

DATA-DRIVEN EXPLORATION OF MOUSE BRAIN TRANSCRIPTOME

by

YUJIE LI

(Under the Direction of Tianming Liu)

ABSTRACT

The mammalian brain is the most complex organ. Modern genetics has shown that the complexity of brain structures and functions is ultimately encoded in the genome. As the primary functional interpretation of genome, a systematic study of transcriptome promises to enlighten how structures and functions are supported from the molecular scale. Fast advance in genomic information and throughput of technologies allows large-scale survey of transcriptome. The technique of *in situ hybridization* offers direct visualization of gene expression at cellular resolution. The spatial correlation among genes is closely associated with different phenotypes of anatomic regions. On the other hand, the correlations among transcripts allow us to investigate how sets of genes act in collaboration to control biological processes. However, how to unbiasedly derive the genetic-neuroanatomic correlations from the high-dimensional transcriptome data remains challenging. This thesis focuses on developing methods to connect genetics to neuroanatomy. To answer whether gene expression patterns can refine the architecture of the brain, I proposed dictionary learning and sparse coding (DLSC) as a tool because it considers the sparse structure of gene expressions. Voxels with similar coexpression patterns form tight clusters. Many clusters correspond well to neuroanatomy while others revealed finer delineation of regions previously considered homogeneous. Regionalized

expressions in fiber tracts and ventricular systems have been discovered and reported for the first time. DLSC is also proven effective in grouping genes into gene coexpression networks (GCNs). The GCNs are crucial to understanding how genes act jointly in defining the anatomy of the brain. Gene ontologies and comparisons with curated gene lists with known functions confirmed the functional roles of these networks. One standing issue for the above-mentioned work is incomplete data. To address the problem, I designed a volume completion network accompanied with customized training scheme. The network successfully completed the large missing region on a slice as well as one or two consecutive missing slices. On the completed data, I seek out a probabilistic-based model Restricted Boltzmann Machine and its extension, deep belief network, to construct a hierarchical transcriptome anatomy. A fine-to-coarse organization emerges from the network layers, providing a multi-resolution transcriptome architecture.

INDEX WORDS: transcriptional architecture, gene coexpression networks, mouse brain, unsupervised learning, deep learning

DATA-DRIVEN EXPLORATION OF MOUSE BRAIN TRANSCRIPTOME

by

YUJIE LI

BS, Huazhong University of Science and Technology, China, 2010

MPhil, University of Cambridge, UK, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2018

© 2018

YUJIE LI

All Rights Reserved

DATA-DRIVEN EXPLORATION OF MOUSE BRAIN TRANSCRIPTOME

by

YUJIE LI

Major Professor: Tianming Liu
Committee: Hanchuan Peng
Suchendra Bhandarkar
Yi Hong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2018

DEDICATION

This dissertation is dedicated to my parents Guorong Li and Yiping Tong, and my husband Hanbo Chen, for their continuous encouragement and support.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Prof Tianming Liu, for his fundamental role in my doctoral work. Without his tremendous efforts, guidance, and support throughout my PhD, I am not able to make through a big transformation from a Chemistry background to the field of computer science.

Second, I want to thank my committee members, Dr. Hanchuan Peng, Prof. Suchendra Bhandarkar and Prof. Yi Hong for their insightful feedback and interactions. Specifically, I want to thank Dr. Hanchuan Peng for a valuable opportunity to contribute to the exciting research projects at Allen Institute for Brain Science. His supervision and encouragement inspired me to widen my research areas from various perspectives.

Third, I would also like to thank my collaborators Prof. Joe Tsien from Augusta University, Prof Franck Polleux and Daniel Iascone from Columbia University. I am grateful for the opportunity to work with them on various exciting projects.

Next, I would like to thank my colleagues in the CAID lab and Allen Institutes for Brain Sciences, along with my friends in UGA. I cannot make this achievement without their help in both research and life.

Finally, I am deeply grateful to my family for their encouragement, love, support, and sacrifices. Without them, I cannot go as far.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER	
1 INTRODUCTION	1
What Is Transcriptome.....	1
Why is Transcriptome Important for Neuroscience.....	1
Technologies that Quantify Transcriptome	5
Allen Mouse Brain Atlas	9
Dissertation Outline	10
2 TRANSCRIPTOME ARCHITECTURE OF ADULT MOUSE BRAIN REVEALED BY SPARSE CODING OF GENOME-WIDE IN SITU HYBRIDIZATION IMAGES.....	13
Abstract.....	14
Background and Motivation	14
Methods.....	17
Transcriptomic Anatomy	21
Comparison with Principal and Independent Component Analysis	27
Online Informatics Portal.....	30

	Discussion and Conclusion	33
3	DISCOVER MOUSE GENE CO-EXPRESSION LANDSCAPES USING DICTIONARY LEARNING AND SPARSE CODING	36
	Abstract	37
	Background and Motivation	37
	Slice-wide GCN Construction and Validation.....	39
	Brain-wide GCN Construction	49
	Slice-wide GCN Analysis.....	51
	Brain-wide GCN Ontology and Spatial Analysis	63
	Discussion and Conclusion.....	68
4	VOLUME COMPLETION OF 3D <i>IN SITU</i> HYBRIDIZATION GRID USING FULLY CONVOLUTIONAL NEURAL NETWORK.....	71
	Abstract	72
	Background and Motivation	72
	Volume Recovery Network	74
	Comparison with 3D CAE and MEN	78
	Importance of Including Partial Slice Training	80
	Discussion and Conclusion.....	81
5	EXPLORING CO-OCCURRENCES OF ADULT MOUSE BRAIN VIA RESTRICTED BOLTZMANN MACHINE AND DEEP BELIEF NETWORK	83
	Abstract	84
	Background and Motivation	84
	Methods.....	86

RBM to Infer Single-Level Transcriptome Architecture.....	90
DBN to Infer a Hierarchy of Transcriptome Architecture.....	92
Discussion and Conclusion.....	93
6 CONCLUSIONS AND FUTURE WORK.....	96
Summary of Contributions.....	96
Future Directions	97
REFERENCES	99

LIST OF TABLES

	Page
Table 3.1: Reconstruction errors on slice 27 using different λ and gene-dictionary ratios.	46
Table 3.2: AUCs between the obtained dictionaries and the annotation map on slice 27 using different λ and gene-dictionary ratios	46
Table 3.3: The percentage of non-zero entries in the coefficient matrix obtained from DLSC on slice 27 using different λ and gene-dictionary ratios	47
Table 3.4: 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.	66
Table 4.1: Performance comparison between VRN, CAE and MEN.....	80

LIST OF FIGURES

	Page
Figure 2.1: Computational pipeline of the deriving transcriptome architecture using DLSC	18
Figure 2.2: Visualization of selected 3D spatial maps of the coefficient matrix	22
Figure 2.3: A comparison of transcriptome anatomy obtained for different dictionary numbers .	23
Figure 2.4: Hippocampal formation related dictionaries obtained by using 100, 200 and 400 dictionaries	24
Figure 2.5: 3D renderings of spatial patterns of field CA3 related components obtained using different dictionary numbers	25
Figure 2.6: Slice-based views of the spatial distribution of components that correspond to the fiber tracts (dictionary 17) and ventricular system (dictionary 71).	27
Figure 2.7: Visualizations of top 36 modes obtained using PCA	28
Figure 2.8: Visualization of 33 components corresponding to anatomical regions using ICA	29
Figure 2.9: Illustration of anatomic and genetic information of a dictionary component on the informatics portal	32
Figure 3.1: Computational pipeline for constructing slice-wide GCNs and brain-wide GCNs	40
Figure 3.2: Comparison between spatial distributions of GCNs and eigen-genes of WGCNA modules on slice 27	53
Figure 3.3: Comparisons of genes in GCN18 and module 15 on slice 27	54
Figure 3.4: Comparisons of genes in GCN17 and module 20 on slice 27	58
Figure 3.5: Visualization of the first 26 modes obtained from principal component analysis	60

Figure 3.6: Representative anatomical divisions based on the GCN features62

Figure 3.7: Visualization of the spatial distribution of brain-wide GCNs significantly enriched or major cell types, particular brain regions, and biological functions67

Figure 4.1: Volume recovery network architecture75

Figure 4.2: Comparison of the volume completion by VRN, CAE and MEN79

Figure 4.3: Illustration of the importance of incorporating partial slice corruption81

Figure 5.1: Illustration of Restricted Boltzmann Machine and deep belief network87

Figure 5.2: Visualization of weight maps learned by RBMs91

Figure 5.3: Visualization of a hierarchy of transcriptome architecture learned by DBN94

CHAPTER 1

Introduction

1 What is Transcriptome

Transcriptome refers to the entire set of RNA molecules, including mRNA, rRNA, tRNA, and other noncoding RNAs. Among them, mRNA plays the most important role because they encode proteins. Virtually all cells have the same copies of DNAs. It is the level of gene expressions that differentiate cells into different types and functions. For a set of instructions encoded in DNA to execute, these DNAs need to be transcribed, or in other words, readout, into transcripts. As a result, an analysis of the entire collection of transcripts gives us information about the gene activity inside the cells, including which transcripts are active, their expression levels and the novel splicing sites.

2 Why is Transcriptome Important in Neuroscience

2.1 Understand cortical organization

Transcriptome in brain plays a crucial role in understanding the cortical organization and the development of brain structure. It provides a relatively stable platform for the research of cortical organization because the transcriptome is not altered much by behavioral manipulations or cognitive states, but varies strongly between anatomical locations, cell types, and developmental stages (Lein, Ed S. et al., 2007). Previous research has revealed extensive regional heterogeneity of transcriptome. A number of molecular markers, such as calcium-binding proteins and growth factors, showed distinct patterns that can be utilized to distinguish

between field CA1 and field CA3 in adult mouse and rat brains (Woodhams, Celio, Ulfig, & Witter, 1993). Tole et al (Tole, Christian, & Grove, 1997) further discovered that two field-specific genes display unique patterns distinguishable between CA1 and CA3 a week before the distinctions in morphology are displayed. Later, with the improvement of DNA microarray and *in situ* hybridization (ISH), a large number of gene expression patterns were reported to mirror the gross anatomical partitioning in hippocampus and some region-specific gene expression patterns can delineate the brain into finer subdivisions (Ed S. Lein, Zhao, & Gage, 2004; X. Zhao et al., 2001). As the current preeminent methodology in transcriptomics, the explorative single-cell RNA sequencing (RNA-Seq) (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) showed its power by classifying cells in the mouse somatosensory cortex and hippocampal CA1 region into 47 subclasses (Zeisel et al., 2015). The discovery of the region-specific and cell-type related genes lays the foundation for the elucidation of the detailed mechanism that controls the specification and differentiation of areas and brain development and functioning, as well as genetic dysregulation in large.

2.2 Common ‘default gene network’

In contrast to Section 2.1, whose goal is to understand the spatial distribution of genes as a function of structures, this line of research aims to identify the core transcriptional machinery conserved across individuals. Faced with complex patterns of expression of thousands of genes, a Gene Coexpression Network (GCN), representing the interactions among genes, is often used. Previous studies have shown that genes displaying similar expression profiles are very likely to be involved in the same transcriptional regulatory program (Allocco, Kohane, & Butte, 2004; Mody et al., 2001), encode interacting proteins (Ge, Liu, Church, & Vidal, 2001) or participate in the same biological processes (Tavazoie, Hughes, Campbell, Cho, & Church, 1999).

An establishment of ‘default gene network’ allows us to focus on the highly conserved features. For example, one pioneering work by Stuart et al (Stuart, Segal, Koller, & Kim, 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 that are essential to viability. Knowledge of these key groups is an essential step to understanding the overall design of genetic pathway. Efforts also went toward deriving common GCNs in the human brain (Michael 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. The common clusters of genes can also be compared on a temporal direction. Kang et al. (Kang et al., 2011) investigated human brain transcriptome over 15 periods from embryo to late adulthood collected from 57 human brains. They identified two temporally regulated clusters of genes. By comparing with the curated list of genes that are the indicators of various neurodevelopmental processes, they discovered common neurodevelopmental trajectory patterns (Kang et al., 2011).

With a common template set up, the differences due to developmental stages, species, or between healthy and diseased brains (Michael Hawrylycz et al., 2015; Kang et al., 2011) are also available for study. Bakken et al (Bakken et al., 2016) compared the expression trajectories in the frontal cortex among rat, rhesus monkey and human. They reported a human-specific developmental trajectory that is attributed to 9% of genes, featuring prolonged maturation. In another study by Voineagu and colleagues (Voineagu et al., 2013), comparisons were also made

between transcriptome of brain tissues of 19 autism patients and 17 healthy controls. The results showed distinct regional patterns at frontal and temporal cortex between patients and controls. The highly correlated clusters of genes are enriched for autism susceptible genes. Further, the following RNA-Sequencing of selected gene candidates demonstrate the splicing dysregulation as the underlying mechanism of the disorder.

2.3 Integration with other modalities

The complexity of brain structures and function is ultimately encoded in genes. As a result, transcriptome serves as a central modality that complements other meso-scale or macroscale modalities. One good example is neuronal cell-type classification. Cell type classification of neurons has been a fundamental, long-standing task in the neuroscience field because a clear taxonomy is the foundation for understanding how brain works or fail to work (Migliore & Shepherd, 2005; Seung & Sumbul, 2014; Zeng & Sanes, 2017). Yet the diverse molecular, morphological and physiological properties of neurons make it extremely challenging (Migliore & Shepherd, 2005; Zeng & Sanes, 2017). Traditionally, neurons are classified by a single feature, such as morphology or electrophysiology. Later on, it is proved that an integration of critical features can better capture the complex functional phenotypes (Migliore & Shepherd, 2005). In a study of somatosensory cortex of juvenile rat, the morpho-electrical characterization assigned all excitatory neurons into one single category (Markram, 2015). In contrast, single cell RNA-Sequence profiling of genes from mouse visual cortex successfully classified 19 subtypes (Tasic et al., 2016). Thus, transcriptomic profiling of cells brings information from a different dimension to the classification task.

Other researchers referred to gene expressions as a means of understanding how the molecular underpinnings correlate with the neural properties such as connectivities and electrical

profiles. One of the heavily studied topics is how the functional connectivities are supported from molecular scale (Fakhry & Ji, 2015; French & Pavlidis, 2011; Wolf, Goldberg, Manor, Sharan, & Ruppin, 2011). Multiple researchers reported that a significant correlation exists between gene expressions and neuroanatomical tracing profiles (Fakhry & Ji, 2015; French & Pavlidis, 2011; Wolf et al., 2011) in rodent as well as functional resting state activity and transcriptome in human (Michael Hawrylycz et al., 2015; G. Wang et al., 2015). Fakhry et al. (Fakhry & Ji, 2015) went a step further and showed that the gene expressions can accurately predict the connectivity with an accuracy of 93% on the voxel level. The ontologies of the genes that contribute most to the prediction are related to neuroanatomical connectivity. Relatedly, Toledo and colleagues (Toledo-Rodriguez et al., 2004) reported that single-cell gene expression can also predict the electrical properties of neurons. Since the transcriptome difference is the root cause for the difference in the phenotypes, linking them at various properties is a fruitful avenue.

3 Technologies that quantify transcriptome

Most of the commonly used transcriptome technologies fall into the categories of hybridization-based methods and sequence-based methods. For hybridization-based methods, it relies on the hybridization reactions between two DNA strands, where one strand is specifically matched to the complementary nucleic acid strand via hydrogen bonds. Sequence-based methods, on the other hand, directly sequence the transcripts.

3.1 Microarray

Microarray has been one of the dominant transcriptional technologies since its first introduction in the 1990s. The ability of processing tens of thousands of transcripts simultaneously at a relatively low cost make it the most widely used tool for large-scale transcriptome analysis. Microarray has been used for various studies, including characterizing

differentially expressed genes across different conditions, exploring genetic abnormalities in a range of tumors (Veltman, Fridlyand, Pejavar, & et.al., 2003), profiling physiological functions of unknown genes and uncovering transcriptional regulation in human and animal model (Su et al., 2002).

Microarray is a typical example of the hybridization-based methods. The sample containing mRNA of interest is first isolated, reversed transcribed and copied into stable double stranded complementary DNA (ds-cDNA). Then these ds-cDNA samples are fragmented, fluorescently labelled and incubated with the probes on the microarray. Each probe on the array is a fragment of predefined complementary oligonucleotides of known DNA or RNA sequence. The abundance of the transcripts in the sample is determined by the fluorescence intensity bound to the probe.

The key limitation of this technology is that it relies on previous knowledge of genome sequence. The quantification of transcriptome is only restricted to the genes arrayed in a probe and any sequence beyond the pre-defined genomic sequence is not detected. Another concern is the high background noise because of cross-hybridization, which undermines the accuracy of microarray. The high level of noise specifically poses a challenge to accurately measure the transcript in low abundance. Relatedly, due to the high sensitivity to the experiment environment, comparisons across different experiment set-up, time points, different laboratories or on different microarray platforms is often difficult and requires complex normalization methods (Irizarry et al., 2005; Jaksik, Iwanaszko, Rzeszowska-Wolny, & Kimmel, 2015; Johnson, Li, & Rabinovic, 2007).

3.2 RNA-Sequencing (RNA-Seq)

Sequence-based approach, as the name suggests, measure the cDNA or RNA sequence. With the advance in next generation sequencing technology (Metzker, 2010), it is possible to quantitatively analyze the RNA molecules through sequencing on a large scale and this technology is named RNA-Seq. The methods for RNA-Seq is different from microarray. Rather than incubate with the probes on the microarray, the samples that contain RNA or ds-cDNA are sequenced. The level of transcripts is reported as the number of reads within the gene bounds, normalized by sequencing depth.

As a sequence-based transcriptome profiling technology, RNA-Seq enjoys many advantages. First, RNA-Seq is not constrained to pre-defined genome sequence. In the scenario where no reference genome is given, it can assemble the sequence *de novo* and provide both the sequence as well as the expression counts for each transcript. This property makes RNA-Seq very appealing to transcriptome studies of non-model organisms and research that characterize alternative splicing patterns (Trapnell, Pachter, & Salzberg, 2009) or gene fusion (Maher et al., 2009). Second, RNA-Seq quantifies the expression by counting the exact match between a DNA sequence and the region of genome. Therefore, it has a much broader detection range given little background noise or signal saturation. Multiple studies that compare the results obtained using microarray and RNA-Seq showed that RNA-Seq provide better estimates to the absolute transcript levels (S. Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014).

It is worth mentioning that single-cell RNA-Seq (sc RNA-Seq) has been particularly useful for neuronal cell type classification (E. Lein, Borm, & Linnarsson, 2017; Zeng & Sanes, 2017). Cells dissected from neuronal tissues of a specified brain region are sorted and further

sequenced. The inventory of the transcriptome profiles of cells provide a key component useful for cell typing.

3.3 *In situ* hybridization

In situ hybridization (ISH) is a powerful technique to localize a specific DNA or RNA sequence in a tissue. It is achieved by hybridizing pre-designed complementary oligonucleotide probes, which are fluorescently or calorimetrically labeled. A visualization of these hybridized probes shows the spatial distribution along with the level of expressions of the sequence of interest. Unlike the non *in-situ* approaches that require samples removed from the native environment, ISH is an image-based approach and thus naturally preserves the spatial location of RNAs within a cell and the organization of cells within tissue.

The spatial information of transcriptome is crucial for probing many biological questions. For example, to understand the molecular underpinning of the different phenotypes of subregions in hippocampus, Lein et al (E. S. Lein, 2004) and colleagues first used microarray to locate the genes that are regionally expressed in the hippocampus regions. Further with ISH, they visualize the spatial distributions of the selected gene expressions and reported that gene expressions respect the cytoarchitectural boundaries in hippocampus. ISH is also a powerful tool to understand the linkage of presence of candidate genes in a brain region to a brain disease because many brain diseases show strong spatial organizations. In the work by Cohen et al (Cohen, Golde, Usiak, Younkin, & Younkin, 1988), a comparison of the ISH of brain tissues from Alzheimer diseased (AD) patients and controls confirmed that the increased level of β -amyloid mRNA, which plays a role in AD, comes from nucleus basalis neurons in the Broadman area 21 (Cohen et al., 1988).

4 Allen Mouse Brain Atlas

The central dataset used in the thesis is the publicly available Allen Mouse Brain Atlas (AMBA). The AMBA (Lein, Ed S. et al., 2007) provides genome-wide *in situ* hybridization image data for approximately 20,000 genes in 56-day-old male C57Bl/6J mouse brain. The inbred mouse strain is used to reduce the animal-to-animal variation in brains. Processed brain tissues were first cut into slices and a set of 2-dimensional (2D) ISH images were generated for each transcript tested. These ISH images were then processed in an informatics pipeline to obtain a three-dimensional (3D) expression grids (Lau et al., 2008). To enable three-dimensional volumetric representations from the acquired coronal or sagittal series images, a common coordinate space of the 3D reference atlas (H. Dong, 2008) was first created so that the ISH images of each gene can be consistently registered to the same space and aligned. To enable quantification, each image was divided into a 200 μm isotropic grid and pixel-based statistics were collected. The output is a 3D summary of the gene expression statistics for each transcript. The resulted voxelized expression grids encode the important spatial information of 4,345 genes in coronal sections and 21,718 genes in sagittal sections. They make up the key components of the AMBA. In the paper, expression energy metric was used for all analyses. As seen in equations (1.1-1.3), this metric is correlated with the total transcript count incorporating both area occupied by expressing pixels as well as pixel intensity.

$$\text{expression density} = \text{sum of expressing pixels} \div \text{sum of all pixels in division} \quad (1.1)$$

$$\text{expression intensity} = \text{sum of expressing pixel intensity} \div \text{sum of expressing pixels} \quad (1.2)$$

$$\text{expression energy} = \text{expression intensity} \times \text{expression density} \quad (1.3)$$

Throughout the thesis, coronal sections are chosen for analysis because they registered more accurately to the reference model than the counterparts of sagittal sections. 4,345 3D

volumes of expression energy of coronal sections were downloaded from the website of ABA (<http://mouse.brain-map.org/>) to perform our analysis. A 3D volume of brain anatomical annotation based on the ARA (Version 3) was also downloaded. The dimension of all 3D volumes is 67 (posterior-anterior) by 41(inferior-superior) and by 58 (right-left).

5 Dissertation Outline

In this thesis, I focus on developing data-driven methods to relate gene expression patterns to neuroanatomy.

Chapter 1 starts with an introduction of the concept of transcriptome and explanation of why the study of transcriptome is important for advancing our understanding in how brain shapes and works by a brief survey of recent literatures. Next, I give a summary of different methods for characterizing transcriptome and their pros and cons. In section 4, a detailed description on the central dataset for the thesis projects is given. The final section is the outline of each chapter.

In chapter 2 and chapter 3, I showed that dictionary learning and sparse coding (DLSC) is a useful data-driven method in relating spatially resolved gene expression data to neuroanatomy. Chapter 2 focuses on building a whole-brain transcriptome architecture of adult mouse using DLSC. The key idea is to consider gene expressions as features of each voxel. Those voxels using the same dictionary for representation should share similar features and therefore clustered to the same region. Multiple components were found corresponding to the canonical neuroanatomical subdivisions. Other components revealed finer anatomical delineation of domains previously considered homogeneous. An informatics portal was built as an open-access resource for result visualization and further explorations.

In chapter 3, using DLSC, I demonstrated a new way of constructing gene coexpression networks (GCNs). GCNs is an effective and efficient representation of gene-gene interactions.

They are useful when we need to make comparisons across species, or over developmental time points. The key assumption is that if two genes use the same dictionary for representation, these two genes should share similar coexpression patterns and thereby belong to the same network. Following the assumption, 50 GCNs were constructed. To verify the constructed GCNs, I compared them with the ones generated by a most widely used method weighted gene coexpression network analysis (WGCNA) (Langfelder & Horvath, 2008). The comparative analysis showed a very good consistency between the GCNs generated by the two methods while DLSC provides a complementary perspective when different gene assignment arises. Further, to interpret GCNs biologically, I performed gene ontologies and compared with published gene lists of known functions. A set of GCNs were found significantly enriched for major cell types, anatomical regions, biological pathways and/or brain diseases.

In chapter 4, I proposed a volume recovery network (VRN) that completes 3D volume data. A major issue encountered to the work presented in chapter 2 and 3 is data missing of the *in-situ* hybridization data. In chapter 2, about 30% of data was removed from further analysis due to a lack of data. In chapter 3, I worked around the data missing problem by only considering the gene-gene interaction on slices with data. However, this is only a temporary solution for a specific task. In chapter 4, I provide a complete solution to this problem. The rationale for the design of VRN is analogous to denoising autoencoders (Vincent, Larochelle, Bengio, & Manzagol, 2008). Instead of feeding the network data with manually added noises and teaching the network to undo noises, we hide a portion of each training sample so that the network can learn to recover missing voxels from the context. A comparison with different training schemes shows the importance of designing the right strategy that fits the missing data patterns.

In chapter 5, I continued the work on analyzing the transcriptional patterning of mouse brain. In chapter 2, we have successfully delineated brain regions based on the differentially regulated gene. However, most of the existing methods, including the work in chapter 2, assumes linear shallow mappings and inadequate for inferring complex non-linear structures of data. In this chapter, I seek out to a probabilistic model known as a restricted Boltzmann machine (RBM) and the deeper model deep belief network, which is a stack of RBMs. It is demonstrated RBM can derive meaningful delineations from the transcriptome. Then we stack multiple RBMs to form a DBN. With DBN, a fine-to-coarse organization emerges from the network layers. This organization incidentally corresponds to the anatomical structures, suggesting a close link between structures and the genetic underpinnings.

In chapter 6, I summarize the thesis and discuss future work.

CHAPTER 2

TRANSCRIPTOME ARCHITECTURE OF ADULT MOUSE BRAIN REVEALED BY SPARSE CODING OF GENOME-WIDE IN SITU HYBRIDIZATION IMAGES¹

¹ Yujie Li, Hanbo Chen, Xi Jiang, Xiang Li, Jinglei Lv, Meng Li, Hanchuan Peng, Joe Z. Tsien, Tianming Liu, 2017, Transcriptome Architecture of Adult Mouse Brain Revealed by Sparse Coding of Genome-Wide *in Situ* Hybridization Images, *Neuroinformatics*. 15:285–295.
Reprinted here with permission of the publisher.

1 Abstract

In this chapter, I demonstrate how dictionary learning and sparse coding can be used to derive transcriptome organization from gene expression data from mouse brain.

Highly differentiated brain structures with distinctly different phenotypes are closely correlated with the unique combination of gene expression patterns. Using a genome-wide *in situ* hybridization image dataset released by Allen Mouse Brain Atlas, we present a data-driven method of dictionary learning and sparse coding. Our results show that sparse coding can elucidate patterns of transcriptome organization of mouse brain. A collection of components obtained from sparse coding display robust region-specific molecular signatures corresponding to the canonical neuroanatomical subdivisions including fiber tracts and ventricular systems. Other components revealed finer anatomical delineation of domains previously considered homogeneous. We also build an open-access informatics portal that contains the detail of each component along with its ontology and expressed genes. This portal allows intuitive visualization, interpretation and explorations of the transcriptome architecture of a mouse brain.

2 Background and Motivation

Highly differentiated brain structures with distinctly different phenotypes are closely correlated with the unique combination of gene expression patterns (Jiang, Tsien, Schultz, & Hu, 2001; Mody et al., 2001). Many studies have reported that transcriptomes can serve as important, informative modalities to classify cell types and reveal deeper organization of brain structures (Mike Hawrylycz et al., 2010; Heintz, 2004; Nelson, Sugino, & Hempel, 2006; Winden et al., 2009). These results, together with many others (Belgard et al., 2011; Heintz, 2004; Molyneaux, Arlotta, Menezes, & Macklis, 2007), provide strong evidence that gene expression patterns are useful features in revealing the cellular makeup of different brain regions.

Led by the exciting discoveries revealed by gene expression studies, a global systematic study on a wide range of cellular markers with fine resolutions is essential to make quantitative associations between genetic and anatomical architecture of the entire brain. One enormous effort is the openly available Allen Mouse Brain Atlas (AMBA) (Lein, Ed S. et al., 2007), which provides genome-wide *in situ* hybridization (ISH) image series of the adult mouse brain at cellular resolution. To investigate the differences between the “transcriptome fingerprints” of different brain locations, ISH image series for each mRNA is registered to a common atlas space, the Allen Reference Atlas (ARA) (H. Dong, 2008) so that a global comparison across regions and against the classical neuroanatomy is possible. (Mike Hawrylycz et al., 2010; Ng et al., 2009; Thompson et al., 2008).

Multiple tools and methods have been developed for mining the ISH dataset. The Anatomic Gene Expression Atlas (AGEA) (Ng et al., 2009), for instance, is a publicly available computational tool specifically designed to visualize the spatial correlations of gene expression patterns in the mouse brain. In AGEA, gene expression patterns are features of each voxel and Pearson correlation metric is used to measure the similarity between voxels. Based on the calculated similarity, a hierarchical clustering was applied to parcellate apparent anatomical subdivision. Yet the tool requires regions defined for enrichment a-priori. On the other hand, Bolhand and colleagues (Bohland et al., 2010) have shown that singular value decomposition (SVD) was able to reveal structures in rough concordance with classical anatomy, yet finer structures were not resolved and an extra step of K-means clustering was required to cluster voxels with similar gene expression profiles. Relatedly, a modified non-negative matrix factorization (mNMF), was also used to study ~2600 genes expressed in hippocampus and led to the identification of a large groups of regionally enriched transcripts (Thompson et al., 2008).

Inspired by the above promising findings, we proposed to apply dictionary learning and sparse coding (DLSC) on genomic data. DLSC is a data-driven method aiming at obtaining parsimonious representation of data. The popularity of applying DLSC on images derived from the observations that neurons encode sensory information using a small number of active neurons at any given point in time (Olshausen & 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 & Field, 2004). Due to these properties, DLSC has found great success in applications such as image denoising, demosaicing and inpainting (Elad & Aharon, 2006; Mairal, Elad, & Sapiro, 2008). In the context of revealing the transcriptome organization based on gene expression profiles, we assume that if multiple voxels use the same dictionary atom for sparse representation, then these voxels must share the features described by the shared dictionary atom and thereby should belong to the same subregion. 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, Ed S. 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. Thus, DLSC can serve as a useful tool that learns the internal transcriptome architecture from the ISH dataset without any prior knowledge.

In this study, we performed a comprehensive analysis on the genome-wide *in situ* hybridization data of the mouse brain and showed that DLSC can effectively elucidate patterns of transcriptome organization. Several components obtained from sparse coding display robust regional specific molecular signatures corresponding to the canonical neuroanatomical subdivisions. Other components revealed finer anatomical delineation of domains previously considered homogeneous. An informatics portal was built as an open-access resource for result

visualization and further explorations. The webpages contain the spatial distribution of the components and the corresponding ARA ontology of neuroanatomical structures, as well as the genes that are regionally enriched. The links to the original dataset affords a direct comparison and a convenient interpretation.

3 Methods

The computational pipeline is outlined as follows (Figure 2. 1). First, images of gene expression patterns were downloaded from AMBA dataset (Lein, Ed S. et al., 2007). Based on the corresponding annotation map, foreground voxels were extracted for analysis. Those voxels with missing data were either excluded from analysis or estimated from the neighboring voxels (Figure 2. 1 (a)). Then the 3D expression energies for one gene were flattened out into one line so that all gene expression data can be arranged into a big matrix where each row corresponds to one gene and each column corresponds to one voxel. The matrix was next decomposed into a fixed number of dictionaries and its corresponding coefficient matrix (Figure 2. 1(b)). Due to the sparse constraints on the energy function, the coefficient matrix is sparse and encodes the spatial distribution of each dictionary. Finally, we compared the spatial patterns of the learned dictionary components with the manual annotation atlas from ARA (Figure 2. 1(c)). An informatics portal was built to present the whole mouse brain's transcriptome architecture (Figure 2. 1(d)).

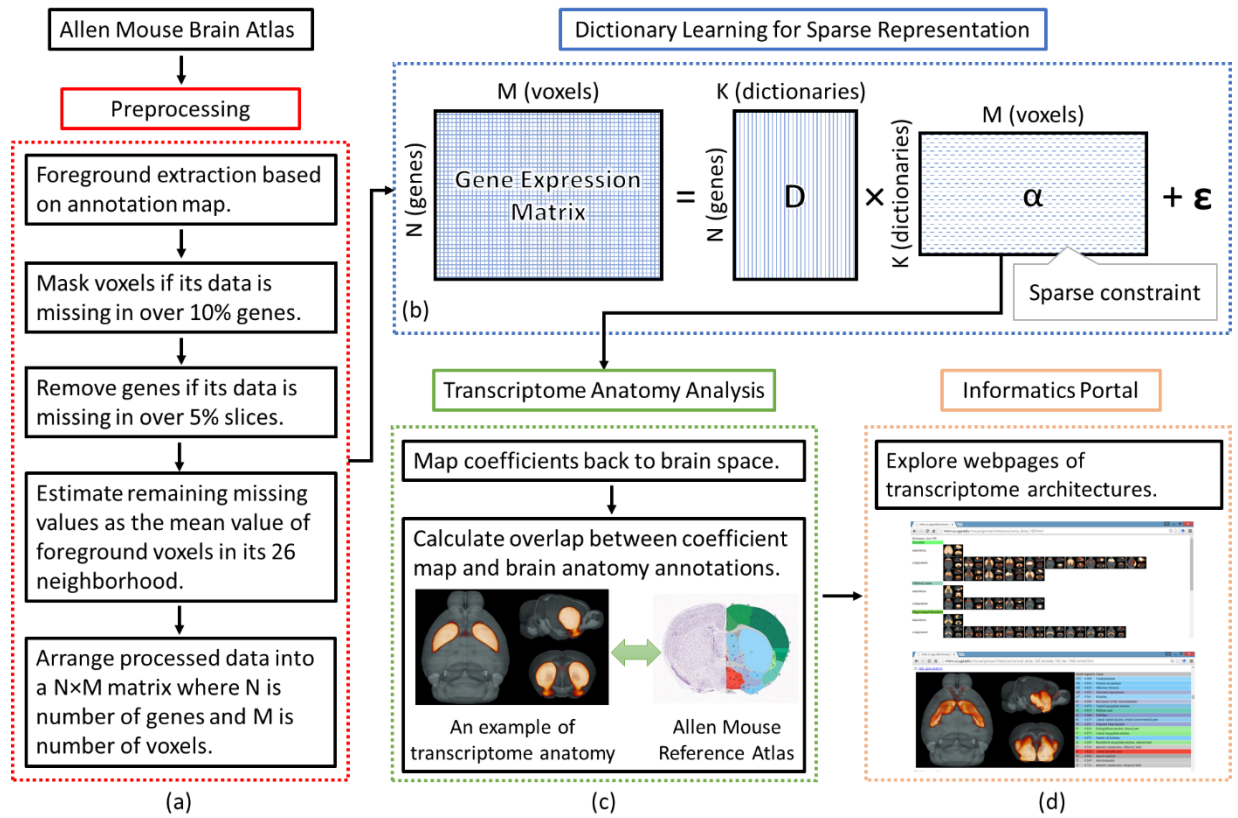


Figure 2. 1. Computational pipeline of the deriving transcriptome architecture using DLSC. (a) Preprocessing steps for ISH data from Allen Mouse Brain Atlas. (b) Dictionary learning and sparse coding of ISH matrix. (c) Comparisons between transcriptome spatial patterns with the neuroanatomy. (d) Informatics portal to facilitate the exploration of transcriptome architecture.

3.1 Data preprocessing

Based on the 3D annotation, a mask of brain volume was generated and applied to extract foreground voxels (62529 voxels). By observation, data were missing for many foreground voxels (-1 in expression energy). The lack of data was assumed mostly due to problems during data acquisition such as missing slices, broken tissues, and slice misalignment. Mainly the missing data were categorized into three groups: 1) An entire slice was lost; 2) Part of a slice was lost; 3) A few voxels were missing. To reduce the impact of missing data, two filtering steps and an estimation step were performed at the preprocessing stage. First, a filtering step was applied to mask out “unreliable” voxels. A foreground voxel with gene expressions missing in over 10%

of the total transcripts was removed. In this step, about 7% of foreground voxels were eliminated. Second, a filtering step was applied to filter out “unreliable” transcripts. A transcript with expressions missing for an entire slice was excluded. After this step, 67% (2905/4345) transcripts were retained for further analysis. Most missing values were resolved in the two filtering steps. The remaining missing values were estimated as the mean of foreground voxels in its 26 neighborhood. Recursive mean calculations were performed on the images until all missing values were filled. Eventually, 2905 transcripts on 60904 foreground voxels were sent to the DLSC module.

3.2 Dictionary learning and sparse coding

Dictionary learning and sparse coding is a useful tool that can extract meaningful patterns from signals. Given a matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, it can be approximated by the matrix factorization such that:

$$\mathbf{X} = \mathbf{D} \times \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (2.1)$$

where $\mathbf{D} \in \mathbb{R}^{N \times K}$ is the dictionary matrix, $\boldsymbol{\alpha} \in \mathbb{R}^{K \times M}$ is the corresponding coefficient matrix, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{N \times M}$ is the reconstruction error. This matrix decomposition problem is solved with a sparse constraint on $\boldsymbol{\alpha}$, which limits the number of dictionaries used to reconstruct the original signals.

The factorization can be formulated as the following optimization problem:

$$\langle \mathbf{D}, \boldsymbol{\alpha} \rangle = \operatorname{argmin} \frac{1}{2} \|\mathbf{X} - \mathbf{D} \times \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \quad (2.2)$$

where $\|\cdot\|_2$ is the summation of ℓ_2 norm of each column and $\|\cdot\|_1$ is the summation of ℓ_1 norm of each column. λ regulates the tradeoff between the sparsity of $\boldsymbol{\alpha}$ and the reconstruction error.

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 2.5 should be minimized (Mairal, Bach, Ponce, & Sapiro, 2010).

In practice, we arranged the gene expression energies into a single matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, such that N rows correspond to N genes and M columns correspond to M foreground voxels. Then, each column of the matrix was centered and then normalized by the standard deviation of the elements in each column. After normalization, the publicly available online dictionary learning and sparse coding package was applied to solve the matrix factorization problem proposed in equation 2.5 (Mairal et al., 2010). Eventually, the gene expression energy matrix \mathbf{X} was decomposed into a dictionary matrix \mathbf{D} and a sparse coefficient matrix $\boldsymbol{\alpha}$.

The key idea of applying sparse coding to the ISH dataset is that if multiple voxels use the same dictionary atom for sparse representation, then these voxels share the features described by the shared dictionary atom and thereby should form a subregion. The major assumptions of applying sparse coding to the ISH data is that each gene is expressed in a limited number of voxels in the brain. This assumption is supported by the fact 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, Ed S. et al., 2007). The other assumption is that the gene expression energies can be linearly combined because in DLSC each dictionary is a linear combination of gene expressions. If the integration of two gene expression follows a non-linear relationship, DLSC would not be able to reconstruct the original signals correctly. The similarities between the reconstructions and the raw signals validate that this assumption holds here.

The degree of sparsity of α is controlled by the regularization parameter λ . Too large of a λ will result in very sparse networks, potentially losing important patterns, while a small λ will introduce more irrelevant features into the results. In addition to λ , the number of dictionaries can also impact the sparsity of α and the decomposition accuracy. As no gold standard exists for parameter selection, we proposed three criteria, the reconstruction error, the density of α matrix and the mutual information with the reference atlas, to evaluate the performance of DLSC and then carried out a grid search on the optimized parameters. $\lambda=1.5$ was selected and different dictionary sizes were tested fixing the λ . By visual check, the parameter combinations resulted in meaningful brain delineations.

4 Transcriptomic Anatomy

Based on the method proposed, gene expression energy signals of a whole mouse brain were decomposed into multiple components. After mapping the coefficient matrix back to 3D volume space, different spatial patterns were observed for different dictionary atoms. A visual inspection showed that voxels with high coefficients smoothly distributed in 3D space and forms tight clusters. The formed clusters correspond to various canonical anatomical regions spanning the entire brain - ranging from isocortex, olfactory area, striatum to thalamus, midbrain and cerebellum etc., conceptually validating sparse coding as a useful data-driven approach to extract region-specific gene signatures from transcriptome and obtain meaningful brain divisions (Figure 2. 2). This clustering patterning agrees with the brain's organizational principle that transcriptome similarities are strongest between spatial neighbors, both between cortical areas and between cortical layers (Bernard et al. 2012), which has been seen in a range of methods including unsupervised hierarchical clustering, analysis of variance (ANOVA) and etc.

Interestingly, multiple white fiber pathways, as well as the ventricular system, were also extracted by DLSC.

Different numbers of dictionaries (100, 200, 400, 600, 800, and 1000) were tested for matrix decomposition (Figure 2. 3, Figure 2. 4). Intuitively, larger numbers of dictionaries would be expected to result in finer parcellation of the mouse brain. It should be noted that when the dictionary number is set to 200 or below, the gene expression based laminar structures are not obvious (Figure 2. 3). With a growing number of dictionaries, the coarsely parcellated subcortical areas were further parcellated into subregions and more details of layered and laminar architectures of neocortex were observed (Figure 2. 3).

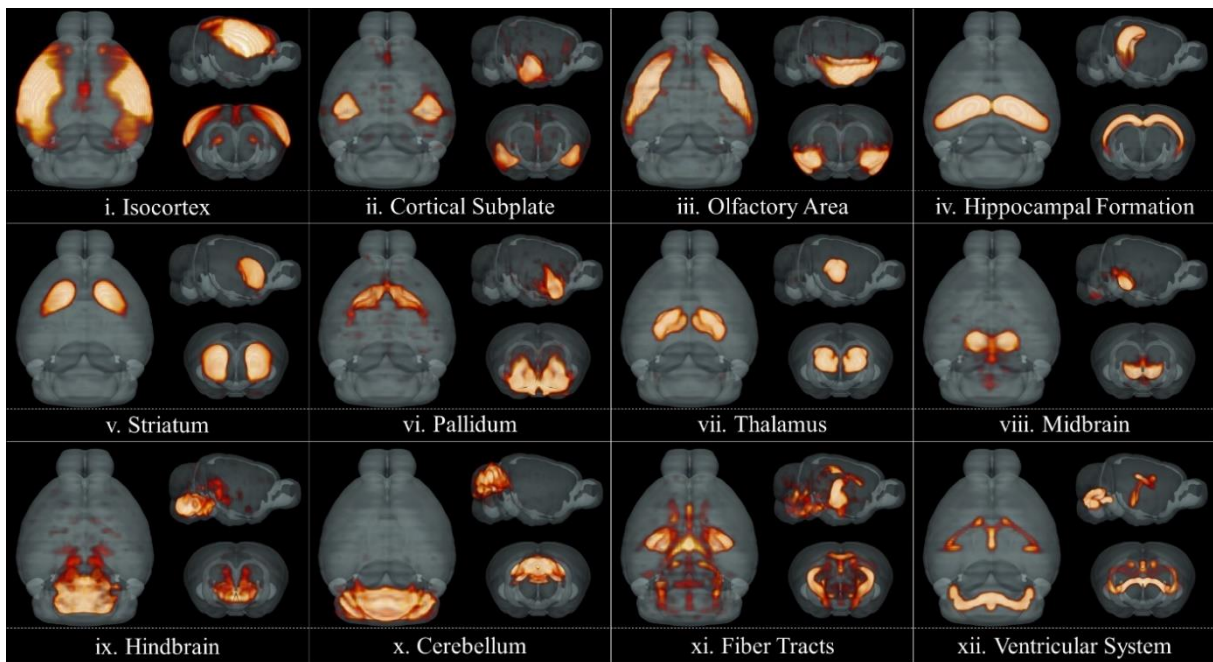


Figure 2. 2. Visualization of selected 3D spatial maps of the coefficient matrix. Results were obtained using 200 dictionaries. 12 dictionaries corresponding to 12 major canonical regions were selected.

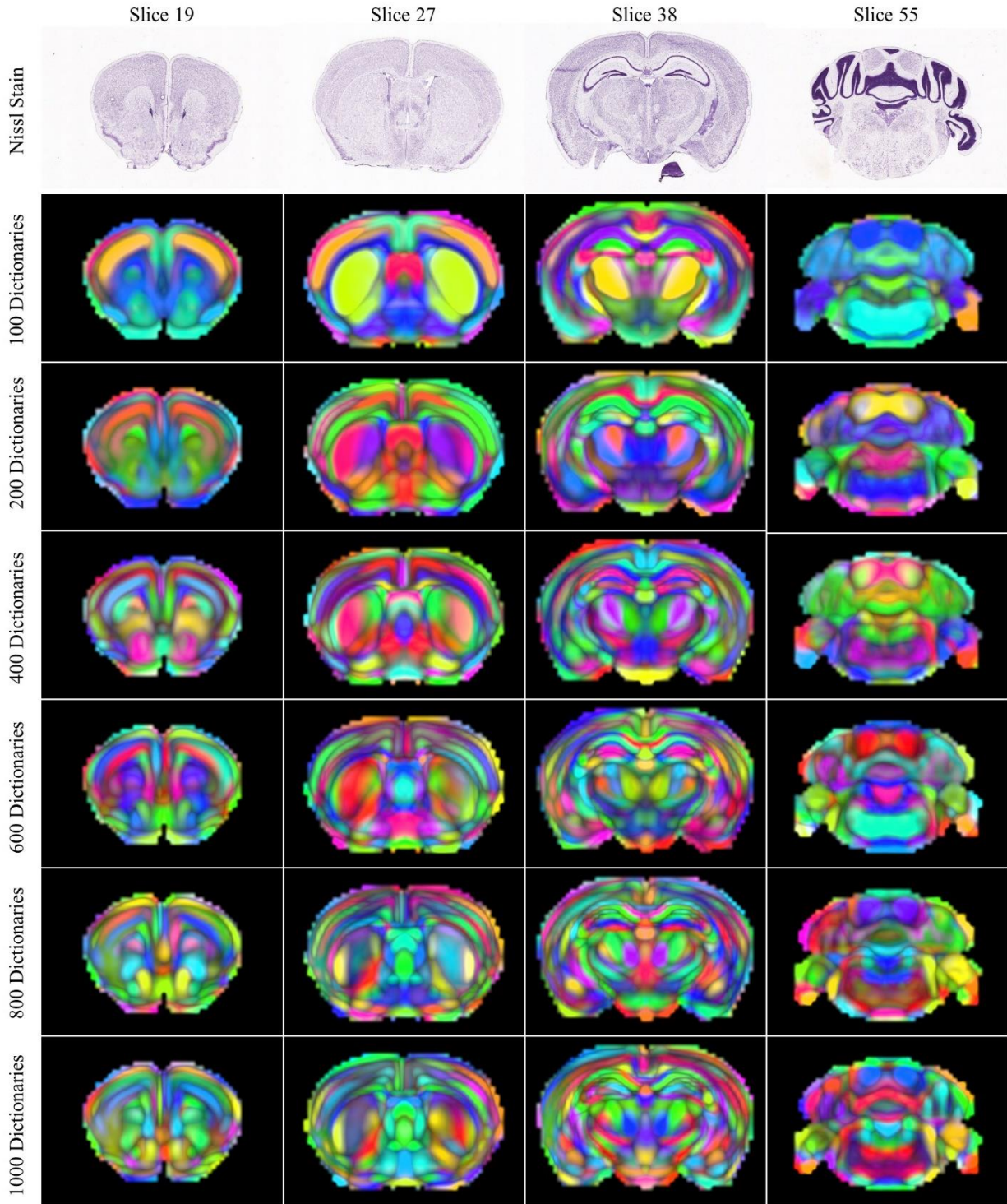


Figure 2. 3. A comparison of transcriptome anatomy obtained for different dictionary numbers. A random color was chosen for each dictionary and the intensity was scaled by dictionary coefficients. 4 coronal slices were selected for visualization. The corresponding Nissl stain image was shown in the first row. From top to bottom, finer delineations of the mouse brain were shown.

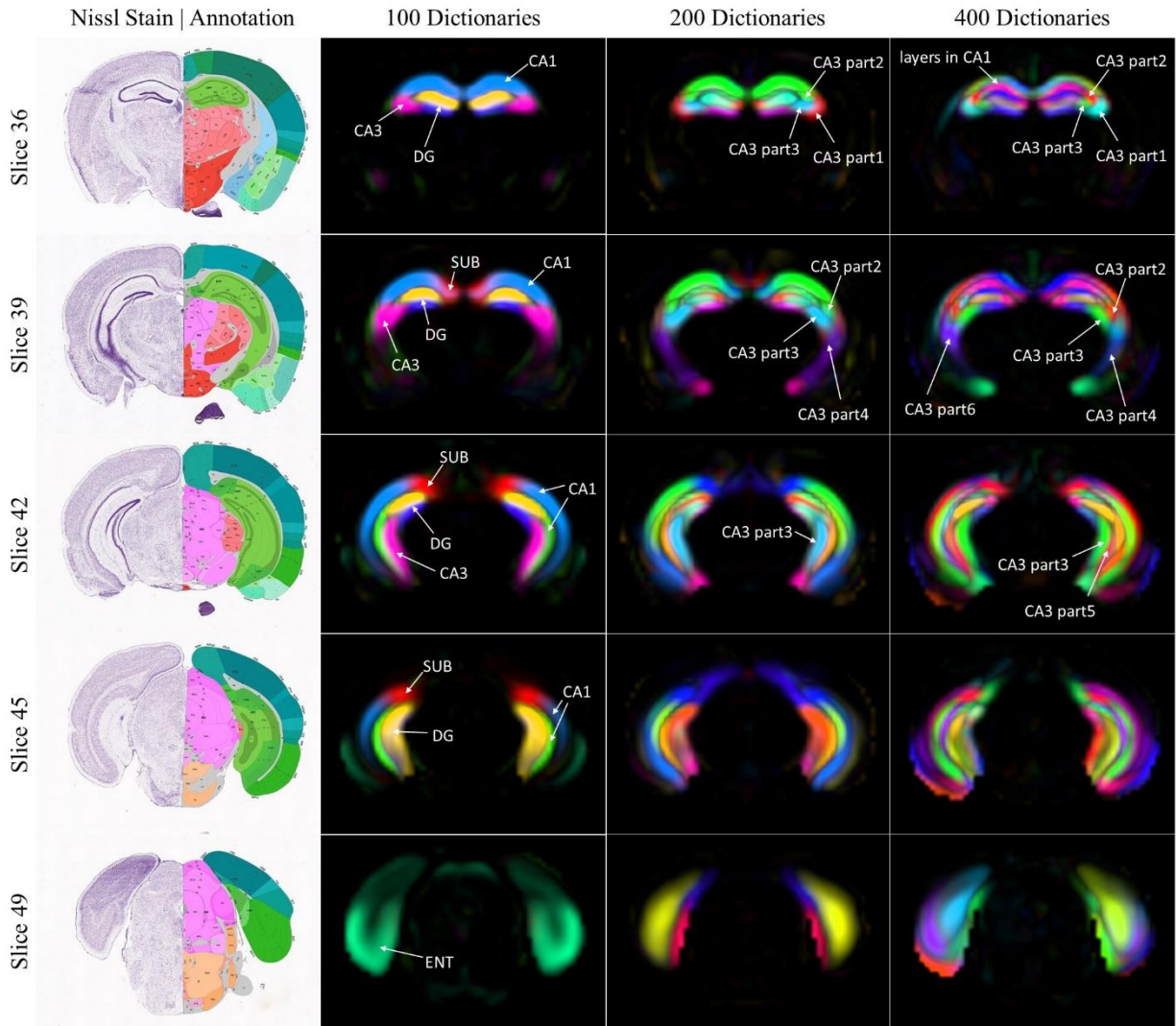


Figure 2. 4. Hippocampal formation related dictionaries obtained by using 100, 200 and 400 dictionaries. A random color was chosen for each dictionary and the intensity is scaled by dictionary coefficients. Here 5 coronal planes of sections were selected for visualization and the corresponding Nissl stained image as well as anatomical annotation downloaded from ARA were shown on the left.

4.1 Hippocampal formation

To show as an example, we analyzed the hippocampus-related components. The components obtained from 100, 200 and 400 dictionaries were identified by overlapping measurement with ARA (Figure 2. 4). With 100 dictionaries, the proposed method successfully separated major anatomical structures in hippocampus including field CA1, field CA3, dentate

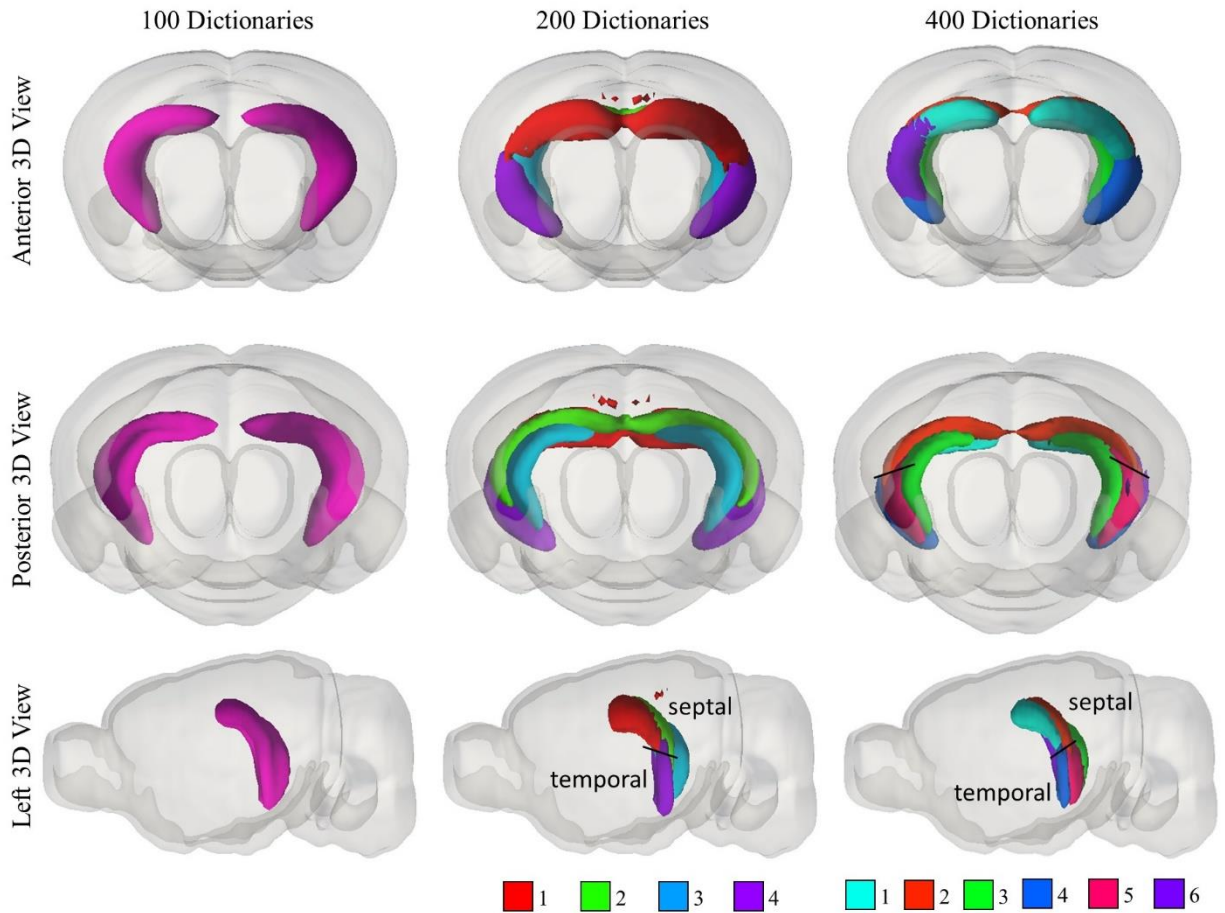


Figure 2. 5. 3D renderings of spatial pattern of field CA3 related components obtained using different dictionary numbers. The color scheme of each region is listed at the bottom of subfigure and is the same as Figure 2. 4.

gyrus (DG), subiculum (SUB), and entorhinal area (ENT). With more dictionaries, layered structures of these regions gradually emerge. Specifically, as shown in Figure 2. 5, field CA3 was identified as a complete piece when 100 dictionaries were used. When 200 dictionaries were used, field CA3 was decomposed into 4 sub-components including 2 frontal components and 2 posterior components. When 400 dictionaries were used, 6 finer components related to field CA 3 were identified. For the lateral components, field CA3 was completely separated into septal and temporal parts as highlighted in Figure 2. 5. These components might be associated the

various pyramidal neurons that send and receive signals with other parts of the hippocampus and reflect the distribution of intrahippocampal projections (Ishizuka et al. 1990). A non-symmetric component was shown on the right hemisphere only. Having examined the ISH images, the unilateral component was a result of artefacts during image acquisition and preprocessing.

4.2 Fiber tracts and ventricular system

One of the most interesting findings is that the DLSC can extract expression patterns that correspond to fiber tracts and ventricular system. One example is dictionary 17 that corresponds to the white matter pathways. Specifically, the fiber tracts observed here are mainly corpus callosum (Figure 2. 6a-c white arrows), internal capsule (Figure 2. 6b yellow arrows) and fimbria (Figure 2. 6c blue arrows). Even though the signals at other regions are relatively strong, the distinctly high expressions at corpus callosum and internal capsules agree well with the reference atlas for fiber tracts. Many transcripts that showed enhanced signals at these regions are also markers for oligodendrocyte (Cahoy et al., 2004). The two presented transcripts *Mbp*, *Cdn11* encode myelin basic proteins (Figure 2. 6g-i, j-l). Other transcripts that heavily use the dictionary for representation such as *Plp1* and *Cnp* are also related to myelination, which is a featured function for oligodendrocyte. The increased myelin level is presumed the reason for the enhanced signals in white matter in comparison with other regions because it is known that oligodendrocytes produces myelin membranes in the white matter. Another example is Dictionary 71, which features enhanced expression patterns at lateral ventricle (Figure 2. 6A-C white arrows), third (Figure 2. 6B-C yellow arrows) and fourth ventricles (Figure 2. 6C blue arrows). As seen in Figure 2. 6, both transcripts *Cd63* and *Slc38a3* showed prominent signals at these regions (Figure 2. 6I-P), corroborating the spatial map of dictionary 71. Notably, both transcripts are markers for astrocyte (Cahoy et al., 2004; Ng et al., 2009). The significantly high

expressions at the ventricles is reminiscent of that the subventricular zone is rich with astrocytes ((Quinones-Hinojosa and Chaichana, 2007)). The abundance of astrocytes is likely the reason for the enriched and restricted expression at ventricular regions. The above two examples demonstrate that DLSC can extract expression patterns that are restricted to white matter and ventricular systems possibly via cell-type markers that are enriched at these regions.

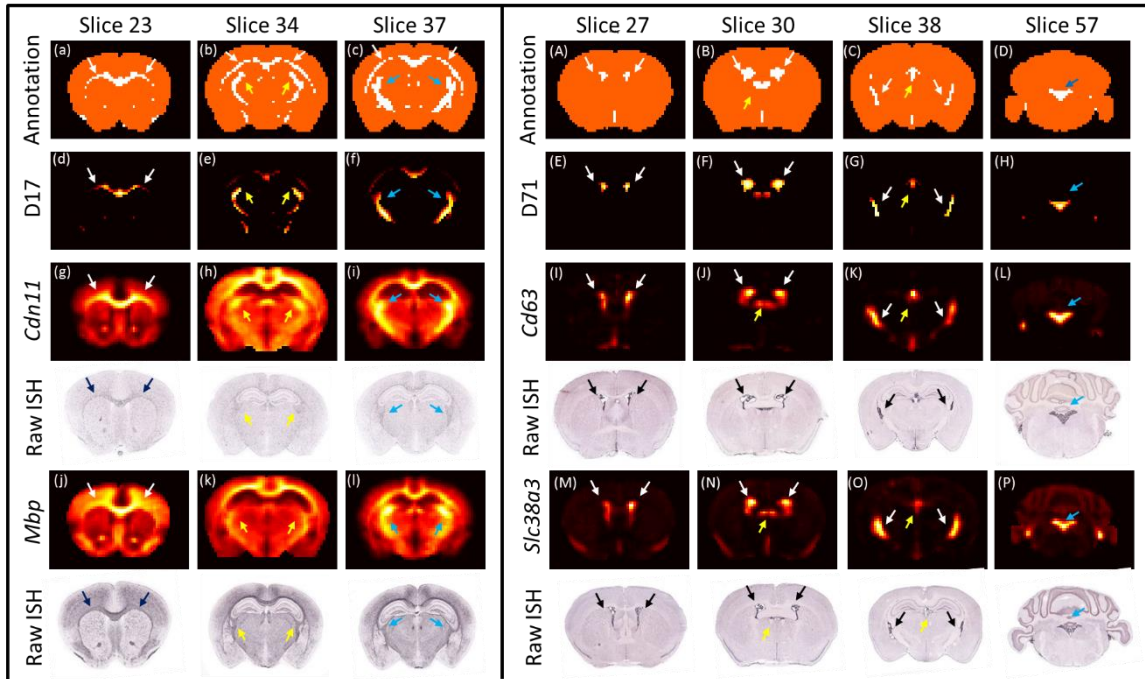


Figure 2. 6. Slice-based views of the spatial distribution of components that correspond to the fiber tracts (dictionary 17) and ventricular system (dictionary 71). Each column is a different slice. First row are the reference atlases for fiber tracts (left) and ventricular system (right). Second row are the spatial distribution of the components. Third and fifth rows are the normalized energy expression of selected genes. Fourth and sixth rows are the raw ISH data for the selected genes. Gene acronyms are on the left of ISH images.

5 Comparison with Principal and Independent Component Analysis

To benchmark with the alternative matrix factorization methods, we performed principal component analysis (PCA) and independent component analysis (ICA) on the same gene expression matrix. For PCA, data was first centered and then whitened. Singular value

decomposition algorithm was used as the solver. To visualize the spatial distributions, we projected each individual mode back to the brain space (Figure 2. 7). The top four modes account for over 80% of variance. The first two modes have a very broad distribution across the brain. The third mode is also broadly distributed with enhanced specificity for the cerebellum, and the

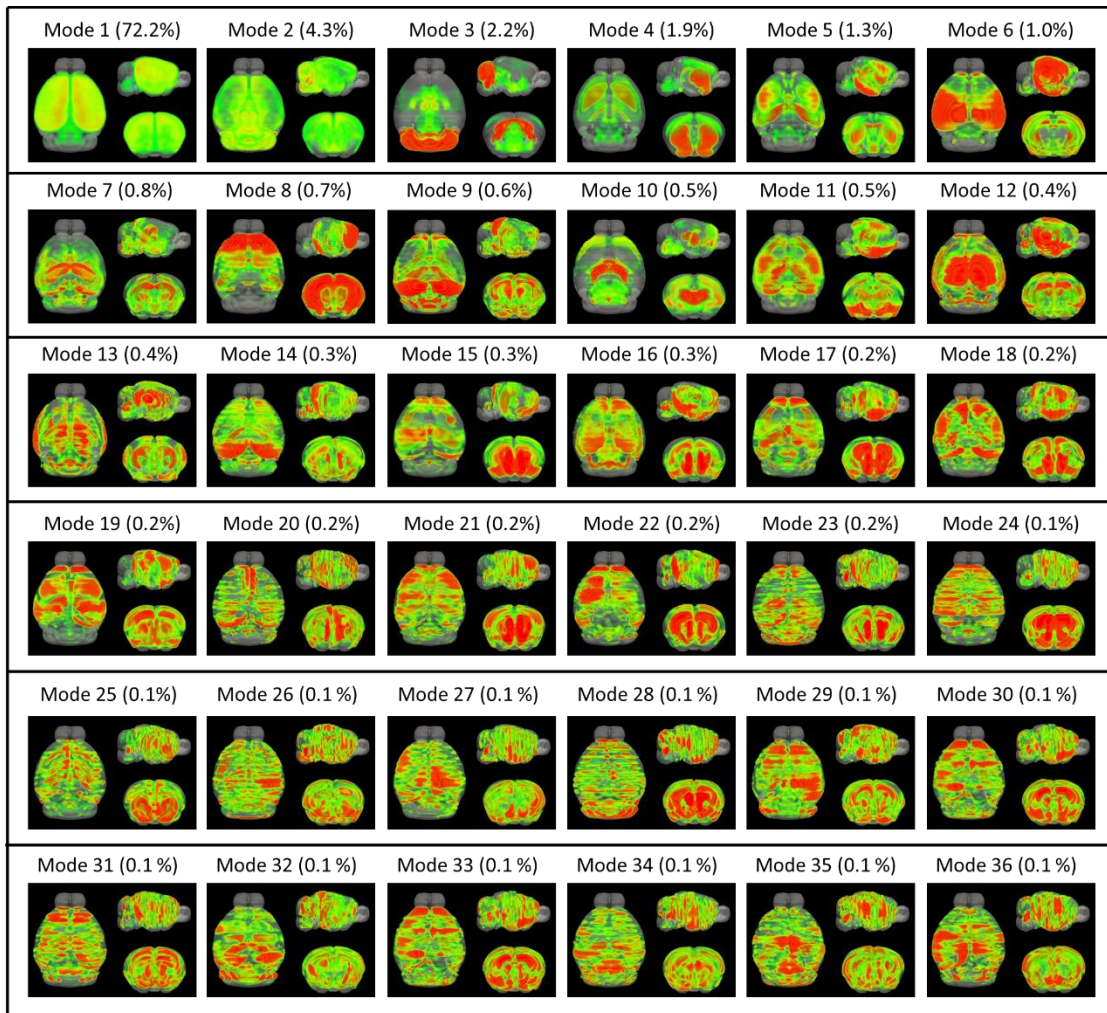


Figure 2. 7. Visualizations of top 36 modes obtained using PCA. The values in the parentheses are the percentage of variance explained by the mode.

fourth mode is particularly prominent in striatum and CA3. For modes that account for less variance, the spatial distributions span the entire brain and the agreement to the anatomy is less

obvious. In summary, the first few modes contain spatial structures in rough concordance with classical anatomy. However, it is also apparent that finer structure cannot be revealed by PCA.

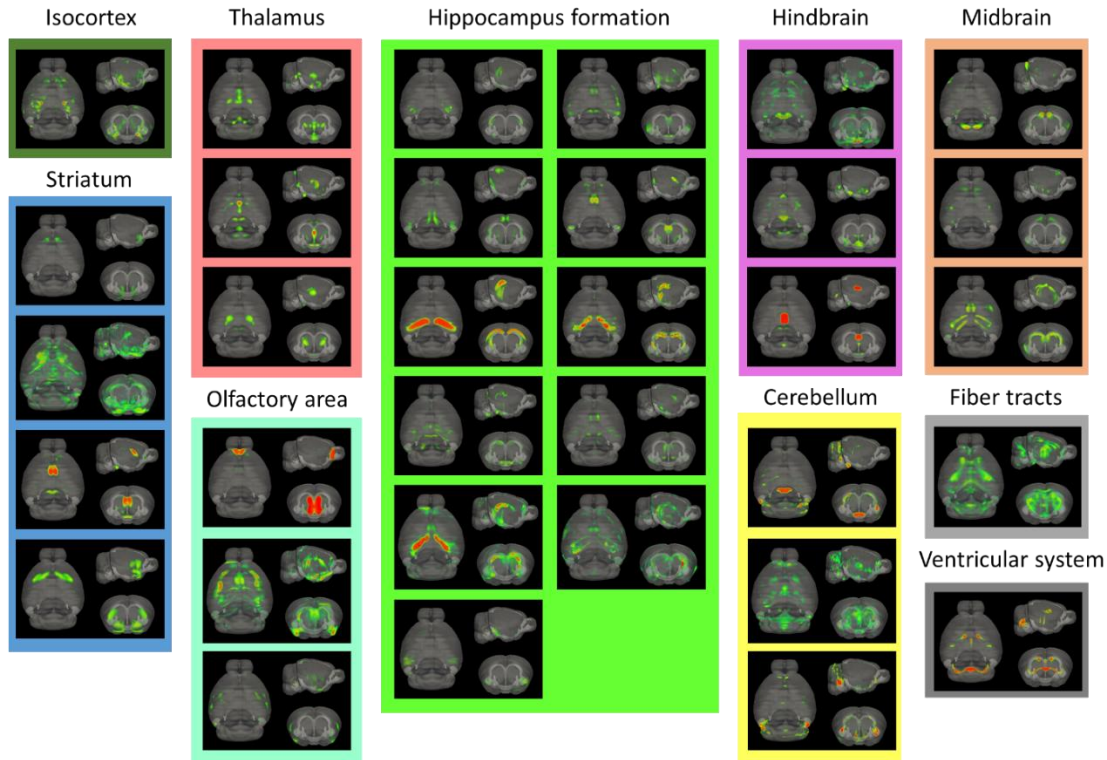


Figure 2. 8. Visualization of 33 components corresponding to anatomical regions using ICA. The components were assigned to 10 brain regions by calculating the overlaps between the reference atlas and the spatial maps. The brain regions were color-coded. Results were obtained using 100 components.

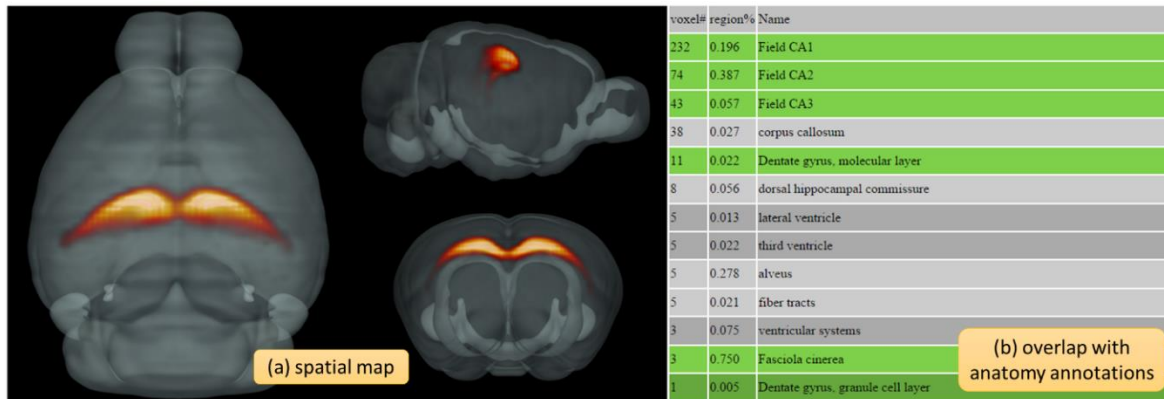
A comparison with the results from the application of ICA also confirmed that DLSC is a better fit to the context of deriving the transcriptome organizations. The basic goal of ICA is to determine a transformation so that the transformed components are statistically as independent from each other as possible. The goal is realized by finding a direction that maximizes the negentropy (Comon, 1994). Therefore, ICA requires a strong assumption that the components are independent. In comparison, DLSC minimizes the total loss of reconstruction error and the ℓ_1 penalty of the coefficient matrix, without imposing assumptions on the relationship between

components. To ensure a fair comparison, 100 components were generated using ICA. The algorithm used in ICA was FastICA (Hyvärinen, 1999). Spatial maps were obtained by projecting the coefficient matrix to the brain space and then classified into 10 major brain regions (Figure 2. 8). The biggest difference observed between DLSC and ICA is that DLSC is able to produce components that cover most parts of major anatomical brain regions including thalamus, striatum, midbrain, olfactory area etc. (Figure 2. 2). In comparison, almost all components generated by ICA were in concordance with only a small portion of the major brain regions. Such example components were seen in thalamus, hindbrain, midbrain, cerebellum etc. A few exceptions were ventricular system, field CA3, field CA1 and dentate gyrus. The lack of components that correspond to the complete brain regions is probably a result of unsupported assumptions. ICA assumes the components to be independent and solves the matrix factorization by maximizing the statistical independence of the estimated components. However, it is likely that two genes are regulated by the same transcription factors and thereby their expressions are dependent. In comparison, the assumption of DLSC is the sparsity of the coefficient matrix and supported by that 70% genes are expressed in a limited number of cells. The advantage of sparse coding over ICA has also been demonstrated in other data modality such as functional magnetic resonance imaging (Lv et al., 2015).

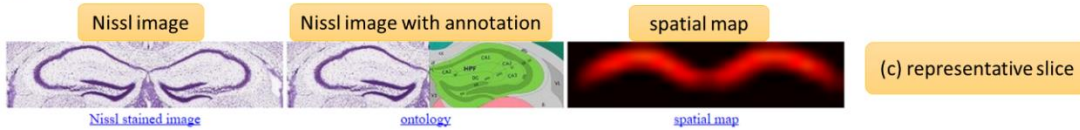
6 Online Informatics Portal

To allow other researchers to explore the comprehensive transcriptome architecture identified by the proposed framework, all the information is organized into web pages and can be easily accessed at: http://mbm.cs.uga.edu/mouse/transcriptome_architecture. To facilitate the exploration of components, the portal provides two main ways to view the transcriptome architecture - by dictionary number and by anatomical brain regions. Altogether, there are 6

levels of brain delineations by sparse coding with the dictionary number varying from 100, 200 to 1000 and 13 canonical brain divisions. In each component, a comprehensive webpage consisted of both the anatomical and genomic information of the component has been generated (Figure 2. 9). As to the anatomical information, in addition to the selected Nissl stained image and its ontology that afford the context for interpretation, a 3D spatial map corresponding to its coefficient matrix (Figure 2. 9 (a)) is visualized. To quantify the composition of the obtained component, the percentage of overlapping volume between the component and ARA is calculated and the top 20 regions along with the number of voxels occupied by the component and the overlap percentage were tabulated (Figure 2. 9 (b)). Each of the obtained components can be downloaded as a zip file for further investigation. With respect to the genetic information, the regionally enriched and restricted transcripts were retrieved and the related ISH raw data are shown alongside, offering a direct link to the original data in the database. For the convenience of comparison, we only visualized the slice with the highest expressions of the component (Figure 2. 9(c)). The differentially expressed transcripts were not determined from the absolute expression levels but ranked by the average expression energy within each component weighted by the dictionary coefficients. Transcripts with the top two highest (lowest) expression energies in a specific component were taken as a relatively expressed (non-expressed) gene in this component. In addition to the differentially expressed transcripts, we also include the transcripts that heavily used the dictionary for signal reconstructions (Figure 2. 7 (e)). To evaluate the importance of a dictionary for a particular transcript, we first calculated the error changes in reconstructions of each transcript after removing this particular dictionary and then weighted the

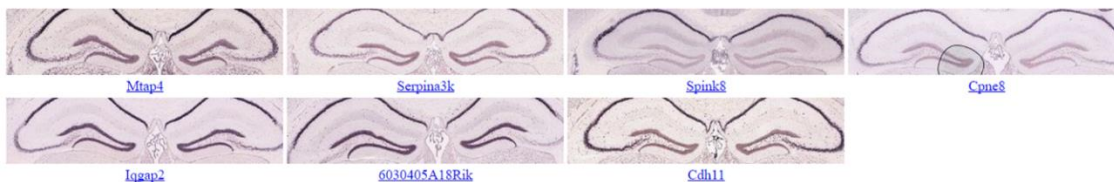


anatomy

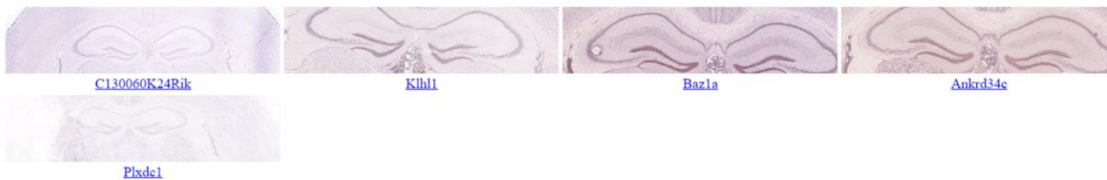


relatively highly expressed genes

(d) Transcripts ranked by average expressions



relatively none expressed genes



Genes that use this dictionary

(e) Transcripts ranked by the importance of dictionary for reconstruction

*value in the parentheses show the weight of the dictionary

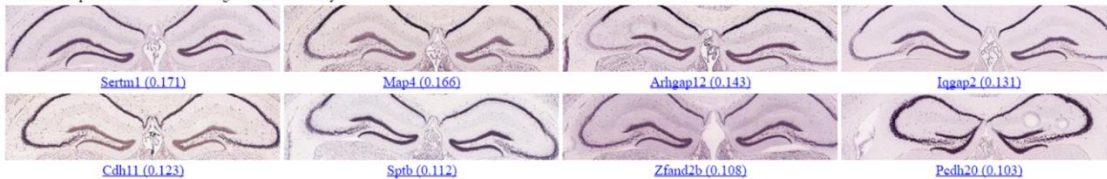


Figure 2. 9. Illustration of anatomic and genetic information of a dictionary component on the informatics portal. (a) 3D spatial map of the component. (b) The 20 regions that showed the highest overlaps with the spatial distribution of the component. (c) Nissl stained image, reference atlas and spatial distribution of the coronal slice that showed major expressions. (d) ISH raw images of transcripts that showed high and low expressions regionally. (e) ISH raw images of transcripts that use the dictionary for signal reconstructions.

changes by the ℓ_2 norm of the raw signals considering that transcripts with higher signals overall tend to use more dictionaries for representation. The obtained scores were the indicator of the importance of this dictionary for each transcript. Accompanying the above-mentioned two ways to examine the components, a slice-by-slice view (Figure 2. 3) is also enabled for more detailed comparisons on each slice between the components obtained from different dictionary numbers.

7 Discussion and Conclusion

We have presented a data-driven DLSC framework that delineates the entire mouse brain into multiple components based on the whole-genome transcriptome. Visualization of the components reveals meaningful patterns spanning the entire brain. When the input dictionary number is low, most of the obtained components correspond to the classical anatomical regions while other components, intriguingly, accord well with the white matter pathways and ventricular systems. At higher dictionary number, a deeper and more detailed parcellation is seen, reflecting a more complex nature of brain organizational principle. However, one caveat is that a higher dictionary number does not always result in a more intricate parcellation. A main cause is the artifacts associated with tissue handling, image acquisition and registration integrity. Although DLSC has proved a robust analytical method and can de-noise images (Elad & Aharon, 2006), some of the obtained components were clearly identified as products of artifacts by visual inspection. The other reason is concerning to the limited resolution of current ISH image mapping. The voxel size is 200 microns on a side and exceedingly large to discern cells of different types and classes. Nonetheless, we have shown that the parcellation of fiber tracts and the ventricular systems is probably via markers for oligodendrocytes and astrocytes that are enriched in these regions.

As mentioned earlier, the two key assumptions of the DLSC framework are 1) each gene is expressed in a limited number of cells in the brain. 2) The integration of two gene expression follows a linear relationship. The second assumption is necessary for all matrix factorization methods. The comparative analysis with the results generated from ICA and PCA showed that DLSC was able to produce localized components that correspond to the major brain regions. In contrast, the modes obtained from PCA usually span multiple brain regions and finer structures cannot be directly resolved. Most of the components obtained from ICA either distribute across multiple brain regions or correspond to a small portion of major brain regions. The explanation to these components is attributed to the unsupported assumption that gene expressions were independent from one another. Interestingly, the ventricular system is also revealed by ICA.

In addition to the proposed framework, we have contributed a comprehensive transcriptome architecture of the adult mouse brain. It is comprehensive on two levels. First, the input of the framework is the whole-genome ISH data of the entire mouse brain. Second, the components generated by the framework are brain-wide, covering not only the canonical anatomical areas but also white matter pathways and ventricular systems. Further work will include a detailed analysis of the relationship between the mouse brain connectomes and the revealed white matter pathways, as well as the involved functioning genes. Another focus will be a comprehensive characterization of co-expressed gene networks of the whole mouse brain. A deeper knowledge of these networks is an essential step toward understanding protein interactions, regulatory pathways and, ultimately, brain organization, structures and functions. Additionally, the genetic architecture, especially when it is coupled with systematic profiling in various stages of brain development and aging processes (Jiang et al., 2001; Mody et al., 2001),

can serve as an informative and complementary approach to the on-going, large-scale brain mapping and decoding efforts (H. Chen et al., 2015; Tsien et al., 2013).

CHAPTER 3

DISCOVER MOUSE GENE CO-EXPRESSION LANDSCAPES USING DICTIONARY LEARNING AND SPARSE CODING²

² Yujie Li, Hanbo Chen, Xi Jiang, Xiang Li, Jinglei Lv, Meng Li, Hanchuan Peng, Joe Z. Tsien, Tianming Liu, 2017, Discover Mouse Gene Co-expression Landscapes Using Dictionary Learning and Sparse Coding, *Brain Structure and Function*, 222(9), 4253-4270.
Reprinted here with permission of the publisher.

1 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 (ISH) 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 “co-expressed”. 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.

2 Background and Motivation

Gene coexpression patterns carry rich information about enormously complex cellular processes (Brown, Johnson, & Sidow, 2007; Eisen, Spellman, Brown, & Botstein, 1999; Grange et al., 2014; Lee, Hsu, Sajdak, Qin, & Pavlidis, 2004; Oldham, Horvath, & Geschwind, 2006; Peng et al., 2007; Stuart et al., 2003). Previous studies have shown that genes displaying similar expression profiles are very likely to participate in the same biological processes (Tavazoie et al., 1999). Gene coexpression networks (GCNs), offering an integrated and effective representation of gene interactions, has shown advantages in deciphering the biological and genetic mechanisms across species and during evolution. 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 a result of dysfunction of complex network of molecular interactions (Bando et al., 2013; Carter, Hofree, & Ideker, 2013; Gaiteri, Ding, French, Tseng, & Sibille, 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, Langfelder, & Horvath, 2012). Given the high dimensions of genetic data and the urgent need in revealing the differences or the consensus between subjects or species, one common theme of all these methods is dimension reduction. Instead of analyzing the interactions across over tens of thousands of genes, the grouping of genes by their co-expression 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. 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. 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 “co-expressed”. 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, Ed S. 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, Ed S. 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.

3 Slice-Wide GCN Construction and Validation

The computational pipeline consists of two parts: the slice-based GCN construction and validation (Figure 3. 1a-d) and global GCN construction and analysis (Figure 3. 1e). We discuss

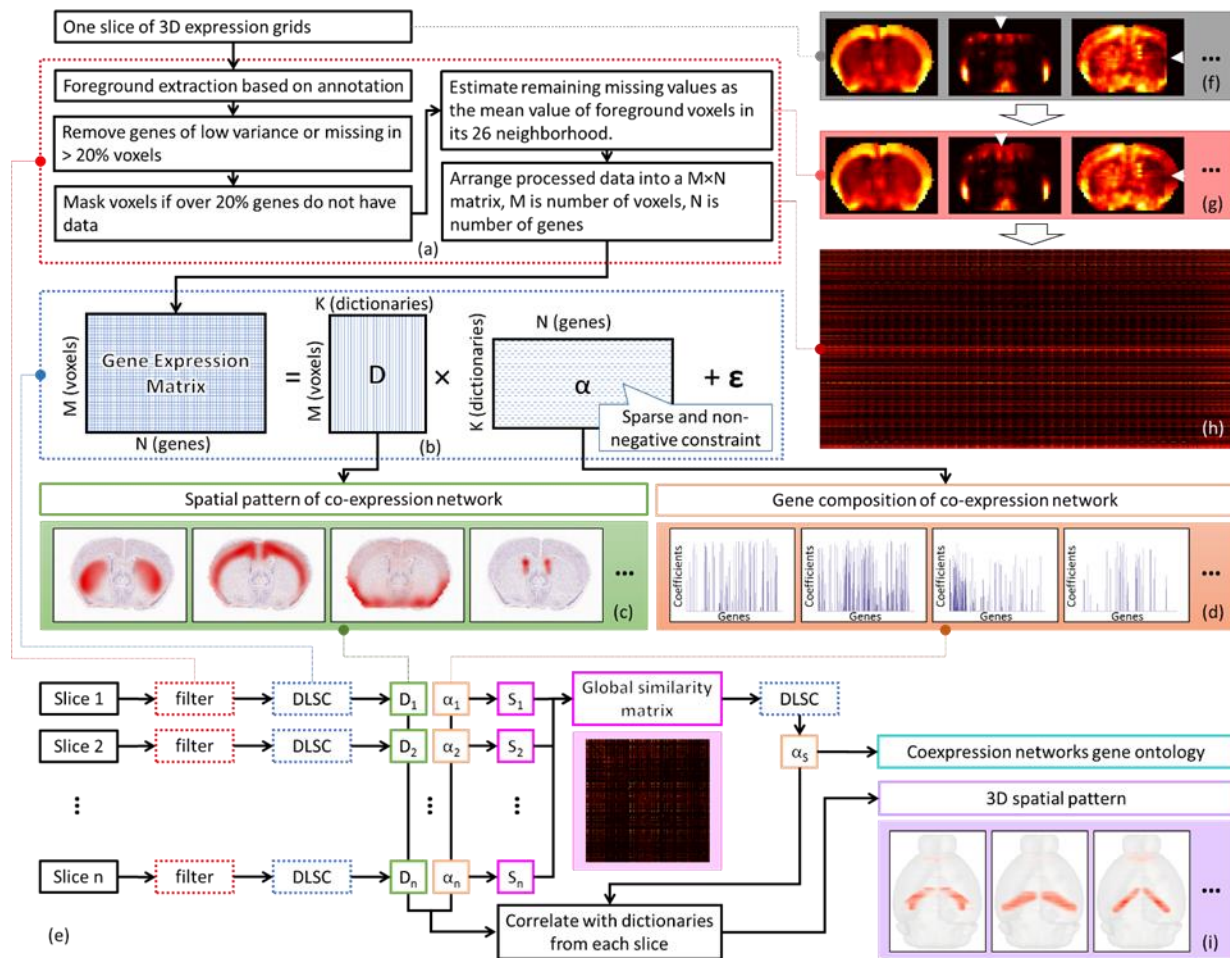


Figure 3. 1. Computational pipeline for constructing slice-wide GCNs (a)-(d) and brain-wide GCNs (e). (a) Raw ISH data preprocessing step that removes unreliable genes and voxels and estimates the remaining missing data. (b) Dictionary learning and sparse coding of ISH matrix with sparse and non-negative constraints on α matrix. D is the dictionary matrix and α is the coefficient matrix. ϵ is the reconstruction error. (c) Visualization of spatial distributions of slice-based GCNs (d) Visualizations of co-expression 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 grids were arranged in an M by N matrix. (i) Visualizations of 3D spatial patterns of global GCNs.

the first part here. The major obstacle to a global analysis of ISH data on all coronal slices is the number of missing data observed on each slice. 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 was 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. 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.

3.1 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 3. 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 the artefacts during ISH including missing slices, broken tissue and image processing steps such as 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

suggest that the estimated data is reasonable. After preprocessing, the cleaned expression energies were organized into a matrix and sent to DLSC (Figure 3. 1b). In DLSC (section 3.2), 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 (Figure 3. 1c-d) and were further analyzed and validated against the raw data and existing methods.

3.2 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 rows correspond to M foreground voxels and columns correspond to N genes (Figure 3. 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 equation 3.2 (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 \mathbf{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 $\mathbf{X}=[x_1, \dots, x_N]$ in $\mathbb{R}^{M \times N}$, learning a fixed number of dictionaries for sparse representation of \mathbf{X} can be accomplished by solving the following optimization problem:

$$\langle \mathbf{D}, \boldsymbol{\alpha} \rangle = \operatorname{argmin} \frac{1}{2} \|\mathbf{X} - \mathbf{D} \times \boldsymbol{\alpha}\|_2^2 \text{ s.t. } \|\boldsymbol{\alpha}\|_1 \leq \lambda \quad (3.1)$$

where $\mathbf{D} \in \mathbb{R}^{N \times K}$ 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, $\|\cdot\|_2$ is the summation of ℓ_2 norm of each column and $\|\cdot\|_1$ is the summation of ℓ_1 norm of each column. $\|\mathbf{X} - \mathbf{D} \times \boldsymbol{\alpha}\|_2^2$ 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 3.1 should be minimized (Mairal et al., 2010).

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 3.2, can be handled by homotopy method presented in Efron et al (Efron, Hastie, Johnstone, & Tibshirani, 2004).

$$\langle \mathbf{D}, \alpha \rangle = \operatorname{argmin} \sum_{i=1}^N \frac{1}{2} \|x_i - \mathbf{D} \times \alpha_i\|_2^2 \text{ s.t. } \|\alpha\|_1 \leq \lambda, \forall i, \alpha_i \geq 0 \quad (3.2)$$

The key assumption of enforcing the sparsity is that each gene is involved in a 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.

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 learned 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 \mathbf{D} back to the ARA space, we can visualize the spatial patterns of the coexpressed networks. Combining information from both \mathbf{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 the network, these two terms will be used interchangeably.

3.3 Parameter Selection

The choice of the number of dictionaries and the regularization parameter λ are crucial for effective sparse representation. As there exists no gold 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 criterion is the reconstruction error. It is defined as the square difference between the original signal matrix and the reconstruction from sparse representation (equation 3.3). A high reconstruction error indicates a less accurate representation.

$$error_y = \frac{1}{2} \|\mathbf{X} - \mathbf{D}\alpha\|_F^2 \quad (3.3)$$

The second evaluation metric is the average uncertainty coefficient (AUC) between the obtained dictionaries and the reference atlas. The uncertainty coefficient, defined in equation 3.5, 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 that the two spatial distributions are independent whereas values close to one suggesting that 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:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (3.4)$$

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:

$$U(X, Y) = \frac{2 * I(X, Y)}{H(X) + H(Y)} \quad (3.5)$$

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 for comparison.

Another important measurement to examine the DLSC performance is the degree of density measured by the percentage of non-zero-valued elements in the coefficient matrix. As we are search 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 section 2.2.2, the density is

Table 3. 1. Reconstruction errors 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-dictionary ratio \ λ	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 3. 2. AUCs between the obtained dictionaries and the annotation map on slice 27 using different λ and gene-dictionary ratios.

gene-dictionary ratio \ λ	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

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.

Table 3. 3. The percentage of non-zero entries in the coefficient matrix obtained from DLSC on slice 27 using different λ and gene-dictionary ratios.

gene-dictionary ratio \ λ	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

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 were 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. 55 combinations of λ and gene-dictionary ratios were considered with 5 choices of λ and 11 different gene-dictionary ratios (Table 3. 1, Table 3. 2, Table 3. 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 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 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.

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

WGCNA was applied on the same dataset to validate findings generated by DLSC. WGCNA (Langfelder & 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 β , 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: $\beta = 18$, $\text{deepSplit} = 4$ (highest) and $\text{minModuleSize} = 15$. Also, we changed the number of GCNs from the optimal 29 to 26 to ensure a fair comparison.

Then the number of shared genes were counted for groups identified by both methods. Besides quantification, another intuitive way to compare the two methods is by comparing the obtained spatial maps (Figure 3. 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 eigen-gene of that module.

4 Brain-Wide GCN Construction

4.1 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, was calculated from the coefficient matrix α with the coefficients as the feature of each gene. Let v_1 , v_2 be the coefficient vectors of gene1 and gene 2. The gene similarity measure is defined as the overlap rate OR, as below:

$$OR(v_1, v_2) = 2 \frac{|\min(v_1, v_2)|}{|v_1| + |v_2|} \quad (3.6)$$

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 over the original α matrix and also reflects 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 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 α .

4.2 Parameter selection

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

4.3 Fuse 3D spatial pattern of GCNs

As described in section 3.2, 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 section 4.1. 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.

4.4 Gene ontology analysis of brain-wide GCNs

Brain-wide GCN characterization was made based on common gene ontology (GO) 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.

5 Slice-Wide GCN Analysis

First, 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.

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.

5.1 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 vary 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. 14 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 co-expressed 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 eigen-genes of WGCNA modules (Figure 3. 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.

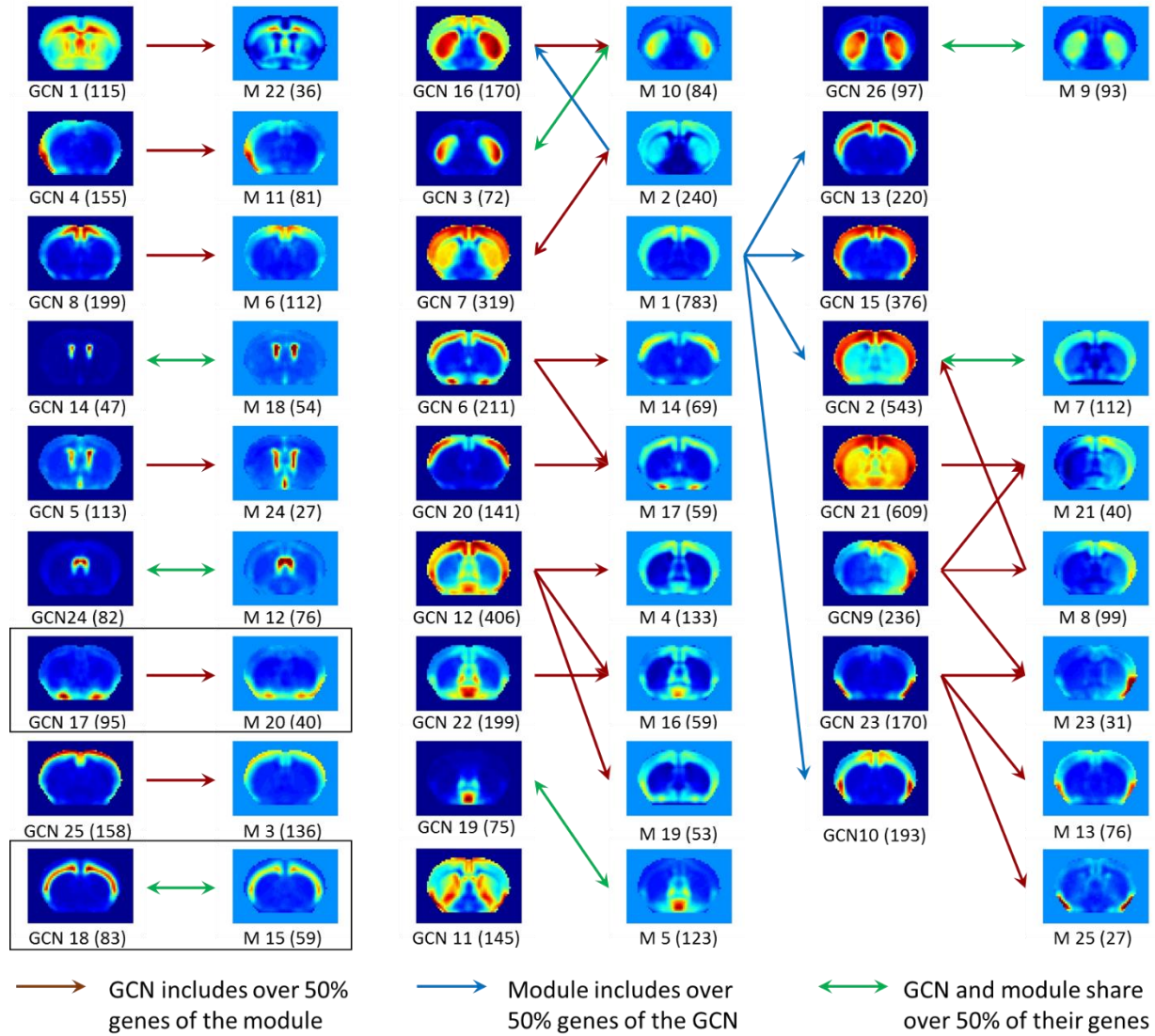


Figure 3. 2. Comparison between spatial distributions of GCNs and eigen-genes of WGCNA modules on slice 27. For clarity, the groups identified by WGCNA and DLSC are denoted as modules and GCNs respectively. The number of overlapping genes between a GCN and a module was counted. At the bottom of each image is the name of the networks/modules. ‘M’ represents a module generated by WGCNA and GCN represents a co-expression network generated by DLSC. The number in the parentheses are 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 Figure 3. 3 and Figure 3. 4. The background color for modules and GCNs are fixed to -0.05 and 0.

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 (Figure 3. 2). Relatedly, there are many modules small in size given that roughly half of the genes are assigned to module 1 and module 2.

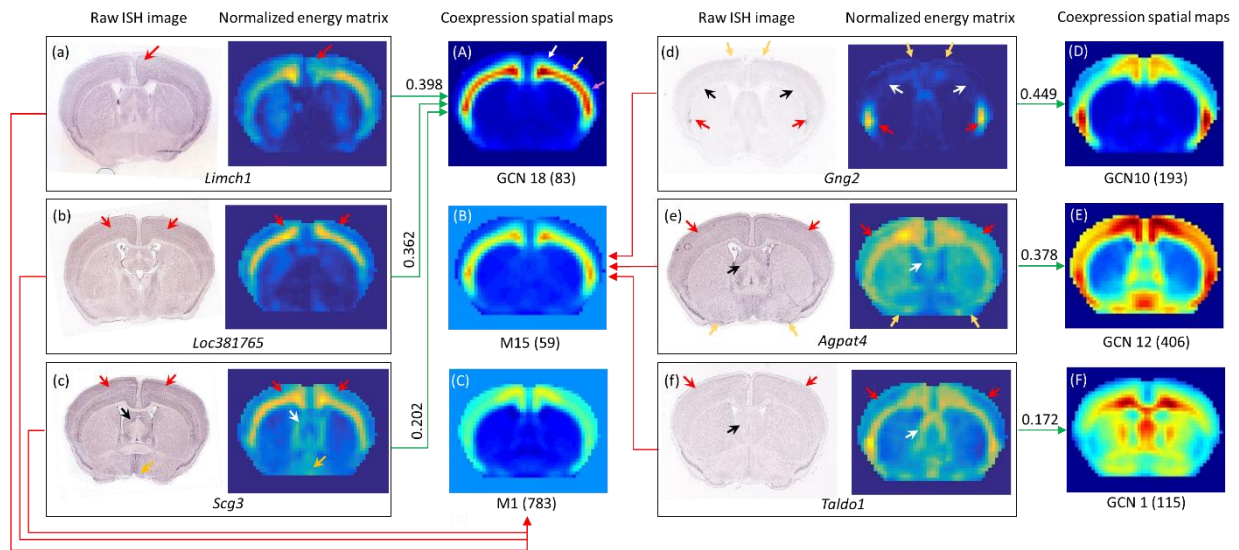


Figure 3. 3. Comparisons of genes in GCN18 and module 15 on slice 27. For each gene (a-f) we showed the raw ISH image together with the normalized energy matrix. On the left are three representative genes only found by DLSC. On the right are the three genes only found by WGCNA. (A-C, D-F) are the spatial distributions of selected GCNs and the eigen-genes 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.

There are multiple pairs that share over 50% of their genes (Figure 3. 2 green arrows). One example is GCN 18 and module 15, whose spatial patterns are quite similar (Figure 3. 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 module15. 31 genes were found only by GCN18 and 7 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 (Figure 3. 3A white arrow), motor area (Figure 3. 3A yellow arrow) and somatosensory area (Figure 3. 3A pink arrow). Three genes were selected for illustration from those only found by DLSC (Figure 3. 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 layer 5 and 6 and agree with the overall shape of the GCN 18 (Figure 3. 3 red arrows). However, *Scg3* displayed additional activations at medial preoptic area (Figure 3. 3c, f yellow arrow) and lateral septal nucleus (Figure 3. 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 but not assigned to Module 15, including the *Limch1*, *Loc381765* and *Scg3*, were assigned to module 1 (Figure 3. 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 decreases 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 (Figure 3. 3a-f red arrows) suggest the expression pattern a better consistency to GCN18.

We also looked at the genes found only by WGCNA (Figure 3. 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 example, *Gng2* was assigned to GCN10 (Figure 3. 3D) with a high weight of 0.449. The peak expressions at the endopiriform nucleus (Figure 3. 3d, red arrows) and relatively weaker expressions at the cortex regions (Figure 3. 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 shows enhanced signals at the medial preoptic nucleus (Figure 3. 3e black arrows), the piriform area (Figure 3. 3e yellow arrows), as well as all outer layers of cingulate areas (Figure 3. 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 (Figure 3. 3f red arrows) and the enhanced signals in septal nucleus (Figure 3. 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 eigen-gene 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 (Figure 3. 4D,E red arrows) and extends to piriform area with lower expressions (Figure 3. 4D-E white arrows). There were 95 genes in GCN17 and 40 genes in M20. Among

them, 35 genes were shared. 5 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 (Figure 3. 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 (Figure 3. 4c yellow arrows). In contrast, the high activations at the lateral preoptic area is more consistent with GCN17 (Figure 3. 4c red arrows). *Spp1* and *sgpp2* both showed broad activations in addition to the enhanced signals at the lateral preoptic area (Figure 3. 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 (Figure 3. 4f red arrows) and was more consistent to GCN 22 (Figure 3. 4F) than M20. *Dner* showed enhanced signals at piriform areas (Figure 3. 4g yellow arrows) and extended further to isocortex (Figure 3. 4g green arrow), thalamus (Figure 3. 4g black/white arrow) and hypothalamus (Figure 3. 4g red arrow) with lower expressions. The expression pattern was captured by both GCN 23 (Figure 3. 4r) and GCN 22 (Figure 3. 4s) with the degree of consistency of around 0.2. *Stc1* showed strong signals at piriform area (Figure 3. 4h yellow arrows), but not as strong at lateral preoptic area (Figure 3. 4h red arrows). This pattern was more consistent with GCN23 (Figure 3. 4G). A similar case was also seen in *Dmwd* (Figure 3. 4i). Finally, the expressions of *Slc25a3* almost spanned the entire slice, with enhanced signals at the cortex (Figure 3. 4j yellow arrows) and preoptic areas (Figure 3. 4j red arrows). The expression pattern was better captured by GCN21 (Figure 3. 4I).

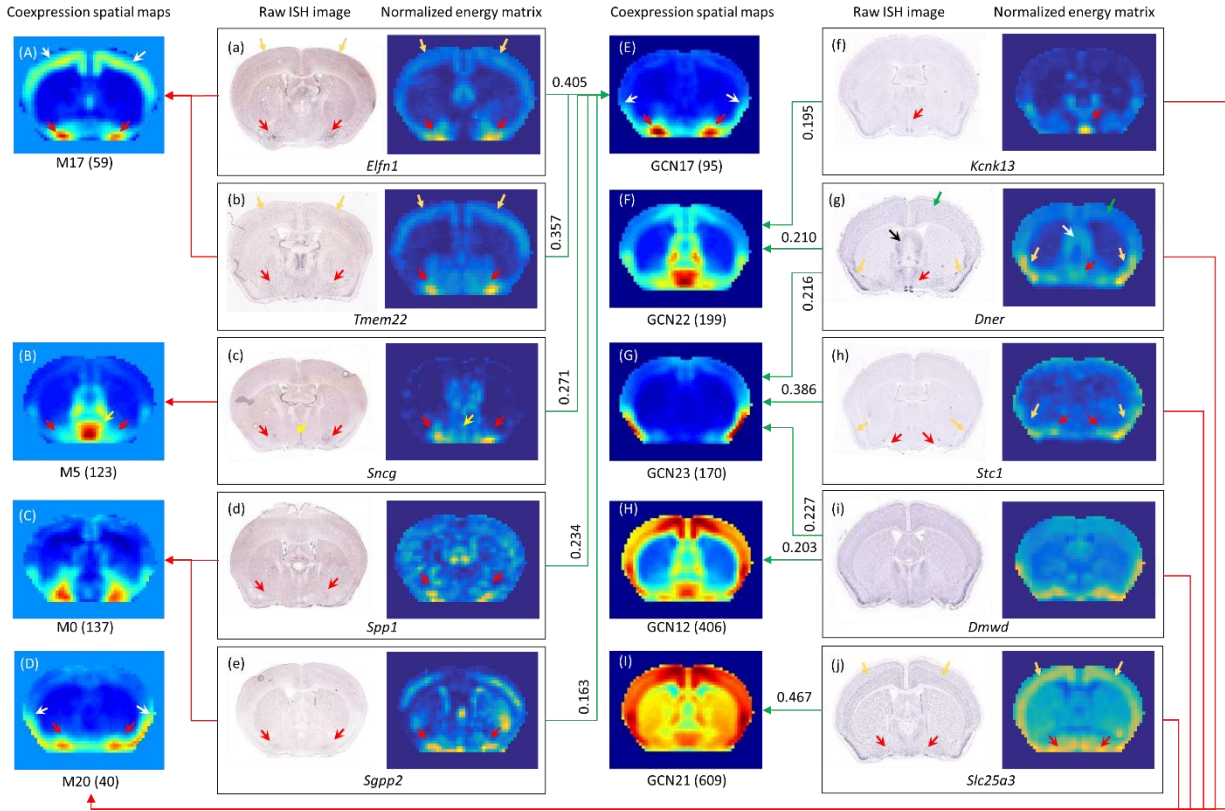


Figure 3. 4. Comparisons of genes in GCN17 and module 20 on slice 27. For each gene (a-j), the raw ISH image together with the normalized energy matrix were shown. (a-e) are five representative genes only found by DLSC. (f-j) are five genes only found by WGCNA. (A-D) are the spatial distributions of the eigen-genes 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.

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.

5.2 Comparative analysis with Principal Component Analysis

To compare with other matrix factorization method, we performed principal component analysis (PCA) 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 (Figure 3. 5a). The second mode is also broadly distributed with distinctly high amplitude in caudoputamen (Figure 3. 5b). In contrast, the third mode features an absence of caudoputamen and is prominent in the hypothalamus (Figure 3. 5c). Overall, PCA is able to extract correlated structures that correspond to the broad anatomical regions, such as caudoputamen (Figure 3. 5b) and isocortex (Figure 3. 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 (Figure 3. 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 (Figure 3. 5c). Another example is GCN 3 (Figure 3. 2), GCN 26 (Figure 3. 2) and GCN16 (Figure 3. 2), which show distinct patterns at caudoputamen. All 3 GCNs were identified by both DLSC and WGCNA, while for PCA only mode 2 is most related to caudoputamen (Figure 3. 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.

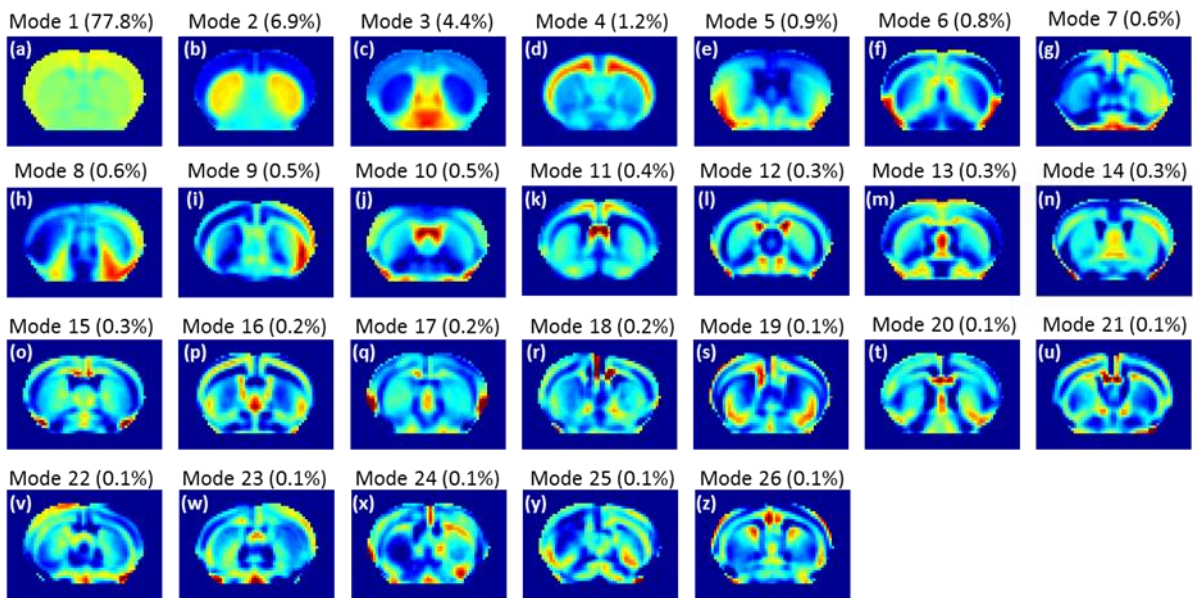


Figure 3. 5. Visualization of the first 26 modes obtained from principal component analysis. The values in the parentheses are the percentage of total variance explained by the mode.

5.3 Gene Coexpression Network and Brain Parcellation

Existing literature have shown that transcriptional profiles reflect the gross brain anatomical structures (Ed S. 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. 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 Figure 3. 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 (Figure 3. 6 blue arrows), hypothalamus (Figure 3. 6 red arrows), thalamus (Figure 3. 6 magenta arrows) etc. The good correspondence is also reflected in the high normalized mutual information. The values are comparable to 0.6 which is the mutual information obtained from slice 27 (Figure 3. 6), 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 (Figure 3. 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 (Mike Hawrylycz et al., 2010) and human and is known crucial to the formation of specialized brain anatomical and functional areas (O'Leary, Stocker, & Zembrzycki, 2007). Within a dominant layered organization, layer-specific areal patterning is

also apparent. For instance, isocortex layers are further divided into motor areas (Figure 3. 6 green arrows), somatosensory area (Figure 3. 6 orange arrows), piriform area (Figure 3. 6 pink arrows), retrosplenial area (Figure 3. 6 dark green arrows), auditory area (Figure 3. 6 purple arrows), and visual area (Figure 3. 6 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

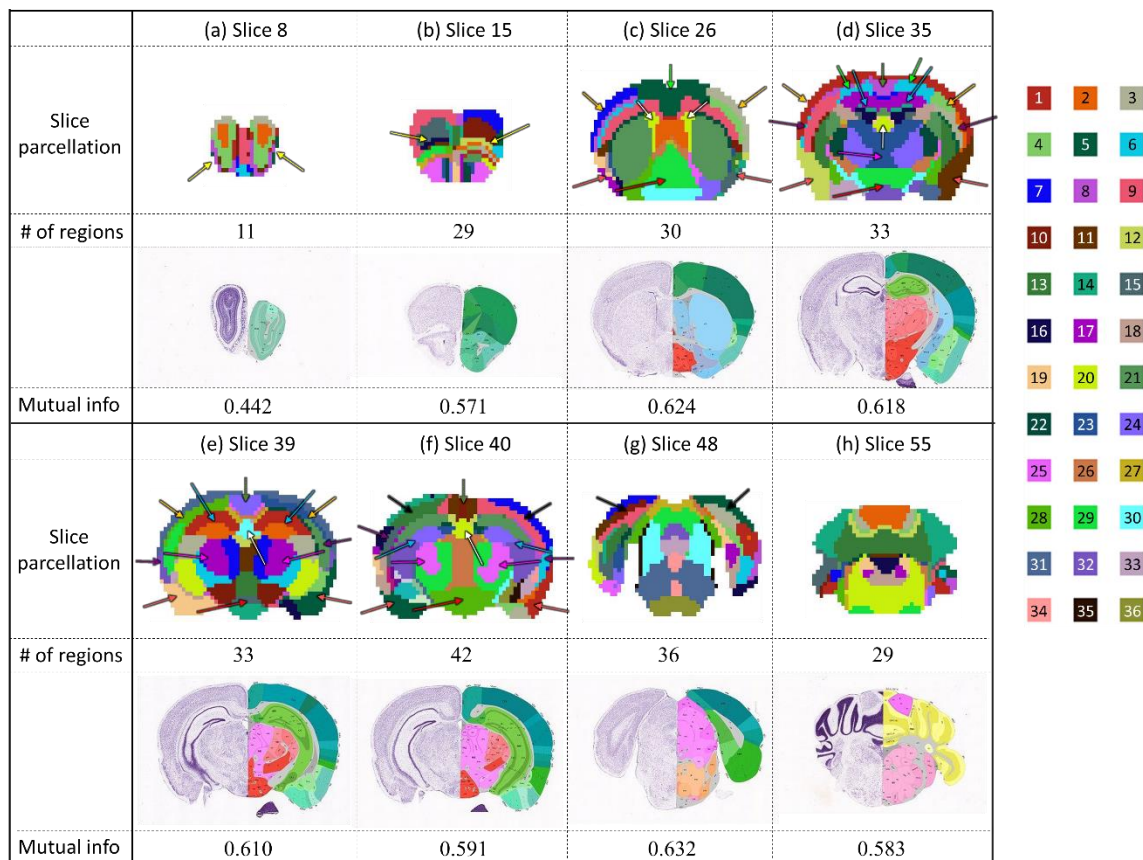


Figure 3. 6. Representative anatomical divisions based on the GCN features. Eight panels correspond to eight selected slices. In each panel, top row: slice number; second row: brain parcellation obtained from spectral clustering with dictionaries as feature vector; third row: number of regions in the 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 code of each region is shown on the right.

example is slice 39 and slice 40. Some major canonical regions such as ventricles (Figure 3. 6 e-f white arrows), hippocampus (Figure 3. 6 e-f blue arrows), thalamus (Figure 3. 6 e-f magenta arrows), retrosplenial area (Figure 3. 6 e-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.

6 Brain-Wide GCN Ontology and Spatial Analysis

Having established the slice-wide GCNs, this section 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 the published gene lists.

Comparisons with the published lists of genes related to cell type markers, specific biological functions and known 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 enriched in a 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 (Cahoy et al., 2004) 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 labeling at different regions of adult mouse forebrain (Sugino et al., 2006). Other networks such as GCN 11, 15, 20 and GCN 12, 41 describe mitochondrial, 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).

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* (proteolipid 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 3. 4). 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 (Figure 3. 7). For example, the fact that GCN 5 (Figure 3. 7ii) locates at ventricle, where the subventricular zone is rich with astrocytes (Quinones-Hinojosa & Chaichana, 2007), confirms its enrichment in astrocyte markers. GCN 7 (Figure 3. 7v) is mainly distributed in the deeper layers of neocortex, which is reminiscent of the distribution of glutamatergic projection neurons in layer V (Molyneaux et al., 2007). GCN 23, located mainly at cerebellar region (Figure 3. 7vi) and the indicated enrichment in GABAergic neurons pointed to a potential enrichment of GABAergic subtype neuron - the Purkinje cells. Quite a number of genes found in GCN 23 coincided with the genes that only labeled Purkinje cells (Wright, Ng, & Guillozet-Bongarts, 2007), 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 (Wright et al., 2007) are also found in GCN 23.

In addition to cell-type specific GCNs, we also found some GCNs remarkably selective for particular brain regions, such as GCN 27 (Figure 3. 7x) in field CA1, GCN 4 (Figure 3. 7xi) in field CA3, GCN 38 (Figure 3. 7xii) in Dentate gyrus, GCN 45 (Figure 3. 7xiii) in cerebellum, GCN 21 (Figure 3. 7xiv) in medulla, GCN 1 (Figure 3. 7xv) in thalamus, and GCN 28 (Figure 3. 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.

Table 3. 4. 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, Ed S. 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, Horvath, & Geschwind, 2010)	5, 30,43
Oligodendrocyte (Ed S. 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, Ed S. 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

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 3. 7). One example is GCN 20. The top weighted gene *Ptp4a1* (protein tyrosine phosphatase 4a1) of GCN is not a marker for pyramidal neuron, but a marker for a neuron. As seen in Table 3. 4, besides pyramidal neuron markers, this network is also enriched for neuron markers and mitochondrial-related genes. In other cases where the top weighted genes were not involved in any of the characterized functions, these genes might suggest potential direct or

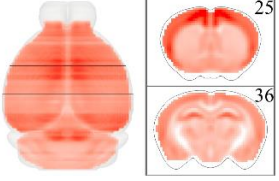
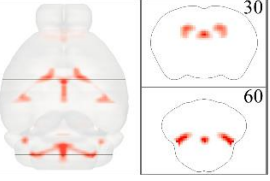
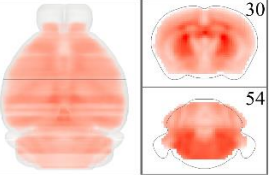
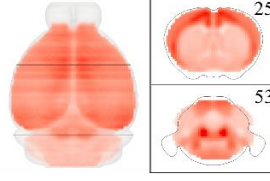
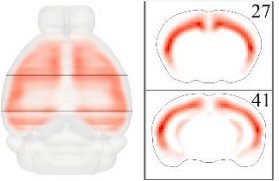
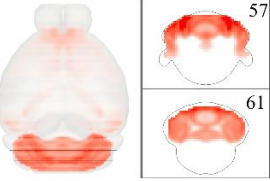
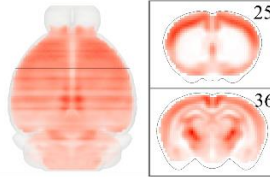
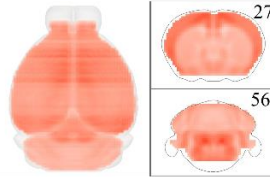
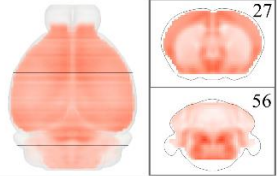
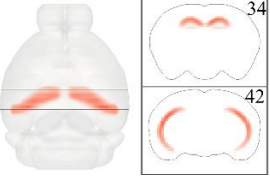
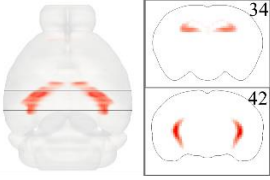
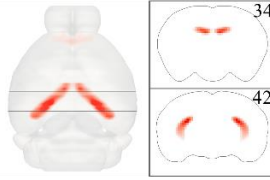
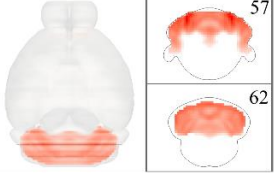
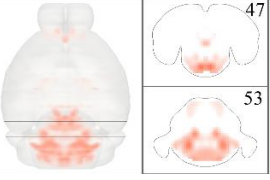
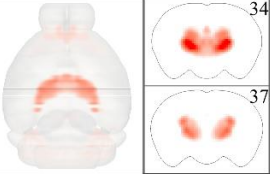
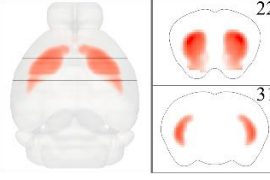
(i) 18	(ii) 5	(iii) 24	(iv) 20
			
Neuron	Astrocyte	Oligodendrocyte	Pyramidal Neuron
<i>Rab6a</i> (3.859) <i>Eid1</i> (3.746) <i>Gpr162</i> (3.523)	<i>Tgfb2</i> (3.828) <i>Bdh2</i> (2.560) <i>Acaa2</i> (2.453)	<i>S100a16</i> (6.032) <i>Cldn11</i> (5.947) <i>Arhgef10</i> (5.910)	<i>Ptp4a1</i> (2.624) <i>Npab</i> (1.203) <i>Arf1</i> (0.946)
(v) 7	(vi) 23	(vii) 10	(viii) 11
			
Glutamatergic neuron	GABAergic neuron	Interneuron	Mitochondrial
<i>Tbr1</i> (3.832) <i>Gng12</i> (3.665) <i>B3galt2</i> (2.744)	<i>Tspan11</i> (4.209) <i>Creg1</i> (3.146) <i>Ptprz1</i> (3.119)	<i>Scn1a</i> (2.870) <i>Asb13</i> (2.819) <i>Nefh</i> (2.733)	<i>Psm11</i> (2.996) <i>Actr1a</i> (2.691) <i>Atp5h</i> (2.597)
(ix) 41	(x) 27	(xi) 4	(xii) 38
			
Ribosomal	Field CA1	Field CA3	Dentate gyrus
<i>Tmx4</i> (3.189) <i>Wbp5</i> (3.150) <i>Rpl8</i> (1.901)	<i>Spink8</i> (3.720) <i>Arl15</i> (3.679) <i>Pantr1</i> (3.413)	<i>Crls1</i> (5.500) <i>Pkp2</i> (5.037) <i>Klk8</i> (4.925)	<i>Crlf1</i> (6.246) <i>Rasl10a</i> (6.216) <i>Cyp7b1</i> (6.126)
(xiii) 45	(xiv) 21	(xv) 1	(xvi) 28
			
Cerebellum cortex	Medulla	Thalamus	Caudoputamen
<i>Gng13</i> (7.881) <i>Syndig1</i> (7.822) <i>Ptpr</i> (7.612)	<i>Acan</i> (4.350) <i>Acyp2</i> (2.929) <i>Ddt</i> (2.670)	<i>Gjc1</i> (5.538) <i>Rgs16</i> (5.013) <i>Vangl1</i> (4.810)	<i>Mme</i> (5.040) <i>Cd4</i> (5.030) <i>Adora2a</i> (4.367)

Figure 3. 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.

indirect link with the known functions. For instance, *Tgfb2* (transforming growth factor, beta receptor II) is not an astrocyte marker. Research has shown that TGF β pathway is relevant to the optic nerve head astrocyte migration (Miao, Crabb, Hernandez, & Lukas, 2010).

7 Discussion and Conclusion

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 (Elad & Aharon, 2006). 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 small percentage of cells (Lein, Ed S. 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 sparseness. 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 neurons. 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 input 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 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 with multiple neuronal cell types. Other GCNs are significantly associated with biological 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, aging, environment, and diseases (S. Dong, Li, Wu, Tsien, & Hu, 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.

CHAPTER 4

VOLUME COMPLETION OF 3D *IN SITU* HYBRIDIZATION GRID USING FULLY CONVOLUTIONAL NEURAL NETWORK³

³ Yujie Li, Heng Huang, Hanbo Chen, Tianming Liu, 2018, Deep Neural Networks for Exploration of Transcriptome of Adult Mouse Brain, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.

© 2018 IEEE. Reprinted with permission from publisher.

1 Abstract

Missing value estimation for microscopy images becomes important when the subsequent analysis depends on complete data. Here we present a novel training scheme that successfully adapts the U-net architecture to the problem of volume recovery. By analogy to denoising autoencoder, we hide a portion of each training sample so that the network can learn to recover the missing voxels from the context. With Allen Mouse Brain Atlas (AMBA), we show that the volume recovery network is successful in completing the large missing region on a slice as well as one or two consecutive missing slices, visually and quantitatively. A comparison with different training schemes showed the importance of designing the right strategy that fits to the missing data patterns. The completed spatially resolved AMBA enables many following statistical and analytical tools that rely on complete data.

2 Background and Motivation

Incomplete data has been a problem frequently encountered to image data analysis (Criminisi, Perez, & Toyama, n.d.; Pathak, Krahenbuhl, Donahue, Darrell, & Efros, 2016; Sun, Yuan, Jia, & Shum, 2005). This is a much severe yet more important problem for microscopic images because of the challenging and time-consuming data acquisition process. On one hand, any mistreatment of tissue slice, loss of focus during imaging, or misalignment during registration could result in the corresponding data loss. On the other hand, the acquired data are often limited and thus requires that we make use of data as much as we can in further data analysis. Therefore, there is a great need for algorithms that can complete microscopic images.

The simplest solution to incomplete data is to ignore them. In one of our prior works (Li, Chen, Jiang, Li, Lv, Peng, et al., 2017), we worked around the problem by first studying the coexpression networks slice-wise and then infer the gene-gene interactions by considering only

the slices with data. Using two steps, we focus on the known interactions. Yet the strategy is only applicable to a specific problem and an extra step is required for data integration.

Alternatively, it is possible to use image inpainting methods for missing value estimation because ISH volumes are directly image structures. Classical inpainting methods (Criminisi et al., n.d.; Sun et al., 2005) restore the image based on either local or non-local information. Most existing methods require continuous textures or contours across the known and missing region. However, this assumption is often not true, especially when the missing region is large and in arbitrary shape. Other methods resort to external database for a possible match for the missing region (Hays & Efros, 2008). Failures occur when the test image is significantly different from the database. Recently, learning-based methods have shown superior performance in image completion problem (Mairal et al., 2008; Xie, Xu, & Chen, 2012). Instead of hand-designing features for patch editing or matching, dictionaries or a neural network are learned from data (Mairal et al., 2008; Xie et al., 2012). Deep neural networks have shown great promise in filling large missing regions in images, a more challenging task that requires a deeper understanding of the image. These models provide a plausible completion by learning the semantic meaning (Pathak et al., 2016; Yeh et al., 2017). However, all the above-mentioned methods are limited to two dimensional images and not directly applicable for 3D volume completion.

Inspired by the rapid rise and revolutionary performance of deep learning algorithms, we propose volume recovery network (VRN), a convolutional neural network that completes 3D volume data. VRN borrows the idea from denoising autoencoder (DA) (Thomas, Price, Paine, & Richards, 2002). Instead of feeding the network data with manually added noises and teaching the network to undo noises, we hide a portion of each training sample so that the network can learn to recover missing voxels from the context.

The architecture of VRN follows that of the U-net (Ronneberger, Fischer, & Brox, 2015). It is essentially an autoencoder with skip connections between mirrored layers of encoder and decoder. The addition of skip connections between encoder and decoder is crucial for localizing high-resolution features. The training strategies of which part of volume to hold out is key to the performance of VRN. Using Allen Mouse Brain Atlas (AMBA) as an example dataset, we showed that the tailored strategy is effective in training the network to learn from adjacent slices as well as the surrounding voxels. The completed spatially resolved AMBA enables a holistic investigation for many statistical and analytical tools that depend on complete data.

3 Volume Recovery Network

3.1 Training strategy

VRN borrows the idea from DAs. Instead of adding noise to the inputs to teach the network to undo the noise, we hide some data of each training data to teach the network to recover the missing voxels from the context. By observation, the data loss for AMBA are usually one or more slices along coronal axis. This loss pattern is a result of the acquisition step when the brain tissues were sectioned along coronal axis and then digitally processed, stitched, registered, gridded, and quantified (Lein, Ed S. et al., 2007).

Based on the patterns of missing data, we designed three strategies on which portion of data to hold out. First, we hide a random slice by setting all voxels on the slice to -1, which represents missing values. Second, we randomly pick two consecutive slices to hide. Third, we random pick a slice and sample from existing missing data patterns on that slice and mask out part of the slice. To make sure the third strategy does not overlap with the first situation, we set a range of 10 to 80 to the percentage of allowed missing data on that slice. The three ways of

simulation of missing data render the network to learn to recover missing data from the previous and the subsequent slices as well as the same slice.

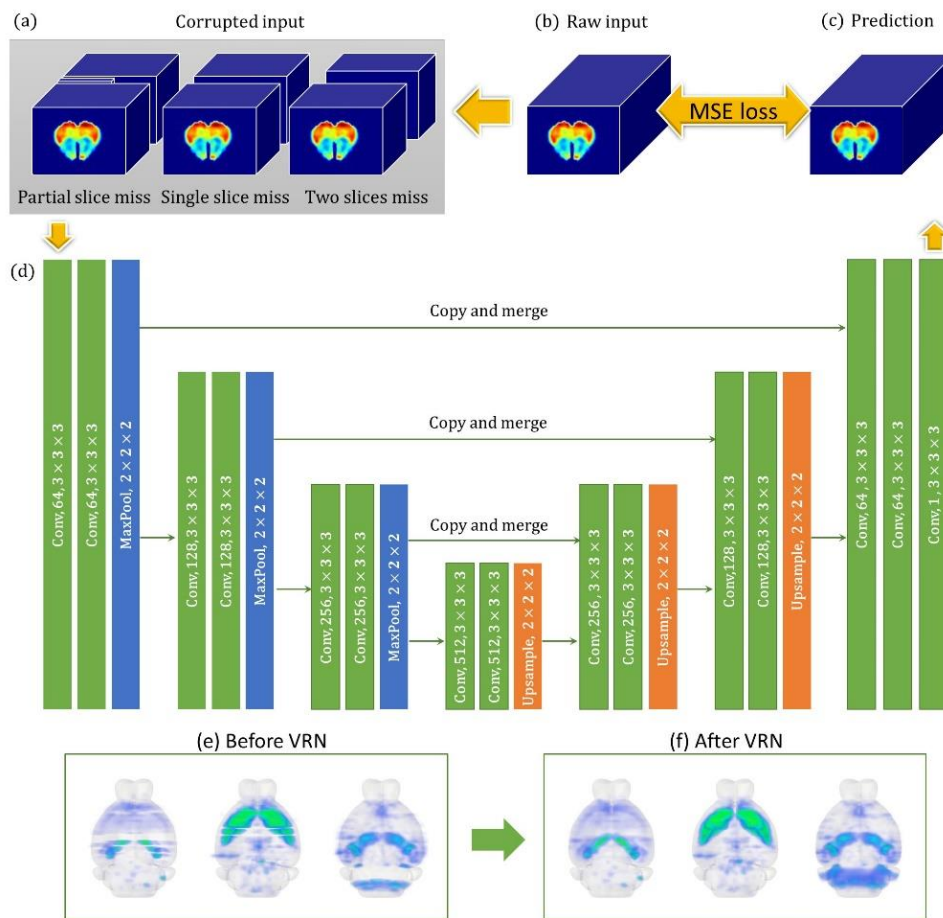


Figure 4. 1. Volume recovery network architecture. Each training sample is a volume of size $72 \times 48 \times 64$. (a) We designed three schemes of holding out data, hide partial slice, hide a single slice and hide two consecutive slices. (b) The volume is first corrupted by one of the three ways before it is inputted to the network. (c) The output is the predicted volume of the same size of the input. MSE loss is calculated between the predicted volume and the raw input. (d) VRN consists of an encoder and decoder and the mirrored layers are connected via skip layers. The type of the layer, number and size of kernels are denoted in the box. (e,f) axial view of three raw ISH volumes and respective completed volumes.

The choice for partial slice mask is important. Previously the frequently used image masks for face/natural image completion are central square mask or random blocks (Pathak et al., 2016). Here the AMBA brings dozens of existing missing slice masks. Instead of arbitrarily

generating the partial slice masks, these masks are sampled from existing data because they come from the exact distribution of the data to be completed. On average, each slice has about 135 different masks with 10-80 percent of voxels of the slice missing. As we will show later, the inclusion of the partial masks is essential to prevent the network from learning the low-level features latching onto the boundaries. Additionally, the limitation on the percent of missing voxels is also essential in preventing the slice training from degrading to strategy 1.

3.2 Network Architecture

Figure 4. 1 shows the architecture of our network, a 3D U-net architecture (Ronneberger et al., 2015). It consists of an encoder and a decoder. In the contracting path, repeated convolutions using $3 \times 3 \times 3$ filters followed by a Rectified Linear Unit (ReLU) and $2 \times 2 \times 2$ pooling layers are used to aggregate features and increase the size of the receptive field. In the expansive path, the $2 \times 2 \times 2$ deconvolutional layers are used to propagate context information to higher layers. The skip layers between the encoding and decoding path ensure that high resolutions features are retained and localized (Ronneberger et al., 2015). The architecture is fully convolutional, which means the network allows the input volume in arbitrary shape.

The training is achieved by regressing to the ground truth content of the entire volume, including the held-out region. Mean squared error (MSE) loss is used as our reconstruction loss function. As the ground truth volume might contain missing values, only the losses of the voxels with ground truth were counted.

The model was implemented using Keras package (François Chollet, 2015). The initial learning rate was 10^{-6} and decay rate is 10^{-6} . Adam (Kingma & Ba, 2015) was used as the optimizer. To ensure a seamless tiling of the output, each training sample is padded on each side and the full volume is of size $72 \times 48 \times 64$. The number of training samples is 3300 and the number

of validation samples is 330, which consists of 85% of the data. The remaining 15% of data is used for testing. During training, we consider all three strategies for each training sample and each time only one strategy is applied. Assuming that each volume has 57 coronal slices in use, then the first two strategies generate $57 \times 2 = 114$ ways of corruption. For the third strategy, the average number of partial mask for each slice is 135 and with 57 coronal slices there are $135 \times 57 = 7695$ ways of corruption. Putting it altogether, for each volume we can generate ~ 7800 new samples. Therefore, no data augmentation is required. Each epoch took about 40 hours on a 12GB Nvidia Geforce GPU.

3.3 Evaluation

In addition to MSE, we use two more metrics to evaluate the quality of the predictions. The first metric is structural similarity index (SSIM) (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004). SSIM estimates the holistic similarity between two images and has been used as a useful metric for evaluating algorithms designed for image compression, image reconstruction, denoising and super-resolution. The second one is the peak signal-to-noise ratio (PSNR) which directly measures the difference in pixel values. The evaluation scheme corresponds to the training strategy, which consists of conditions of one slice missing, two consecutive slices missing and partial slice missing. Instead of using entire volume, we only evaluate the metrics on the completed slice(s) against the ground truth. For partial slice missing, to reduce computations, we estimate the performance by using the same randomly selected ten missing masks for each slice.

3.4 Volume recovery by mean estimation from neighbors (MEN)

Missing values were estimated as the mean of the foreground voxels in its 26 neighborhood. In each iteration, mean calculations were performed for each missing voxel. For a

voxel whose surrounding voxels are all missing, it is skipped from filling for the current iteration. The estimation stops when all missing values were filled. MEN is an effective simple method and it is used as the baseline model.

3.5 Volume recovery by three-dimensional convolutional autoencoder (3D CAE)

We also compared the results obtained from 3D CAE (Du et al., 2017). The network of a 3D CAE is same to that of U-net except for that all the skip connections are removed. All hyperparameters, loss functions and the training strategies remain the same to those of VRN.

3.6 Experimental materials

The dataset used for the experiment is the ISH volumes from AMBA. Please refer to Section 1.4 in Chapter 1 for details.

4 Comparison with 3D CAE and MEN

To demonstrate that VRN is able to complete missing data, we manually hide a portion of volume and then compare the results predicted by VRN, CAE and MEN. In the first and second experiment, part of slice 24 were masked (Figure 4. 2b,c). The masks were sampled from existing missing data patterns of the AMBA. As shown, the missing regions predicted by VRN and CAE preserves the gradient of expressions at isocortex layers. The gradient is smoothed out by MEN. The visual differences are also reflected in MSE, SSMI and PSNR in both cases (Figure 4. 2b,c). Then we tested on masking out the entire slice of 24 (Figure 4. 2d). A blurry isocortex layers remained for MEN. In contrast, the performance by VRN and CAE did not deteriorate. The predicted slice 24 emulates the patterns of the raw energies, suggesting the deep models use the previous and subsequent slices for prediction. For CAE, the middle gradient in the isocortex generated is not as sharp as that of VRN (Figure 4. 2d white arrows). Next, we

evaluated the performance of both methods on predicting slice 24 when the previous slice (Figure 4. 2e) or the subsequent slice (Figure 4. 2f) is also missing. In both cases, the missing slice filled by VRN are sharper and the layer gradient remains clear. In contrast, the slice filled by CAE and MEN lost the details around somatosensory and motor regions. Overall, the VRN is able to preserve the high-resolution details that are not captured by either MEN or CAE during missing voxels recovery.

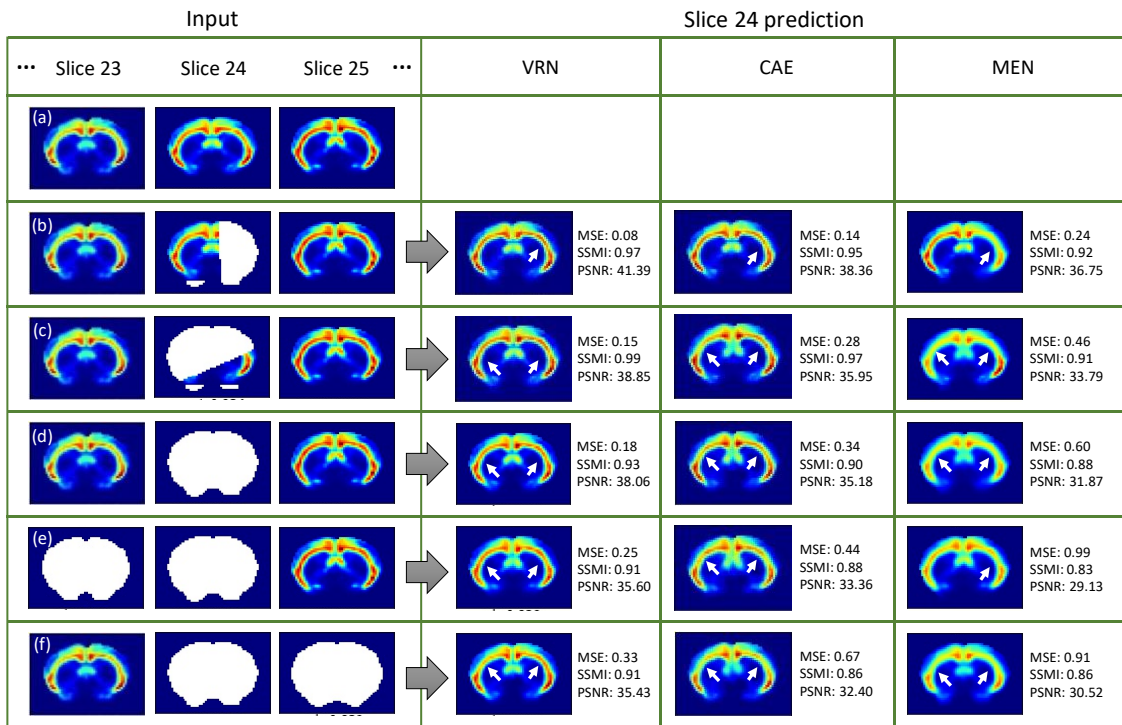


Figure 4. 2. Comparison of the volume completion by VRN, CAE and MEN. The gene acronym is *syt7*. The input is the entire volume. Only slice 23, 24 and 25 are shown. White parts indicate regions with missing values. In all cases, slice 24 is predicted. (a) Ground truth for slice 23, 24 and 25. Input volume is corrupted by hiding (b)(c) part of slice 24, (d) the entire slice 24. (e) slice 23 and 24, (f) slice 24 and 25.

Then we also made comparison among the three methods for all transcripts (Table 4. 1). In all three conditions where one or two slices miss or partial slice miss, the predictions made by VRN is consistently better than MEN as well as CAE. The performance improvement is

confirmed by all evaluation metrics. The comparison also confirmed the importance of skipping connections in maintaining the details in the volume data.

Table 4. 1. Performance comparison between VRN, CAE and MEN. The average of three metrics on all slices are presented.

		one slice missing			two slices missing			partial slice missing		
		MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR
VRN	train	1.42	0.87	23.30	2.07	0.81	20.78	0.54	0.94	29.28
	test	1.32	0.88	23.21	1.98	0.82	20.29	0.50	0.94	28.88
CAE	train	2.10	0.81	21.77	3.30	0.75	19.59	0.82	0.93	27.80
	test	1.82	0.83	21.39	2.87	0.79	19.28	0.73	0.92	27.96
MEN	train	2.71	0.82	18.80	4.71	0.75	18.09	1.02	0.93	26.97
	test	2.31	0.86	18.78	3.81	0.79	17.82	0.88	0.94	27.36

5 Importance of Including Partial Slice Training

It is worth noting the importance of incorporating the partial slice training into the training strategy. To demonstrate this point, we trained a network without the strategy of hiding only part of a slice. As a baseline we first compare the performance of both networks when the entire slice is missing. As seen in Figure 4. 3f,g,m,n, the completed slices by both networks are close to the ground truth and they are very similar to each other. Quantitatively, the SSIM are both 0.98 and PSNR are about 33. The MSEs are ~0.6. The performance deviates in the case where only part of the slice is missing. The performance of the network with partial slice training shows improvements (Figure 4. 3j) or remains at a similar level (Figure 4. 3c), probably due to the extra information on the slice. In contrast, for the network trained without partial slice training, the voxels near the boundaries of the slice mask showed obvious cracks (Figure 4. 3d, k black arrows). The discontinuities are also reflected in a much higher MSE and a lower SSIM

and PSNR. For slice 32, even though over half of the slice is given, the MSE is almost the same as that of the case where the entire slice is hidden. The performance on slice 22 is even worse because the MSE of the prediction on a portion of slice is over five times higher (3.46) than that of the prediction on the entire slice (0.63). The results here indicate that without partial slice training, the information on the same slice does not help the network complete the missing region but makes it more complex. The incorporation of partial slice training is essential as it helps train the filters to integrate the information from the previous and subsequent slice as well as the information on the slice.

Ground truth	Partial slice corruption	With partial slice training	No partial slice training	Entire slice corruption	With partial slice training	No partial slice training
(a) Slice 22	(b)	(c) MSE: 0.63 SSMI: 0.98 PSNR: 33.20	(d) MSE: 3.46 SSMI: 0.93 PSNR: 25.84	(e)	(f) MSE: 0.56 SSMI: 0.98 PSNR: 33.08	(g) MSE: 0.63 SSMI: 0.98 PSNR: 33.23
(h) Slice 32	(i)	(j) MSE: 2.60 SSMI: 0.97 PSNR: 28.78	(k) MSE: 6.82 SSMI: 0.93 PSNR: 24.58	(l)	(m) MSE: 7.01 SSMI: 0.91 PSNR: 22.95	(n) MSE: 7.16 SSMI: 0.91 PSNR: 22.66

Figure 4. 3. Illustration of the importance of incorporating partial slice corruption. Slice 22 (row 1) and 32 (row 2) of gene *Itfg1* are presented. First image(a,h) is the ground truth. Partial slice (b,i) is corrupted with sampled masks. Missing region is denoted in white. Then we show the completed slices with(out) partial slice training when part of slice is missing (c,d,j,k) or the entire slice is corrupted (f,g,m,n).

6 Discussion and Conclusion

The architecture of VRN is a convolutional autoencoder with skip layers linking the encoder and decoder. The skip layers are essential for maintaining high-resolution details as they pass image details from the convolutional layers to deconvolutional layers. These high-resolution details that are important for gene expression data. For example, the expression gradient is a key

feature associated with how genes regulate brain functions. Many of the differences reported between functionally distinct cortical regions is not due to the selective expression in functionally discrete regions but rather discontinuous sampling across a gradient (Sansom & Livesey, 2009).

The right training strategy is also essential for the performance of VRN. The scheme of hiding a single slice or partial slice can effectively train the convolutional filters learn from previous and subsequent slices as well as the surrounding voxels on the same slice. The strategy of hiding two consecutive slices trains the network to integrate higher-level semantic meanings from regions further apart. Due to these novel and effective designs, the performance of VRN is superior in comparison with mean estimation from neighbors and CAE. A full dataset is usually the prerequisite for many statistical and analytical tools. The completed spatially resolved AMBA enables many subsequent potent computational approaches that depend on complete data and offers more possibilities to understanding the cortex at the level of its underlying genetic code.

CHAPTER 5

EXPLORING TRANSCRIPTOME ARCHITECTURE OF ADULT MOUSE BRAIN VIA
RESTRICTED BOLTZMANN MACHINE AND DEEP BELIEF NETWORK ⁴

⁴ Yujie Li, Heng Huang, Hanbo Chen, Tianming Liu, 2018, Deep Neural Networks for Exploration of Transcriptome of Adult Mouse Brain, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.
© 2018 IEEE. Reprinted with permission from publisher.

1 Abstract

Transcriptome in brain plays a crucial role in understanding the cortical organization and the development of brain structure and function. In this work, I show that Restricted Boltzmann Machines (RBMs) can be used to infer co-occurrences among voxels, providing foundations for dividing the cortex into discrete subregions. As we stack multiple RBMs to form a deep belief network (DBN), we progressively map the high-dimensional raw input into abstract representations and create a hierarchy of transcriptome architecture. A fine-to-coarse organization emerges from the network layers. This organization incidentally corresponds to the anatomical structures, suggesting a close link between structures and the genetic underpinnings. Thus, we demonstrate a new way of learning transcriptome-based hierarchical organization using RBM and DBN.

2 Introduction

Transcriptome in brain plays a crucial role in understanding the cortical organization. Previous research has revealed extensive regional heterogeneity of transcriptome. For instance, laminar-specific genes have been identified through gene expression studies comparing subregions of neocortex (Liu, Dwyer, & O’Leary, 2000; Zhong et al., 2004). Relatedly, microarray analysis of purified populations of neuronal subtypes identified many cell-type enriched genes (Cahoy et al., 2004; Winden et al., 2009). These findings have also been confirmed by large-scale *in situ hybridization* studies (Ed S. Lein et al., 2004), where the restricted expressions due to specific cell populations are directly visualized. However, how to analyze the genomic-neuroanatomic relationship remains challenging because of the combinatorial complexity of gene expression patterns.

Unsupervised machine learning methods have shown advantages in discovering transcriptome architecture. Similarity-based clustering (Bohland et al., 2010; Ng et al., 2009) and non-linear embedding techniques (Mahfouz et al., 2015) have been applied on gene expression data, revealing areal and laminar structures in the mouse neocortex (Mike Hawrylycz et al., 2010). Singular value decomposition (Grange et al., 2014) and matrix factorization methods are also popular approaches in identifying genes with similar coexpression patterns (Li, Chen, Jiang, Li, Lv, Li, et al., 2017; Li, Chen, Jiang, Li, Lv, Peng, et al., 2017; Thompson et al., 2008). As helpful they are in reducing the dimensionality, they are all linear shallow mappings and inadequate for inferring complex non-linear structures of data. For example, it is reported that features captured by principal components can sometimes degrade cluster quality (Yeung & Ruzzo, 2001). Relatedly, Barnes-Hut Stochastic Neighbor Embedding(BH-SNE) is shown superior than linear methods because of its ability of capturing non-linear relations (Mahfouz et al., 2015). However, the BH-SNE still failed to produce voxel clusters corresponding to brain structures when no prior dimension reduction was performed (Mahfouz et al., 2015).

Deep models such as deep neural networks (DNNs) show a larger representation power through composition of many nonlinearities. For instance, DAs have been used to learn a compact representation of yeast microarray expression profiles (Gupta, Wang, & Ganapathiraju, 2015). The clusters obtained from the learned codes is more consistent with the pre-assigned ground truth labels in comparison with those obtained via clustering the raw data. Relatedly, an ensemble of DAs proves effective in extracting stable expression signatures from public gene expression data with diverse experiments (Tan et al., 2017). In addition to the expression power, DNNs have a hierarchical structure in which higher-level features are obtained by composing lower-level ones. This compositional hierarchies are also seen in many natural signals such as

signaling systems of cells (Xu & Lan, 2015). In several recent publications, DNNs have been successfully used to simulate the cellular signaling system (L. Chen, Cai, Chen, & Lu, 2015, 2016).

For the above reasons, we consider Restricted Boltzmann Machines (RBMs) and deep belief network (DBN), for modelling the transcriptome data of adult mouse brain. RBMs are fit for inferring the transcriptome-based brain parcellation because through training it learns which units in the visible layer tend to co-occur and then record the significant activation in the hidden layer. The co-occurrences of voxels provide foundations for dividing the cortex into discrete subregions. As we stack multiple RBMs to form a DBN, we progressively map the high-dimensional raw input into abstract representations and create a hierarchical data-driven transcriptome architecture, presumably revealing how brain subregions interact with one another in a hierarchical manner.

DBN is a multilayer generative model. It is formed by stacking multiple layers of Restricted Boltzmann Machines (RBMs). Both RBMs and DBNs have been demonstrated effective in extracting features from various modalities including text (Srivastava & Salakhutdinov, 2012), video, audio (Ngiam et al., 2011), gene microarray data (L. Chen et al., 2016) and neuroimages (Devon et al., 2015).

3 Methods

3.1 Experimental materials

The dataset used for the experiment is the ISH volumes from AMBA. Please refer to Section 4 in Chapter 1 for details.

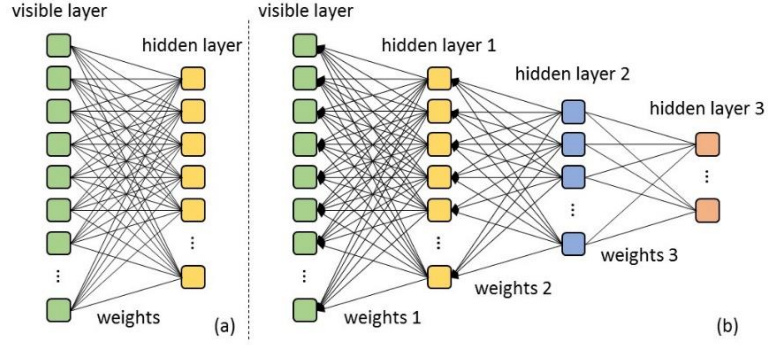


Figure 5. 1. Illustration of (a) Restricted Boltzmann Machine and (b) deep belief network.

3.2 Restricted Boltzmann Machine

An RBM (G. E. Hinton, 2002) is a probabilistic energy-based model and the objective is to fit a probability distribution model over a set of visible random variables to the observed data. As shown in the definition, the model restricts the interactions in the Boltzmann energy function to only those between visible neurons and hidden neurons. An RBM can be graphically represented as a bipartite graph (Figure 5. 1a). For binary visible units $\mathbf{v} \in \{0,1\}$ and $\mathbf{h} \in \{0,1\}$, the energy function is defined as follows:

$$E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = - \sum_{i \in \text{visible}} a_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (5.1)$$

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ are the model parameters, weights \mathbf{W} connect the visible units (\mathbf{v}) and the hidden units (\mathbf{h}) and \mathbf{a} and \mathbf{b} are their biases.

The joint distribution over the visible and hidden units can be obtained via the energy function:

$$P(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})} \quad (5.2)$$

where Z is the normalization term. It is given by summing over all possible pairs of visible and hidden vectors and thus intractable.

The probability that the network assigns to a visible vector, \mathbf{v} , is given by summing over all possible hidden vectors (5.3) and the objective of a RBM is to maximize the log likelihood of all data points.

$$P(\mathbf{v}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})} \quad (5.3)$$

Following equation (5.3), the model updates can be obtained from the derivatives of the log likelihood given a set of observations \mathbf{v} . As shown in the equation, there are two ways to adjust the probability that the network assigns to a training image. First is by adjusting the weights and biases to lower the energy of that image. Alternatively, we can raise the energy of other images, which results an increase to the partition function.

$$\frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(v_n; \boldsymbol{\theta})}{\partial W_{ij}} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \quad (5.4)$$

where the angle brackets denote the expectation with respect to the specified distribution.

Sampling unbiasedly from $\langle v_i h_j \rangle_{data}$ is easy because there are no direct connections between hidden units or visible units. However, getting an unbiased sample of $\langle v_i h_j \rangle_{model}$ is difficult because of the long computation for Gibbs sampling. This problem is solved by ‘contrastive divergence’ (G. E. Hinton, Osindero, & Teh, 2006), whose key ideas are to 1) initialize the Markov chain with a distribution close to the training data and 2) use samples from a few steps of Gibbs sampling as a close approximation.

In our work, the observed expression energies are real-valued $\mathbf{v} \in \mathbb{R}^D$. We use a variant of RBM, the Gaussian-binary RBM (G. Hinton, 2010), for modelling real-valued vectors. The model assumes that each visible unit have independent Gaussian noise. Given real-valued visible units $\mathbf{v} \in \mathbb{R}^D$, and $\mathbf{h} \in \{0,1\}^F$, the energy function is defined as:

$$E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = \sum_{i \in \text{visible}} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \quad (5.5)$$

where σ_i is the standard deviation of the Gaussian noise for visible unit i .

The conditional probability of the visible unit given the hidden units is modeled by a Gaussian distribution, whose mean is a function of the hidden units.

$$P(\mathbf{v}|\mathbf{h}; \boldsymbol{\theta}) = \prod_{i=1}^D p(v_i|h) \quad , \text{ with } v_i|h \sim N(b_i + \sigma_i \sum_{j=1}^F W_{ij} h_j, \sigma_i^2) \quad (5.6)$$

The update of model parameters takes a very similar form to the RBM with binary visible unit.

$$\frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(v_n; \boldsymbol{\theta})}{\partial W_{ij}} = \langle \frac{v_i}{\sigma_i} h_j \rangle_{data} - \langle \frac{v_i}{\sigma_i} h_j \rangle_{model} \quad (5.7)$$

3.3 Deep Belief Network

Unlike RBM with a single layer, DBN captures the features using multiple layers in a stochastic manner (G. E. Hinton et al., 2006). The top two layers with undirected connections form an RBM and the lower layers have directed connections (Figure 5. 1b). The training for each RBM was performed layer-wise in a greedy manner. The hidden units in the first RBM (hidden layer 1) are taken as the visible units of the second RBM. The hidden units of the second RBM (hidden layer 2) are fed as the visible units into the third RBM. It has been shown that each addition of a RBM can improve the variational lower bound on the log probability of the training data (G. E. Hinton et al., 2006).

In this work, the deepnet (<https://github.com/nitishsrivastava/deepnet>), a public available package, was applied to train the RBM and DBN. To handle real-valued data, Gaussian visible units were used. Each Gaussian visible unit was set to have unit variance ($\sigma_i = 1$) which was kept fixed and not learned. The DBN for exploring the voxel co-occurrences consists of 60144

Gaussian visible units followed by 1024 binary hidden units in the first hidden layer, 256 hidden units in the second hidden layer, 64 in the third hidden layer. Each layer of weights was trained using CD with the number of times running CD, i.e. $k=1$ (G. E. Hinton et al., 2006).

The learning rates in all hidden layers are set initially to 0.001 and decrease as the inverse of time with a decay half-life of 5000. The activation function used is tangent hyperbolic function. The initial and final momentum are 0.5 and 0.9 with the change step to be 5000. The batch size is 100. The weights are initialized as a zero-mean Gaussian distribution with a standard deviation of 0.01. We want the weights to be sparse because most genes are expressed in a small percentage of cells (Lein, Ed S. et al., 2007). Thus, we add a regularization term, the ℓ_1 norm of the weights to induce the weight sparsity in each RBM. We found that the ℓ_1 value of 0.1 works well in practice over multiple experiments. It has been reported that the ℓ_1 constraint does not dominate RBM learning results (Devon et al., 2015).

4 RBM to infer single-level transcriptome architecture

An RBM serves as a helpful learning tool as it models the density of visible variables by introducing a set of conditionally independent latent variables. In the context of unveiling the transcriptome organization based on gene expression profiles, it is trained to learn which foreground voxels in the visible layer tend to co-occur for the given set of gene expression patterns and then record the significant activation in the hidden layer. We assume that the co-occurred voxels should belong to the same region and these co-occurring patterns are fundamental and intrinsic to the transcriptome data. For the convenience of future discussion, we denote the spatial map of each presented weight as a weight map.

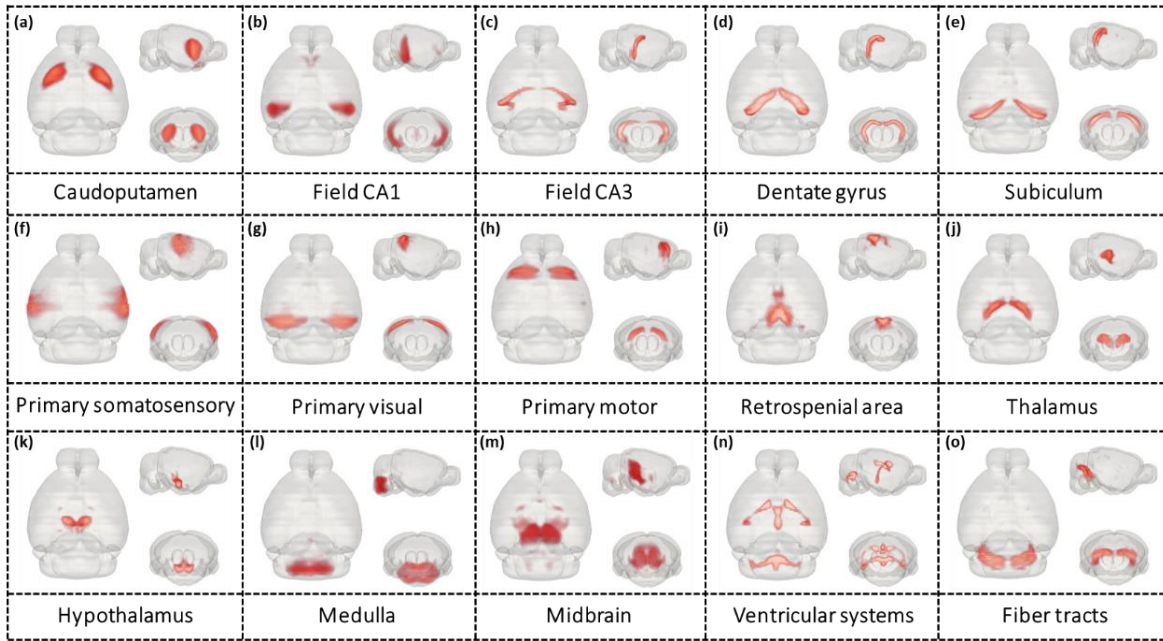


Figure 5. 2. Visualization of weight maps learned by RBMs. Results were obtained from an RBM with 1024 hidden units. In each subfigure, three views including axial view, sagittal and coronal view were shown.

To discover voxel co-occurrence, we set the number of hidden units to 1024. The trained weights were linearly projected to the input space for intuitive interpretation of the representations. The weight maps help us understand how much each foreground voxel contributes to a specific activation pattern. By visual inspection, most spatial distributions of these weights form tight continuous clusters. This clustering patterning agrees with the brain's organizational principle that transcriptome similarities are strongest between spatial neighbors, both between cortical areas and between cortical layers (Bernard et al., 2012). The delineations are in general symmetric and match major canonical brain regions including caudoputamen (Figure 5. 2a), hippocampus (Figure 5. 2b), isocortex (Figure 5. 2f-i), thalamus (Figure 5. 2j), hypothalamus (Figure 5. 2k), medulla (Figure 5. 2l), midbrain (Figure 5. 2m) as well as ventricular systems (Figure 5. 2n) and fiber tracts (Figure 5. 2o). Interestingly, the features

learned often correspond to a finer breakdown of known brain regions. For instance, field CA1, field CA3, and dentate gyrus (DG) and subiculum are identified individually (Figure 5. 2b-e). The isocortex is further divided into primary somatosensory area (Figure 5. 2g), primary visual area (Figure 5. 2h) and primary motor area (Figure 5. 2i). We also made comparisons with the clustering results obtained using K-means and hierarchical clustering (results not shown here). The clusters obtained by RBM and DBN are more coherent and robust to noise.

5 DBN to infer a hierarchy of transcriptome architecture

One superior advantage of deep neural networks over other shallow models is that they can capture the hierarchy of features. As we stack multiple RBMs to form a DBN, we create a hierarchy of transcriptome architecture. We use Figure 5. 3 to demonstrate this hierarchy. A visualization of the weight maps of DBN shows that the features learned in the first levels are generally localized and clustered and, in this case, mostly correspond to subregions of caudoputamen or the nearby regions such as olfactory tubercle and striatum (Figure 5. 3 green shadow). Yet the differences among the patterns are discernable. As we go to the second layer, these fine anatomical subregions learned at the first layer started to merge into larger areas (Figure 5. 3 blue shadow). Intuitively, the weight maps are more likely to merge with those with similar spatial distributions because the second RBM learns the co-occurrences of subregions identified in the first RBM. Indeed, all weight maps that were combined show spatial overlaps. For example, the weight map 292 (Figure 5. 3f) in the first layer shows strong signals at medial caudoputamen and it is combined with weight map 764 (Figure 5. 3g) that is activated at caudal part of caudoputamen. In addition to the merge of subregions, it is common to see the merge of regions spatially adjacent. The weight map 776 (Figure 5. 3m) features high values at olfactory tubercle and the weight map 519 (Figure 5. 3l) shows higher weights at rostral nucleus

accumbens. Both weight maps are summarized by the weight map 227 (Figure 5. 3d) in the second hidden layer. It is worth noting that one lower layer weight map can have connections with multiple higher-level weight maps. For example, the layer-1 weight map 519 (Figure 5. 3l) that shows high values at the rostral nucleus accumbens is used by both layer-2 weight map 199 (Figure 5. 3e) and 227 (Figure 5. 3d). At the third layer, we saw further combinations of weight maps in layer 2, obtaining a layer-3 weight map spanning the entire caudoputamen, nucleus accumbens, olfactory tubercle as well as piriform area and substantia innominate, etc. (Figure 5. 3a).

6 Discussion and Conclusion

The objective of the work is to understand the organization of brain structure from the transcriptome's point of view. This objective is achieved by exploring a public dataset called Allen Mouse Brain Atlas using data-driven methods RBM and DBN. To handle the high dimensionality of data, we showed that RBM and DBN are effective tools in studying the transcriptome architecture. Specifically, RBM can learn the co-occurrences patterns among voxels, providing a transcriptome-based anatomy. The 3D visualizations of the weight maps show that the voxel clusters match well with the classical neuroanatomy. This provides strong evidence of a close link between transcriptome and brain structure. It should be noted that the input to RBMs are not constrained to image data. If each visible node is a transcript, then RBM can learn the co-occurrence among genes, i.e. the coexpression patterns.

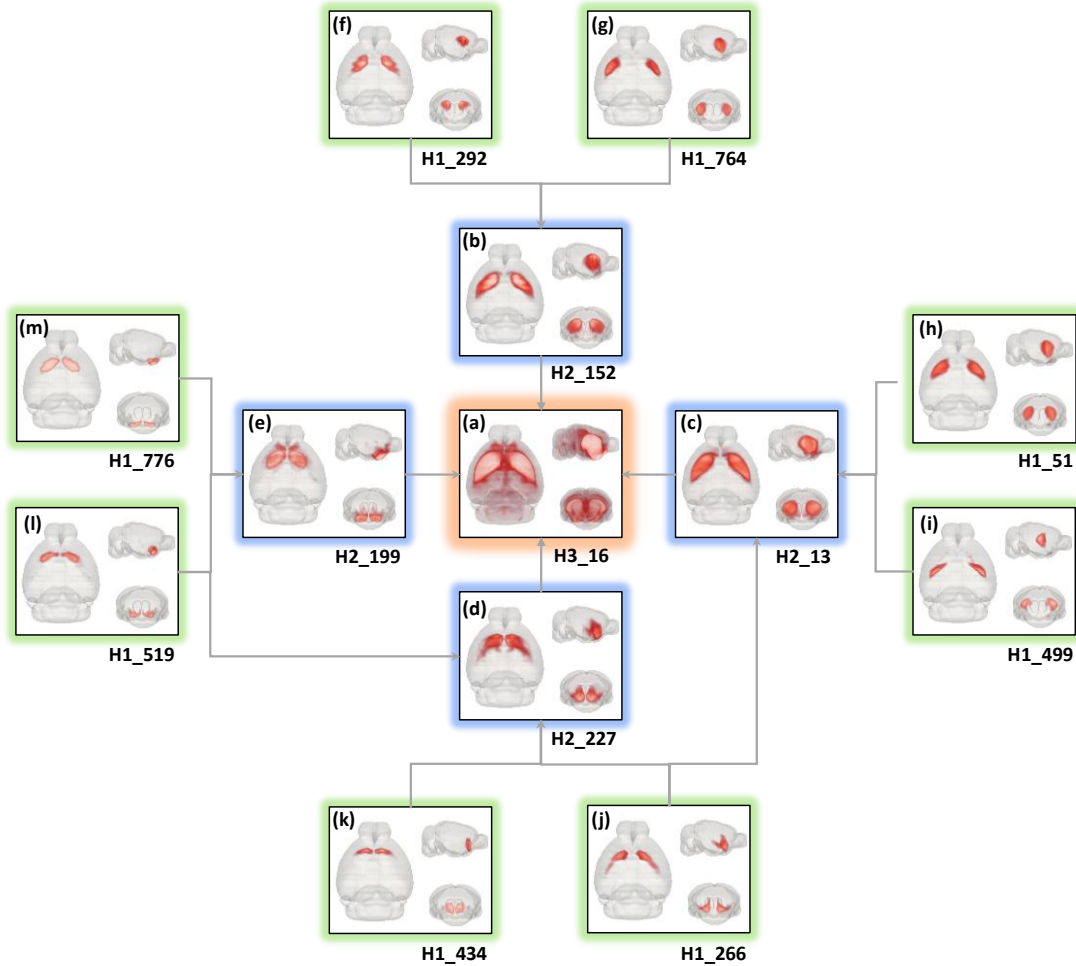


Figure 5. 3. Visualization of a hierarchy of transcriptome architecture learned by DBN. In each subfigure, three views including axial view (left), sagittal (right top) and coronal view (right bottom) are shown. The weight maps of hidden layer 1, hidden layer 2, hidden layer 3 are colored in green, blue and orange respectively. The index of the weight in the respective layer is noted under each subfigure. For the hidden units, their weights were visualized as a weighted linear combination of the weights of the Gaussian RBM. Here four layer-2 weight maps were selected and for each of the layer-2 weight map, two layer-1 maps were selected for presentation. The arrows indicate a compositional relationship. The weight in the next layer is a sum of weighted linear combination of the weights in the previous layers.

Another reason for the choice of RBM is that RBMs are the building block for more complex deeper models like DBN. These deep models can capture features that are not possible for a shallow model. Having validated the clusters obtained from RBMs, we stack the RBMs into

a three-layer DBN and create a hierarchy. The components learned on lower levels are localized and match with subregions of canonical neuroanatomy. It agrees with the principle of brain organization that transcriptome similarities are the strongest between anatomical neighbors. At higher levels, these localized features merged and interacted with adjacent groups. Overall, a fine-to-coarse organization emerges from the network layers, and we show how the subregions merge and interact with one another. It is found that the organization incidentally correspond to the anatomical structures well, suggesting a close link between brain structures and the genetic underpinnings. Thus, we demonstrated a new way of learning transcriptome-based hierarchical organization of mouse brain using RBM and DBN.

The transcriptomic similarity provide useful hints for the similarities in structures (Mike Hawrylycz et al., 2010; Ed S. Lein, Belgard, Hawrylycz, & Molnár, 2017; Thompson et al., 2008) and functions (Toledo-Rodriguez et al., 2004) as well as brain connectivity (Fakhry & Ji, 2015; French & Pavlidis, 2011). In future work, we will extend current framework by including other image modalities such as diffusion tensor image, neuronal tracing data or functional magnetic resonance image. A joint representation of micro-scale gene expression and macro-scale neuroimages can possibly reveal the correlations across different scales and modalities, thus providing deeper understanding of the organization architecture of the brain.

CHAPTER 6

CONCLUSION AND FUTURE WORK

1 Summary of contribution

In this thesis, I systematically study the relationship between gene expressions and neuroanatomy on mouse brain. My contributions are summarized as follows.

To answer whether the gene expressions can refine the anatomic regions, I put forward two data-driven methods that derive the spatial correlations among gene expressions. The first method formats the voxel clustering problem as a sparse representation problem (Chapter 2). The key assumption is that voxels that use the same dictionary for representation should belong to the same brain region. The resulted clusters align well with the classic neuroanatomy. Genes that are enriched in fiber tracts and ventricular systems have been reported for the first time. The second method considers a probability-based model RBM and its extension DBN, which consists of three layers of RBMs (Chapter 5). The RBM fits a distribution model to the voxels, which summarizes the co-occurrence patterns. A visualization of the co-occurrence patterns showed tight clusters corresponding to neuroanatomy. Further with DBN, we build a hierarchical data-driven transcriptome architecture. A fine-to-coarse organization emerges from the network, revealing how brain subregions interact with one another in a hierarchical manner.

The other side of the genomic-neuroanatomy relationship is the relations among the genes that are spatially correlated. These genes are usually analyzed via networks. I provide a new way of generating the coexpression networks using DLSC. The motivation of the method lies in the observations that neurons encode sensory information using a small number of active

neurons. We presume gene expressions also follows a sparse coding strategy. By comparing with the most popular method of deriving the coexpression networks -- WGCNA, we have shown that DLSC is able to summarize the structures that best represent the gene expression features into dictionaries so that much finer and more balanced coexpression networks in comparison to WGCNA can be achieved. In addition to a finer gene parcellation, another benefit that DLSC offers is soft clustering, i.e. multiple assignment is possible for one gene. Such set-up considers the multiple roles of genes in the regulatory domains. Rather than simply assigning each gene to a network, DLSC also provides a quantitative measure on the extent to which a gene conforms to the coexpression patterns. The quantification can serve as a useful feature in identifying important genes via regression with external data modalities.

A standing problem for the above work is missing data. In chapter 4, I proposed a deep learning based network for 3D volumetric data completion. The proposed network is fully convolutional and takes a 3D volume as input and outputs a completed 3D volume. The rationale of the network for volume completion is similar to a denoising autoencoder. By covering up a portion of slice of choice, we teach the network to reconstruct the concealed regions from context. We show that designing a training strategy fit for the data is essential to a good completion. Quantitatively and visually, the completed 3D volumes resemble the ground truth with high-resolution details preserved.

2 Future directions

The future work of this research topic has two general directions. The first direction is multi-modal studies. An integration of transcriptome data with other modalities including various properties of neurons and macroscale imaging modalities like magnetic resonance images (MRI), will facilitate a thorough understanding of the functional circuit from multiple perspectives. For

example, adding a systematic examination of transcriptome data to the investigation of many neurological disorders, along with the medical images, can shed light on the causes of the abnormalities. The key challenge of multimodal learning lies in the distinct form of representations and correlation structure for different modalities. It makes it very difficult to discover relationship across modalities. One ongoing effort is cell typing of neurons. Existing evidence shows that cell classification requires consideration of features from different perspectives. Yet how to link the cell types defined by transcriptome to the types defined by morphology, electrophysiology and/or connectivity, remains a grand challenge and awaits solutions.

The second direction is the spatiotemporal analysis. Transcriptome only provides a snapshot of the cell status at a time point. How the cell machinery changes over developmental stages or during evolution can only be revealed by adding the temporal axis. The observations over time is specifically useful for studying transcription factors, whose role varies by time. Many transcription factors are reported to participate in different biological processes at different developmental stages. The development process offers us a window to investigate numerous important events such as neurogenesis, neuron differentiation, cell migration and differentiation, synaptogenesis etc. as well as how these events are regulated via transcriptome (Borodinsky et al., 2015).

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. <http://doi.org/10.1186/1471-2105-5-18>
- Bakken, T. E., Miller, J. A., Ding, S., Sunkin, S. M., Smith, K. A., & Ng, L. (2016). Comprehensive transcriptional map of primate brain development. *Nature*, 535(7612), 367–375. <http://doi.org/10.1038/nature18637>. Comprehensive
- Bando, S. Y., Silva, F. N., Costa, L. D. F., Silva, A. V., Pimentel-Silva, L. R., Castro, L. H., ... Moreira-Filho, C. A. (2013). Complex network analysis of CA3 transcriptome reveals pathogenic and compensatory pathways in refractory temporal lobe epilepsy. *PLoS One*, 8(11), e79913. <http://doi.org/10.1371/journal.pone.0079913>
- Belgard, T. G., Marques, A. C., Oliver, P. L., Abaan, H. O., Sirey, T. M., Hoerder-Suabedissen, A., ... Ponting, C. P. (2011). A transcriptomic atlas of mouse neocortical layers. *Neuron*, 71(4), 605–616. <http://doi.org/10.1016/j.neuron.2011.06.039>
- Bernard, A., Lubbers, L. S., Tanis, K. Q., Luo, R., Podtelezhnikov, A. A., Finney, E. M., ... Lein, E. (2012). Transcriptional Architecture of the Primate Neocortex. *Neuron*, 73(6), 1083–1099. <http://doi.org/10.1016/j.neuron.2012.03.002>. Transcriptional
- 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. *Proceedings of the National Academy of Sciences*, 101(7), 2173–2178. <http://doi.org/10.1073/pnas.0308512100>
- Bohland, J. W., Bokil, H., Pathak, S. D., Lee, C.-K., Ng, L., Lau, C., ... Mitra, P. P. (2010). Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods*, 50(2), 105–12. <http://doi.org/10.1016/j.ymeth.2009.09.001>
- Borodinsky, L. N., Belgacem, Y. H., Swapna, I., Visina, O., Olga, A., Sequerra, E. B., ... Shim, S. (2015). Spatiotemporal integration of developmental cues in neural development. *Developmental Neurobiology*, 75(4), 349–359. <http://doi.org/10.1002/dneu.22254>. Spatiotemporal
- Brown, C. D., Johnson, D. S., & 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., ... Barres, B. (2004). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *Journal of Neuroscience*, 28(1), 264–278. <http://doi.org/10.1523/JNEUROSCI.4178-07.2008>

- Carter, H., Hofree, M., & Ideker, T. (2013). Genotype to phenotype via network analysis. *Current Opinion in Genetics & Development*, 23(6), 611–621. <http://doi.org/10.1016/j.gde.2013.10.003>
- Chen, H., Liu, T., Zhao, Y., Zhang, T., Li, Y., Li, M., ... Liu, T. (2015). Optimization of large-scale mouse brain connectome via joint evaluation of DTI and neuron tracing data. *NeuroImage*, 115, 202–213. <http://doi.org/10.1016/j.neuroimage.2015.04.050>
- Chen, L., Cai, C., Chen, V., & Lu, X. (2015). Trans-species learning of cellular signaling systems with bimodal deep belief networks. *Bioinformatics*, 31(18), 3008–3015. <http://doi.org/10.1093/bioinformatics/btv315>
- Chen, L., Cai, C., Chen, V., & Lu, X. (2016). Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics*, 17(S1), 9. <http://doi.org/10.1186/s12859-015-0852-1>
- Cohen, M. L., Golde, T. E., Usiak, M. F., Younkin, L. H., & Younkin, S. G. (1988). In situ hybridization of nucleus basalis neurons shows increased beta-amyloid mRNA in Alzheimer disease. *Proceedings of the National Academy of Sciences*, 85(4), 1227–1231. <http://doi.org/10.1073/pnas.85.4.1227>
- Comon, P. (1994). Independent Component Analysis: A new Concept. *IEEE Transactions Signal Processing*, 36(November), 287–314. [http://doi.org/Doi 10.1016/0165-1684\(94\)90029-9](http://doi.org/Doi 10.1016/0165-1684(94)90029-9)
- Criminisi, A., Perez, P., & Toyama, K. (n.d.). Object removal by exemplar-based inpainting. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2, II-721-II-728. <http://doi.org/10.1109/CVPR.2003.1211538>
- 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 Biology*, 4(5), P3. <http://doi.org/10.1186/gb-2003-4-5-p3>
- Devon, H., Vince, C., Ruslan, S., Elena, A., Tulay, A., & Sergey, P. (2015). Restricted Boltzmann Machines for Neuroimaging: an Application in Identifying Intrinsic Networks. *Neuroimage*, 344(6188), 1173–1178. <http://doi.org/10.1126/science.1249098.Sleep>
- Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M. L., Carlson, S., ... Schadt, E. E. (2009). Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biology*, 10(5), R55. <http://doi.org/10.1186/gb-2009-10-5-r55>
- Dong, H. (2008). *Allen Reference Atlas. A Digital Color Brain Atlas of the C57BL/6J Male Mouse - by H. W. Dong. Genes, brain, and behavior* (Vol. 9). John Wiley & Sons. <http://doi.org/10.1086/596246>
- Dong, S., Li, C., Wu, P., Tsien, J. Z., & Hu, Y. (2007). Environment enrichment rescues the neurodegenerative phenotypes in presenilins-deficient mice. *European Journal of*

Neuroscience, 26(1), 101–112. <http://doi.org/10.1111/j.1460-9568.2007.05641.x>

Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L., & Tao, D. (2017). Stacked Convolutional Denoising Auto-Encoders for Feature Representation. *IEEE Transactions on Cybernetics*, 47(4), 1017–1027. <http://doi.org/10.1109/TCYB.2016.2536638>

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2), 407–499.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1999). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(22), 12930–12933. <http://doi.org/10.1073/pnas.95.25.14863>

Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12), 3736–45. <http://doi.org/10.1109/ICIG.2009.101>

Fakhry, A., & Ji, S. (2015). High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods*, 73, 71–78.

François Chollet. (2015). Keras. *GitHub Repository*.

French, L., & Pavlidis, P. (2011). Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Computational Biology*, 7(1). <http://doi.org/10.1371/journal.pcbi.1001049>

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, and Behavior*, 13(1), 13–24. <http://doi.org/10.1111/gbb.12106>

Ge, H., Liu, Z., Church, G. M., & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics*, 29(4), 482–6. <http://doi.org/10.1038/ng776>

Grange, P., Bohland, J. W., Okaty, B. W., Sugino, K., Bokil, H., Nelson, S. B., ... Mitra, P. P. (2014). Cell-type-based model explaining coexpression patterns of genes in the brain. *Proceedings of the National Academy of Sciences*, 111(14), 5397–402. <http://doi.org/10.1073/pnas.1312098111>

Gupta, A., Wang, H., & Ganapathiraju, M. (2015). Learning structure in gene expression data using deep architectures, with an application to gene clustering. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1328–1335. <http://doi.org/10.1109/BIBM.2015.7359871>

Hawrylycz, M., Bernard, A., Lau, C., Sunkin, S. M., Chakravarty, M. M., Lein, E. S., ... Ng, L. (2010). Areal and laminar differentiation in the mouse neocortex using large scale

- gene expression data. *Methods*, 50(2), 113–21. <http://doi.org/10.1016/j.ymeth.2009.09.005>
- Hawrylycz, M., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., ... Lein, E. (2015). Canonical genetic signatures of the adult human brain. *Nature Neuroscience*, 18(12). <http://doi.org/10.1038/nn.4171>
- Hays, J., & Efros, A. A. (2008). Scene completion using millions of photographs. *Communications of the ACM*, 51(10), 87. <http://doi.org/10.1145/1400181.1400202>
- Heintz, N. (2004). Gene Expression Nervous System Atlas (GENSAT). *Nature Neuroscience*, 7(5), 483. <http://doi.org/10.1038/nn0504-483>
- Hinton, G. (2010). A Practical Guide to Training Restricted Boltzmann Machines. *Momentum*, 9, 1. http://doi.org/10.1007/978-3-642-35289-8_32
- Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8), 1771–1800. <http://doi.org/10.1162/089976602760128018>
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–54. <http://doi.org/10.1162/neco.2006.18.7.1527>
- Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3), 626–634. <http://doi.org/10.1109/72.761722>
- Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., ... Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5), 345–349. <http://doi.org/10.1038/nmeth756>
- Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, 10(1), 1–14. <http://doi.org/10.1186/s13062-015-0077-2>
- 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. *Proceedings of the National Academy of Sciences*, 98(4), 1930–1934. <http://doi.org/10.1073/pnas.98.4.1930>
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 118–127. <http://doi.org/10.1093/biostatistics/kxj037>
- Kang, H. J., Kawasawa, T. I., Cheng, F., Zhu, Y., Xu, X., Li, M., ... Sestan, N. (2011). Spatiotemporal transcriptome of the human brain. *Nature*, 478(7370), 483–489. <http://doi.org/10.1038/nature10523>. Spatiotemporal
- Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations* (pp. 1–15). <http://doi.org/http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503>

- Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. <http://doi.org/10.1186/1471-2105-9-559>
- Lau, C., Ng, L., Thompson, C., Pathak, S., Kuan, L., Jones, A., & Hawrylycz, M. (2008). Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, *9*, 153. <http://doi.org/10.1186/1471-2105-9-153>
- Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research*, *(14)*, 1085–1094. <http://doi.org/10.1101/gr.1910904.1>
- Lein, Ed S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., ... Jones, A. R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, *445*(7124), 168–176. <http://doi.org/10.1038/nature05453>
- Lein, E., Borm, L. E., & Linnarsson, S. (2017). The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, *358*(6359), 64–69. <http://doi.org/10.1126/science.aan6827>
- Lein, E. S. (2004). Defining a Molecular Atlas of the Hippocampus Using DNA Microarrays and High-Throughput In Situ Hybridization. *Journal of Neuroscience*, *24*(15), 3879–3889. <http://doi.org/10.1523/JNEUROSCI.4710-03.2004>
- Lein, E. S., Belgard, T. G., Hawrylycz, M., & Molnár, Z. (2017). Transcriptomic Perspectives on Neocortical Structure, Development, Evolution, and Disease. *Annual Review of Neuroscience*, *40*(1), 629–652. <http://doi.org/10.1146/annurev-neuro-070815-013858>
- 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. *The Journal of Neuroscience*, *24*(15), 3879–89. <http://doi.org/10.1523/JNEUROSCI.4710-03.2004>
- Li, Y., Chen, H., Jiang, X., Li, X., Lv, J., Li, M., ... Liu, T. (2017). Transcriptome Architecture of Adult Mouse Brain Revealed by Sparse Coding of Genome-Wide In Situ Hybridization Images. *Neuroinformatics*, *15*(3), 285–295. <http://doi.org/10.1007/s12021-017-9333-1>
- Li, Y., Chen, H., Jiang, X., Li, X., Lv, J., Peng, H., ... Liu, T. (2017). Discover Mouse Gene Co-expression Landscapes Using Dictionary Learning and Sparse Coding. *Brain Structure and Function*, *11*, 1–18.
- Liu, Q., Dwyer, N. D., & O’Leary, D. D. (2000). Differential expression of COUP-TFI, CHL1, and two novel genes in developing neocortex identified by differential display PCR. *The Journal of Neuroscience*, *20*(20), 7682–7690. <http://doi.org/20/20/7682> [pii]
- 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*(June), 883–891. <http://doi.org/10.1038/nature02618.1>

- Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., ... Liu, T. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Medical Image Analysis*, 20(1), 112–34. <http://doi.org/10.1016/j.media.2014.10.011>
- Maher, C. a, Kumar-sinha, C., Cao, X., Kalyana-, S., Han, B., Jing, X., ... Chinnaiyan, A. M. (2009). Transcriptome Sequencing to Detect Gene Fusions in Cancer Christopher. *Nature*, 458(7234), 97–101. <http://doi.org/10.1038/nature07638>. Transcriptome
- Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M. J., & Lelieveldt, B. P. F. (2015). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods*, 73, 79–89. <http://doi.org/10.1016/j.ymeth.2014.10.004>
- Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11, 19–60.
- Mairal, J., Elad, M., & Sapiro, G. (2008). Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1), 53–69. <http://doi.org/10.1109/TIP.2007.911828>
- Markram, H. et al. (2015). Reconstruction and Simulation of Neocortical Microcircuitry. *Cell*, 163(2), 456–492. <http://doi.org/10.1016/j.cell.2015.09.029>
- Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1), 31–46. <http://doi.org/10.1038/nrg2626>
- Miao, H., Crabb, A. W., Hernandez, M. R., & Lukas, T. J. (2010). Modulation of factors affecting optic nerve head astrocyte migration. *Investigative Ophthalmology and Visual Science*, 51(8), 4096–4103. <http://doi.org/10.1167/iovs.10-5177>
- Migliore, M., & Shepherd, G. M. (2005). An integrated approach to classifying neuronal phenotypes. *Nature Reviews Neuroscience*, 6(10), 810–818. <http://doi.org/10.1038/nrn1769>
- 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(1), 322. <http://doi.org/10.1186/1471-2105-12-322>
- Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences*, 107(28), 12698–12703. <http://doi.org/10.1073/pnas.0914257107>
- Mody, M., Cao, Y., Cui, Z., Tay, K. Y., Shyong, A., Shimizu, E., ... Tsien, J. Z. (2001). Genome-wide gene expression profiles of the developing mouse hippocampus. *Proceedings of the National Academy of Sciences*, 98(15), 8862–8867. <http://doi.org/10.1073/pnas.141244998>
- Molyneaux, B. J., Arlotta, P., Menezes, J. R. L., & Macklis, J. D. (2007). Neuronal subtype specification in the cerebral cortex. *Nature Reviews Neuroscience*, 8(6), 427–37.

<http://doi.org/10.1038/nrn2151>

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628. <http://doi.org/10.1038/nmeth.1226>

Nelson, S. B., Sugino, K., & Hempel, C. M. (2006). The problem of neuronal cell types: a physiological genomics approach. *Trends in Neurosciences*, 29(6), 339–45. <http://doi.org/10.1016/j.tins.2006.05.004>

Ng, L., Bernard, A., Lau, C., Overly, C. C., Dong, H.-W., Kuan, C., ... Hawrylycz, M. (2009). An anatomic gene expression atlas of the adult mouse brain. *Nature Neuroscience*, 12(3), 356–62. <http://doi.org/10.1038/nn.2281>

Ng, L., Pathak, S. D., Kuan, C., Lau, C., Dong, H., Sodt, A., ... Hawrylycz, M. (2007). Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3), 382–392. <http://doi.org/10.1109/TCBB.2007.1035>

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning. *Proceedings of The 28th International Conference on Machine Learning (ICML)*, 689–696. <http://doi.org/10.1145/2647868.2654931>

O’Leary, D. D. M., Stocker, A. M., & Zembrzycki, A. (2007). Area Patterning of the Mammalian Cortex. *Neuron*, 56(2), 252–269. <http://doi.org/10.1016/B978-0-12-397265-1.00021-6>

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, 103(47), 17973–8. <http://doi.org/10.1073/pnas.0605938103>

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. *Nature Neuroscience*, 11(11), 1271–1282. <http://doi.org/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 Systems Biology*, 6(1), 63. <http://doi.org/10.1186/1752-0509-6-63>

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context Encoders: Feature Learning by Inpainting. In *Conference on Computer Vision and Pattern Recognition*. <http://doi.org/10.1109/CVPR.2016.278>

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 Biology*, 8(1), S7. <http://doi.org/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. *Experimental Neurology*, 205(2), 313–324. <http://doi.org/10.1016/j.expneurol.2007.03.016>

Rampon, C., Jiang, C. H., Dong, H., Tang, Y. P., Lockhart, D. J., Schultz, P. G., ... Hu, Y. (2000). Effects of environmental enrichment on gene expression in the brain. *Proceedings of the National Academy of Sciences*, 97(23), 12880–4. <http://doi.org/10.1073/pnas.97.23.12880>

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (pp. 1–8). http://doi.org/10.1007/978-3-319-24574-4_28

Sansom, S. N., & Livesey, F. J. (2009). Gradients in the Brain: The Control of the Development of Form and Function in the Cerebral Cortex. *Cold Spring Harbor Perspectives Biology*, 1(2), 16. <http://doi.org/10.1101/cshperspect.a002519>

Seung, H. S., & Sumbul, U. (2014). Neuronal cell types and connectivity: lessons from the retina. *Neuron*, 83(6), 1262–1272. <http://doi.org/10.1016/j.immuni.2010.12.017>. Two-stage

Srivastava, N., & Salakhutdinov, R. (2012). Multimodal Learning with Deep Boltzmann Machines. *Advances in Neural Information Processing Systems (NIPS)*, 2222–2230. <http://doi.org/10.1109/CVPR.2013.49>

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643), 249–255. <http://doi.org/10.1126/science.1087447>

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., ... Hogenesch, J. B. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99(7), 4465–70. <http://doi.org/10.1073/pnas.012025199>

Sugino, K., Hempel, C. M., Miller, M. N., Hattox, A. M., Shapiro, P., Wu, C., ... Nelson, S. B. (2006). Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature Neuroscience*, 9(1), 99–107. <http://doi.org/10.1038/nn1618>

Sun, J., Yuan, L., Jia, J., & Shum, H.-Y. (2005). Image completion with structure propagation. *ACM Transactions on Graphics*, 24(3), 861. <http://doi.org/10.1145/1073204.1073274>

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., ... Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6), 2907–2912. <http://doi.org/10.1073/pnas.96.6.2907>

Tan, J., Doing, G., Lewis, K. A., Price, C. E., Chen, K. M., Cady, K. C., ... Greene, C. S. (2017). Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell Systems*, 5(1), 63–71.e6.

<http://doi.org/10.1016/j.cels.2017.06.003>

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., ... Zeng, H. (2016). Adult Mouse Cortical Cell Taxonomy by Single Cell Transcriptomics. *Nature Neuroscience*, 19(2), 335–346. <http://doi.org/10.1038/nn.4216>.Adult

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22(3), 281–285. <http://doi.org/10.1038/10343>

Thomas, P., Price, B., Paine, C., & Richards, M. (2002). Remote electronic examinations: student experiences. *British Journal of Educational Technology*, 33(5), 537–549. <http://doi.org/10.1111/1467-8535.00290>

Thompson, C. L., Pathak, S. D., Jeromin, A., Ng, L. L., MacPherson, C. R., Mortrud, M. T., ... Lein, E. S. (2008). Genomic Anatomy of the Hippocampus. *Neuron*, 60(6), 1010–1021. <http://doi.org/10.1016/j.neuron.2008.12.008>

Tole, S., Christian, C., & Grove, E. A. (1997). Early specification and autonomous development of cortical fields in the mouse hippocampus. *Development*, 124(24), 4959–4970.

Toledo-Rodriguez, M., Blumenfeld, B., Wu, C., Luo, J., Attali, B., Goodman, P., & Markram, H. (2004). Correlation maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in rat neocortex. *Cerebral Cortex*, 14(12), 1310–1327. <http://doi.org/10.1093/cercor/bhh092>

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105–1111. <http://doi.org/10.1093/bioinformatics/btp120>

Tsien, J., Li, M., Osan, R., Chen, G., Lin, L., Wang, P., ... Kuang, H. (2013). On initial Brain Activity Mapping of episodic and semantic memory code in the hippocampus. *Neurobiology of Learning and Memory*, 105, 200–210. <http://doi.org/10.1016/j.nlm.2013.06.019>.On

Veltman, J. A., Fridlyand, J., Pejavar, S., & et.al., A. B. O. (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Research*, 63, 2872–2880.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, 1096–1103. <http://doi.org/10.1145/1390156.1390294>

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Mill, J., ... Daniel, H. (2013). Transcriptomic Analysis of Autistic Brain Reveals Convergent Molecular Pathology. *Nature*, 474(7351), 380–384. <http://doi.org/10.1038/nature10110>.Transcriptomic

- Wang, G., Belgard, T. G., Mao, D., Chen, L., Berto, S., Preuss, T. M., ... Konopka, G. (2015). Correspondence between resting state activity and brain gene expression. *Neuron*, 88(4), 659–666. <http://doi.org/10.1016/j.neuron.2015.10.022>.Correspondence
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <http://doi.org/10.1109/TIP.2003.819861>
- Winden, K. D., Oldham, M. C., Mirnics, K., Ebert, P. J., Swan, C. H., Levitt, P., ... Geschwind, D. H. (2009). The organization of the transcriptional network in specific neuronal classes. *Molecular Systems Biology*, 5(291), 1–18. <http://doi.org/10.1038/msb.2009.46>
- Wolf, L., Goldberg, C., Manor, N., Sharan, R., & Ruppin, E. (2011). Gene expression in the rodent brain is associated with its regional connectivity. *PLoS Computational Biology*, 7(5). <http://doi.org/10.1371/journal.pcbi.1002040>
- Woodhams, P., Celio, M., Ulfing, N., & Witter, M. (1993). Morphological and functional correlates of borders in the entorhinal cortex and hippocampus. *Hippocampus*, 3(S1), 303–312. <http://doi.org/10.1002/hipo.1993.4500030733>
- Wright, E., Ng, L., & Guillozet-Bongarts, A. (2007). *Annotation report on cerebellar cortex, pukinje cell layer. Allen Bran Atlas Mouse Brain.*
- Xie, J., Xu, L., & Chen, E. (2012). Image Denoising and Inpainting with Deep Neural Networks. *Advances in Neural Information Processing Systems*, 1–9.
- Xu, J., & Lan, Y. (2015). Hierarchical feedback modules and reaction hubs in cell signaling networks. *PLoS ONE*, 10(5), 1–25. <http://doi.org/10.1371/journal.pone.0125886>
- Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2017). Semantic Image Inpainting with Deep Generative Models. In *Conference on Computer Vision and Pattern Recognition*. <http://doi.org/10.1109/CVPR.2017.728>
- Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9), 763–774. <http://doi.org/10.1093/bioinformatics/17.9.763>
- Zeisel, A., Machado, A. B. M., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., ... Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226), 1138–42. <http://doi.org/10.1126/science.aaa1934>
- Zeng, H., & Sanes, J. R. (2017). Neuronal cell-type classification: Challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9), 530–546. <http://doi.org/10.1038/nrn.2017.85>
- Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1).

<http://doi.org/10.1371/journal.pone.0078644>

Zhao, X., Lein, E. S., He, A., Smith, S. C., Aston, C., & Gage, F. H. (2001). Transcriptional profiling reveals strict boundaries between hippocampal subregions. *The Journal of Comparative Neurology*, *441*(3), 187–196. <http://doi.org/10.1002/cne.1406>

Zhong, Y., Takemoto, M., Fukuda, T., Hattori, Y., Murakami, F., Nakajima, D., ... Yamamoto, N. (2004). Identification of the genes that are expressed in the upper layers of the neocortex. *Cerebral Cortex*, *14*(10), 1144–1152. <http://doi.org/10.1093/cercor/bhh074>