Brain Structure and Function

Discover Mouse Gene Co-expression Landscapes Using Dictionary Learning and Sparse Coding

--Manuscript Draft--

Manuscript Number:	
Full Title:	Discover Mouse Gene Co-expression Landscapes Using Dictionary Learning and Sparse Coding
Article Type:	Original Article
Keywords:	Gene coexpression network, sparse coding, transcriptome
Corresponding Author:	Tianming Liu University of Georgia UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Georgia
Corresponding Author's Secondary Institution:	
First Author:	Yujie Li
First Author Secondary Information:	
Order of Authors:	Yujie Li
	Hanbo Chen
	Xi Jiang
	Xiang Li
	Jinglei Lv
	Hanchuan Peng
	Joe Tsien
	Tianming Liu
Order of Authors Secondary Information:	
Funding Information:	National Institutes of Health (DA033393) Dr. Tianming Liu National Institutes of Health (US) (AG042599) Dr. Tianming Liu
Abstract:	Gene coexpression patterns carry a rich amount of valuable information regarding enormously complex brain structures and functions. Characterization of these patterns in an unbiased, integrated and anatomically comprehensive manner will illuminate the higher order transcriptome organization and offer genetic foundations of functional circuitry. Here we demonstrate a data-driven method to extract coexpression networks from transcriptome profiles in the Allen Mouse Brain Atlas dataset. For each of the obtained networks, both the genetic compositions and the spatial distributions in brain volume are learned. A simultaneous knowledge of spatial distributions of a specific gene, the networks in which the gene plays and the weights it carries, can bring insights into the molecular mechanism of brain formation and functions. Gene ontologies and the comparisons with published gene lists reveal biologically identified coexpression networks, some of which correspond to major cell types, biological pathways and/or anatomical regions.
Suggested Reviewers:	Shuiwang Ji Associate Professor, Washington State University - Spokane sji@eecs.wsu.edu

	Heng Huang Professor, University of Texas at Arlington heng@uta.edu
	Guorong Wu Assistant Professor, University of North Carolina at Chapel Hill guorong_wu@med.unc.edu
	Jing Zhang Assistant Professor, Georgia state university jing.maria.zhang@gmail.com
Opposed Reviewers:	

Discover Mouse Gene Co-expression Landscapes Using Dictionary Learning and Sparse Coding

Yujie Li[1,*], Hanbo Chen[1,*], Xi Jiang[1], Xiang Li[1], Jinglei Lv[1,2], Hanchuan Peng[3,**], Joe Z. Tsien[4,**], Tianming Liu[1,**]

[1]Cortical Architecture Imaging and Discovery Lab, Department of Computer Science and Bioimaging Research Center, The University of Georgia, Athens, GA, USA; [2]School of Automation, Northwestern Polytechnical University, Xi'an, China; [3]Allen Institute for Brain Science, Seattle, WA, USA; [4]Brain and Behavior Discovery Institute, Medical College of Georgia at Augusta University, USA; [*]Co-first Authors, [**]Joint Corresponding Authors.

Abstract: Gene coexpression patterns carry a rich amount of valuable information regarding enormously complex brain structures and functions. Characterization of these patterns in an unbiased, integrated and anatomically comprehensive manner will illuminate the higher order transcriptome organization and offer genetic foundations of functional circuitry. Here we demonstrate a data-driven method to extract coexpression networks from transcriptome profiles in the Allen Mouse Brain Atlas dataset. For each of the obtained networks, both the genetic compositions and the spatial distributions in brain volume are learned. A simultaneous knowledge of spatial distributions of a specific gene, the networks in which the gene plays and the weights it carries, can bring insights into the molecular mechanism of brain formation and functions. Gene ontologies and the comparisons with published gene lists reveal biologically identified coexpression networks, some of which correspond to major cell types, biological pathways and/or anatomical regions.

Key words: Gene coexpression network, sparse coding, transcriptome

1 INTRODUCTION

Gene coexpression patterns carry rich amount of valuable information about enormously complex cellular processes (Peng et al., 2007). Previous studies have shown that genes displaying similar expression profiles are very likely to be involved in the same transcriptional regulatory program (Allocco et al., 2004; Mody et al., 2001), encode interacting proteins (Ge et al., 2001) or participate in the same biological processes (Tavazoie et al., 1999). A Gene Coexpression Network (GCN), represents the interactions among genes and is often used to study biological and genetic mechanisms across species and during evolution. For example, one pioneering work by Stuart et al (Stuart, 2003) is a comparative study on the microarray data of humans, flies, worms, and yeast. The results showed that multiple groups of conserved genes are associated with core biological functions. Knowledge of these key groups is an essential step to understand the overall design of genetic pathway. Efforts also went toward deriving common GCNs in the human brain (Hawrylycz et al., 2015; Oldham et al., 2008). Despite significant variations between individuals, preserved clusters of genes corresponding to discrete neuronal subtypes emerged from the comparisons of GCNs in different subjects. These consensus groups of genes consistently found in different subjects across brain regions provide strong evidence of a link between conserved gene expression and functionally relevant circuitry. In addition to revealing the intrinsic transcriptome organizations, GCNs have also demonstrated superior performance when they are used to generate novel hypotheses for molecular mechanisms of diseases because many disease phenotypes are not caused by one gene - or even a few genes or proteins, but as a result of dysfunction of a complex network of molecular interactions (Bando et al., 2013; Carter et al., 2013; Gaiteri et al., 2014).

Various proposals have been made to identify the GCNs. The most common and useful class of approach is clustering. Many clustering variants including hierarchical clustering and k-means clustering, have demonstrated good capability in identifying genes that share common roles in cellular processes (Bohland et al., 2010; Eisen et al., 1999; Tamayo et al., 1999). The alternative group of methods is to apply network concepts and models, which offers more descriptive power to the complicated molecular system, to describe gene-gene interactions (Oldham et al., 2012). Given the high dimension of genetic data and the urgent need in making comparisons to unveil the changes or the consensus between subjects or species, one common theme of all of these methods is dimension reduction. Instead of analyzing the interactions between over tens of thousands of genes, the groupings of genes by their co-expression patterns can considerably reduce the complexity of comparisons from tens of thousands of genes to dozens of networks or clusters, while preserving the original interactions.

Along the line of data-reduction, we proposed dictionary learning and sparse coding (DLSC) algorithm for GCN construction. DLSC is an unbiased data-driven method that learns a set of new bases (denoted as dictionaries) from the signal matrix so that the original signals can be represented in a sparse and linear manner. Because of the sparsity constraint, the dimension of genetic data can be significantly reduced. The grouping by co-expression patterns is encoded in the sparse coefficient matrix with the assumption that if two genes use the same dictionary to represent their original signals, then their gene expressions must share similar patterns, thereby considering them as “co-expressed”. The proposed method overcomes the potential issues of overlooking multiple roles of regulatory domains in different networks that are seen in many clustering methods (Gaiteri et al., 2014) because DLSC does not impose the bases be orthogonal and that one gene can be claimed by multiple networks. More importantly, for each of the obtained networks, both genetic compositions and spatial distributions in brain volume are learned. A simultaneous knowledge of the distributions of a specific gene, the networks in which the gene functions and the weights it carries can bring insights into the genetic mechanism of brain formation and functions.

Most of the GCNs were constructed from the microarray data and *in situ* hybridization (ISH) data. One major advantage of ISH over microarray data is that ISH preserves the precise spatial distribution of genes. One of the most valuable ISH resources is the openly available Allen Mouse Brain Atlas (AMBA) initiated by the Allen Institute for Brain Sciences (Lein et al. 2007), which surveyed over 20,000 genes expression patterns in 56-day-old C57BL/6J mouse brain using ISH. This valuable dataset, featured by the whole-genome scale, cellular resolution and anatomically comprehensive coverage, allows systematic and holistic investigation of the molecular underpinnings and related functional circuitry. Using AMBA, the GCNs identified by DLSC showed significant enrichment for major cell types, biological functions, anatomical regions, and/or brain disorders. The identified GCNs holds promises to serve as foundations to explore different cell types and functional processes in diseased and healthy brains.

2 METHODS

The computational pipeline of proposed framework is illustrated in Figure 1. The pipeline consists of two parts: the slice-based GCN construction and validation (Figure 1a-d) and global GCN construction and analysis (Figure 1e).

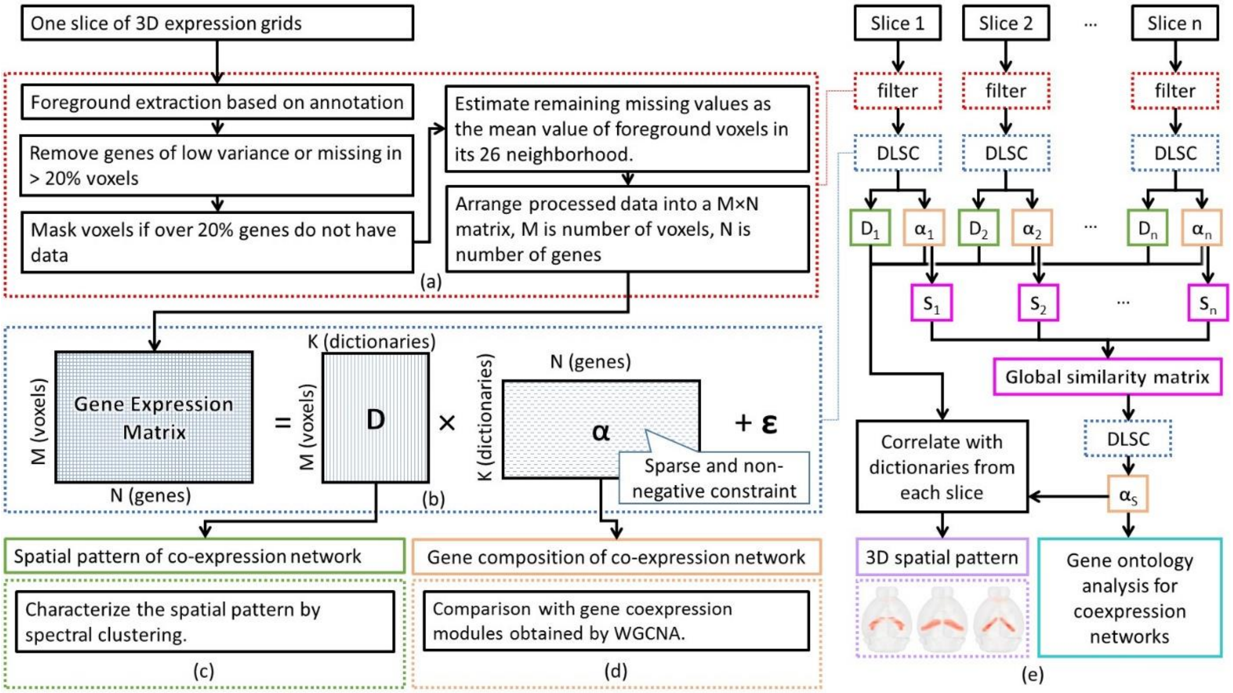
One slice of 3D expression grids
Foreground extraction based on annotation
Remove genes of low variance or missing in > 20% voxels
Mask voxels if over 20% genes do not have data
Estimate remaining missing values as the mean value of foreground voxels in its 26 neighborhood.
Arrange processed data into a M×N matrix, M is number of voxels, N is number of genes
(a)
M (voxels)
Gene Expression Matrix
N (genes)
=
K (dictionaries)
M (voxels)
D
×
K (dictionaries)
N (genes)
α
+ ε
Sparse and non-negative constraint
(b)
Spatial pattern of co-expression network
Characterize the spatial pattern by spectral clustering.
(c)
Gene composition of co-expression network
Comparison with gene coexpression modules obtained by WGCNA.
(d)
Slice 1
Slice 2
...
Slice n
filter
DLSC
D1
α1
D2
α2
Dn
αn
S1
S2
Sn
Global similarity matrix
DLSC
αS
Correlate with dictionaries from each slice
3D spatial pattern
Gene ontology analysis for coexpression networks
(e)

Figure 1. Computational pipeline for constructing slice-wide GCNs (a)-(d) and brain-wide GCNs (e). (a) Raw ISH data preprocessing step that removes unreliable genes and voxels and estimates the remaining missing data. (b) Dictionary learning and sparse coding of ISH matrix with sparse and non-negative constraints on coefficient $\boldsymbol{\alpha}$ matrix. $\mathbf{D}$ is the dictionary matrix and $\boldsymbol{\alpha}$ is the coefficient matrix. $\boldsymbol{\varepsilon}$ is the reconstruction error. (c) Visualization of spatial distributions of slice-based GCNs and brain delineation by spectral clustering using dictionaries as feature vector. (d) Characterization of gene composition of GCN and validation by comparing with WGCNA. (e) Integrating slice-based GCNs into global GCNs and global GCN gene ontology.

2.1 Experiment Material

AMBA is a genome-wide cellular-resolution map of gene expressions using ISH that offers brain-wide anatomical coverage of mouse brain. The inbred mouse strain is used to reduce the animal-to-animal variation in brains. For each tested gene, the mouse brain was sectioned into series of tissues in coronal or sagittal planes and then imaged. To enable three-dimensional volumetric representations from the acquired coronal or sagittal series images, a common coordinate space of the three-dimensional (3D) reference atlas was first created so that the ISH images of each gene can be consistently registered to the same space and aligned. Later each image was uniformly divided into 200×200 um grids and gene-expression statistics were computed from the detected signals for each voxel. The resulted voxelized expression grids encoding the important spatial information of over 4,345 genes in coronal sections and 21,718 genes in sagittal sections make up the key components of the AMBA.

We downloaded the 4,345 3D volumes of expression energy of coronal sections as well as the corresponding reference atlas from the website of ABA (http://mouse.brain-map.org/) to perform our analysis. Coronal sections are chosen because they registered more accurately to the reference model than the counterparts of sagittal sections. The dimension of all 3D volumes applied in this study is 67×41×58.

2.2 Slice-wide GCN Construction and Validation

The major obstacle to a global analysis of ISH data on all coronal slices is the number of missing data observed on each slice (Supplementary Figure S1). Since each slice has its own missing genes, in order to obtain a common set of genes on all slices will require roughly 33% of the genes removed from analysis, resulting in a significant amount of information loss. Additionally, as the ISH data is acquired by each coronal slice before they were stitched and aligned into a complete 3D volume, despite extensive preprocessing steps (Ng et al., 2007) such as a global adaptive thresholding method and morphological filtering employed to remove noise and connect broken segments, quite significant changes in average expression levels of the same gene between slices are observed (Supplementary Figure S1). Considering these problems, studying the coexpression networks slice by slice enables leveraging off the information loss and alleviation of the artifacts due to slice handling and preprocessing. Yet additional efforts are needed to integrate gene-gene interactions on each slice.

2.2.1 Data preprocessing

For slice-wide analysis, the input of the pipeline are the expression grids of one of 67 coronal slices. A preprocessing module (Figure 1a) is first applied to handle the foreground voxels with missing data (-1 in expression energy). The lack of data is assumed mostly due to problems during ISH and image processing steps such as missing slices, broken tissue, and slice alignment. Specifically, this module includes an extraction step, a filtering step and an estimation step. First, the foreground voxels of the slice based on the annotation map from ARA were extracted. Then the genes of low variance (standard deviation <0.5) or genes with missing values in over 20% of foreground voxels were excluded from further analysis because they provide little information to network construction. A similar filtering step is also applied to remove voxels in which over 20%

genes do not have data. Most missing values were resolved in the two filtering steps. The remaining missing values were recursively estimated as the mean of foreground voxels in its 8 neighborhood until all missing values were filled. The maximum number of iterations is 4 with most values using 2 or 3 iterations. The low number of iterations suggest that the estimated data is reasonable. After preprocessing, the cleaned expression energies were organized into a matrix and sent to DLSC (Figure 1b). In DLSC (section 2.2.2), the gene expression matrix is factorized into a dictionary matrix **D** and a coefficient matrix **α**. These two matrices encode the distribution and composition of GCN (Figure 1c-d) and will be further analyzed and validated against the raw data and existing method.

2.2.2 Dictionary Learning and Sparse Coding

DLSC is a popular method to achieve a compressed and succinct representation for ideally all signal vectors. Given a set of M-dimensional input signals X=[x1,…,xN] in $\mathbb{R}^{M\times N}$, learning a fixed number of dictionaries for sparse representation of X can be accomplished by solving the following optimization problem:

$$< \mathbf{D}, \boldsymbol{\alpha} >= \operatorname{argmin} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{D} \times \boldsymbol{\alpha}\|_2^2 \ \ s.t \ \|\boldsymbol{\alpha}\|_1 \ \leq \ \lambda \qquad (1)$$

where $\boldsymbol{D} \in \mathbb{R}^{N\times K}$is the dictionary matrix, $\boldsymbol{\alpha} \in \mathbb{R}^{K\times M}$ is the corresponding loading coefficient matrix, λ is a sparsity constraint factor and indicates each signal has fewer than λ items in its decomposition, $\|*\|_2$ is the summation of ℓ_2 norm of each column and $\|*\|_1$ is the summation of ℓ_1 norm of each column. $\|\boldsymbol{X} - \boldsymbol{D} \times \boldsymbol{\alpha}\|_2^2$ denotes the reconstruction error.

In efficient sparse coding algorithm, the optimization problem is solved by an alternating minimization procedure through lasso and least-square steps that iteratively updates to improve the estimate of the sparse codes while keeping the dictionaries fixed and then updating dictionaries that fit the sparse codes best. At all times, the energy function in equation (1) should be minimized.

As will be discussed later that each entry of **α** indicates the degree of conformity of a particular gene to a coexpression network, a non-negative constraint was added to the ℓ_1-regularization. This additional prior, included in equation (2), can be handled by homotopy method presented in Efron et al (Efron et al., 2004).

$$< \mathbf{D}, \boldsymbol{\alpha} >= \operatorname{argmin} \sum_{i=1}^{N} \frac{1}{2} \|x_i - \boldsymbol{D} \times \alpha_i\|_2^2 \ \ s.t \ \|\boldsymbol{\alpha}\|_1 \ \leq \ \lambda \, , \forall \, \mathrm{i}, \alpha_i \geq 0 \qquad (2)$$

In practice, the gene expression grids are arranged into a single matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$, such that *M* rows correspond to *M* foreground voxels for analysis and *N* columns correspond to *N* genes (Figure 1(b)). Then, each column of the matrix (gene signal in a voxel) was normalized by its Frobenius norm. After normalization, the publicly available online DLSC package was applied to solve the matrix factorization problem proposed in equation (2) (Mairal et al., 2010). Eventually, the gene expression energy matrix ***X*** was represented as sparse combinations of learned dictionary atoms $\boldsymbol{D}$. Each column in D is one dictionary consisted of a set of voxels. Each row in $\boldsymbol{\alpha}$ corresponds to one dictionary and details the coefficient of each gene in a particular dictionary.

The key assumptions of enforcing the sparseness is that each gene is involved in a very limited number of gene networks. The non-negativity constraint on **α** matrix imposes that no genes with the opposite expression patterns will be placed in the same network.

One mathematical interpretation of the DLSC is that a set of dictionaries are learned and used as new bases so that the original matrix can be described by a sparse matrix **α**. In the context of deriving GCNs, we consider that if two genes use the same dictionary to represent the original signals, then the two genes are coexpressed in this dictionary. There are several benefits of this set-up. First, both the dictionaries and coefficients are learnt from the data and therefore should reflect the intrinsic organization of transcriptome. Second, the level of co-expressions is quantifiable, and the level is not only comparable within one dictionary, but the entire **α** matrix.

Further, if we consider each dictionary as one network, the corresponding row of **α** matrix contains all the genes that use this dictionary for sparse representation, or that are ‘coexpressed’. Additionally, each entry of **α** measures the extent to which this gene conforms to the coexpression pattern described by the dictionary atom. Therefore, this network, denoted as the coexpression network, is formed. Since the dictionary atom is composed of multiple voxels, by mapping each atom in **D** back to the ARA space, we can visualize the spatial patterns of the coexpressed networks. Combining information from both **D** and **α** matrices, we would obtain a set of intrinsically learned GCNs with the knowledge of both their anatomical patterns and gene compositions. As the dictionary is the equivalent of network, these two terms will be used interchangeably.

2.2.3 Parameter Selection

The choice of number of dictionaries and the regularization parameter λ are crucial for effective sparse representation. As there exists no golden standard for parameter selection, we first proposed three criteria to evaluate the performance of DLSC and then carried out a grid search on the optimized parameters using one example slice.

The first criteria is the reconstruction error. It is defined as the least square difference between the original signal matrix and the reconstruction from sparse representation [equation (3)]. A high reconstruction error indicates a less accurate representation.

$$error_y = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D\alpha}\|_F^2 \tag{3}$$

The second evaluation metric is the average uncertainty coefficient (AUC) between the obtained dictionaries and the reference atlas. The uncertainty coefficient, defined in [equation (5)] is a normalized variant of mutual information (MI). Many studies have shown that different combinations of gene expression profiles mirror the gross anatomical partitioning (Dobrin et al., 2009; Oldham et al., 2008). We thus assume the set of the parameters that result in the highest correspondence between the transcriptome patterns and canonical anatomical structures are the optimal parameters. MI, as a powerful criterion to measure the dependencies between variables, can be used to characterize how well the transcriptome patterns match with the canonical neuroanatomical divisions, thereby a good estimate on how meaningful the components are. The advantage of using the normalized MI, is that it varies between 0 and 1 with values close to zero indicating the two spatial distributions are independent whereas values close to one suggesting knowledge of one spatial pattern can reduce uncertainty of the other and thereby being used to predict the other one.

In specific, MI is first calculated between the spatial distribution of each gene network and the reference atlas. Given a continuous variable X that contains the spatial distribution of one gene network, discretization is performed via histogram with an empirically selected 32 equally divided bins. Let categorical variable Y represent the labels in the reference atlas. The MI can be calculated as:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{4}$$

where *p(x,y)* is the joint probability distribution function of *X* and *Y*, and *p(x)* and *p(y)* are the marginal probability distribution functions of *X* and *Y* respectively.

Then the uncertainty is obtained from:

$$U(X,Y) = \frac{2 * \ I(X,Y)}{H(X) + H(Y)} \tag{5}$$

where H(X) and H(Y) are the marginal entropies. For a particular λ and number of dictionaries, the average AUC of all GCNs is used to compare.

Another important measurement to examine the DLSC performance is the degree of density measured by the percentage of none-zero-valued elements in the coefficient matrix. As we are trying to find a set of dictionaries that are rich in representation power so that a compact code can be achieved, a relatively low value is expected. As discussed in DLSC section, the density is regulated by λ. In most cases, increasing λ will give rise to more zero entries in the coefficient matrix. It should be noted that there is no exact monotonic relation between λ and the density of the solution (Mairal et al., 2010). Therefore it would be helpful to monitor λ during the parameter selection process.

Having set up the three criteria, a grid search was performed on slice 27. This slice is chosen due to its good anatomical coverage of various brain regions. As different number of genes are expressed in different slices, the number of dictionaries for each slice should change accordingly. Instead of a fixed set number of dictionaries, a gene-dictionary ratio is used to determine the optimal ratio between the number of genes expressed and the number of dictionaries required to achieve a good representation. 55 combinations of λ and gene-dictionary ratios are considered with 5 choices of λ and 11 different gene-dictionary ratios (Supplementary table S1-3). The results using 55 different combinations of parameters are available at http://mbm.cs.uga.edu/mouse/gcn/para_select/slice.html. As the final goal of parameter selection is to choose a set of parameters that result in a sparse and accurate representation of the original signal, which is translated to a low reconstruction error, a high AUC and a low density, λ=0.5

and gene-dictionary ratio of 100 is the best option among 55 parameter combinations and chosen as the optimal parameters.

2.2.4 Brain parcellation using DLSC

The decomposition of gene expression matrix on each slice results in a dictionary matrix **D** and a coefficient matrix **α**. Each row of **D** describes which dictionary and how much weight one voxel participates in that dictionary. It is assumed that if two voxels have similar dictionary features, i.e. two voxels are involved in the same dictionary (network) and carry similar weights, then these voxels are considered highly similar. With the dictionaries as the feature vector of a voxel, Pearson correlation is employed to calculate the similarity between voxels. Then the voxels on the slice are clustered into groups by spectral clustering (Chen et al., 2013; Luxburg, 2007). The number of clusters is adapted to the data and determined by normalized cut with a threshold of 0.7. (Chen et al., 2013; Luxburg, 2007)

2.2.5 Comparative analysis with Weighted Gene Correlation Network Analysis (WGCNA)

WGCNA is applied on the same dataset to validate findings generated by DLSC. WGCNA (Langfelder and Horvath, 2008) is an unbiased, unsupervised framework to identify coexpressed gene modules. In the framework, genes are viewed as nodes in a weighted network. To achieve a robust and sensitive measure of the interaction between genes, the proximity measure between genes, - namely Topological Overlap Measure (TOM), considers not only the direct connection strength between two genes, but also the connection strengths these two genes share with other "third party" genes. Then based on TOM, genes are clustered into multiple modules using average linkage hierarchical clustering. The module eigengene, defined as the first principal component of the standardized expression profiles of the module is used as a succinct representation of the gene expression profiles of the module. In this study, signed network is used to avoid the “anti-reinforcing” connection strength that might occur in unsigned network. Default parameters were used (β=12, height cut to merge=0.15). As the size of the smallest dictionary is 45 in DLSC, we set the minimum cluster size to 40 in WGCNA.

For clarity, the groups identified by WGCNA and DLSC are denoted as modules and GCNs respectively. To quantitatively compare the found networks, both methods were applied on the gene expressions of the same slice – slice 27. Then the number of shared genes were counted between groups identified by both methods and then normalized by DLSC dictionary length (Figure 2a) or by WGCNA module length (Figure 2b). The resulted overlap percentage is between 0 and 1 with 1 indicating the group is exactly included by the groups identified by the other method and 0 indicating no shared gene found. Besides quantification, another intuitive way to compare the two methods is by comparing the obtained spatial maps. Similar gene groups are very likely to display similar spatial maps. In DLSC, the dictionary atom encodes the network spatial patterns. In WGCNA, the spatial distributions are represented by the spatial pattern of the eigengene of that module.

2.3 BRAIN-WIDE GCNS CONSTRUCTION AND ANALYSIS

2.3.1 Brain-wide GCNs construction

To construct brain-wide coexpression networks, we need to consider the gene interactions on all coronal slices. First, gene similarity on each slice, denoted as the local similarity, is calculated from the coefficient matrix $\boldsymbol{\alpha}$ with the coefficients as the feature of each gene. Let $v1$, $v2$ be the coefficient vectors of gene1 and gene 2. The gene similarity measure is defined as the overlap rate OR, as shown below:

$$OR(v1, v2) = 2\frac{|\min(v1, v2)|}{|v1| + |v2|} \tag{6}$$

where |*| is the ℓ_1 norm of feature vector.

As mentioned above that each slice has missing data for different genes, the interactions of these missing genes on the particular slice should not be considered in the global similarity matrix construction. Therefore, the global gene similarity, i.e., the similarity measure that considers interactions on all slices, is measured by the median of the local similarities of genes with sufficient data. The rationale of adopting a global similarity matrix instead of simply aggregating the coefficients matrices on each slice is to mitigate the influence of missing data as well as the artifacts generated during data acquisition.

In the constructed global similarity matrix, 91 genes show zero similarity to any other genes. The very low similarity is caused by the lack of data, evidenced by that these 91 genes are present in at most 5 out of 67 slices. The separation of these genes that suffer from heavy data loss demonstrates the effectiveness of similarity matrix over the original $\boldsymbol{\alpha}$ matrix, and also reflects the OR as an appropriate measure for gene similarity in this situation.

4254 out of 4345 genes were used to derive the brain-wide GCNs. The global similarity matrix is the input to the subsequent DLSC. The goal of performing DLSC on the similarity matrix is to assign network membership to genes by their associations to all the other genes. We assume that if two genes display similar relationship to all the other genes, these genes should belong to the same group. The network memberships are encoded in the resulted sparse coefficient matrix $\boldsymbol{\alpha}$.

2.3.2 Parameter selection

The parameter selection of decomposing the global similarity matrix is guided by the knowledge from the slice-based study that each network is consisted of on average 185 genes and each gene participates in 1.85 networks. Using these criteria, we performed a grid search of λ and dictionary numbers (Supplementary table S4-5) and selected λ as 0.3 and dictionary number 50, which resulted in an average of 189 genes per network and a slightly larger 2.21 networks for one gene.

2.3.3 Fuse 3D spatial pattern of GCNs

As described in section 2.2.2, the dictionaries trained in each slice encode rich information of the spatial distribution of GCNs. Intuitively, we can fuse the dictionaries of each slice to study the 3D spatial pattern of brain-wide GCNs. First, the similarities between brain-wide GCNs and slice-wide GCNs were calculated. Then, we scaled slice-wide dictionaries based on similarity and accumulated them to generate a 3D volume. Specifically, the similarity was calculated based on the OR of coefficient matrix defined in section 2.3.1. Slightly different from previous definition, here the similarity was calculated between GCNs instead of genes. Also, before comparison, each feature vector were normalized so that the maximum value equals to 1.

2.3.4 Gene ontology analysis of brain-wide GCNs

Brain-wide GCN characterization were made based on common GO gene ontology categories (Molecular Function, Biological Process, Cellular Component), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003). Enrichment analysis was performed by cross-referencing with published lists of genes (Miller et al., 2011) related to cell type markers, known and predicted lists of disease genes, specific biological functions etc. Significance was assessed using one-sided Fisher's exact test with a threshold of 0.01.

3 RESULTS

The organization of the result section is as follows. In section 3.1, we constructed GCNs on each slice. With slice 27 as an example, the slice-based GCNs were validated first by a visual inspection against raw ISH data where the GCNs are derived and then by a comparative study with one of the most widely used methods - WGCNA. On the side as one application, we demonstrated that the learnt dictionaries, 100 fold shorter in length than the gene expressions, can be a relevant and compact feature for brain parcellation. Having established the slice-wide GCNs, section 3.2 will focus on the construction of global GCNs by integrating the gene-gene interactions on all slices. Along with the spatial distribution of the GCNs, we will show that the obtained GCNs are biologically meaningful by comparing with the known gene ontologies and published gene lists.

3.1 SLICE-WIDE GCN ANALYSIS

To show as an example, slice 27 is analyzed due to its good anatomical coverage of various brain regions. Results of all other slices are available at http://mbm.cs.uga.edu/mouse/gcn/allslices/all_slice_anatomy_overview.html. The detailed information including the genes and spatial distributions of modules identified by WGCNA can also be found at http://mbm.cs.uga.edu/mouse/gcn/wgcna_s27_signed_40/overview.html.

3.1.1 Comparative analysis with WGCNA

Both DLSC and WGCNA were applied on gene expressions data of slice 27. WGCNA organized genes into 11 modules while 29 networks were found by DLSC. In Figure 2a, all WGCNA modules were projected to the DLSC GCN space. As WGCNA found fewer numbers of groups, for each WGCNA module, there exists one or more overlapping GCNs. For example, as indicated by the first row of Figure 2a, WGCNA module 1, consisting over 1000 genes, corresponds to several GCNs such as 13, 15, 25 and 27. An inspection of the spatial distributions of these networks confirms the correspondence (Supplementary Figure S2). Also, WGCNA module 2 corresponds to GCN 3 and 26. On the other hand, by projecting the DLSC networks to the WGCNA module space (Figure 2b), some GCNs are found covering the entire or most of elements in WGCNA modules. For instance, the genes in GCN 24 includes almost all genes found in module 11. Similarly, GCN 9 finds about 90% of genes found in module 9 in addition to other genes. The similarity in the networks identified by two methods is not only evident in the gene constituents, but also reflected in the spatial maps illustrated in Figure 3.

To further understand the similarities and differences of the networks identified by the two methods, we performed a more detailed analysis on WGCNA module 2, 4 and 11 and the corresponding GCN 3, 26, 19, 22, and 24 (Figure 3). To first demonstrate that within a GCN the expression patterns are consistent, along with the spatial map, three ISH raw data of the representative genes in the network where the spatial pattern is derived are also displayed. As seen in Figure 3, the ISH raw data match well with the respective spatial map. In GCN 3, the expression peaks at caudaputamen. All three genes showed strong signals in the caudaputamen in the raw data. Similarly, the expression patterns of GCN 24, as well as the three genes showed significantly enhanced signals in the septal region compared with the others. Relatedly, the weight in the parentheses indicates the extent to which a gene conforms to the coexpression pattern. A higher weight indicates a stronger resemblance of the raw data to the spatial map. One example is a comparison among *Dlx1, Col6a3* and *A930038C07Rik* (Figure 3e). The GCN 24 coexpression pattern covers the rostral and caudal part of lateral septal nucleus and septofimbrial nucleus. In *A930038C07Rik* and *Dlx1*, weaker signals were found in caudal part of lateral septal nucleus and septofimbrial nucleus (Figure 3e blue arrows). In *Dlx1*, enhanced signals that are inconsistent with the coexpression pattern were also seen in the bed nuclei of stria terminalis (Figure 3e red arrow). The decreased similarity agrees well with the declining weights.

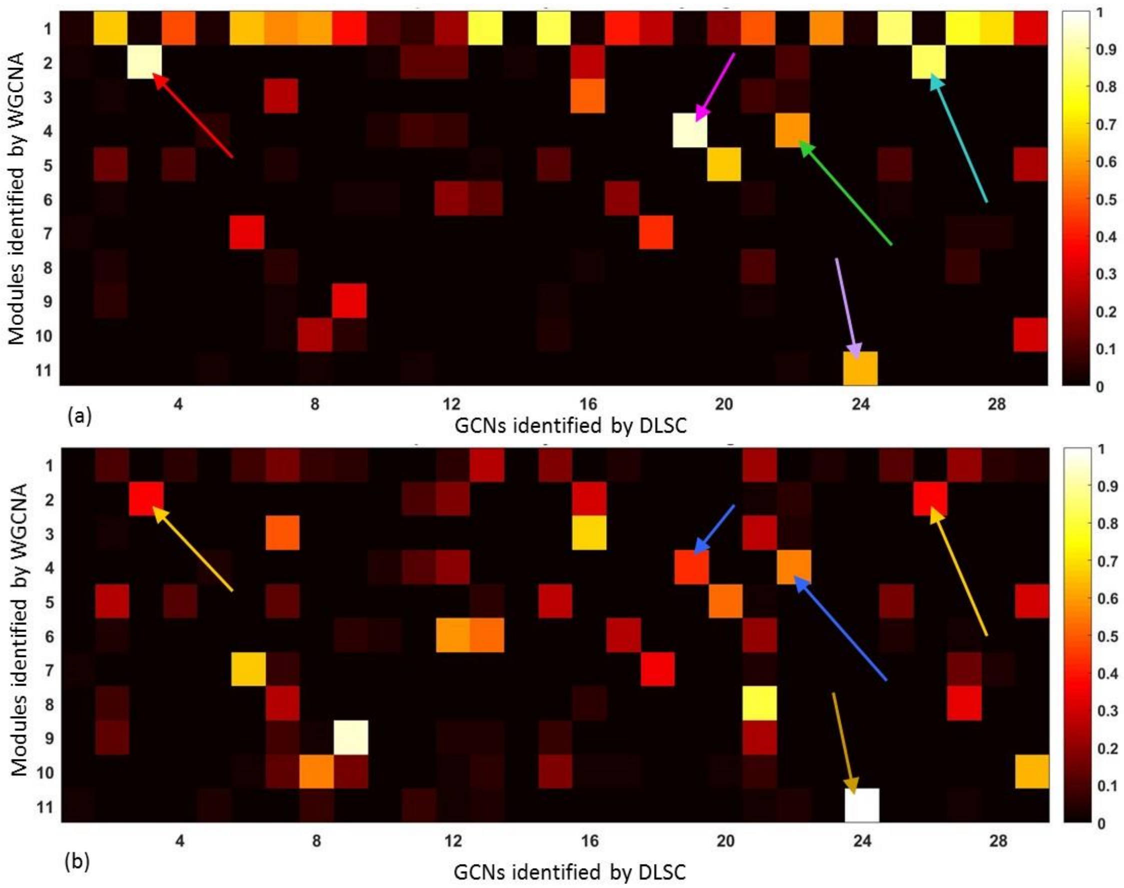

Figure 2. Comparisons between GCNs and modules identified by DLSC and WGCNAs. The number of shared genes were counted between two networks obtained by WGCNA and DLSC and then normalized by the number of genes in the corresponding network obtained by (a) DLSC or (b) WGCNA. The arrows highlighted the networks shown in Figure 3 of corresponding colors.

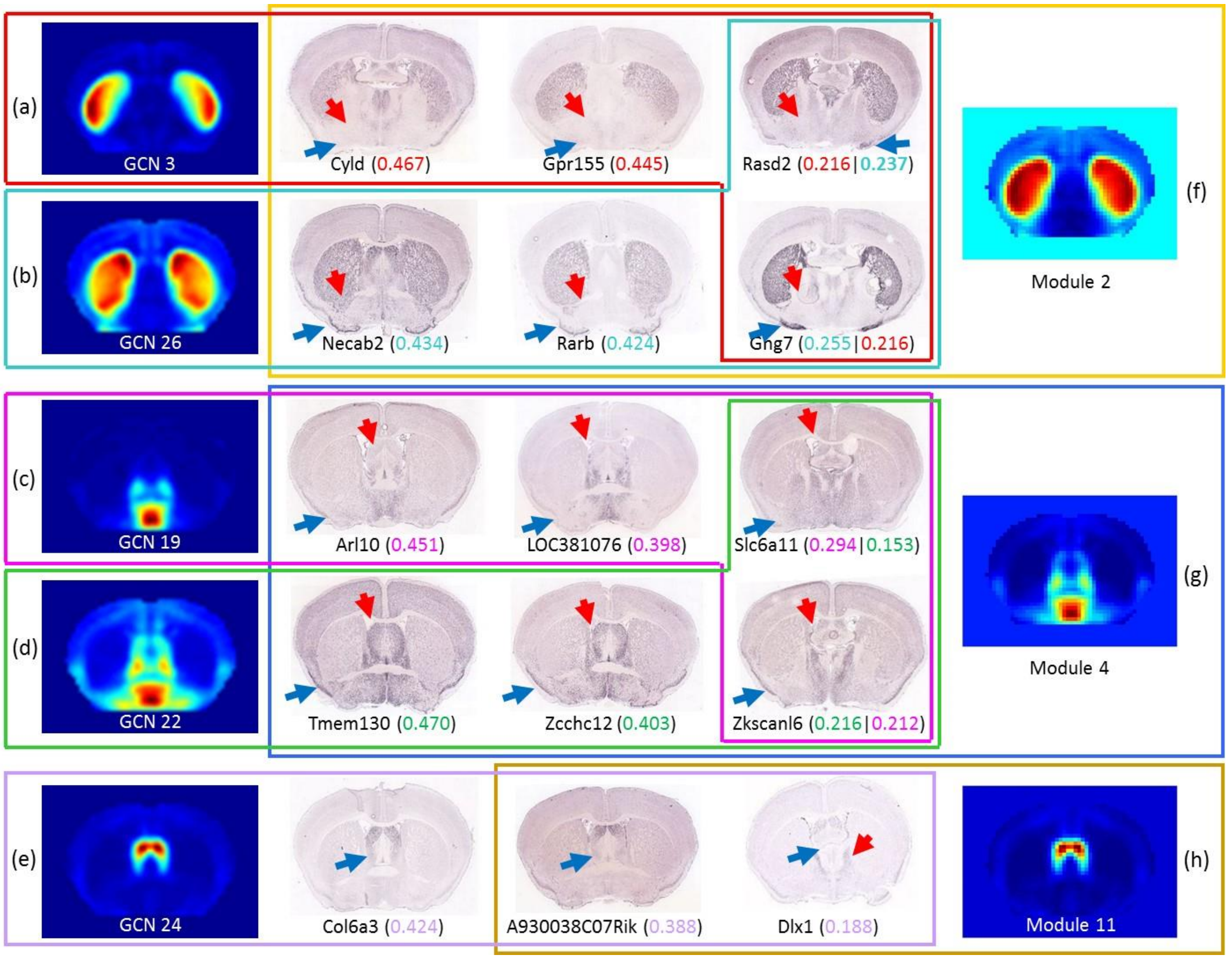

Figure 3. Comparisons between networks found by DLSC and WGCNA on slice 27. The first column are the spatial maps of networks found by DLSC and last column are the spatial maps of WGCNA module eigen-genes. The three columns in the middle are the ISH raw data of representative genes of the networks. Gene acronyms and the weights are listed at the bottom. A higher weight indicates a higher similarity to the coexpression patterns of this network. Genes circled by the same colored box are in the same network/module. In spatial maps, red and blue indicates high and low gene expression levels respectively.

Next we investigated some of the selected modules and networks obtained by the two methods. The WGCNA module 2 is broken down into GCN 3 and 26. As illustrated in Figure 3a-b, these two GCNs display very similar spatial patterns to each other and to the counterpart of module 2, which confirms the correspondence between the two methods. However, irrespective of the major covering regions, the differences between the spatial distributions of two GCNs are still evident. GCN 3 shows a higher expression at the lateral caudoputamen while the expression patterns in GCN 26 peak at the dorsal and medial part of caudoputamen. Another distinct feature is the simultaneous coverage of olfactory tubercle in GCN 26 that is evident in *Necab* and *Rarb* ISH image. This feature, pointed out by *Cyld* and *Gpr155* (Figure 3a), is absent in GCN 3. It is worth mentioning that as in our method genes can have multiple assignment, 19 genes, including *Rasd2* (Figure 3a) and *Gng7* (Figure 3b) are involved in both networks. Given the similarity of the two spatial maps, it should not be a surprise that these genes function in both networks.

Similarly, module 4 corresponds to two GCNs 19 and 22. Like the above example, both GCNs show similar spatial patterns in general with innegligible differences. The difference lies in that GCN 19 centers at bed nuclei of the stria terminalis and preoptic area while GCN 22 extends much further to substantia innominate and part of striatum. It is known that substantia innominate has wide projections to the neocortex via nucleus basalis. The coexpression of the two regions might be associated with the distribution of specific cells or communications via synapses corroborated by that many genes in this GCN are functionally related to synapses and protein transportation.

Another interesting observation is the strong overlap between module 11 and GCN 24. Notably, all 46 genes in this module are included in GCN 24. This overlap explains the almost identical network spatial map (Figure 3e-f). In addition to the shared 46 genes, DLSC also finds 27 other genes. Yet all of them have relatively low weights (<0.224), indicating a lower level involvement in the GCN. In other words, these genes will not be considered as the contributing genes if a higher threshold is chosen during the determination of the participating genes. On the other hand, 24 of these 27 genes were not assigned to any module by WGCNA and 3 were assigned to module 1.

Overall, we have demonstrated that genes within a GCN show common coexpression patterns and the level of similarity of a particular gene to the coexpression patterns is correctly measured by the weights. Then with a detailed analysis of slice 27 we showed a good agreement between the DLSC and WGCNA by both the number of overlapping genes and spatial distributions, which verifies the GCNs generated by DLSC.

3.1.2 Gene Coexpression Network and Brain Parcellation

Existing literatures have shown that transcriptional profiles reflect the gross brain anatomical structures (Lein et al., 2004). Since DLSC is also a dimension reduction step that reduces the transcriptional profile consisting of ~3500 features into a feature vector composed of ~35 dictionaries for a single voxel, we hypothesized that the learned dictionaries can preserve the (dis)similarities between two regions defined by their transcriptional profiles, thus serving as a very relevant and compact feature for brain delineation. As seen in Figure 4, voxels resulted from spectral clustering form a set of spatially contiguous clusters partitioning the slice. The formation of these single tight clusters agrees with the previously identified brain's organizational principle that transcriptome similarities are strongest between anatomical neighbors. The delineations are in general symmetric and match major canonical brain regions including hippocampus (blue arrows), hypothalamus (red arrows), thalamus (magenta arrows) etc. The most striking and principal features are the laminar and areal patterning that are seen in almost all slices (highlighted by yellow and orange arrows in Figure 4(a)-(e)). The patterning - defined by the abrupt changes in gene expression, has been discovered in mammalian brains such as mouse (Hawrylycz et al., 2010) and human (Miller, 2014) and is known crucial to the formation of specialized brain anatomical and functional areas (O'Leary et al., 2013). Within a dominant layered organization, layer specific areal patterning is also apparent. For instance, isocortex layers are further divided into motor areas (green arrows), somatosensory area (orange arrows), piriform area (pink arrows), retrosplenial area (dark green arrows), auditory area (purple arrows), and visual area (black arrows). It is worth mentioning the level of coherence in the partitioning across slices. Some subregions with potentially stable gene expression patterns are consistently found in adjacent slices despite of the slice-to-slice variations in anatomical structures and that DLSC and spectral clustering are performed separately on each slice. One example is slice 39 and slice 40. Some major canonical regions such as ventricles (white arrows), hippocampus (blue arrows), thalamus (magenta arrows), retrosplenial area (dark green arrows) are consistently identified in both slices. The consistent and legitimate segmentations not only demonstrate the validity of DLSC in succinctly representing the transcriptome profile, but also provides strong evidence that the observed networks are reproducible and that there exist unique and robust genetic signatures for different brain structures.

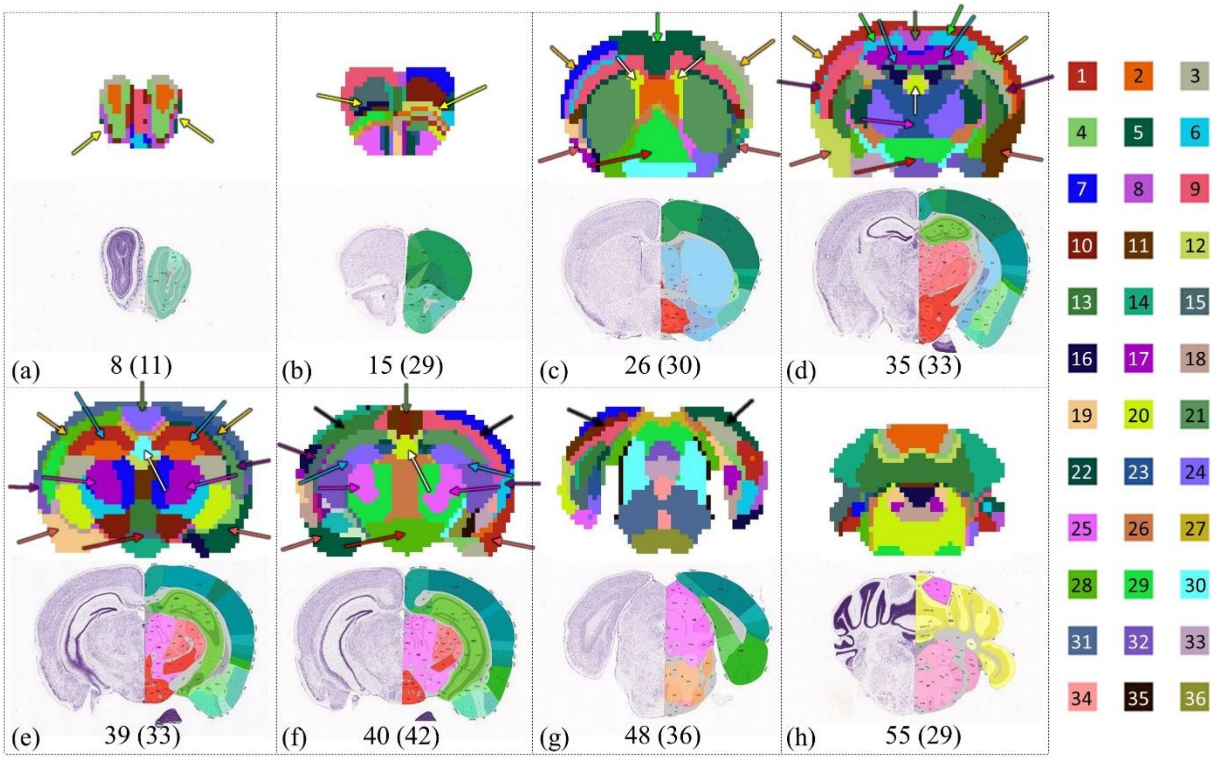

Figure 4. Representative anatomical divisions based on the GCN features. Eight panels correspond to eight selected slices. In each panel, top row: brain parcellation obtained from spectral clustering with dictionaries as feature vector; second row: visualization of Nissl stain image (left) and brain ontology (right) of the corresponding slice downloaded from ABA. Bottom: Slice number and the number of division. Color code of each region is shown on the right.

3.2 BRAIN-WIDE GCN ONTOLOGY AND SPATIAL PATTERN

Comparisons with the published lists of genes related to cell type markers, specific biological functions and known and predicted lists of disease genes reveal exciting biological insights for the constructed GCNs. A complete summary of each brain-wide GCN is available at http://mbm.cs.uga.edu/mouse/gcn/globalGCN/Global_GCNs_overview.html. Multiple brain-wide GCNs are consistently identified to be enriched in certain functional category by several distinct studies using different types of data and different methods for analysis. For example, a comparison with the gene lists generated using purified cellular population indicates that GCN 5, 16, 23, 30, 43, 45 are enriched with markers of astrocyte. Among them, GCN30 and GCN43 are consistently confirmed as astrocyte-enriched by the lists generated using WGCNA on microarray data and gene lists generated using Anatomic Gene Expression Atlas (AGEA) (Ng et al., 2009) on ISH data. Similarly, the significant enrichment of markers of oligodendrocyte is reproducibly identified in GCN 24 and GCN 12, 18, 20, and 22 are significantly enriched with markers of neuron. The consistency of the biological interpretations of the obtained GCNs corroborated by studies using different data types and different analysis methodologies indicate that the GCNs reflect the intrinsic transcriptome organization instead of data-specific or method-specific patterns. Among the major cell types, several GCNs are identified to be enriched in neuron subtypes including pyramidal neurons, GABAergic neurons and Glutamatergic neurons (Sugino et al., 2006). The gene lists for these neuron subtypes are derived from separated populations using retrograde tracing and fluorescent labelling at different regions of adult mouse forebrain (Sugino et al., 2006). Other networks such as GCN 11, 15, 20 and GCN 12, 41 described mitochondrial, ribosomal functions. Literatures suggested that the upregulated or downregulated expressions in these networks can be associated with aging and brain diseases (Blalock et al., 2004; Lu et al., 2004).

The biological meaning of the GCNs have been not only confirmed by existing literatures but also corroborated by the GO terms using DAVID. For example, two significant GO terms in GCN24 are myelination ($p=7.7\times10^{-7}$) and axon ensheathment ($p=2.5\times10^{-8}$), which are featured functions for oligodendrocyte, with established markers including *Plp1* (proteiolipid protein), *Mbp* (myelin basic protein), *Pmp22* (peripheral myelin protein 22), and *Ugt8a* (UDP galactosyltransferase 8A). DAVID also suggests that GCN41 are significantly enriched in the KEGG ribosome pathway ($p=2.5\times10^{-6}$), agreeing with the other studies on human and mouse (Table 1). Also consistent with the enrichment of mitochondrial function, DAVID suggests that GCN 11 is highly enriched in the KEGG oxidative phosphorylation pathway ($p=4.9\times10^{-7}$) and significant BPs include generation of precursor metabolites and energy (1.2×10^{-6}) and ATP metabolic process (5.1×10^{-6}).

A visualization of the spatial map also offers a useful complementary information source (Figure 5). For example, the fact that GCN 5 (Figure 5ii) locates at ventricle, where the subventricular zone is rich with astrocytes (Quinones-Hinojosa and Chaichana, 2007), confirms its enrichment in astrocyte markers. GCN 7 (Figure 5v) is mainly distributed in the deeper layers of neocortex, which is reminiscent of the distribution glutamatergic projection neuron in layer V (Molyneaux et al., 2007). GCN 23, located mainly at cerebellar region (Figure 5vi) and the indicated enrichment in GABAergic pointed to a potential enrichment of GABAergic subtype neuron - the Purkinje cells. Comparing with the gene that only labelled Purkinje cells (Wright et al., 2007), quite a number of genes were found in GCN 23, including *Id2*, *Creg1*, *Cpne2*, *Pcsk6*, *0610007P14Rik*, *Grid2*, *Itpr1*, *Baiap2* etc. The presence of a considerable number of genes with restricted expressions in Purkinje cell layer provided strong evidence for the enrichment of Purkinje cells markers in this GCN. Additionally, genes that are enriched in interneurons and Bergmann Glia cells within Purkinje Cell Layer are also found (Wright et al., 2007).

In addition to cell-type specific GCNs, we also found some GCNs remarkably selective for particular brain regions, such as GCN 27 (Figure 5x) in field CA1, GCN 4 (Figure 5xi) in field CA3, GCN 38 (Figure 5xii) in Dentate gyrus, GCN 45 (Figure 5xiii) in cerebellum, GCN 21 (Figure 5xiv) in medulla, GCN 1 (Figure 5xv) in thalamus, and GCN 28 (Figure 5xvi) in caudaputaman. The region-specific GCNs presumably reflect unique and coherent expression responsible for the functions of specific neuronal types in these regions. The unique expression signatures are the foundation of inferring brain genoarchitecture. Since the 3D GCN patterns are derive from multiple 2D slice-wide GCNs, the smooth and continuous 3D patterns in turn validates the reliability of slice-wide GCNs.

Table 1. Brain-wide GCN enrichment analysis based on cross-referencing with published lists of genes related to cell type markers, known and predicted lists of disease genes, specific biological functions etc. GCNs that are reproducibly identified enriched in certain category across references are bolded.

Categories of cell type markers and biological functions	GCNs (p-value<0.01)
Astrocyte (Lein et al., 2007)	13,24,**30**,35,**43**
Astrocyte (Cahoy et al., 2004)	5,16,23,**30**,**43**,45
Astrocyte (Oldham et al., 2008)	**30**,**43**
Astrocyte (Miller et al., 2010)	5,**30**,**43**
Oligodendrocyte (Lein et al., 2004)	**24**
Oligodendrocyte (Cahoy et al., 2004)	**24**
Oligodendrocyte (Oldham et al., 2008)	**24**
Oligodendrocyte (Miller et al., 2010)	**24**
Neuron (Lein et al., 2007)	3,**12**,17,**18**,**20**,**22**,26,29,35,41
Neuron (Oldham et al., 2008)	**12**,**18**,**20**,**22**,37
Neuron (Miller et al., 2010)	3,10,11,**12**,13,17,**18**,**20**,**22**,26,29, 36,37,40,41,50
Pvalb Interneurons (Oldham et al., 2008)	1,10,33
Pyramidal Neurons (Winden et al., 2009)	3,20,22,29,37
GABAergic Neurons (Sugino et al., 2006)	23,33,41
Glutamatergic Neurons (Sugino et al., 2006)	2,7,44
Mitochondria Human (Miller et al., 2010)	3,**11**,13,18,**20**,22,29,41,**50**
Mitochondria Mouse (Miller et al., 2010)	**11**,**20**,29,37,40,41,**50**
Mitochondria down in AD patients (Blalock et al., 2004)	3,**11**,12,18,**20**,22,29,37,40,41,**50**
Mitochondria down in aging human brains (Lu et al., 2004)	2,**11**,17,18,**20**,26,44,**50**
Ribosome Human (Miller et al., 2010)	**12**,**41**
Ribosome Mouse (Miller et al., 2010)	**12**,**41**,50
Ribosome (Oldham et al., 2008)	**41**

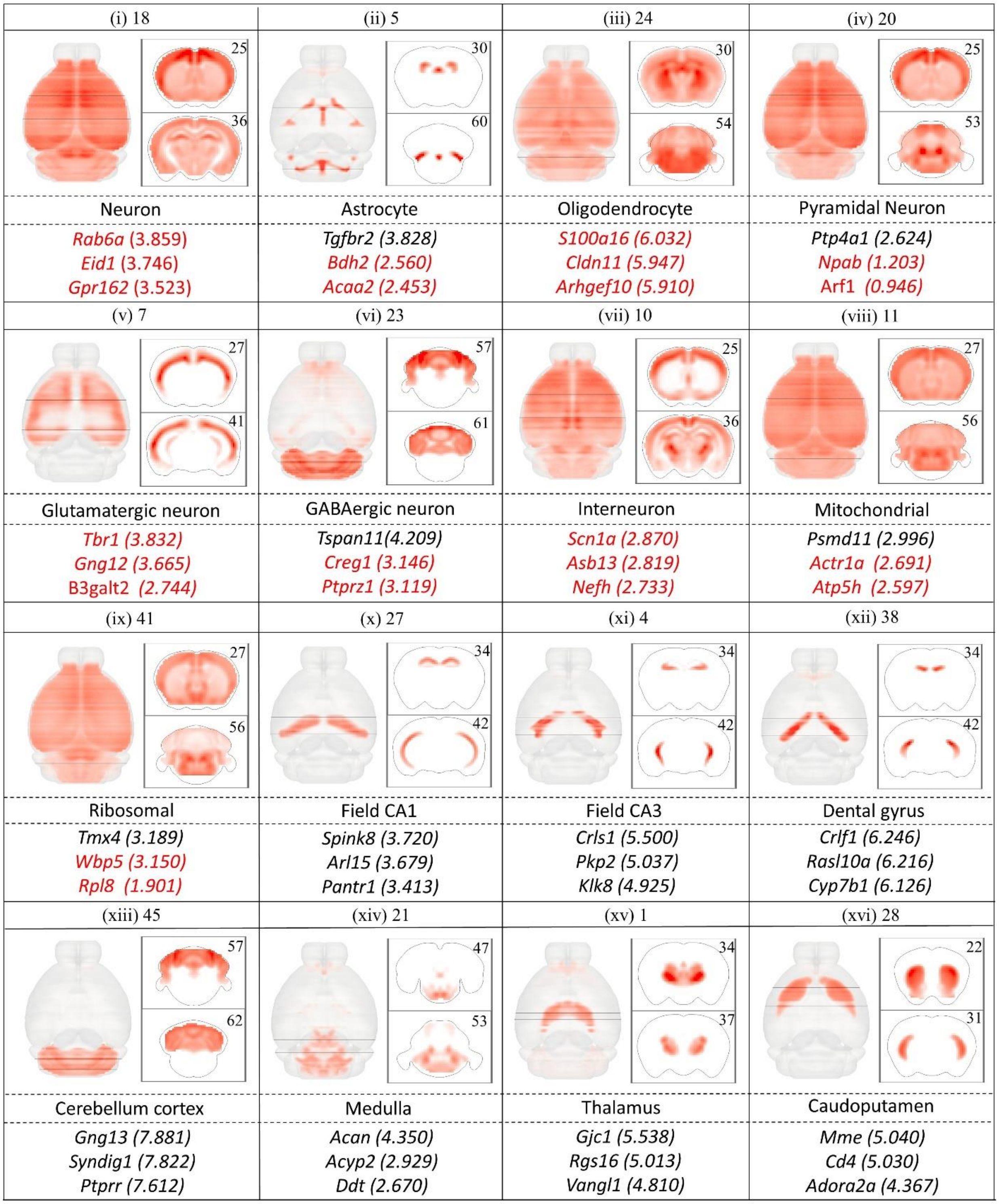

Figure 5. Visualization of spatial distribution of brain-wide GCNs significantly enriched for major cell types, particular brain regions and biological functions. In each sub-figure, top row: sub-figure index and brain-wide GCN ID. Second row: 3D spatial maps of axial (left) and two selected coronal slices (right) of GCN. The location of each slice are high-lighted in 3D spatial map and the slice index is listed in the top right corner. Third row: sub-category. Fourth row: highly weighted genes in the sub-category following the DLSC weight. The functional enriched genes previously reported in literature are highlighted in red.

It should be mentioned that there is no one-to-one mapping between the GCNs and the cell types or biological functions. In fact, many GCNs are enriched in multiple categories and that explains why the top weighted gene is sometimes not the known markers of the listed function (Figure 5). One example is GCN 20. As seen in Table 1, besides pyramidal neuron markers, this network is also enriched for neuron markers and mitochondrial-related genes. The top weighted gene *Ptp4a1* (protein tyrosine phosphatase 4a1) is a neuron marker. In other cases where the top weighted genes were not involved in any of the characterized functions, these genes might suggest potential direct or indirect link with the known functions. For instance, *Tgfbr2* (transforming growth factor, beta receptor II) is not an astrocyte marker. Research has shown that TGFβ pathways is relevant to the optic nerve head astrocyte migration (Miao et al., 2010).

4 DISCUSSIONS

We have presented a data-driven framework that can derive biologically meaningful GCNs from the gene expression data. Using the rich and spatially-resolved ISH AMBA data, we have shown that a set of networks significantly enriched for major cell type markers, specific brain regions and biological functions. The major contribution of the work are threefold. First, the DLSC method is capable of visualizing the spatial distributions of the GCNs while knowing the gene constituents and the weights they carry in the network. The precise gene distribution carry complementary information that helps identify, visualize and in the future manipulate different types of neuron cells. Second, in comparison to most clustering approaches where a single gene can only be assigned to one network, DLSC allows multiple assignments for one gene. This design can accommodate the scenarios that genes such as transcription factors play multiple roles in different networks. Third, we find that the learnt dictionaries can serve as a very relevant and compact feature representing transcriptome profile for each voxel. The brain parcellations based on the learnt dictionaries match well with the canonical neuroanatomy.

In contrast to many approaches that requires inputs of gene-gene similarity matrix, DLSC can take both the gene expression profiles and gene-gene similarity matrix as inputs. In this paper, we have demonstrated the applicability of DLSC on both inputs. We first constructed slice-based GCNs using the gene expression profiles. Then during the brain-wide GCN construction, the global similarity matrix was first calculated by integrating the local similarity matrices on all slices and then input to DLSC. The extra step of slice-based GCNs is to resolve the potential loss of information in genes with missing values and the artefacts associated with data acquisition. Ideally, if gene information are complete and the data acquisition is perfect, this method can be directly applied on the gene expression profiles consisted of all slices to form the brain-wide GCN. The capability of taking two common types of inputs affords more flexibility and robustness to handle noisy data and to incorporate/be integrated in promising methods since many GCN constructions methods are based on gene-gene associations.

The GCNs outputted by DLSC are not traditional networks with nodes and edges. In the slice-wide GCNs, nodes are the tested genes and the edges are not explicitly indicated. In DLSC, a set of coexpression patterns are learnt from the data. At the same time, we also obtain a coefficient matrix detailing how similar the expression patterns of each gene to each of these coexpression patterns although no information is provided on the association between any of the two genes in the network. However, the pairwise gene-gene similarity can still be readily estimated from the coefficients using various metrics. One example is the successful construction of global similarity matrix from the slice-wide GCNs.

In addition to the presented GCNs that reflect neuronal diversity and region specificity, many GCNs are much more difficult to interpret. Comparisons with the published lists show that numerous GCNs are enriched in multiple neuronal cells. Other GCNs are significantly associated with several functions. One explanation to the challenges of GCN interpretation is that the coexpression relationship can come from multiple biological sources such as mechanisms that synchronously regulate transcriptions of multiple genes and mRNA degradation as well as non-biological sources such as batch processing effects (Gaiteri et al., 2014). The changes brought by these sources are not mathematically distinguishable. Additionally, it is widely known that gene coexpression can be dynamically regulated by neural development, ageing, environment, and diseases (Dong et al., 2007; Jiang et al., 2001; Rampon et al., 2000). Since the gene expression profiles used is limited to one set of conditions, we should be cautious when interpreting the GCNs biologically.

The DLSC method described here may contribute to numerous applications including understanding brain evolution across species and brain development and formation. When the GCNs are correlated with neuroimaging measurements as brain phenotypes, we are able to overlay the neuroimage and GCN distribution patterns and narrow the search of genes that might cause the structural and functional differences with a final goal of advancing our understanding of how genetic functions regulate and support brains structures and functions, as well as finding new genetic variants that might account for the variations in brain structures and functions.

ACKNOWLEDGEMENT

T. Liu was supported by NIH R01 DA033393, NSF CAREER Award IIS-1149260, NIH R01 AG-042599, NSF BME-1302089, NSF BCS-1439051 and NSF DBI-1564736.

REFERENCES

Allocco, D.J., Kohane, I.S., Butte, A.J., 2004. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 5, 18. doi:10.1186/1471-2105-5-18

Bando, S.Y., Silva, F.N., Costa, L.D.F., Silva, A. V, Pimentel-Silva, L.R., Castro, L.H., Wen, H.-T., Amaro, E., Moreira-Filho, C.A., 2013. Complex network analysis of CA3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. PLoS One 8, e79913. doi:10.1371/journal.pone.0079913

Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., Landfield, P.W., 2004. Incipient Alzheimer’s disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. U. S. A. 101, 2173–

2178. doi:10.1073/pnas.0308512100

Bohland, J.W., Bokil, H., Pathak, S.D., Lee, C.-K., Ng, L., Lau, C., Kuan, C., Hawrylycz, M., Mitra, P.P., 2010. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods 50, 105–12. doi:10.1016/j.ymeth.2009.09.001

Cahoy, J., Emery, B., Kaushal, A., Foo, L., Zamanian, J., Christopherson, K., Xing, Y., Lubischer, J., Krieg, P., Krupenko, S., Thompson, W., Barres, B., 2004. A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. J. Neuronscience 28, 264–278. doi:10.1523/JNEUROSCI.4178-07.2008

Carter, H., Hofree, M., Ideker, T., 2013. Genotype to phenotype via network analysis. Curr. Opin. Genet. Dev. 23, 611–621. doi:10.1016/j.gde.2013.10.003

Chen, H., Li, K., Zhu, D., Jiang, X., Yuan, Y., Lv, P., Zhang, T., Guo, L., Shen, D., Liu, T., 2013. Inferring group-wise consistent multimodal brain networks via multi-view spectral clustering. IEEE Trans. Med. Imaging 32, 1576–86. doi:10.1109/TMI.2013.2259248

Dennis, G., Sherman, B.T., Hosack, D. a, Yang, J., Gao, W., Lane, H.C., Lempicki, R. a, 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3. doi:10.1186/gb-2003-4-5-p3

Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M.L., Carlson, S., Allan, M.F., Pomp, D., Schadt, E.E., 2009. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. Genome Biol. 10, R55. doi:10.1186/gb-2009-10-5-r55

Dong, S., Li, C., Wu, P., Tsien, J.Z., Hu, Y., 2007. Environment enrichment rescues the neurodegenerative phenotypes in presenilins-deficient mice. Eur. J. Neurosci. 26, 101–112. doi:10.1111/j.1460-9568.2007.05641.x

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least Angle Regression. Ann. Stat. 32, 407–499.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1999. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. 95, 12930–12933. doi:10.1073/pnas.95.25.14863

Gaiteri, C., Ding, Y., French, B., Tseng, G.C., Sibille, E., 2014. 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

Ge, H., Liu, Z., Church, G.M., Vidal, M., 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet. 29, 482–6. doi:10.1038/ng776

Hawrylycz, M., Bernard, A., Lau, C., Sunkin, S.M., Chakravarty, M.M., Lein, E.S., Jones, A.R., Ng, L., 2010. Areal and laminar differentiation in the mouse neocortex using large scale gene expression data. Methods 50, 113–21. doi:10.1016/j.ymeth.2009.09.005

Hawrylycz, M., Miller, J.A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A.L., Jegga, A.G., Aronow, B.J., Lee, C.-K., Bernard, A., Glasser, M.F., Dierker, D.L., Menche, J., Szafer, A., Collman, F., Grange, P., Berman, K.A., Mihalas, S., Yao, Z., Stewart, L., Barabási, A.-L., Schulkin, J., Phillips, J., Ng, L., Dang, C., Haynor, D.R., Jones, A., Van Essen, D.C., Koch, C., Lein, E., 2015. Canonical genetic signatures of the adult human brain. Nat. Neurosci. 18. doi:10.1038/nn.4171

Jiang, C.H., Tsien, J.Z., Schultz, P.G., Hu, Y., 2001. The effects of aging on gene expression in the hypothalamus and cortex of mice. PNAS 98, 1930–1934. doi:10.1073/pnas.98.4.1930

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. doi:10.1186/1471-2105-9-559

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T.-M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T. a, Donelan, M.J., Dong, H.-W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B. a, Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R. a, Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau, C., Lemon, T. a, Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S. a, Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N. V, Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K. a, Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.-R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Varnam, L.R., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Yuan, X.F., Zhang, B., Zwingman, T. a, Jones, A.R., 2007. Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–76. doi:10.1038/nature05453

Lein, E.S., Zhao, X., Gage, F.H., 2004. Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. J. Neurosci. 24, 3879–89. doi:10.1523/JNEUROSCI.4710-03.2004

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., Yankner, B. a, 2004. Gene regulation and DNA damage in the ageing human brain. Nature 429, 883–891. doi:10.1038/nature02618.1.

Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416. doi:10.1007/s11222-007-9033-z

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2010. Online Learning for Matrix Factorization and Sparse Coding. J. Mach. Learn. Res. 11, 19–60.

Miao, H., Crabb, A.W., Hernandez, M.R., Lukas, T.J., 2010. Modulation of factors affecting optic nerve head astrocyte migration. Investig. Ophthalmol. Vis. Sci. 51, 4096–4103. doi:10.1167/iovs.10-5177

Miller, J., 2014. Transcriptional Landscape of the Prenatal Human Brain. Nature 508, 199–206.

doi:10.1038/nature13185.Transcriptional

Miller, J. a, Horvath, S., Geschwind, D.H., 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc. Natl. Acad. Sci. U. S. A. 107, 12698–12703. doi:10.1073/pnas.0914257107

Miller, J.A., Cai, C., Langfelder, P., Geschwind, D.H., Kurian, S.M., Salomon, D.R., Horvath, S., 2011. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics 12, 322. doi:10.1186/1471-2105-12-322

Mody, M., Cao, Y., Cui, Z., Tay, K.Y., Shyong, A., Shimizu, E., Pham, K., Schultz, P., Welsh, D., Tsien, J.Z., 2001. Genome-wide gene expression profiles of the developing mouse hippocampus. PNAS 98, 8862–8867. doi:10.1073/pnas.141244998

Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., Macklis, J.D., 2007. Neuronal subtype specification in the cerebral cortex. Nat. Rev. Neurosci. 8, 427–37. doi:10.1038/nrn2151

Ng, L., Bernard, A., Lau, C., Overly, C.C., Dong, H.-W., Kuan, C., Pathak, S., Sunkin, S.M., Dang, C., Bohland, J.W., Bokil, H., Mitra, P.P., Puelles, L., Hohmann, J., Anderson, D.J., Lein, E.S., Jones, A.R., Hawrylycz, M., 2009. An anatomic gene expression atlas of the adult mouse brain. Nat. Neurosci. 12, 356–62. doi:10.1038/nn.2281

Ng, L., Pathak, S.D., Kuan, C., Lau, C., Dong, H., Sodt, A., Dang, C., Avants, B., Yushkevich, P., Gee, J.C., Haynor, D., Lein, E., Jones, A., Hawrylycz, M., 2007. Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. IEEE/ACM Trans. Comput. Biol. Bioinforma. 4, 382–392. doi:10.1109/TCBB.2007.1035

O'Leary, D.D.M., Stocker, A.M., Zembrzycki, A., 2013. Area Patterning of the Mammalian Cortex. Compr. Dev. Neurosci. Patterning Cell Type Specif. Dev. CNS PNS 61–85. doi:10.1016/B978-0-12-397265-1.00021-6

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

Oldham, M.C., Langfelder, P., Horvath, S., 2012. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst. Biol. 6, 63. doi:10.1186/1752-0509-6-63

Peng, H., Long, F., Zhou, J., Leung, G., Eisen, M.B., Myers, E.W., 2007. Automatic image analysis for gene expression patterns of fly embryos. BMC Cell Biol. 8 Suppl 1, S7. doi:10.1186/1471-2121-8-S1-S7

Quinones-Hinojosa, A., Chaichana, K., 2007. The human subventricular zone: A source of new cells and a potential source of brain tumors. Exp. Neurol. 205, 313–324. doi:10.1016/j.expneurol.2007.03.016

Rampon, C., Jiang, C.H., Dong, H., Tang, Y.P., Lockhart, D.J., Schultz, P.G., Tsien, J.Z., Hu, Y., 2000. Effects of environmental enrichment on gene expression in the brain. Proc. Natl. Acad. Sci. U. S. A. 97, 12880–4. doi:10.1073/pnas.97.23.12880

Stuart, J.M., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic

Modules. Science (80-.). 302, 249–255. doi:10.1126/science.1087447

Sugino, K., Hempel, C.M., Miller, M.N., Hattox, A.M., Shapiro, P., Wu, C., Huang, Z.J., Nelson, S.B., 2006. Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat. Neurosci. 9, 99–107. doi:10.1038/nn1618

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. U. S. A. 96, 2907–2912. doi:10.1073/pnas.96.6.2907

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285. doi:10.1038/10343

Winden, K.D., Oldham, M.C., Mirnics, K., Ebert, P.J., Swan, C.H., Levitt, P., Rubenstein, J.L., Horvath, S., Geschwind, D.H., 2009. The organization of the transcriptional network in specific neuronal classes. Mol. Syst. Biol. 5, 291. doi:10.1038/msb.2009.46

Wright, E., Ng, L., Guillozet-Bongarts, A., 2007. Cerebellar cortex, pukinje cell layer, Allen Bran Atlas Mouse Brain. doi:10.1038/npre.2008.2200.1

Supplemental Materials

1. Image artefacts during acquisition

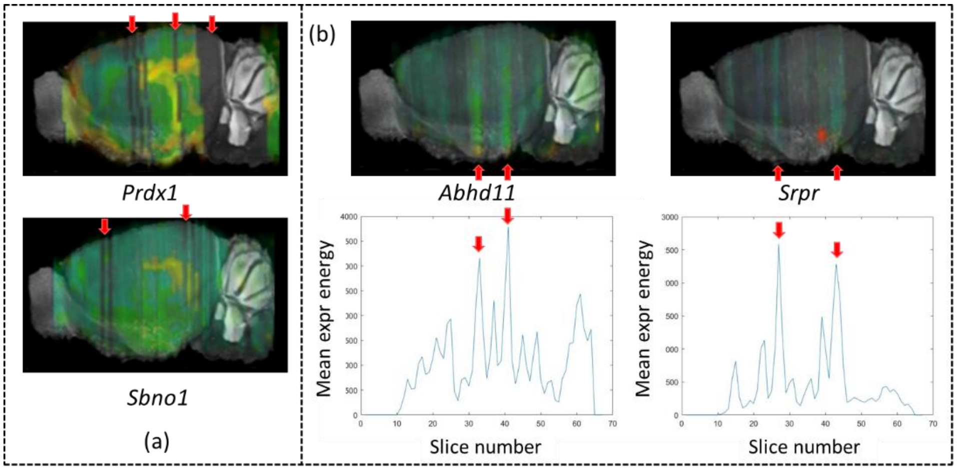

Figure S1 Examples of image artefacts during acquisitions including missing slices (a) and discontinuous changes in the average expression energies between adjacent slices (b). The arrows in (a) highlighted the missing slices. The plots in (b) shows the average expression energy as a function of slice number. The arrows in (b) indicate the correspondence between the 3D visualization and the average expression energy in the plot.

Due to the image artefacts during acquisition such as missing slices (Figure S1(a)) and the discontinuous changes in the average expression energies between adjacent slices (Figure S1(b)), we proposed to study the GCNs slice by slice and then fuse them instead of using 3D data directly as input.

2. Parameter selection for slice-wide GCN construction

As seen in Table S1, the reconstruction errors gradually increase as λ and gene-dictionary ratio grow. This trend is expected because both an enforced sparser coefficient matrix and a fewer number of dictionaries can decrease the description power of the representation and result a further deviation from the original matrix. According to equation 3, the maximum reconstruction error is 0.5. When λ reaches 0.7 and above, the reconstruction errors are over 50%. The big deviation from the original signal matrix narrows the considerations of λ of 0.5 and below. With respect to the AUC, interestingly, the higher values occur when λ is set to 0.5 and 0.7. The increase in the gene-dictionary ratio is coincided with a gradual growth in the AUC in general. The highest AUC occurs at λ of 0.5 and gene-dictionary ratio of 100, which makes this combination the best candidate. Lastly, we checked the density of this parameter combination. The density is 6.4% and acceptable. As the final goal of parameter selection is to choose a set of parameters that result in a sparse and accurate representation of the original signal, which is translated to low reconstruction error, high AUC and low density, λ=0.5 and gene-dictionary ratio of 100 is the best option among 55 parameter combinations and chosen as the optimal parameters.

Table S1 Reconstruction errors of DLSC on slice 27 using different λ and gene-dictionary ratios. The number in parentheses in the first column is the corresponding number of dictionaries

gene- λ dictionary ratio	0.1	0.3	0.5	0.7	0.9

10(284)	0.031	0.086	0.174	0.299	0.453
20(142)	0.038	0.096	0.185	0.310	0.459
30(95)	0.043	0.101	0.19	0.314	0.461
40(71)	0.047	0.105	0.194	0.318	0.463
50(57)	0.050	0.107	0.196	0.321	0.464
60(48)	0.052	0.110	0.198	0.323	0.465
70(41)	0.054	0.112	0.200	0.324	0.466
80(36)	0.056	0.113	0.202	0.326	0.466
90(32)	0.058	0.115	0.204	0.327	0.467
100(29)	0.059	0.116	**0.205**	0.329	0.467
110(26)	0.061	0.118	0.206	0.329	0.368

Table S2 AUCs between the obtained dictionaries and the annotation map on slice 27 using different λ and gene-dictionary ratios.

gene- λ dictionary ratio	0.1	0.3	0.5	0.7	0.9
10(284)	0.303	0.332	0.351	0.365	0.366
20(142)	0.309	0.354	0.372	0.384	0.382
30(95)	0.328	0.375	0.392	0.400	0.394
40(71)	0.339	0.384	0.395	0.406	0.398
50(57)	0.353	0.395	0.404	0.402	0.399
60(48)	0.359	0.399	0.413	0.400	0.395
70(41)	0.358	0.399	0.421	0.412	0.401
80(36)	0.364	0.396	0.417	0.418	0.411
90(32)	0.372	0.398	0.426	0.414	0.411
100(29)	0.379	0.400	**0.434**	0.433	0.424
110(26)	0.377	0.408	0.419	0.423	0.427

Table S3 The percentage of none-zero entries in the coefficient matrix obtained from DLSC on slice 27 using different λ and gene-dictionary ratios.

gene- λ dictionary ratio	0.1	0.3	0.5	0.7	0.9
10(284)	0.026	0.011	0.007	0.004	0.003
20(142)	0.050	0.023	0.014	0.009	0.005
30(95)	0.070	0.033	0.021	0.014	0.007
40(71)	0.089	0.043	0.028	0.018	0.009
50(57)	0.106	0.053	0.034	0.023	0.011
60(48)	0.121	0.062	0.040	0.027	0.013
70(41)	0.136	0.071	0.047	0.032	0.015
80(36)	0.150	0.078	0.052	0.036	0.017
90(32)	0.164	0.086	0.058	0.040	0.019
100(29)	0.176	0.094	**0.064**	0.044	0.020

110(26)	0.190	0.104	0.071	0.049	0.023

2. Parameter selection for brain-wide GCN construction

The parameter selection of decomposing the global similarity matrix is guided by the knowledge from the slice-based study that each network consists of on average 185 genes, and each gene participates in 1.85 networks.

Table S4 The average number of networks one gene participates in after applying DLSC on the global similarity matrix using different combinations of λ and dictionary numbers.

Number λ of dictionaries	0.1	0.3	0.5	0.7
40	3.95	2.58	1.96	1.61
50	4.63	2.90	**2.21**	1.80
60	5.03	3.12	2.38	1.94
70	5.49	3.32	2.55	2.07
80	5.85	3.52	2.65	2.17
90	6.14	3.68	2.79	2.26

Table S5 The average number of genes per network after applying DLSC on the global similarity matrix using different combinations of λ and dictionary numbers.

Number λ of dictionaries	0.1	0.3	0.5	0.7
40	420.225	274.375	208.05	171.55
50	393.82	246.68	**188.24**	153.30
60	356.82	221.07	169.08	137.58
70	333.47	201.80	155.10	125.64
80	311.28	187.20	141.05	115.21
90	290.39	174.14	131.71	106.83

3. Comparisons of spatial distribution of WGCNA module 1 and the corresponding GCNs

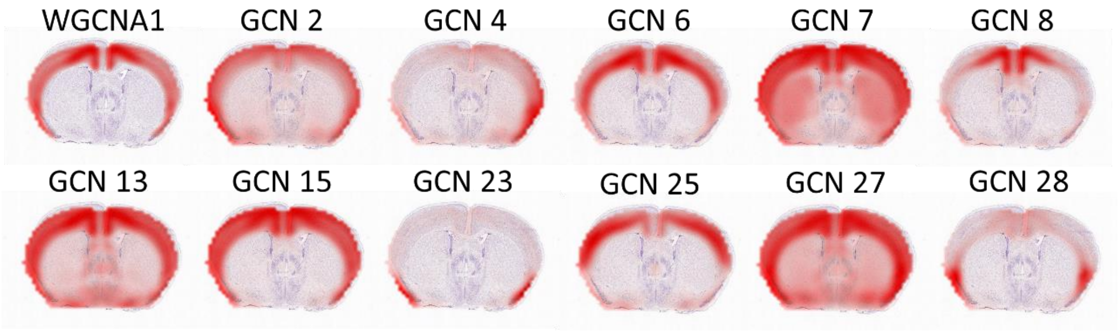

Figure S2 Spatial distributions of WGCNA module 1 and the corresponding GCNs on slice 27.

As seen above, the spatial distribution of each GCN show overlaps with that of WGCNA module 1. The DLSC framework enables finer breakdowns of the isocortex into layers and subregions.