Brain Struct Funct (2017) 222:4253–4270
DOI 10.1007/s00429-017-1460-9

ORIGINAL ARTICLE

Discover mouse gene coexpression 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]

Received: 31 July 2016 / Accepted: 13 June 2017 / Published online: 29 June 2017

Abstract Gene coexpression patterns carry rich 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 using dictionary learning and sparse coding, we derived coexpression networks from the space-resolved anatomical comprehensive in situ hybridization data from Allen Mouse Brain Atlas dataset. The key idea is 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 "coexpressed." For each network, we have simultaneous knowledge of spatial distributions, the genes in the network and the extent a particular gene conforms to the coexpression pattern. 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.

Keywords Gene coexpression network · Sparse coding · Transcriptome

Yujie Li and Hanbo Chen: co-first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00429-017-1460-9) contains supplementary material, which is available to authorized users.

✉ Hanchuan Peng
hanchuanp@alleninstitute.org

✉ Joe Z. Tsien
jtsien@augusta.edu

✉ Tianming Liu
tliu@cs.uga.edu

[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, Augusta, USA

Introduction

Gene coexpression patterns carry rich information about enormously complex cellular processes (Brown et al. 2007; Eisen et al. 1999; Grange et al. 2014; Lee et al. 2004; Oldham et al. 2006; Peng et al. 2007; Stuart et al. 2003). Previous studies have shown that genes displaying similar expression profiles are very likely to involve 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. (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 or 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 a 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 a more descriptive power to the complicated gene–gene interactions (Oldham et al. 2012). Given the high dimensions of genetic data and the urgent need in unveiling the differences or the consensus between subjects or species, one common theme of all of these methods is dimension reduction. Instead of analyzing the interactions across over tens of thousands of genes, the groupings of genes by their coexpression patterns can considerably reduce the complexity to dozens of networks or clusters, while preserving the original interaction relationships.

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. The popularity of applying sparse coding and dictionary learning on images is derived from the observations that neurons encode sensory information using a small number of active neurons at any given point in time (Olshausen and Field 2004). It is reported that sparsification can "weed out" those basis functions not needed to describe a given image structure, thus obtaining an easier interpretation (Olshausen and Field 2004). Unlike decompositions based on principal component analysis and its variants, sparse learned models do not impose that the basis vectors be orthogonal, allowing more flexibility to adapt the representation to the data (Mairal et al. 2010). An equally important feature is that sparse coding can model inhibition between the bases by sparsifying their activations. Due to these properties, DLSC has found great success in applications such as image denoising, demosaicing, and inpainting (Mairal et al. 2008). In the context of extracting coexpression patterns, we assume 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 "coexpressed." On the other hand, it is reported that most genes are expressed in a fairly small percentage of cells (70.5% of genes are expressed in less than 20% of total cells in the ISH dataset) (Lein et al. 2007). We assume this notion can be captured by imposing a sparsity constraint that limits the number of voxels that a gene can be active on. The added sparse constraint will also encourage the dictionary to capture the most common gene coexpression patterns so that a parsimonious representation is possible. Thus, DLSC can serve as a useful tool for GCN construction.

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 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 hold promises to serve as foundations to explore different cell types and functional processes in diseased and healthy brains.

Methods

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

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

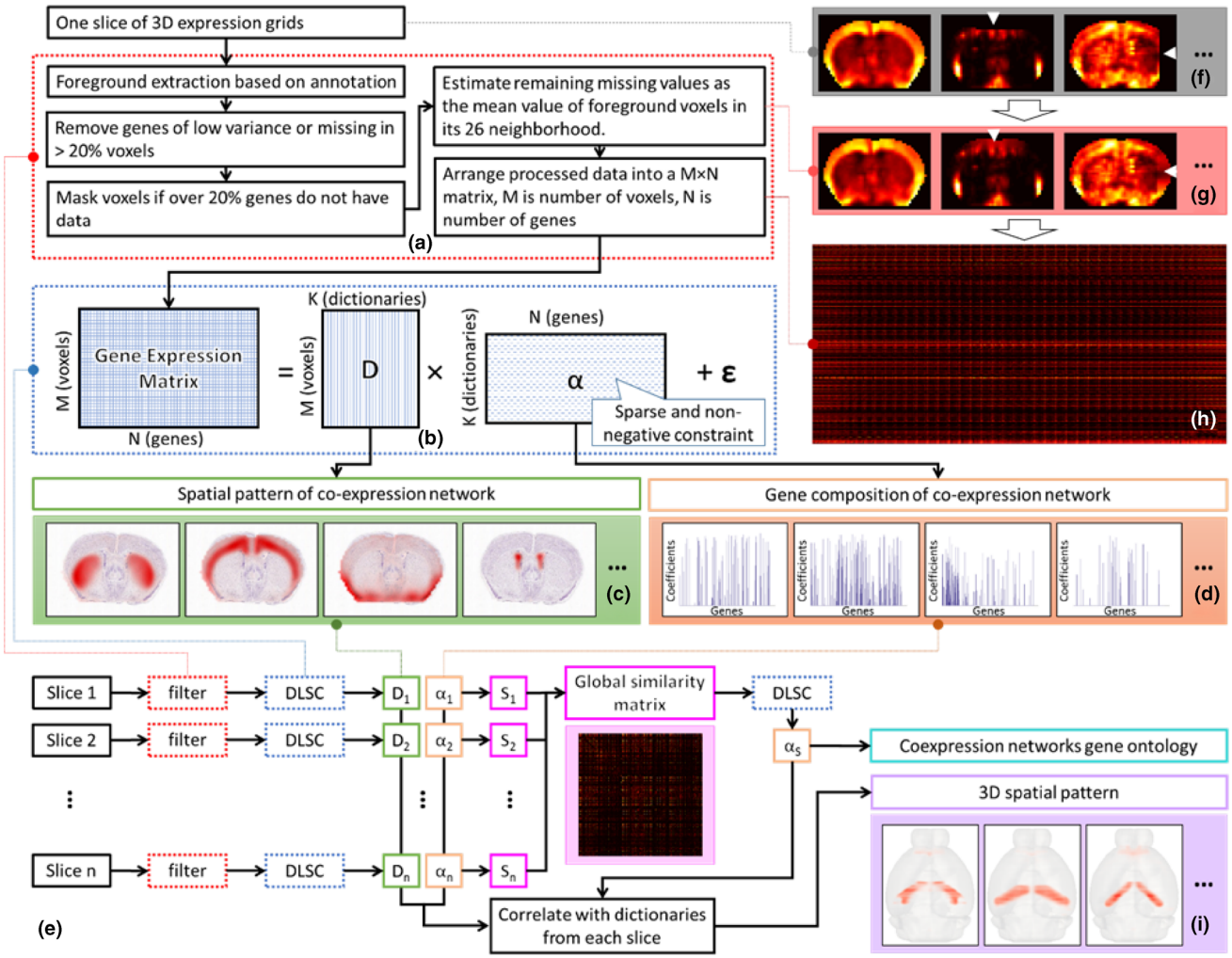

Fig. 1 Computational pipeline for constructing slice-wide GCNs (**a**–**d**) and brain-wide GCNs (**e**). **a** Raw ISH data preprocessing step that removes unreliable genes and voxels and estimates the remaining missing data. **b** Dictionary learning and sparse coding of ISH matrix with sparse and nonnegative constraints on α matrix. **D** is the dictionary matrix and $\boldsymbol{\alpha}$ is the coefficient matrix. ε is the reconstruction error. **c** Visualization of spatial distributions of slice-based GCNs. **d** Visualizations of coexpression networks. **e** Integrating slice-based GCNs into global GCNs and global GCN gene ontology. **f** Visualization of slices of raw expression grids before preprocessing. **g** Visualization of slices of raw expression grids after preprocessing. Some missing data were estimated. **h** Expression girds were arranged in an *M* by *N* matrix. **i** Visualizations of 3D spatial patterns of global GCNs

200 × 200 μm 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 4345 genes in coronal sections and 21,718 genes in sagittal sections make up the key components of the AMBA.

We downloaded the 4345 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 were 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.

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, obtaining a common set of genes on all slices would require roughly 33% of the genes removed from analysis, resulting in a significant amount of information loss. Additionally, as the ISH data were 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 were 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 (Supplementary Figure S2). Yet additional efforts are needed to integrate gene–gene interactions on each slice.

Data preprocessing

For slice-wide analysis, the input of the pipeline is the expression grids of one of 67 coronal slices. A preprocessing module (Fig. 1a) was 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 error. 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 provided little information for network construction. A similar filtering step was 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 suggests that the estimated data are reasonable. After preprocessing, the cleaned expression energies were organized into a matrix and sent to DLSC (Fig. 1b). In DLSC (Sect. 2.2.2), the gene expression matrix was factorized into a dictionary matrix $\mathbf{D}$ and a coefficient matrix $\boldsymbol{\alpha}$. These two matrices encode the distribution and composition of GCN (Fig. 1c–d) and were further analyzed and validated against the raw data and existing methods.

Dictionary learning and sparse coding

The gene expression grids were arranged into a single matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$, such that M rows correspond to M foreground voxels for analysis and N columns correspond to N genes (Fig. 1b). Then, each column of the matrix (gene signal in a voxel) was normalized by the L2-norm of the column. After normalization, the publicly available online DLSC package was applied to solve the matrix factorization problem proposed in Eq. (2) (Mairal et al. 2010). Eventually, the gene expression energy matrix $\mathbf{X}$ was represented as sparse combinations of learned dictionary atoms $\mathbf{D}$. 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.

Formally, given a set of M-dimensional input signals $X = [x1,\ldots,xN]$ 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:

$$\langle \mathbf{D}, \boldsymbol{\alpha} \rangle = \text{argmin} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{D} \times \boldsymbol{\alpha}\|_2^2 \text{s.t} \|\boldsymbol{\alpha}\|_1 \leq \lambda \quad (1)$$

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 Eq. (1) should be minimized.

As will be discussed later that each entry of $\boldsymbol{\alpha}$ indicates the degree of conformity of a particular gene to a coexpression network, a nonnegative constraint was added to the ℓ_1-regularization. This additional prior, included in Eq. (2), can be handled by homotopy method presented in Efron et al. (2004):

$$\langle \mathbf{D}, \boldsymbol{\alpha} \rangle = \text{argmin} \sum_{i=1}^{N} \frac{1}{2} \|x_i - \boldsymbol{D} \times \alpha_i\|_2^2 \text{s.t} \|\boldsymbol{\alpha}\|_1 \leq \lambda, \forall \text{i}, \alpha_i \geq 0. \quad (2)$$

The key assumption of enforcing the sparsification is that each gene is involved in a very limited number of gene networks. The nonnegativity constraint on $\boldsymbol{\alpha}$ matrix imposes that no genes with the opposite expression patterns will be placed in the same network.

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 two benefits of this setup. First, both the dictionaries and coefficients are learned from the data and therefore should reflect the intrinsic organization of transcriptome. Second, the level of coexpressions is quantifiable, and the level is not only comparable within one dictionary, but the entire $\boldsymbol{\alpha}$ matrix.

Further, if we consider each dictionary as one network, the corresponding row of $\boldsymbol{\alpha}$ matrix contains all the genes that use this dictionary for sparse representation, or that are "coexpressed." Additionally, each entry of $\boldsymbol{\alpha}$ 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 $\mathbf{D}$ back to the ARA space, we can visualize the spatial patterns of the coexpressed networks. Combining information from both $\mathbf{D}$ and $\boldsymbol{\alpha}$ 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 the network, these two terms will be used interchangeably.

Parameter selection

The choice of the number of dictionaries and the regularization parameter λ are crucial for effective sparse representation. As no gold standard exists 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 criterion is the reconstruction error. It is defined as the square difference between the original signal matrix and the reconstruction from sparse representation [Eq. (3)]. A high reconstruction error indicates a less accurate representation:

$$\text{error}_y = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{D\alpha}\|_F^2. \tag{3}$$

The second evaluation metric is the average uncertainty coefficient (AUC) between the obtained dictionaries and the reference atlas. The uncertainty coefficient, defined in [Eq. (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 that measures 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 the uncertainty of the other and thereby dependent.

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:

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

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

Another important measurement to examine the DLSC performance is the degree of density measured by the percentage of nonzero-valued elements in the coefficient matrix. As we are searching for 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 Sect. 2.2.2, 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 was chosen due to its good anatomical coverage of various brain regions. As different number of genes was expressed in different slices, the number of dictionaries for each slice should change accordingly. Instead of fixing the number of dictionaries, a gene dictionary ratio was used to determine the optimal ratio between the number of genes expressed and the number of dictionaries required to achieve a good representation. Fifty-five combinations of λ and gene dictionary ratios were considered with 5 choices of λ and 11 different gene dictionary ratios (Supplementary table S1–3). The results obtained from 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 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 coefficient density, $\lambda = 0.5$ and gene dictionary ratio of 100 is the best option among 55 parameter combinations and chosen as the optimal parameters.

Brain parcellation using DLSC

The decomposition of gene expression matrix on each slice results in a dictionary matrix $\mathbf{D}$ and a coefficient matrix $\boldsymbol{\alpha}$. Each row of $\mathbf{D}$ 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 was employed to calculate the similarity between voxels. Then the voxels on the slice were clustered into groups by spectral clustering (Chen et al. 2013; Luxburg 2007). The number of clusters was adapted to the data and determined by the normalized cut using an empirically selected threshold of 0.7. (Chen et al. 2013; Luxburg 2007).

Comparative analysis with weighted gene correlation network analysis (WGCNA)

WGCNA was 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, a signed network is used to avoid the "anti-reinforcing" connection strength that might occur in the unsigned network. 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. Default parameters of WGCNA resulted in 14 modules, while the DLSC gave 29 GCNs. To get a more balanced comparison between the two methods, we increased the number of modules extracted by WGCNA by tuning three parameters: the soft thresholding power beta, deepSplit, and minModuleSize. Multiple combinations of these parameters have been tested and the highest number of modules WGCNA was able to get was 25 modules with one additional module for unassigned genes. The parameters used in the experiment were: beta = 18, deepSplit = 4 (highest) and minModuleSize = 15. Also, we changed the number of GCNs from the optimal 29–26 to ensure a fair comparison.

Then the number of shared genes was counted for groups identified by both methods. Besides quantification, another intuitive way to compare the two methods is by comparing the obtained spatial maps (Fig. 2). Similar gene groups are likely to show 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.

Brain-wide GCNs construction and analysis

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, was calculated from the coefficient matrix $\boldsymbol{\alpha}$ with the coefficients as the feature of each gene. Let $v1$ and $v2$ be the coefficient vectors of gene 1 and gene 2. The gene similarity measure is defined as the overlap rate OR, as below:

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

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

As each slice has missing data for different genes, the interactions of these missing genes on a 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 showed zero similarity to any other genes. The very low similarity was caused by the lack of data, evidenced by that these 91 genes were present in at most 5 out of 67 slices. The separation of these genes that suffered from heavy data loss demonstrates the effectiveness of similarity matrix

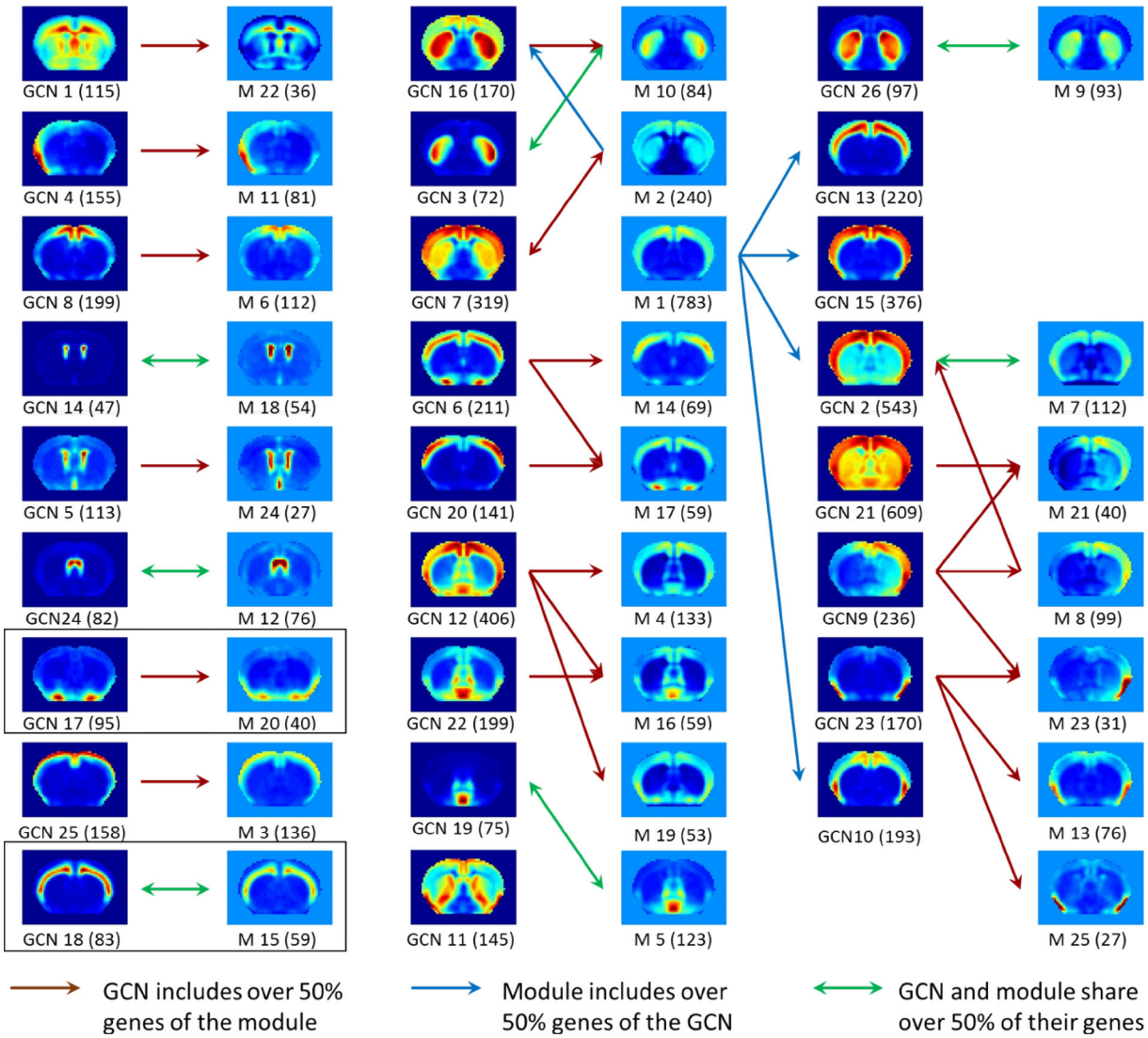

Fig. 2 Comparison between spatial maps of GCNs and eigengenes of WGCNA modules on slice 27. For clarity, the groups identified by WGCNA and DLSC are denoted as modules and GCNs, respectively. The number of overlapping genes between a GCN and a module was counted. At the *bottom* of each image is the name of the networks/modules. '*M*' represents a module generated by WGCNA and GCN represents a coexpression network generated by DLSC. The *number in the parentheses* is the number of total genes in that module/network. *Brown arrows* indicate that the GCN includes over 50% genes of that module. *Blue arrows* indicate that the module has over 50% of the same gene of the GCN. *Green double arrows* indicate that the GCN and module share 50% of their own genes. The *black boxes* highlight the GCN/module compared in detail in Figs. 3 and 4. The *background color* for modules and GCNs are fixed to −0.05 and 0

over the original $\boldsymbol{\alpha}$ matrix, and also reflects OR as an appropriate measure for gene similarity in this situation.

A total of 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 a similar relationship to all the other genes, these genes should belong to the same group. The network memberships were encoded in the resulted sparse coefficient matrix $\boldsymbol{\alpha}$.

Parameter selection

The parameter selection of decomposing the global similarity matrix is guided by the knowledge from the slice-based study that each network contains 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.

Fuse 3D spatial pattern of GCNs

As described in Sect. 2.2.2, the dictionaries trained in each slice encode 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 the similarity and integrated them into a 3D volume. Specifically, the similarity was calculated based on the OR of the coefficient matrix defined in Sect. 2.3.1. Slightly different from the previous definition, here the similarity was calculated between GCNs instead of genes. Also, before comparison, each feature vector was normalized so that the maximum value equals to 1.

Gene ontology analysis of brain-wide GCNs

Brain-wide GCN characterization was 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.

Results

The organization of the result section is as follows. In Sect. 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 were derived and then by a comparative study with one of the most widely used methods—WGCNA as well as a matrix factorization method principal component analysis (PCA). On the side as an application, we demonstrated that the learned 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, Sect. 3.2 focuses on the construction of global GCNs by integrating the gene–gene interactions on all slices. Along with the spatial distributions of the GCNs, we showed that the obtained GCNs are biologically meaningful by comparing with the known gene ontologies and published gene lists.

Slice-wide GCN analysis

To show as an example, slice 27 was 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_adj/overview.html.

Comparative analysis with WGCNA

Both DLSC and WGCNA were applied on the gene expressions data of slice 27. Although a larger number of modules (from 14 to 25) were obtained by tuning the parameters of WGCNA, the number of genes in a module varies significantly. Specifically, the top three modules (module 1, 2, and 3) consist of 783, 240, and 136 genes, respectively, and modules 15–25 all contain fewer than 60 genes, indicating the genes were not well separated. The observation of a single large module together with multiple small modules was also seen when the default WGCNA parameters were used. Fourteen modules were obtained with the largest module containing over 1000 genes. In contrast, the number of genes in the GCNs was more balanced. The top three GCNs contain 609, 543, and 406 genes even though some genes have been counted multiple times. In this sense, the DLSC gives better coexpression networks as it is able to separate genes into more balanced groups when the number of groups is relatively large.

To test whether DLSC provides an improved view of coexpressed genes, we measured the correspondence at the level of network/module pairs by quantifying the number of shared genes. We used a brown arrow pointing from a GCN to a module to denote that the GCN containing over 50% of the genes in that module. Similarly, a blue arrow pointing from a module to a GCN indicates a module containing over 50% of genes in that GCN. If the number of shared genes is above 50% of the genes in the module as well as the GCN, a green double arrow was used. By laying out the spatial maps of the GCNs and the eigengenes of WGCNA modules (Fig. 2), it is evident that the spatial maps of GCNs and modules sharing over 50% are either very similar (e.g., GCN17 and M20, GCN4 and M11, GCN8 and M6) or have large spatial overlaps (e.g., GCN22 and M16, GCN19 and M5, GCN7 and M2). Overall, the spatial maps of the groups generated by WGCNA and DLSC are on the same scale. For each spatial map of the module, we can find one or more similar spatial maps of the GCNs.

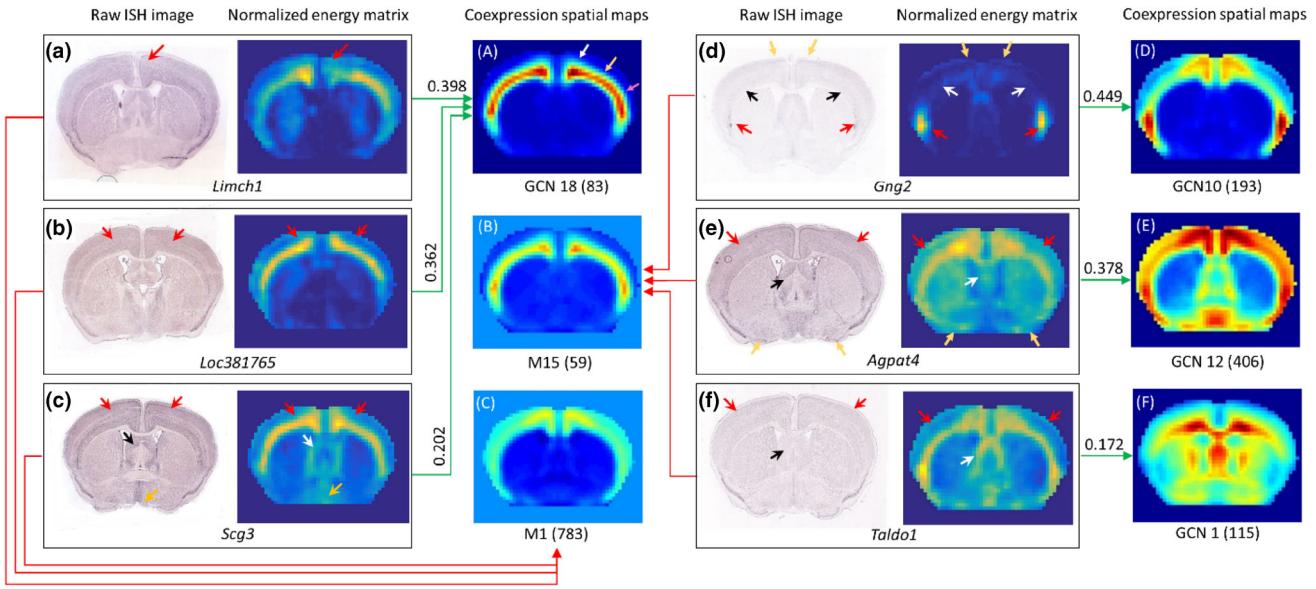

Fig. 3 Comparisons of genes in GCN18 and module 15 on slice 27. For each gene (**a**–**f**) we showed the raw ISH image together with the normalized energy matrix. On the *left* are three representative genes only found by DLSC. On the *right* are three genes only found by WGCNA. **A**–**F** Are the spatial distributions of selected GCNs and the eigengenes of selected modules. The *number in the parentheses* of GCNs/modules denotes the number of genes in the module/GCN. The *long red arrows* show the module assignment made by WGCNA. The *green arrows* show the GCN assignment made by DLSC. DLSC offers a weight that measures the degree to which the gene expression conforms to the coexpression pattern. These weights are the values above the respective *green arrows*

Then we focus on the genes in the GCNs/modules. Most GCNs have more genes than the respective module that share the similar spatial pattern, indicated by the considerably more brown arrows than the blue arrows (Fig. 2). Relatedly, there are many modules small in size given that roughly half of the genes are assigned to module 1 and module 2.

There are multiple pairs that share over 50% of their genes (Fig. 2 green arrows). One example is GCN 18 and module 15, whose spatial patterns are quite similar (Fig. 3A, B). The number of genes in GCN 18 is 83 and module 15 has 59 genes. It turns out 52 out of 83 genes were shared by both GCN18 and module 15. Thirty-one genes were found only by GCN18, and seven genes were found only by module 15. We first examined the raw ISH data of genes that were only found by DLSC. The spatial map of GCN 18 featured high activations at cortex layer 5 and 6, covering the cingulate area (Fig. 3A white arrow), motor area (Fig. 3A yellow arrow) and somatosensory area (Fig. 3A pink arrow). Three genes were selected for illustration from those only found by DLSC (Fig. 3a–c). The weight above the green arrow is a measure of the degree to which a gene conforms to the coexpression patterns. With a decreasing weight, the resemblance of the raw data to the spatial map became weaker. All three genes showed strong signals in layers 5 and 6 and agree with the overall shape of the GCN 18 (Fig. 3 red arrows). However, *Scg3* displayed additional activations at medial preoptic area (Fig. 3c, f yellow arrow) and lateral septal nucleus (Fig. 3c black/white arrow) and thus was assigned a lower weight of 0.202. By examining the normalized energy matrix as well as the raw ISH, we were convinced that these genes have similar spatial distributions to the GCN18 and that the assignment is correct.

Interestingly, 27 out of the 37 genes that were assigned to GCN18 and not assigned to module 15, including the *Limch1*, *Loc381765,* and *Scg3*, were assigned to module 1 (Fig. 3C) by WGCNA, which featured the entire cortex layer from layer 1 to layer 6 and the expression peaks at the anterior cingulate area and the motor area and gradually decreased in the primary and supplementary somatosensory regions. Despite some similarities, the absence of expressions in the outer layer of cortex and the fairly homogeneous expression across cingulate, motor and somatosensory regions (Fig. 3a–f red arrows) suggests the expression pattern a better consistency to GCN18.

We also looked at the genes found only by WGCNA (Fig. 3d–f). These genes were given zero weights by DLSC in GCN18, meaning they were not part of GCN18. It should be noted that the weights are comparable between GCNs because the entire alpha matrix was learned altogether during the matrix factorization. Although the raw data showed some similarities with the spatial map of M15, we believe the assignments made by DLSC a better fit. For

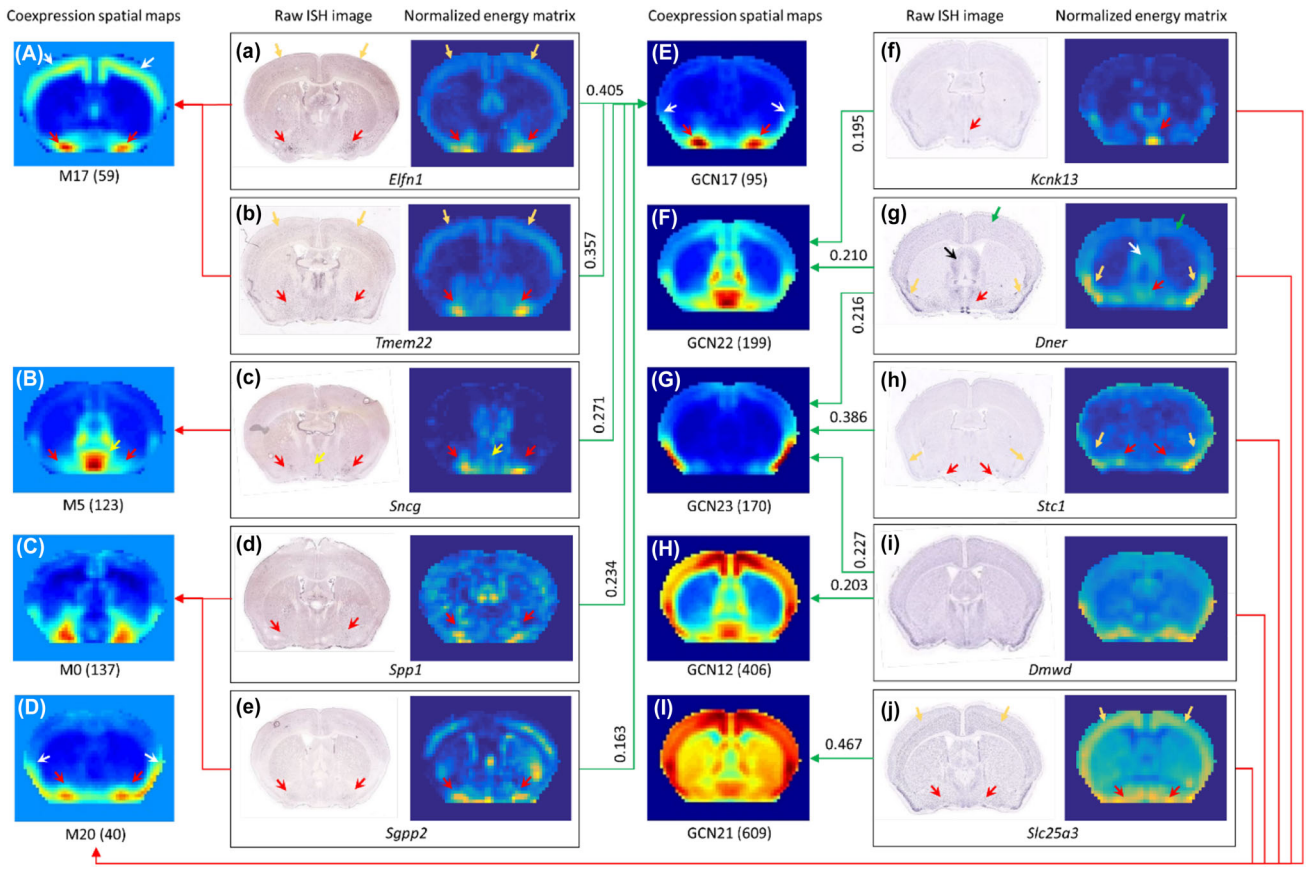

Fig. 4 Comparisons of genes in GCN17 and module 20 on slice 27. For each gene (**a**–**j**), the raw ISH image together with the normalized energy matrix is shown. **a**–**e** are five representative genes only found by DLSC. **f**–**j** are five genes only found by WGCNA. **A**–**D** are the spatial distributions of the eigengenes of selected modules. M0 is the module for unassigned genes. **E**–**I** are the spatial distributions of selected GCNs. The *number in the parentheses* of GCNs/modules denotes the number of genes in the GCN/module. The *long red arrows* show the module assignment made by WGCNA. The *green arrows* show the GCN assignment made by DLSC. DLSC offers a weight that measures the degree to which the gene expression conforms to the coexpression pattern. These weights are the values above the respective *green arrows*

example, *Gng2* was assigned to GCN10 (Fig. 3D) with a high weight of 0.449. The peak expressions at the endopiriform nucleus (Fig. 3d, red arrows) and relatively weaker expressions at the cortex regions (Fig. 3d black arrows) showed more resemblance to the spatial pattern of GCN10 than that of module 15. As to the second gene *Agpat4*, its raw ISH showed enhanced signals at the medial preoptic nucleus (Fig. 3e black arrows), the piriform area (Fig. 3e yellow arrows), as well as all outer layers of cingulate areas (Fig. 3e red arrows). These patterns were absent in M15, but featured in GCN12. The high weight of 0.378 also suggests a good agreement between *Agpat4* and GCN12. The last WGCNA-only gene is *Taldo1*. The similarity to module 15 is low as evidenced by the weak activations in cortex layers (Fig. 3f red arrows) and the enhanced signals in septal nucleus (Fig. 3f black arrows). DLSC assigned the gene to GCN1 which has wider yet lower activations throughout the slice with a low weight of 0.172. The energies from the three WGCNA-only genes were found diverged from the spatial map of represented by the eigengene of M15.

Following the same strategy, we examined another pair of networks where GCN includes over 50% of genes in the corresponding module, GCN17 and M20. This pair displays very similar spatial patterns that feature high expressions at lateral preoptic area and substantia innominata (Fig. 4D, E red arrows) and extends to piriform area with lower expressions (Fig. 4D, E white arrows). There were 95 genes in GCN17 and 40 genes in M20. Among them, 35 genes were shared. Five genes were WGCNA only, and the other 60 genes were DLSC only. Five DLSC-only genes with different weights were presented. With the decreasing weights, the resemblance to the spatial map of GCN17 decreased. Interestingly, both *Elfn1* and *Tmem22* were assigned to M17, which showed a better match at isocortex in comparison with that of GCN 17 (Fig. 4a, b

yellow arrows). *sncq* was assigned to module 5, presumably due to the similarity of the overall activations at hypothalamus although there was a mismatch of the degree of activation at medial preoptic area (Fig. 4c yellow arrows). In contrast, the high activations at the lateral preoptic area were more consistent with GCN17 (Fig. 4c red arrows). *Spp1* and *sgpp2* both showed broad activations in addition to the enhanced signals at the lateral preoptic area (Fig. 4d,e red arrows). They were left unassigned by WGCNA (M0 is the unassigned module).

Then we examined all the WGCNA-only genes. The expression of *kcnk13* peaked at the medial preoptic area (Fig. 4f red arrows) and was more consistent to GCN 22 (Fig. 4F) than M20. *Dner* showed enhanced signals at piriform areas (Fig. 4g yellow arrows) and extended further to isocortex (Fig. 4g green arrow), thalamus (Fig. 4g black/white arrow), and hypothalamus (Fig. 4g red arrow) with lower expressions. The expression pattern was captured by both GCN 23 (Fig. 4r) and GCN 22 (Fig. 4s) with the degree of consistency of around 0.2. *Stc1* showed strong signals at piriform area (Fig. 4h yellow arrows), but not as strong at lateral preoptic area (Fig. 4h red arrows). This pattern was more consistent with GCN23 (Fig. 4G). A similar case was also seen in *Dmwd* (Fig. 4i). Finally, the expressions of *Slc25a3* almost spanned the entire slice, with enhanced signals at the cortex (Fig. 4j yellow arrows) and preoptic areas (Fig. 4j red arrows). The expression pattern was better captured by GCN21 (Fig. 4I).

By analyzing the gene parcellations using WGCNA and DLSC on slice 27 in depth, we showed a very good consistency between the results obtained by WGCNA and DLSC. The discrepancy in the gene assignment was also demonstrated, which arises from different interpretation of the coexpression relationships. Thus, DLSC can provide a complementary perspective to other framework for gene coexpression network construction.

Notably, DLSC is robust to parameter selections as the result shown above were ran using sub-optimal parameters. When dictionary number is reduced from 29 to 26, most spatial patterns remain the same with slight changes to adapt for the reduced number of dictionaries (data not shown). Among 26 GCNs, 24 of them have over 50% the same genes as the counterpart in the GCNs derived using 29 dictionaries.

Comparative analysis with principal component analysis

To compare with other matrix factorization method, we performed principal component analysis on slice 27. Data was first centered by subtracting column means. Singular value decomposition algorithm was used as the solver. For visualization we projected each individual mode back to the brain space. The first 13 modes account for ~95% of variance, while the top 3 modes explain ~90% of the total variance. The first mode has a very broad distribution across the brain, with slightly higher expressions at the isocortex region (Fig. 5a). The second mode is also broadly distributed with distinctly high amplitude in caudoputamen (Fig. 5b). In contrast, the third mode features an absence of caudoputamen and is prominent in the hypothalamus

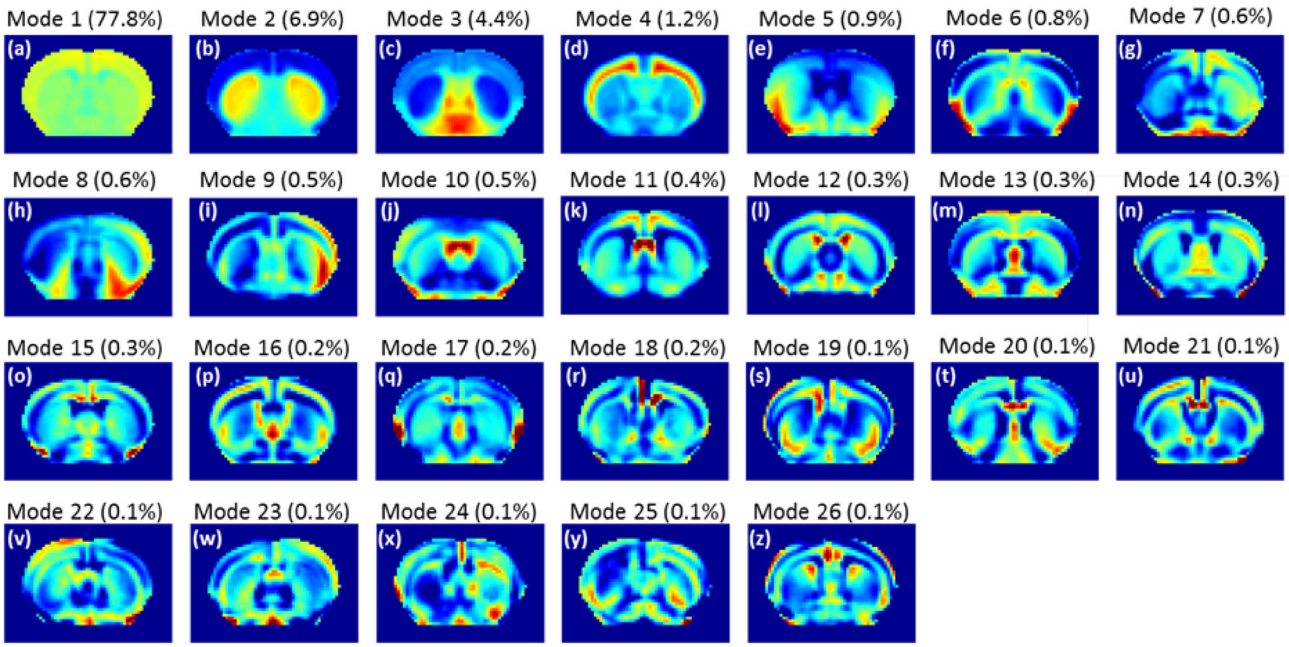

Fig. 5 Visualization of the first 26 modes obtained from principal component analysis. The values in the *parentheses* are the percentage of total variance explained by the mode

(Fig. 5c). Overall, PCA is able to extract correlated structures that correspond to the broad anatomical regions, such as caudoputamen (Fig. 5b) and isocortex (Fig. 5d). Yet with the additional modes that account for much less variance, the correspondence to the classical anatomy becomes increasingly weaker. On the other hand, with the goal of finding the coexpression patterns regardless of directions, PCA is not the best model for the problem because the modes are designed to capture the variance of the data instead of the common patterns of the data. Further, the orthogonal constraint keeps the model from finding meaningful overlapping coexpression patterns. One example is GCN 22 and GCN 19 (Fig. 2). Both GCNs show enhanced activations at the bed nuclei of the stria terminalis and were reported by DLSC and WGCNA. Using PCA, only mode 3 was found (Fig. 5c). Another example is GCN 3 (Fig. 2), GCN 26 (Fig. 2) and GCN16 (Fig. 2), which show distinctly different patterns at caudoputamen. All 3 GCNs were identified by both DLSC and WGCNA, while for PCA only mode 2 is most related to caudoputamen (Fig. 5b). Additionally, since our goal is to cluster genes with similar coexpression patterns, there requires an extra step of clustering analysis for PCA because with no sparsity constraint on the coefficients, the representation for the new bases is dense and the group assignment of genes is not readily available as DLSC. One last disadvantage of using PCA for GCN construction is that PCA generates negative numbers. The interpretation of the negative values does not appear immediately obvious in the context of gene expression patterns.

Gene coexpression network and brain parcellation

Existing literature 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 $\sim$3500 features into a feature vector composed of $\sim$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. Additionally, since parcellation agreement is used as an objective in the parameter optimization that is only performed on slice 27, we want to validate whether the selected parameters can result in good performance on other slices, by examining the features with reduced dimensionality. To quantify the level of correspondence between clustered voxels and the ARA on each slice we used normalized mutual information that is also used in parameter optimization. As seen in Fig. 6, 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 (Bernard et al. 2012). The delineations are in general symmetric and match major canonical brain regions including the hippocampus (Fig. 6 blue arrows), hypothalamus (Fig. 6 red arrows), and thalamus (Fig. 6 magenta arrows). The good correspondence is also reflected in the high normalize mutual information. The values are comparable to 0.6 which is the mutual information obtained from slice 27 (Fig. 6, Supplementary figure S3), suggesting the parameters are close to optimal for other slices. The most striking and principal features are the laminar and areal patterning that are seen in almost all slices (Fig. 6a–e yellow and orange arrows). 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. 2007). Within a dominant layered organization, layer specific areal patterning is also apparent. For instance, isocortex layers are further divided into motor areas (Fig. 6 green arrows), somatosensory area (Fig. 6 orange arrows), piriform area (Fig. 6 pink arrows), retrosplenial area (Fig. 6 dark green arrows), auditory area (Fig. 6 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 (Fig. 6e–f white arrows), hippocampus (Fig. 6e–f blue arrows), thalamus (Fig. 6e–f magenta arrows), and retrosplenial area (Fig. 6e–f 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.

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 a certain functional category by

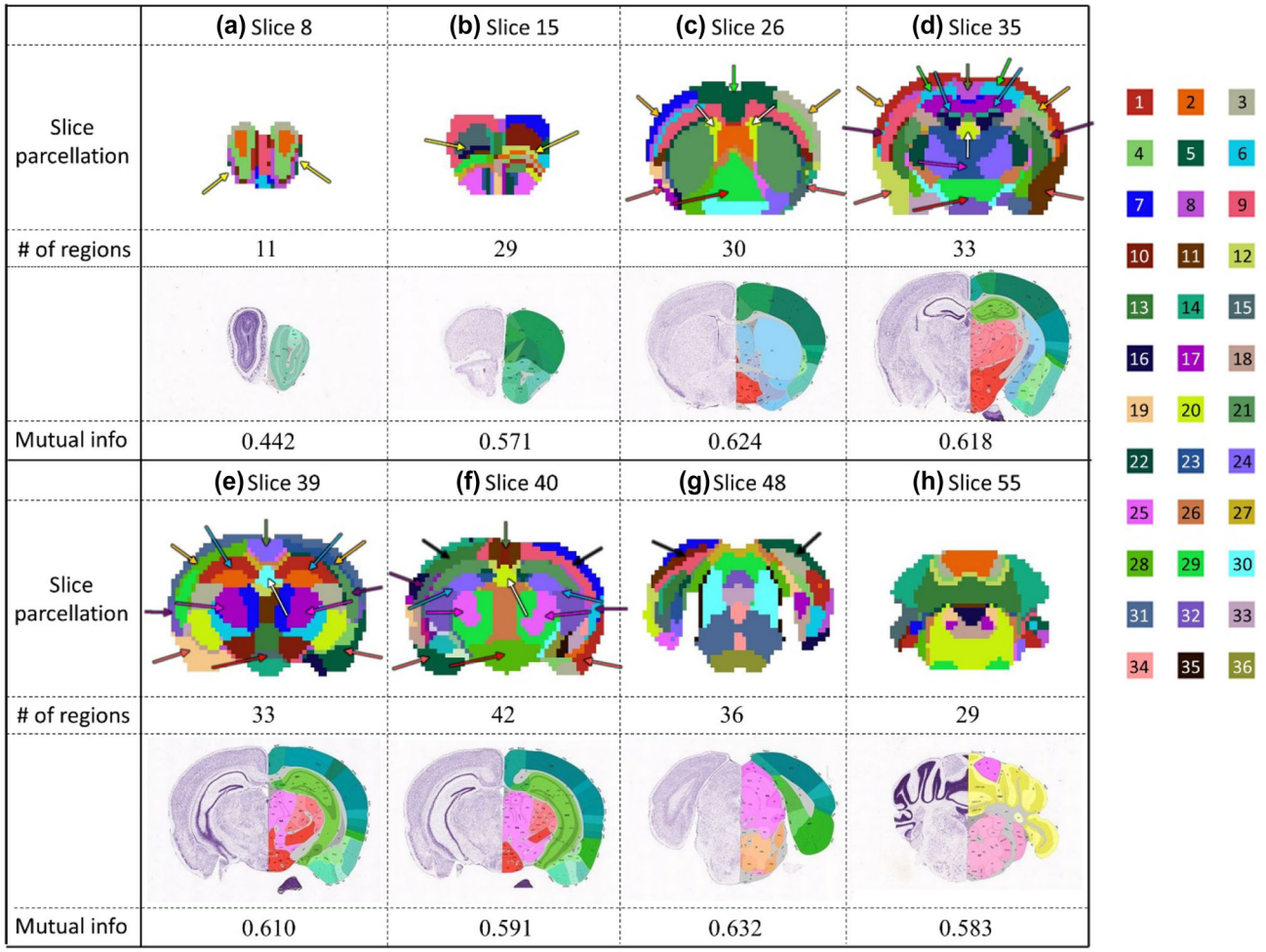

Fig. 6 Representative anatomical divisions based on the GCN features. Eight panels correspond to eight selected slices. In each panel, *top row* slice number; *second row* brain parcellation obtained from spectral clustering with dictionaries as feature vector; *third row* number of regions in the slice obtained by brain parcellation; *fourth row* visualization of Nissl stain image (*left*) and brain ontology (*right*) of the corresponding slice downloaded from ABA. *Fifth row* normalized mutual information between brain divisions and ARA in that slice. *Color* codes of each region are shown on the *right*

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 (Cahoy et al. 2004) indicates that GCN 5, 16, 23, 30, 43, and 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 labeling 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 and ribosomal functions. Literature 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).

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

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**

GCNs that are reproducibly identified enriched in certain category across references are bolded

The biological meaning of the GCNs has been not only confirmed by existing literature but also corroborated by the GO terms using DAVID. For example, two significant GO terms in GCN24 are myelination ($p = 7.7 \times 10^{-7}$) and axon ensheathment ($p = 2.5 \times 10^{-8}$), 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 ($p = 2.5 \times 10^{-6}$), agreeing with the other studies in 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 ($p = 4.9 \times 10^{-7}$) 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 (Fig. 7). For example, the fact that GCN 5 (Fig. 7ii) locates at ventricle, where the subventricular zone is rich with astrocytes (Quinones-Hinojosa and Chaichana 2007), confirms its enrichment in astrocyte markers. GCN 7 (Fig. 7v) 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 (Fig. 7vi) and the indicated enrichment in GABAergic pointed to a potential enrichment of GABAergic subtype neuron—the Purkinje cells. Comparing with the genes that only labeled 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 (Fig. 7x) in field CA1, GCN 4 (Fig. 7xi) in field CA3, GCN 38 (Fig. 7xii) in Dentate gyrus, GCN 45 (Fig. 7xiii) in cerebellum, GCN 21 (Fig. 7xiv) in medulla, GCN 1 (Fig. 7xv) in thalamus, and

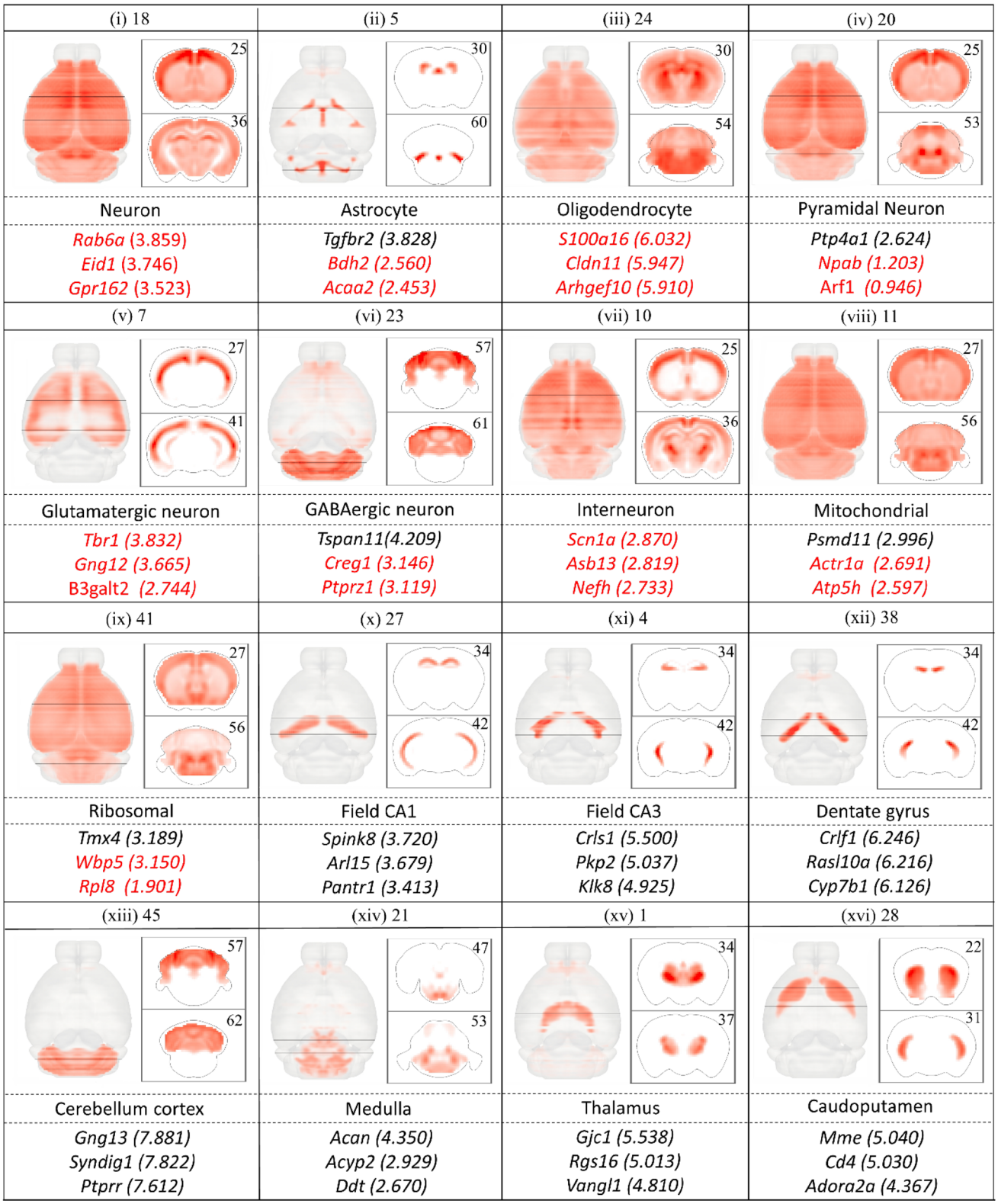

Fig. 7 Visualization of the spatial distribution of brain-wide GCNs significantly enriched for major cell types, particular brain regions, and biological functions. In each sub-figure, *top row* sub-figure index and brain-wide GCN ID. *Second row* 3D spatial maps of axial (*left*) and two selected coronal slices (*right*) of GCN. The location of each slice is highlighted in the 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 functionally enriched genes previously reported in the literature are highlighted in *red*

GCN 28 (Fig. 7xvi) in caudoputamen. 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 derived from multiple 2D slice-wide GCNs, the smooth and continuous 3D patterns, in turn, validates the reliability of slice-wide GCNs.

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 (Fig. 7). 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 are 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 pathway is relevant to the optic nerve head astrocyte migration (Miao et al. 2010).

Discussions

We have presented a data-driven framework that can derive biologically meaningful GCNs from the gene expression data. The motivation of the method comes from the recent success of applying DLSC for image denoising, demosaicing, etc. The sparse constraint on the coefficients can encourage dictionaries to capture the most common structures in images so that a parsimonious representation is possible. On the other hand, it is reported that most genes are expressed in a fairly small percentage of cells (Lein et al. 2007). We assume this notion can be captured by imposing a sparsity constraint that limits the number of voxels that a gene can be active on. To this end, DLSC can serve as a useful tool to extract the coexpression patterns. Using the 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. Thus we have contributed a new way of generating the coexpression networks by considering the transcriptome sparisty. The proposed 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 carries complementary information that helps identify, visualize, and in the future manipulate different types of neuron cells. Besides, we find that the learned dictionaries can serve as a very relevant and compact feature representing transcriptome profile for each voxel. The brain parcellations based on the learned dictionaries match well with the canonical neuroanatomy.

In contrast to many approaches that require 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 artifacts associated with data acquisition. Ideally, if gene information is complete and the data acquisition is perfect, this method can be directly applied to 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 into 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 is learned 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 nonbiological 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, aging, 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.

Acknowledgements T. Liu is supported by NIH R01 DA-033393, NSF CAREER Award IIS-1149260, NIH R01 AG-042599, NSF BME-1302089, NSF BCS-1439051 and NSF DBI-1564736.

References

Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinform 5:18. doi:10.1186/1471-2105-5-18

Bando SY, Silva FN, Costa LDF, Silva AV, Pimentel-Silva LR, Castro LH et al (2013) Complex network analysis of CA3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. PLoS One 8(11):e79913. doi:10.1371/journal.pone.0079913

Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, Finney EM et al (2012) Transcriptional architecture of the primate neocortex. Neuron 73(6):1083–1099. doi:10.1016/j.neuron.2012.03.002.Transcriptional

Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci USA 101(7):2173–2178. doi:10.1073/pnas.0308512100

Bohland JW, Bokil H, Pathak SD, Lee C-K, Ng L, Lau C et al (2010) Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods 50(2):105–112. doi:10.1016/j.ymeth.2009.09.001

Brown CD, Johnson DS, Sidow A (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science 317(September):1557–1560

Cahoy J, Emery B, Kaushal A, Foo L, Zamanian J, Christopherson K et al (2004) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neuronsci 28(1):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(6):611–621. doi:10.1016/j.gde.2013.10.003

Chen H, Li K, Zhu D, Jiang X, Yuan Y, Lv P et al (2013) Inferring group-wise consistent multimodal brain networks via multi-view spectral clustering. IEEE Trans Med Imaging 32(9):1576–1586. doi:10.1109/TMI.2013.2259248

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4(5):P3. doi:10.1186/gb-2003-4-5-p3

Dobrin R, Zhu J, Molony C, Argman C, Parrish ML, Carlson S et al (2009) Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. Genome Biol 10(5):R55. doi:10.1186/gb-2009-10-5-r55

Dong S, Li C, Wu P, Tsien JZ, Hu Y (2007) Environment enrichment rescues the neurodegenerative phenotypes in presenilins-deficient mice. Eur J Neurosci 26(1):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(2):407–499

Eisen MB, Spellman PT, Brown PO, Botstein D (1999) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(22):12930–12933. doi:10.1073/pnas.95.25.14863

Gaiteri C, Ding Y, French B, Tseng GC, 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(1):13–24. doi:10.1111/gbb.12106

Ge H, Liu Z, Church GM, Vidal M (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet 29(4):482–486. doi:10.1038/ng776

Grange P, Bohland JW, Okaty BW, Sugino K, Bokil H, Nelson SB et al (2014) Cell-type-based model explaining coexpression patterns of genes in the brain. Proc Natl Acad Sci USA 111(14):5397–5402. doi:10.1073/pnas.1312098111

Hawrylycz M, Bernard A, Lau C, Sunkin SM, Chakravarty MM, Lein ES et al (2010) Areal and laminar differentiation in the mouse neocortex using large scale gene expression data. Methods 50(2):113–121. doi:10.1016/j.ymeth.2009.09.005

Hawrylycz M, Miller JA, Menon V, Feng D, Dolbeare T, Guillozet-Bongaarts AL et al (2015) Canonical genetic signatures of the adult human brain. Nat Neurosci. doi:10.1038/nn.4171

Jiang CH, Tsien JZ, Schultz PG, Hu Y (2001) The effects of aging on gene expression in the hypothalamus and cortex of mice. PNAS 98(4):1930–1934. doi:10.1073/pnas.98.4.1930

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

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14:1085–1094. doi:10.1101/gr.1910904.1

Lein ES, Zhao X, Gage FH (2004) Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. J Neurosci 24(15):3879–3889. doi:10.1523/JNEUROSCI.4710-03.2004

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A et al (2007) Genome-wide atlas of gene expression in the adult mouse brain. Nature 445(7124):168–176. doi:10.1038/nature05453

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

Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416. doi:10.1007/s11222-007-9033-z

Mairal J, Elad M, Sapiro G (2008) Sparse representation for color image restoration. IEEE Trans Image Process 17(1):53–69. doi:10.1109/TIP.2007.911828

Mairal J, Bach F, Ponce J, Sapiro G (2010) Online learning for matrix factorization and sparse coding. J Mach Learn Res 11:19–60. http://portal.acm.org/citation.cfm?id=1756008

Miao H, Crabb AW, Hernandez MR, Lukas TJ (2010) Modulation of factors affecting optic nerve head astrocyte migration. Invest Ophthalmol Vis Sci 51(8):4096–4103. doi:10.1167/iovs.10-5177

Miller J (2014) Transcriptional landscape of the prenatal human brain. Nature 508(7495):199–206. doi:10.1038/nature13185.Transcriptional

Miller JA, Horvath S, Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci USA 107(28):12698–12703. doi:10.1073/pnas.0914257107

Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S (2011) Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinform 12(1):322. doi:10.1186/1471-2105-12-322

Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E et al (2001) Genome-wide gene expression profiles of the developing mouse

hippocampus. PNAS 98:8862–8867. doi:10.1073/pnas.141244998

Molyneaux BJ, Arlotta P, Menezes JRL, Macklis JD (2007) Neuronal subtype specification in the cerebral cortex. Nat Rev Neurosci 8(6):427–437. doi:10.1038/nrn2151

Ng L, Pathak SD, Kuan C, Lau C, Dong H, Sodt A et al (2007) Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. IEEE/ACM Trans Comput Biol Bioinf 4(3):382–392. doi:10.1109/TCBB.2007.1035

Ng L, Bernard A, Lau C, Overly CC, Dong H-W, Kuan C et al (2009) An anatomic gene expression atlas of the adult mouse brain. Nat Neurosci 12(3):356–362. doi:10.1038/nn.2281

O'Leary DD, Chou SJ, Sahara S (2007) Area patterning of the mammalian cortex. Neuron 56(2):252–269. doi:10.1016/j.neuron.2007.10.010

Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103(47):17973–17978. doi:10.1073/pnas.0605938103

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

Oldham MC, Langfelder P, Horvath S (2012) Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst Biol 6(1):63. doi:10.1186/1752-0509-6-63

Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14(4):481–487. doi:10.1016/j.conb.2004.07.007

Peng H, Long F, Zhou J, Leung G, Eisen MB, Myers EW (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(2):313–324. doi:10.1016/j.expneurol.2007.03.016

Rampon C, Jiang CH, Dong H, Tang YP, Lockhart DJ, Schultz PG et al (2000) Effects of environmental enrichment on gene expression in the brain. Proc Natl Acad Sci USA 97(23):12880–12884. doi:10.1073/pnas.97.23.12880

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255. doi:10.1126/science.1087447

Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C et al (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat Neurosci 9(1):99–107. doi:10.1038/nn1618

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E et al (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96(6):2907–2912. doi:10.1073/pnas.96.6.2907

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nat Genet 22(3):281–285. doi:10.1038/10343

Winden KD, Oldham MC, Mirnics K, Ebert PJ, Swan CH, Levitt P et al (2009) The organization of the transcriptional network in specific neuronal classes. Mol Syst Biol 5(291):291. doi:10.1038/msb.2009.46

Wright E, Ng L, Guillozet-Bongarts A (2007) Annotation report on cerebellar cortex, pukinje cell layer. http://community.brain-map.org/download/attachments/798/cbxpu.pdf?version=1