A). Across the datasets, clusters are labeled using a consistent hierarchical taxonomy (Figure S1). Our strategy is to build co-expression networks based on the known hierarchy of cell types within the BICCN data and to evaluate the co-expression of cell-type markers in networks that control this source of variation. Specifically, consider two genes α and β. Each gene can be seen as a vector of expression values over all cells C. Let C_1, …, C_K be K cell types forming a partition of C, such that $C = C_1 \cup C_2 \cup \ldots \cup C_K$. We can then split vectors α and β on the basis of cell types, for example, $\alpha[C_1]$ is the expression of gene α over cell type 1. We compute the within-cell-type co-expression as the average Pearson correlation (rank normalization omitted for clarity, see STAR Methods):

$$\overline{R} = \frac{1}{K}\sum_{i=1..K} \frac{\sum_{j\in C_i}\left(\alpha_j - \overline{\alpha[C_i]}\right)\left(\beta_j - \overline{\beta[C_i]}\right)}{\sqrt{\sum_{j\in C_i}\left(\alpha_j - \overline{\alpha[C_i]}\right)^2 \cdot \sum_{j\in Ci}\left(\beta_j - \overline{\beta[C_i]}\right)^2}}$$

where α_j is the expression of gene α in cell j, and $\overline{\alpha[C_i]}$ is the average expression of gene α in cell-type i. The final co-expression value can only be driven by within-cell-type correlation, as cell-type-specific trends are effectively removed in the form of the $\overline{\alpha[C_i]}$. At K = 1 (single partition with all cells), co-expression is largely driven by average cell-type-specific trends, while for K >> 1, all these trends have been controlled.

Thus, for example, in the absence of cell-type partitioning, two genes, which are highly expressed in cell-type A relative to cell-type B, will be co-expressed in a network containing both cell types because the genes are co-variable with respect to cell

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

type. Historically, this is the case in bulk expression data, where the co-expression of two marker genes for cell-type A will have been calculated from samples with varying proportions of the two cell types (Figure 1B). The fundamental question of single-cell co-expression is the degree to which novel co-variation is present in cell-type A (or B) individually, which reflects the regulatory interactions rather than compositional effects.

Cell-type-specific co-expression relationships can be described using at least four models: Simpson's paradox, no co-expression, differential co-expression, and multiscale co-expression. Single-cell resolution data now make it possible to quantify the occurrence of these models. The different models make assumptions about the relative direction of within cell-type co-variation versus that across cell types. In the Simpson's paradox model, correlations between gene A and gene B take one sign across all cells but reverse for subsets of cells corresponding to the cell types. Biologically, this would suggest a shared regulatory relationship (e.g., higher gene A expression is associated with lower gene B expression), which is reversed in bulk compositional data due to differential expression of the genes (e.g., expression of genes A and B are systematically higher in the first cell type). The no co-expression model is exhibited when a given gene pair is uncorrelated within each cell type but is co-expressed when considering both cell types together. This would suggest that markers are not directly co-regulated within cell types and are simply differentially expressed across cell types. A differential co-expression model, where one cell type exhibits a significant correlation between two genes, while the other cell type exhibits the opposite or no correlation would suggest co-regulatory network rewiring. If this last model is found to be predominant, then the cell types would be defined by the creation of new gene-gene relationships. Finally, the multiscale model occurs when co-expression is similar in both bulk and single cell data. In this model, gene-gene relationships are consistent within and across cell types, i.e., differential expression patterns align with the co-regulatory relationships, signifying modulation of the degree to which they are used.

We use the terms "co-regulatory" versus "compositional" networks for those which do and do not control cell-type variation, respectively. We use the term "network" to refer to the genome-wide weighted relationships between genes, and we identified robust co-expression relationships using a meta-analytic approach (Figure 1C). At the extreme end of defining co-regulatory gene interactions, we took advantage of "metacell" networks, which measure gene-gene co-variation over statistically similar sets of cells. The metacells are smaller groups of 20–100 cells (Figure S2A) that are significantly more homogeneous than clusters (t test between the distribution of distances for each cell to its respective metacell or cluster centroid, p~6e–23, Figure S2B). By comparing gene-gene relationships that sample from more and more diverse cells, we incorporated increasing compositional effects across the types of cells sampled (e.g., subtypes of inhibitory cells). At the broadest level of analysis are brain-specific bulk co-expression networks, using samples made up of large numbers of cells. For our bulk analysis, we generated meta-analytic networks using 52 datasets of bulk mouse brain data from the Gemma database (Zoubarev et al., 2012; Figure S3). Throughout, we focused on genes that were broadly expressed across cell types and thus open to robust analyses of co-variation across and within cell types.

A consistent topology between compositional and co-regulatory networks

For our first experiment, we compared metacell with bulk RNA-seq co-expression networks in order to capture similarities and differences at the greatest range of the spectrum (see STAR Methods for details on network construction). We first observed that both networks reflect known biology using a guilt-by-association formalism, in which each network is measured for its ability to reconstruct a partially hidden gene list from preferential connectivity within it, outputting an area under the ROC curve (AUROC) (Figure 2A). In the metacell network, the average AUROC across all GO slim and KEGG functional groups were 0.64 and 0.63, respectively, and similarly, the average AUROC of the bulk RNA-seq network was 0.67 for GO slim and 0.70 for KEGG (Figure 2B). We also found that these networks had highly similar topologies. A comparison of coarse hierarchical clustering of both co-expression networks showed large, shared modules between the two networks, visualized as a riverplot in Figure 2C. Moreover, the average AUROC of modules drawn from the metacell network in the bulk network was 0.84, and the same was true in the case of the reverse analysis (Figure 2D). This indicates that modules present in one network are present in the other to a very specific degree, which is surprising, because these two networks were constructed using data that capture vastly different signals—compositional versus co-regulatory.

Persistent co-expression of cell-type markers in compositional and non-compositional networks

To investigate the overlap between compositional and co-regulatory variation more directly, we evaluated the modularity of neuronal subclass markers in each of these two networks, measuring how well network connectivity can reconstruct a partially hidden marker list in cross-validation. As expected, the markers were well connected in meta-analytic networks built from bulk RNA-seq (average AUROC = 0.84, Figure 3A), consistent with the notion that these networks contain cell-type signals. Surprisingly, markers were also well connected in the networks where cell-type variation has been controlled (average AUROC = 0.84, Figure 3B). The performance of the subclass markers in both networks was well correlated ($r = 0.73$, $p = 0.004$, Figure 3C), in agreement with the consistent topology we found in both networks. As another comparison to the bulk data, we created pseudobulk samples from each scRNA-seq dataset by randomly dividing each dataset into 20 pseudobulk samples. Networks created from pseudobulk produced comparable results for the markers and GO to the bulk RNA-seq data (GO AUROC = 0.56, Subclass Marker AUROC = 0.87, Figure S4). This suggests that the regulatory factors that drive differences between cell types remain important as a source of differences within cell types.

The BICCN data offer the unique opportunity to use consistent cell-type labels across independently sampled datasets so that robust analyses can be conducted at varying levels of specificity in the cell-type hierarchy across independent data. We took advantage of the known hierarchy for our next series of experiments. For each of the three levels of increasing specific

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

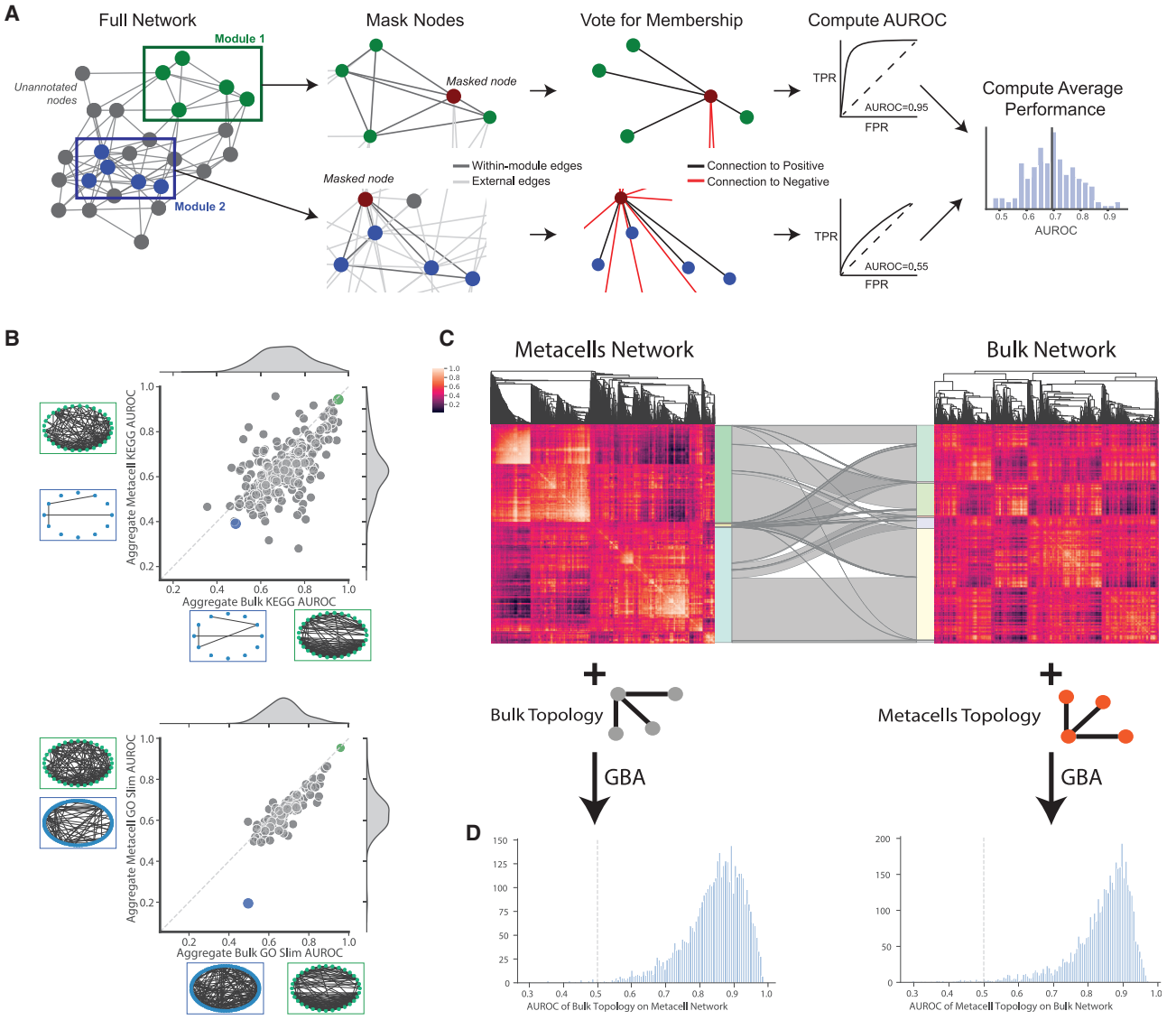


Figure 2. Consistent topology between networks across furthest scales

(A) Guilt-By-Association algorithm assesses a network's ability to reconstruct modules.

(B) Performance of KEGG and GO on both metacell and bulk RNA-seq co-expression networks is well correlated. Network diagrams show the best performing (green) and worst performing (blue) terms in each dataset.

(C) Clustered heatmaps for metacell and bulk networks. The riverplot joining them identifies shared genes across hierarchical clusters.

(D) Prediction of small neighborhoods in one network using the other network's topology shows shared local topology.

cell-type classification (class, subclass, and cluster), we built aggregate networks to capture replicable gene-gene relationships (Figure 2D). In each case, samples were divided into homogeneous groups at the given level of specificity so that only covariation at more specific levels affects co-expression. Therefore, for example, when we evaluate subclass markers, the class network will be compositional with respect to them, but the subclass and cluster networks will be non-compositional and should only capture co-regulatory relationships between the same sets of genes. We found that the class network had the highest performance for subclass markers (average AUROC = 0.94) but that the subclass and cluster networks still performed exceptionally well (subclass: average AUROC = 0.85, cluster: average AUROC = 0.83, Figure 2E). This was also true for subsampled networks that reduce within-cluster heterogeneity, which further strengthened this observation (Figure S5). Thus, genes which are preferentially co-expressed across cell types remain co-expressed within cell types: the same sets of differentially expressed genes, which distinguish cells at the subclass level, continue to vary across cells even when subclass is held constant.

While our focus has been on genes expressed across most cell types, a natural question is whether the same multiscale co-expression is visible for genes selected based on expression in only specific cell types. To test this, we performed our analysis as discussed earlier, but we selected genes based only on their

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

CellPress

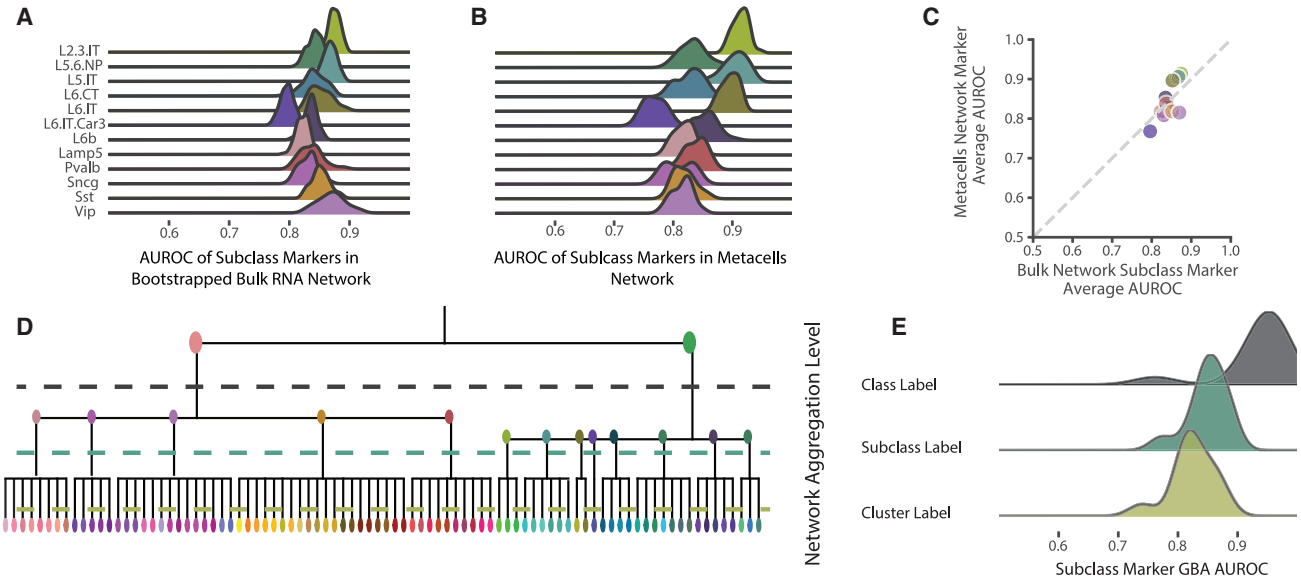


Figure 3. Cell-type marker gene sets show multiscale co-expression

(A) Markers show high performance in bulk RNA-seq networks.

(B) Subclass markers show consistent and high co-expression in metacell networks.

(C) Average performance of bulk and metacell networks is highly correlated.

(D) Dendrogram of cell-type hierarchy showing class, subclass, and clusters used to construct co-expression networks.

(E) Consistent and strong co-expression of markers in networks at each level of the cell-type hierarchy.

expression in pairs of subclasses. We found the same tendency in gene sets that are co-expressed across a given pair of subclasses to be co-expressed when the subclass was held constant (Figure S6A). Global co-expression performance was lower, in this case, likely reflecting the slightly less robust gene expression of the selected genes (Figure S6B).

Proper normalization is an important concern for all scRNA-seq analysis. Most commonly, modeling methods can correct for sampling artifacts. However, most of these methods rely on the existing correlation structure and would induce circularity in the analysis if applied in this study. While we use Pearson's correlation coefficient (PCC) for the computational efficiency, proportionality is an association metric that is agnostic to normalization. This makes it an effective alternative to PCC. Additionally, benchmarking of the various measures of association report similar performance across the board (Skinnider et al., 2019). We built aggregate networks at the class, subclass, and cluster levels using proportionality and co-expression across the levels of classification comparable with the PCC results (class AUROC = 0.92, subclass AUROC = 0.85, cluster AUROC = 0.84, Figures S7 and 3E).

One question is the degree to which the specificity of the co-expression relationships is maintained. For example, it could be that the exact cell-type marker sets are maintained at more specific levels of the cell-type hierarchy, or it could be that new types only sample from within those sets to form new marker sets, creating some novel gene-gene relationships in the process. To investigate this, we first focused on connectivity for two of the GABAergic subclasses: Vip and Sst. In the class network, the subclass markers were extremely modular, with dense connectivity within each gene list and sparser connections between Sst and Vip markers. However, for the subclass and cluster networks, the connections between the modules increased significantly. Despite this increase, the Vip and Sst modules could still be clearly discerned from each other (Figure S8A). We quantified the change in connectivity between modules by measuring how one gene list, the training list, predicts connectivity to another gene list, the testing list. As expected, in the class network, the marker lists were essentially unpredictive of one another (since they mark separate cells, class network AUROC $\sim$ 0.5, Figure S8B). However, the Vip and Sst modules were more highly interconnected in the subclass and cluster networks (subclass: average AUROC = 0.73, cluster: average AUROC = 0.73). Testing of all pairwise combinations of subclass modules showed a consistent trend of increased modularity between subclass modules in the more homogeneous subclass and cluster networks (subclass AUROC = 0.65, and cluster AUROC = 0.66) relative to the class network (AUROC $\sim$ 0.5). The modest cross-module performance of marker sets suggests that there is some "cross-talk" between modules as we move down the hierarchy of cell types because modules are combined in novel ways to define new cell types.

Performance of each subclass market set is consistently high within any subclass-specific networks (Figure S9A). Marker sets, such as the Vip interneuron markers, have extremely low variation in performance across the subclass-specific networks. Diagrams of the networks show consistently dense networks (Figure S6A). To further investigate connectivity of subclass markers in the subclass-specific networks we focused on the consistency and strength of connectivity to individual genes by predicting each gene's connectivity to the rest of the genes in the subclass marker set. The strength of a gene's connection to its marker set does not depend on the data from which it was constructed, remaining high regardless of the subclass

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

network being measured (Figure S9B). This once again highlights the consistency of a core co-regulatory network across the cell types. Individual pairs of genes also exhibit multiscale co-expression. We illustrate the co-expression of Arpp21 and Baiap2, two glutamatergic markers, and Spock3 and Abat, two GABAergic markers (Figure S7). These gene pairs exhibit multiscale co-expression because they are co-expressed at the class, subclass, and cluster level, even in cell types that the genes are not markers of. The scale of the BICCN expression data and cell-type annotations cannot be matched by any other organ system. However, using four human pancreas datasets that are normally analyzed together, we also found consistent co-expression of cell-type markers at multiple scales (full datasets [compositional] AUROC = 0.97, cluster [non-compositional] AUROC = 0.97, Figure S10). All of these analyses show consistent co-expression of gene sets that define cell types from broadest to finest levels of cell-type classification.

Differential co-expression to identify novel gene-gene relationships

Our results provide evidence that the multiscale model of co-expression (differential expression aligns with conserved co-regulatory relationships) plays an important role in regulatory networks. We next evaluated if we can find evidence for the differential co-expression model (change in co-regulatory relationships) by looking for cell-type-specific gene-gene relationships. We took the difference between a single subclass's network and a network of the remaining subclasses to find the edges most specific to a given subclass. In a differential co-expression network between subclass A and the rest of the subclasses, the strongest connections in the network are gene pairs that are only co-expressed in subclass A. This means that if marker genes for subclass A are only co-expressed in subclass A, they will have a high AUROC. However, differential co-expression networks showed minimal connectivity of subclass markers (average AUROC = 0.69, Figure 4A). These low values are particularly notable in contrast to the earlier performance of the multiscale co-expression model, where, using aggregates that assume a purely consistent regulatory architecture, we found strong enrichment of subclass markers (subclass network AUROC = 0.84, cluster network AUROC = 0.84, Figure 4A). GO and KEGG modules are also relatively weakly connected in the differential co-expression networks (Figure S11). These results emphasize the consistent modularity across the cell types in co-regulatory modules.

While the multiscale model explains most of the co-expression signal within cell types, the performance of the subclass markers in the differential co-expression networks, while lower, is non-random. This suggests the potential to identify individual edges as significantly differentially co-expressed and identify novel cell-type-specific co-expression modules. We first considered how the heterogeneity of data affects our ability to confidently call connections as significantly different. When computing differential co-expression between the GABAergic and glutamatergic cell types, we could aggregate the networks at either the class, subclass, or cluster level. We found that the most heterogeneous class networks identified ~10× more edges at a given false discovery rate (FDR) threshold than the subclass networks (Figure 4B). The subclass networks also identified ~10× more edges than the cluster networks. Selecting significant edges from the class network resulted in ~0.1% of edges being significant at an FDR < 0.01, while using the cluster network had no significant edges even at a more permissive FDR < 0.1. These results suggest that, even when aggregating across seven datasets, we were underpowered to detect changes in co-regulation at the cell-type level.

Incorporating more scRNA-seq datasets should provide sufficient power to confidently identify cell-type-specific co-expression relationships. We showed the power gained by aggregating from 2 to 7 of the existing datasets, providing an improvement in power and statistical significance on par with the improvement from the coarsest to finest cell-type definitions (Figure 4C). Using a threshold of 1% of edges being differentially co-expressed, the class-level differential co-expression network is sufficiently powered at an FDR < 0.01 using only 6 of the 7 datasets. Extrapolating from subclass level results, estimates ~11 datasets are required to achieve the same thresholds. Even with all seven datasets, no edges are significantly different in the cluster aggregate network; hence, we cannot extrapolate the number of datasets required directly. (Figure S12A). Within the edges in the class label network that are FDR < 0.01, we found that 82% of them contained at least 1 gene that is a subclass marker (Figure S12B). With 709 marker genes, only 31% of edges are expected to contain a marker gene. Using differential co-expression to identify novel gene-gene relationships will require controlling for composition using networks aggregated at the finest scales. While we are underpowered for differential co-expression, the existence of multiscale co-expression presents a powerful toehold for future analyses, potentially limiting the search space for variability to a much smaller core set of modules. In this view, co-expression is largely maintained within cell types with a major source of variability simply being the dynamic range over each of the genes is operating (Figure 4D).

DISCUSSION

Our meta-analysis of bulk RNA-seq data and the BICCN scRNA-seq data from the mouse brain establishes the importance of a multiscale model of co-expression across neurons. We identified shared topology between compositional and cell-type-specific networks using both reference functional networks, the gene ontology and KEGG, as well as direct comparisons of network topology. Cell-type-level markers for neurons exhibit consistent topologies in networks built at all levels of the cell-type hierarchy.

Our result highlights the existence of a core co-regulatory network that is reused in all cell types of the brain. We note that this result is not likely to be brain specific, or even cell-type specific, as previous research has also shown strong convergence in co-expression across systems. Indeed, while expression levels of genes vary across brain regions, many modules associated with cell types replicate across brain regions and species (Hartl et al., 2020). Outside the brain, drug perturbation experiments using human iPSC-derived cardiac myocytes and fibroblasts have shown that cell identity maintenance factors are usually not tissue specific. Rather, genes that play important functional roles in cell identity maintenance are broadly expressed across tissues (Mellis et al., 2020). The critical role of non-tissue-specific genes to perturbation highlights the important role of core regulatory networks in contexts defined by cell-type specificity.

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

CellPress

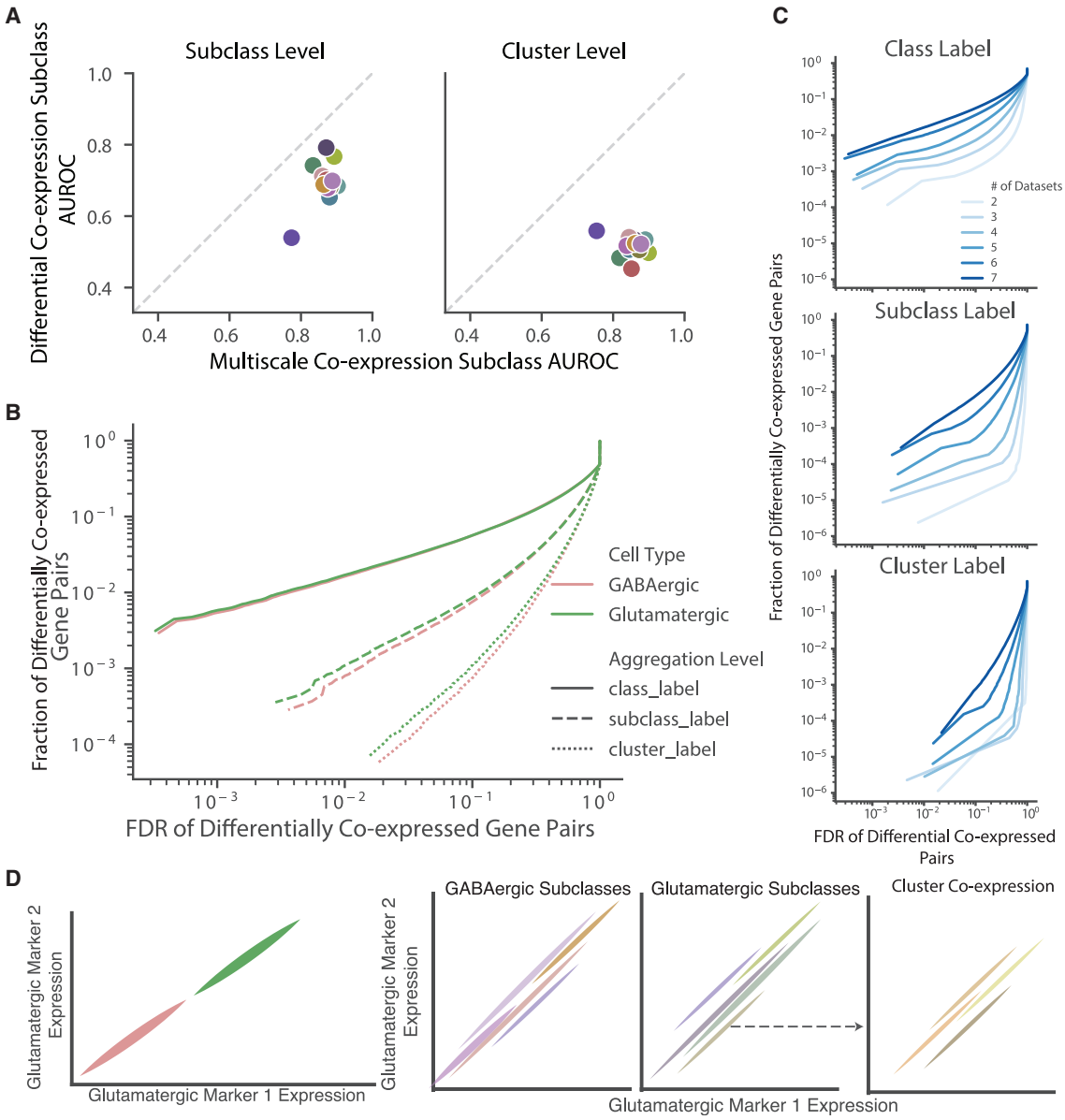


Figure 4. Differential co-expression to identify cell-type-specific gene-gene relationships

(A) The multiscale model outperforms differential co-expression for recapitulating cell-type marker modules. Networks were aggregated at the subclass (left) and cluster (right) level. Differential co-expression was calculated by taking the normalized difference between a given network and all others at the same level in the cell-type hierarchy.

(B) Differential co-expression between GABAergic and glutamatergic cells aggregated at different levels shows limited statistical power.

(C) Statistical power increases as the number of datasets used in the meta-analysis increases.

(D) A multiscale model for co-expression of a pair of glutamatergic markers shows co-expression of the markers through all levels of the hierarchy of cell types.

Given the major role of multiscale co-expression, we expected that finding differences between cell-type-specific co-expression networks will be difficult. We explored the statistical power necessary to identify cell-type-specific co-expression and how heterogeneity within the data influences the power. Despite the large amount of data considered, we were underpowered to identify cell-type-specific co-expression, although networks built at the lowest resolution of cell-type classification were nearly sufficiently powered. As more scRNA-seq data become available, we expect the value of meta-analysis to become increasingly apparent within these data, not just as a mechanism for overcoming experiment specific biases but in generating gold-standard co-expression networks that can be used as a groundwork for exploration in data where some differences are expected (e.g., disease).

A central limitation of our study is our focus on genes that are broadly expressed across cell types. This is a simple necessity for our analysis, since co-expression is undefined if, for example,

one gene shows no variation (is unexpressed) in a given cell type. On the other hand, it may very well be that the binary property of being expressed or not constitutes a large fraction of cell-type variability that we do not explore. While interesting, such variation does not really reflect changes in co-expression, since it can be much more easily explained through the single-gene expression. The multiscale co-expression we see may be most relevant to the growing literature on the importance of gradients in defining cell types, particularly in the brain (Cembrowski and Menon, 2018). The relatively high cross-type marker learning performance similarly suggests a relatively simple continuous axis of co-variation between genes, at least within the well-powered BICCN data. When measuring co-variation within finer scales, such as in the cluster and metacell networks, the proportion of non-biological variance might be higher due to the smaller size and greater homogeneity of each grouping of cells compared with higher levels. We control this by using replicable relative correlations, which tend to be insensitive to global shifts in the correlation, although more complex interactions could still affect results.

Beyond continued evaluation within the BICCN, our results open up two main directions to take future analyses: improved gene function annotation and improved cell-type-specific co-expression. Computational gene function annotation is typically done using orthology or features generated from DNA sequence (Škunca et al., 2012). Our evidence for the multiscale model shows that well-powered co-expression networks built across species should be a valuable addition to methods for computational annotation of gene function. Improved cell-type co-expression should also be a major addition to mechanistic studies. Inferring mechanistic relationships from scRNA-seq alone has proven to be difficult (Qiu et al., 2020), with methods that incorporate ATACseq or ChIPseq data doing only a little better (Burdziak et al., 2019). Using well-powered cell-type-specific co-expression networks should open up both stronger integration with other modalities (e.g., ATAC-seq) and better inference of convergent changes across conditions (Hie et al., 2020). Thus, a major source of utility of the BICCN data is simply the presence of reference data that crosses technologies, labs, and other nuisance variables to permit robust aggregation, a process, which is particularly important and convenient within the co-expression space. The meta-analytic use of the BICCN data sets a standard that we hope can continue into the future, integrating data from outside the BICCN to obtain increasingly high quality and useful reference co-expression networks.

Conclusions

The shared co-expression signal of marker genes and regulatory modules throughout the cell-type hierarchy makes it clear that co-expression is, in part, multiscale. Multiscale co-expression means that while gene expression values are significantly different between groups of cells, the core co-regulatory network remains consistent throughout the highly refined cell-type hierarchy defined within the primary cortex. The sparsity and noise in scRNA-seq data often make co-expression and differential co-expression challenging. Using a meta-analytic framework, we highlight robust methods and significant use cases for co-expression and differential co-expression analysis using scRNA-seq data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Single cell datasets and preprocessing
  - Bulk RNA sequencing data from GEMMA
  - Network construction and aggregation
  - Computing marker genes
  - Measuring network performance with EGAD
  - Computing metacells
  - Computing differential co-expression

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.04.010.

ACKNOWLEDGMENTS

B.D.H. was supported by the CSHL Crick Cray Fellowship. M.C. was supported by NIH K99MH120050. J.G. and S.F. were supported by NIH R01MH113005, R01LM012736, and U19MH114821.

AUTHOR CONTRIBUTIONS

J.G. conceived the study. J.G., M.C., and B.H. conceived the experiments. B.H. and S.F. performed the computational analysis. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The present address for M.C. is Genentech. The authors declare no competing interests.

Received: September 30, 2020
Revised: February 11, 2021
Accepted: April 23, 2021
Published: May 19, 2021

REFERENCES

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. *20*, 206.

Burdziak, C., Azizi, E., Prabhakaran, S., and Pe'er, D. (2019). A nonparametric multi-view model for estimating cell type-specific gene regulatory networks. arXiv, 08138.

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. *48*, e55.

Cembrowski, M.S., and Menon, V. (2018). Continuous variation within cell types of the nervous system. Trends Neurosci. *41*, 337–348.

Crow, M., and Gillis, J. (2018). Co-expression in single-cell analysis: saving grace or original sin? Trends Genet. *34*, 823–831.

Crow, M., Paul, A., Ballouz, S., Huang, Z.J., and Gillis, J. (2016). Exploiting single-cell expression to characterize co-expression replicability. Genome Biol. *17*, 101.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. *15*, 1484–1506.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Farahbod, M., and Pavlidis, P. (2020). Untangling the effects of cellular composition on coexpression analysis. Genome Res. *30*, 849–859.

Feregrino, C., Sacher, F., Parnas, O., and Tschopp, P. (2019). A single-cell transcriptomic atlas of the developing chicken limb. BMC Genomics *20*, 401.

Fiers, M.W.E.J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. Brief. Funct. Genomics *17*, 246–254.

Hartl, C.L., Ramaswami, G., Pembroke, W., Saha, A., Parsana, P., Muller, S., Pintacuda, G., Lage, K., Battle, A., and Geschwind, D.H. (2020). The architecture of brain co-expression reveals the brain-wide basis of disease susceptibility. bioRxiv. https://doi.org/10.1101/2020.03.05.965749.

Hie, B., Cho, H., Bryson, B., and Berger, B. (2020). Coexpression enables multi-study cellular trajectories of development and disease. bioRxiv. https://doi.org/10.1101/719088.

Kelley, K.W., Nakao-Inoue, H., Molofsky, A.V., and Oldham, M.C. (2018). Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nat. Neurosci. *21*, 1171–1184.

Kulkarni, S.R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2018). TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. Nucleic Acids Res. *46*, e31.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes Across many microarray data sets. Genome Res. *14*, 1085–1094.

McCall, M.N., Illei, P.B., and Halushka, M.K. (2016). Complex sources of variation in tissue expression data: analysis of the GTEx lung transcriptome. Am. J. Hum. Genet. *99*, 624–635.

Mellis, I.A., Edelstein, H.I., Truitt, R., Beck, L.E., Symmons, O., Goyal, Y., Dunagin, M.C., Saldana, R.A.L., Shah, P.P., Yang, W., et al. (2020). Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity. bioRxiv. https://doi.org/10.1101/2020.06.11.147207.

Mohammadi, S., Davila-Velderrain, J., and Kellis, M. (2019). Reconstruction of cell-type-specific interactomes at single-cell resolution. Cell Syst. *9*, 559–568.e4.

Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell Syst. *10*, 265–274.e11.

Quinn, T.P., Erb, I., Gloor, G., Notredame, C., Richardson, M.F., and Crowley, T.M. (2019). A field guide for the compositional analysis of any-omics data. GigaScience *8*, giz107.

Skinnider, M.A., Squair, J.W., and Foster, L.J. (2019). Evaluating measures of association for single-cell transcriptomics. Nat. Methods *16*, 381–386.

Škunca, N., Altenhoff, A., and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. PLoS Comput. Biol. *8*, e1002533.

Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell *178*, 714–730.e22.

Song, L., Huang, S.C., Wise, A., Castanon, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science *354*, aag1550.

Tange, O. (2018). GNU Parallel 2018. https://doi.org/10.5281/zenodo.1146014.

Torkamani, A., Dean, B., Schork, N.J., and Thomas, E.A. (2010). Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Res. *20*, 403–412.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

van der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. Comput. Sci. Eng. *13*, 22–30.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Yao, Z., Liu, H., Xie, F., Fischer, S., Booeshaghi, A.S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Pinto-Duarte, A., Bartlett, A., et al. (2020). An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. bioRxiv. https://doi.org/10.1101/2020.02.29.970558.

Zhang, Y., Cuerdo, J., Halushka, M.K., and McCall, M.N. (2021). The effect of tissue composition on gene co-expression. Brief. Bioinform. *22*, 127–139.

Zoubarev, A., Hamer, K.M., Keshav, K.D., McCarthy, E.L., Santos, J.R., Van Rossum, T., McDonald, C., Hall, A., Wan, X., Lim, R., et al. (2012). Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics *28*, 2272–2273.

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Brain Initiative Cell Census Network scRNAseq data	Nemo archive	nemo:dat-ch1nqgb7
Bulk Brain expression data	GEMMA database	GSE65770 GSE75386 GSE54651 GSE77243 GSE104775 GSE70732 GSE63943 GSE108269 GSE95141 GSE100070 GSE91387 GSE79510 GSE39911.2 GSE96627 GSE99791 GSE91396 GSE77005 GSE30617 GSE75858 GSE102014 GSE22131 GSE79790 GSE58523 GSE50809 GSE36025 GSE49379.3 GSE78747 GSE112575 GSE61991.1 GSE104882 GSE107925 GSE65000.2 GSE79702 GSE90806 GSE74985 GSE53380 GSE33912.2 GSE48962 GSE80465 GSE110372 GSE90019 GSE69952 GSE76857 GSE36026 GSE75229 GSE104709 GSE101239 GSE47966.2 GSE99353 GSE96938 GSE103194 GSE119182

(*Continued on next page*)

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pancreas scRNAseq data	Bioconductor scRNAseq package	GSE84133 GSE85241 E-MTAB-5061 GSE86469
KEGG gene lists 2018	genome.jp	KEGG release 87.1
Gene Ontolgoy 2018	Gene Ontology Consortium	GO 2018-09-05
Software and Algorithms		
Scanpy	Wolf et al., 2018	PyPi
Numpy	van der Walt et al., 2011	PyPi
Parallel	Tange, 2015	GNU.org
Metacells	Baran et al., 2019	Bioconductor
propr	Quinn et al., 2019	CRAN

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources should be directed to and will be fulfilled by the Lead Contact, Jesse Gillis (jgillis@cshl.edu)

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These datasets' accession numbers are provided in the key resource table.
- Python and R original code for all of this analysis is publicly available at 10.5281/zenodo.4645997 (www.github.com/bharris12/multiscale_brain)
- Scripts used to generate the figures reported in this paper are available at 10.5281/zenodo.4645997 (www.github.com/bharris12/multiscale_brain)
- Any additional information required to reproduce this work is available from the Lead Contact

METHOD DETAILS

Single cell datasets and preprocessing

We acquired the datasets and associated metadata directly from the BICCN. To work with the expression data, anndata objects were created by filtering the droplets to the whitelist defined by the consortium and merging with all associated metadata. All analyses were done using CPM normalized expression values. To select a shared list of genes we ranked each gene by its average expression and selected the top 7,500 genes in each dataset. Then genes that were in the top 7,500 for at least 6 of the 7 datasets were used in all analysis, leaving us with 4,201 genes. All analyses were done with this list of genes.

Bulk RNA sequencing data from GEMMA

Metadata from the Gemma database was acquired on November 29, 2019. The metadata was filtered to include only mouse bulk RNAseq datasets with at least 20 samples. Then metadata terms were filtered for relevance for the brain, leaving 29 terms (See github for terms and data info). The expression data was then downloaded using the GEMMA R API and filtered to the same genes as the scRNAseq data. Networks were built as detailed below.

Network construction and aggregation

Networks were built by rank standardizing the Pearson correlation matrix of the genes. After ranking, we replace the undefined values with the average of the network. For the bulk data networks are built using an entire dataset. For single cell datasets, a full compositional network is computed using only the labeled neurons in each dataset. When computing class, subclass, cluster, and metacell networks we partition each dataset by the metadata label and build a network for each value. After aggregating networks within each dataset, we aggregate the ranked dataset networks. Aggregating datasets occurs by summing the networks from each dataset and then ranking the sum.

Please cite this article in press as: Harris et al., Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.04.010

When computing network performance of markers in the bulk network we bootstrap the bulk datasets hundred times to create 100 networks. In the down-sampling experiment we compute centroids for each dataset partition and select the 50 closest cells to the centroid within that partition. We exclude any partition with fewer than 100 cells to make sure it is at most 50% of the original data in the partition.

Computing marker genes

Marker gene lists are computed using the Mann-Whitney test in each dataset using a 1vsAll design. Significance is computed with a threshold of log2FC >2 and FDR <0.05. To compute markers across datasets we compute recurrence of each gene by totaling the number of datasets the gene is significantly different in. After sorting genes by recurrence, we sort by average AUROC. We used gene sets of size 100. Subclass-specific markers are computed within classes, e.g. Vip markers are extracted by finding genes that are differentially expressed with respect to all other GABAergic subclasses. For example when computing the markers for the Vip subclass, a GABAergic subclass, we only compare the expression of the Vip cells to the other GABAergic subclasses.

Measuring network performance with EGAD

A python version of the R package EGAD was created by translating the runGBA() function from the R package. It was modified to do cross validation in known splits, instead of randomly partitioning the data. We run it with 3 fold cross-validation. The algorithm uses neighbor voting to compute the sum of ranks of predictions for a given gene set within a network. Using the sum of predicted ranks we calculate an AUROC and/or a p-value as an output. When measuring performance of the subclass marker genes on the scRNAseq networks we create aggregate networks for each combination of 4 datasets and measure the performance using meta-analytic markers using the remaining 3 datasets. In the figures we report either the entire distribution or just the average of these values and in the text we report the average value.

Computing metacells

Metacells are computed using the R metacells package. We set the parameters to encourage extremely small clusters(K=20, m=5,b=1000). Additionally, we used the 4,201 recurrently highly expressed genes as the gene list for the method. While the metacells method, like the original clustering method, is graph based, minor differences in the methods allow for metacell clusters that contain multiple subclasses. To avoid any compositional effects, we filter out all metacell clusters containing cells from multiple subclasses.

Computing differential co-expression

Differential co-expression was computed by subtracting networks within datasets, then ranking the difference. Afterwards the differential networks were averaged across the datasets. To compute an FDR we used a null distribution of the average of 7 networks generated by sampling random uniformly distributed numbers.