Computational analysis of brain transcriptome atlases
Understanding molecular mechanisms

Mahfouz, Ahmed

DOI
10.4233/uuid:e8241c58-a443-46b7-8e98-c91078d338df

Publication date
2016

Document Version
Final published version

Citation (APA)
Mahfouz, A. (2016). *Computational analysis of brain transcriptome atlases: Understanding molecular mechanisms*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:e8241c58-a443-46b7-8e98-c91078d338df

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Propositions

accompanying the dissertation

COMPUTATIONAL ANALYSIS OF BRAIN TRANSCRIPTOME ATLASES
UNDERSTANDING MOLECULAR MECHANISMS

by

Ahmed Mohamed Essam Taha Ahmed MAHFOUZ

1. A high resolution spatially-mapped transcriptome of a single sample is more valuable than low-resolution data from more samples. *(This thesis)*
2. Co-expression networks must be spatially and temporally specific to capture brain complexity. *(This thesis)*
3. "Good" low-dimensional representations are more informative than high-dimensional representations of the same data. *(This thesis)*
4. Studying orchestrated activity of genes in the brains of healthy individuals is essential to understand molecular mechanisms of brain disorders. *(This thesis)*
5. Scooping clearly demonstrates the conflict between a scientist's interest in scientific progress and his/her interest in an academic career.
6. Neuroscience research is doomed by averaging across unknown subpopulations of patient groups as well as cell populations.
7. Machine learning algorithms have a greater potential than models to increase our understanding of the brain.
8. Methods aimed at making predictions based on biological data should focus more on out-of-sample generalization rather than in-sample accuracy.
9. In neuropsychiatric disorders, genotyping should come first.
10. Our lack of statistical intuition further supports the notion that human evolution is an ongoing process.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors prof. dr. ir. M.J.T. Reinders and prof. dr. ir. B.P.F. Lelieveldt.

COMPUTATIONAL ANALYSIS OF BRAIN TRANSCRIPTOME ATLASES

UNDERSTANDING MOLECULAR MECHANISMS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 23 juni 2016 om 12:30 uur

door

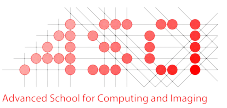
Ahmed Mohamed Essam Taha Ahmed MAHFOUZ

Master of Science in Communication & Information Technology,
Nile University (Egypt),
geboren te Gizeh, Egypte.

This dissertation has been approved by the
promotors: Prof. dr. ir. M.J.T. Reinders and Prof. dr. ir. B.P.F. Lelieveldt

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology
Prof. dr. ir. B.P.F. Lelieveldt,	Delft University of Technology
	Leiden University Medical Center
Independent members:	
Prof. dr. D. Posthuma	Vrije Universiteit Amsterdam
Prof. dr. L. Wessels	Delft University of Technology
	Netherlands Cancer Institute, NKI
Prof. dr. J.J. Goeman	Leiden University Medical Center
Dr. J. Bohland	Boston University, USA
Dr. M. Creyghton	Hubrecht Institute
Prof. dr. R.C.H.J. van Ham,	Delft University of Technology, reserve member



This work was carried out in graduate school ASCI.
ASCI dissertation series number 352.

Parts of this thesis have received funding from The Netherlands Technology Foundation (STW), as part of STW Project 12721 ("Genes in Space") under the IMAGENE perspective program.

Cover. Elements of the human genome projected onto a model at the "Genome: Unlocking Life's Code" exhibition at the Smithsonian's Natural History Museum (Washington, D.C., USA). Photo copyright A. Mahfouz.

ISBN 978-94-6186-678-3
Published by Uitgeverij BOXPress || Proefschriftmaken.nl

An electronic version of this dissertation is available at: `http://repository.tudelft.nl/`

To my parents, Ola and Essam

CONTENTS

1 Introduction **1**

2 Brain Transcriptome Atlases: A Computational Perspective **9**

3 Visualizing the Spatial Gene Expression Organization in the Brain through Non-Linear Similarity Embeddings **33**

4 Comprehensive isoform analysis characterizes dystrophin function in human brain development **51**

5 Shared Pathways among Autism Candidate Genes determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome **67**

6 Genome-wide co-expression of steroid receptors in the mouse brain: identifying signaling pathways and functionally coordinated regions **85**

7 Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex **103**

8 Discussion **121**

Bibliography **129**

Summary **147**

Samenvatting **149**

Acknowledgements **151**

Curriculum Vitæ **155**

List of Publications **157**

Chapter 1
Introduction

If our brains were simple enough for us to understand them,
we'd be so simple that we couldn't.

Ian Stewart

1

1.1. A BRIEF HISTORY OF BRAIN SCIENCE

The human brain is the complex system responsible for our consciousness, perception, learning and actions. Understanding the biological basis of the human brain and these mental processes is one of the ultimate challenges faced by scientists. Fueled by philosophical and medical questions, the quest to understand the human brain dates back to the early days of human history. The Edwin Smith Surgical Papyrus (Middle Kingdom in Ancient Egypt, ca. 1700 BC) contains the earliest written reference to the brain [1] (Figure 1.1). Likely a practical handbook for a battlefield surgeon, the text notes the pulsations of the cerebral cortex, describes the influence of brain injuries on parts of the body (such as paralysis), as well as the effect of crushing injuries of vertebrae on impaired motor and sensory functions [2, 3]. Hippocrates (460–377 BC) was arguably the first physician to assert that the center of intelligence is the brain and not the heart as others, including Aristotle, believed [4]. Describing the symptoms of epilepsy in children in his book *On the Sacred Disease* he wrote: "It is thus with regard to the disease called Sacred: it appears to me to be nowise more divine nor more sacred than other diseases, but has a natural cause like other affections." With this he states that epilepsy is a brain disorder rather than a curse or a prophetic power, as was previously believed. [5] The roots of modern neuroscience can be traced back to the latter part of the nineteenth century when new tools and techniques boosted our ability to study the structure and function of the mammalian nervous system [6]. The development of a method to stain neurons with silver salts to reveal their entire structure under the microscope by the Italian physician and scientist Camillo Golgi (1843–1926) paved the way for the Spanish anatomist Santiago Ramón y Cajal (1852–1934) to stain individual cells, showing that nervous tissue is not one continuous web but a network of discrete cells [7]. This lead to the formation of the neuron doctrine — the principle that individual neurons are the elementary signaling elements of the nervous system. The Nobel Prize in Physiology or Medicine 1906 was awarded jointly to Camillo Golgi and Santiago Ramón y Cajal "in recognition of their work on the structure of the nervous system" [8].

Despite this long history of studying the mammalian brain, fundamental questions, such as the actual number of cell types in the central nervous system (CNS), remain unanswered to date [9]. This is mainly due to the high complexity of the brain which consists of billions of neuronal and glia cells, organized through brain development into distinct functional populations [10]. In humans and other species, deviations from the normal trajectories of development and aging of the CNS can lead to brain disorders, such as autism spectrum disorders (ASD) and Alzheimer's disease [11].

Rapid developments in neuroimaging and electro/magnetoencelphalography (EEG / MEG) have greatly enhanced our understanding of the human brain function. Magnetic resonance imaging (MRI) is widely used to characterize the structural and functional organization of the human brain. Methods such as diffusion MRI and functional MRI (fMRI) are used to map structural and functional connections in the human brain, providing a crucial foundation for understanding how networks of neurons function and dysfunction in the brain [12–14]. Despite the clinical value of these methods they provide little information on the underlying neurobiological mechanisms.

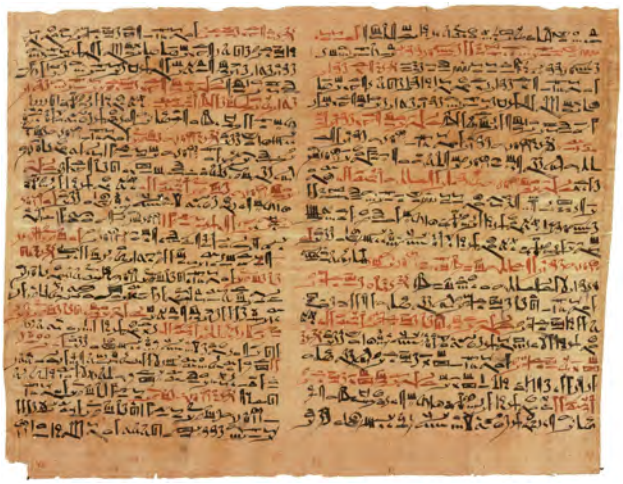


Figure 1.1: **The Edwin Smith Surgical Papyrus** (Middle Kingdom in Ancient Egypt, ca. 1700 BC) [1].

1.2. UNDERSTANDING THE BRAIN THROUGH MOLECULAR MECHANISMS

The high complexity of the brain is largely reflected in the underlying molecular neurobiology of neurons that determine their morphological and electrophysiological properties as well as their connectivity patterns. Variations in the genotype of neurons affect the cognitive and behavioral tasks carried out by the brain through several molecular levels; transcriptomic, proteomic, and epigenomic. Nowadays, several methods are used to study the brain at different levels, including the molecular, cellular, circuit and network levels. At the molecular level, high-throughput technologies such as Next-generation sequencing (NGS) tremendously increases our ability to measure the neuronal molecular profile. This includes various assays such as profiling DNA variations (using exome- and whole genome-sequencing), messenger (mRNA) and micro (miRNA) expression (using microarrays and RNA sequencing – RNA-seq), methylation levels (using bisulfite sequencing), accessible chromatin (using DNase I hypersensitive sites sequencing – DNase-seq), protein-DNA binding and histone modifications (using Chromatin Immunoprecipitation Sequencing – ChIP-seq), long-range interactions (using chromatin conformation capture techniques: 3C, 4C and Hi-C and Chromatin Interaction Analysis by Paired-End Tag Sequencing – ChIA-PET). Proteomic profiling methods have lagged behind due to our inability to amplify amino acids compared to nucleic acids [10]. Despite these limitations, efforts have been made to achieve high-throughput profiling of the human proteome [15, 16] and the brain proteome [17].

Genetic association studies have established a substantial role for genetic etiologies in brain disorders. Evidence for association of generic risk factors to neurological and psychiatric disorders can be revealed through patient-control studies and twin studies in case of heritable risk assessment. Using genome-wide association studies

(GWAS), exome sequencing, and whole genome sequencing, hundreds of variants have been linked to complex neurological disorders, such as autism, schizophrenia, Migraine, and Alzheimer's. Despite these efforts, the identified common and rare variants explain only a small portion of the genetic contribution to brain disorders. Meanwhile, genomic screening in imaging studies including large cohorts of patients have increased the power to detect the influence of genetic variants on brain structure and volume. For example, the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium [18] analyzed the association of single nucleotide variants (SNVs) with the volume of subcortical structures in 30,000 individuals. Yet, linking imaging phenotypes in patient cohorts to the underlying genetic causes remains challenging. A key factor is that genetic variants are identified from blood samples and it is not clear how these variants affect gene regulation in the tissue of interest (brain), due to the scarcity of tissue samples especially from living subjects.

Given the high complexity of the brain, it is important to characterize the different cell types and their distinct molecular profiles across different parts of the brain as well as different developmental stages. This can greatly enhance our ability to identify the mechanisms by which genetic variants affect the brain. Despite recent efforts to sequence the DNA of individual neurons using single-cell sequencing technology [19], it remains yet unfeasible to sequence billions of single cells from the human brain at different stages of development. Alternatively, characterizing the transcriptome of neurons across different brain regions and ages can help us identify the distinct cell populations constituting functional units of the brain. Whole-brain maps of genome-wide gene expression provide an invaluable resource to both neuroscientists and geneticists to study the spatially-localized functional role of genes in normal brain function. Despite not holding information on how diseases and disorders affect gene expression, these atlases provide a rich resource that can be used to study the function of genes in normal conditions. This can offer new insights on how deviations from normal can lead to a CNS disorders and ultimately on how to develop therapies for neurological disorders.

1.3. BRAIN TRANSCRIPTOME ATLASES

Several efforts have been made to build a genome-wide gene expression map of the mammalian brain [20–22]. Yet the most successful of those are made by the Allen Institute for Brain Science. They provide the most comprehensive maps of gene expression in the mouse, macaque, and human brains in terms of the number of genes, the spatial-resolution, and the developmental stages covered [23]. Several atlases have been released that map gene expression in the adult and developing mouse [24, 25] and human [26, 27] brains.

Gene expression (the amount of mRNA) can be measured using a wide range of technologies. Both the adult and the developing Allen Mouse Brain Atlases [24, 25] rely on colorimetric in situ hybridization (ISH) to comprehensively map genome-wide expression throughout the mouse brain. Compared to other high-throughput techniques to measure gene expression, such as microarrays and serial analysis of gene expression (SAGE), ISH can capture the expression of genes at a near cellular level. This is particularly important in a complex structure like the brain with its ultra-high cellular diversity such that neighboring neurons can have different expression profiles. Despite

the power of using ISH to map gene-expression at a cellular resolution, it is not suitable to create a comprehensive map of gene expression across the entire human brain for several reasons. Following the pipeline used to generate the mouse brain atlases, thousands of brains are needed. Moreover, several mouse brain sections can be fit on a single microscope slide, while a single human brain section has to be dissected into multiple subsection before scanning. For these reasons, the Allen Human Brain Atlas [26] and the BrainSpan Atlas [27] rely on high-throughput technologies such as DNA microarrays and RNA-seq of targeted brain regions. These technologies allow quantitative profiling of gene- and exon-level transcriptomes of postmortem tissues extracted from a small number of clinically unremarkable human brains.

Despite these substantial efforts, understanding the role of genetic variation on cognitive and behavioral phenotypes remains challenging due to several factors. These include the complexity of the underlying molecular neurobiology and the poor clinical definition of neurological disorders. In addition, the heterogeneity and high dimensionality of the collected data at different molecular and phenotypic levels requires complex computational models to analyze. Spatial and temporal brain transcriptomes provide an opportunity to solve some of these challenges. They allow associations of genetic variants to specific brain regions and/or or developmental stages. This localization can be used for instance as an intermediate step in linking genetic variables to imaging phenotypes. Instead of analyzing individual genes, spatio-temporal co-expression networks can be used to study the gene regulatory network underlying the structural and functional organization of the brain. Biological networks provide an attractive framework to model interactions between several biomolecules inside the cell. Networks have also been widely used to integrate genetics with transcriptomics, epigenetics and proteomics [28, 29]. However, one fundamental issue is the scale at which to construct a network. Multiple evidence suggests a multi-scale analysis approach for network-based integration of omics data is crucial [30, 31].

1.4. THESIS CONTRIBUTIONS

In this thesis, we study the relationship between gene expression on one hand and the anatomical and functional organization of the mouse and human brains on the other hand in order to gain new insights into the genetic etiology of brain processes and disorders. A better understanding of this relationship in normal and diseased brains can elucidate mechanisms of neurological disorders and is key to develop treatments. We achieve this goal be developing computational methods to analyze brain gene expression atlases in order to answer different biological questions about the brain.

In Chapter 2, we start with a review of the computational challenges presented by brain gene expression atlases and discuss the different methodologies developed to address them. We classify these methods into two main categories. First, a class of methods used to analyze the expression profiles of genes across different brain regions, cell types and developmental stages. Second, methods focusing on the molecular organization and the genetic signature of the brain. In addition, we discuss future perspectives of these methods in terms of potential new approaches to integrate multiple sources of neuro-omics data.

Given the high dimensionality of the brain transcriptomes, there is need for dimen-

sionality reduction methods that can summarize both local and global relationships in the data to allow informative visualizations. In Chapter 3, we quantitatively assess the superiority of t-distributed Stochastic Neighbor Embedding (t-SNE) [32] to classical Multi-Dimensional Scaling (cMDS) and principle component analysis (PCA) in separating neuroanatomical regions in low-dimensional (2D) embeddings of the mouse and human brains. We show the consistency of the low dimensional embedding across 6 human brains of the Allen Human Brain Atlas [26] as well as between the sagittal and coronal sections of the Allen Mouse Brain Atlas [24]. Finally, we used the low-dimensional embeddings to analyze the contribution of different cell-type markers in determining the structural organization of the mammalian brain.

Duchenne and Becker Muscular Dystrophies are X-linked genetic neuromuscular disorders caused by mutations in the *DMD* gene and characterized by severe and progressive muscle weakness. In addition to the muscle pathology, there is high incidence of learning and behavioral problems accompanying both diseases. Yet, the pathophysiology of brain involvement in these disorders remains elusive with a handful of studies analyzing the role of the *DMD* gene in the brain. In Chapter 4, we provide a detailed description of the localization and function of the different isoforms of the *DMD* gene throughout the development of the human brain. Our results provide a first detailed description of the *DMD* gene expression in different regions of the human brain at different stages of development. Moreover, we use co-expression analysis to provide the first genetic link that might explain the comorbidity of neurodevelopmental disorders.

Genetic studies have implicated hundreds of genes in autism spectrum disorder (ASD). However, understanding how these functionally diverse genes can all be associated to ASD has proved challenging. In Chapter 5, we used the Brain Span atlas of gene expression in the developing human brain to identify convergent biological processes between a heterogeneous set of autism-related genes. Using differential co-expression networks of autism-related genes, we show that autism-related genes can be grouped in three modules associated to distinct biological functions during human brain development including synaptogenesis, apoptosis, and GABA-ergic neurons. By building a genome-wide co-expression network from the entire transcriptome, we found that autism-related genes were enriched in modules related to mitochondrial function, protein translation, and ubiquitination. These findings can help our understanding of the disease etiology along with translational work for drug discovery.

In Chapter 6, we tested whether we can identify signaling pathways of steroid receptors through spatial correlation of steroid receptors with genome-wide mRNA expression across different regions in the mouse brain. The ISH-based Allen Mouse Brain Atlas provides us with enough resolution (i.e. enough samples per brain region) to analyze the region-specific co-expression relationships of six nuclear steroid receptors. Using known targets of steroid receptors, we observed high co-expression within brain regions of steroid action. We were able to functionally validate two genes identified as targets of estrogen receptor alpha (*Esr1*) in the hypothalamus; namely *Irs4* and *Magel2*. While the former is a known target of Esr1, Magel2 was previously unknown, highlighting the power of using genome-wide spatial co-expression to identify steroid receptor targets. Furthermore, we provide a method to identify concurrent co-expression between steroid receptors and potential co-regulators in more than one brain region.

The final contribution in the thesis is a chapter on biological data integration. In Chapter 7, we studied the functional role of three dimensional conformation of the genome in the cell nucleus on gene expression regulation. Long-range chromatin interactions arise as a result of the three-dimensional (3D) conformation of chromosomes in the cell nucleus and can result in the co-localization of co-regulated genes. To assess the influence of 3D conformation on gene co-expression, we used chromatin conformation capture (Hi-C) data from the mouse cortex to build a chromatin interaction network (CIN) of genes. We show that by characterizing the topology of the CIN at different scales it is possible to accurately predict spatial co-expression between genes in the mouse cortex.

We conclude the thesis with a discussion of our contributions and potential extensions to our work together with a brief discussion on the future of brain transcriptomes.

CHAPTER 2
BRAIN TRANSCRIPTOME ATLASES: A COMPUTATIONAL PERSPECTIVE

Ahmed Mahfouz
Sjoerd MH Huisman
Boudewijn PF Lelieveldt
Marcel JT Reinders

Submitted.

THE immense complexity of the mammalian brain is largely reflected in the underlying molecular signatures of billions of cells. Brain transcriptome atlases provide a valuable insight into the gene expression patterns across different brain areas throughout development. Such atlases allow researchers to probe the molecular mechanisms which define neuronal identities, neuroanatomy, and patterns of connectivity. Despite the immense effort put into generating such atlases, an even greater effort is needed to develop methods to probe the resulting high dimensional multivariate data in order to answer fundamental questions in neuroscience. We provide a comprehensive overview of the various computational methods used to analyze brain transcriptome atlases. These methods can be grouped into two categories: (1) methods analyzing spatial and temporal expression patterns of gene(s) in the brain and (2) methods analyzing the genetic signatures of anatomical and functional brain regions. We discuss the various methodologies adopted as well the mechanistic insights they provide into neurological processes and disorders. We conclude with a discussion of the contribution of such computational methods as well as directions to improve them, with a focus on integrating data types and how that can further our understanding of the brain at different scales, ranging from molecular to behavioral.

2.1. MAPPING GENE EXPRESSION IN THE BRAIN

The mammalian brain is a complex system consisting of billions of neuronal and glia cells that can be categorized into hundreds of different subtypes. Understanding the organization of these cells, throughout development, into functional circuits carrying out sophisticated cognitive tasks can help us better characterize disease-associated changes. Advances in technology and automation of laboratory procedures have facilitated high-throughput characterization of functional neuronal circuits and connections at different scales [23]. For example, the Human Connectome Project maps the complete wiring of the brain using magnetic resonance imaging [33]. Despite the importance of these imaging modalities in characterizing brain pathologies and development, it is imperative to analyze the molecular structure to gain a better mechanistic understanding of how the brain works. The high complexity of the brain, due to the unknown large number of cell types [9], yields the study of the molecular mechanisms very challenging. Invasive methods such as viral [34] and optogenetic techniques [35] allow functional manipulation of specific cell populations and can potentially lead to the development of cell-type targeted therapeutics.

Characterizing the molecular profile of all the cells across the brain can greatly enhance our understanding of brain function and disease. Ultimately, sequencing all the brain cells and mapping their gene, protein and metabolic expression levels will allow in depth investigation of the role of genomic variation on cell function. The complexity of the brain is largely reflected in the underlying patterns of gene expression that defines neuronal identities, neuroanatomy, and patterns of connectivity. Several experimental approaches have been used to characterize gene expression of different neuronal cell types in the brain including: microarrays [36], RNA-sequencing [37], serial analysis of gene expression (SAGE) [38], bacterial artificial chromosome (BAC) transgenesis [20], In situ hybridization (ISH) [24], and most recently single cell sequencing [39]. With 80% of the 20,000 genes in the mammalian genome expressed in the brain [24], characteriz-

ing spatial and temporal gene expression patterns can provide valuable insights into the relationship between genes and brain function and their role throughout neurodevelopment.

Following earlier progress in other model organisms [40–42], several projects have assessed gene expression in the mouse brain with various degrees of coverage for genes, anatomical regions, and developmental time points [9, 23]. In rodents, the Gene Expression Nervous System Atlas (GENSAT) [20, 21] and GenePaint [43] mapped gene expression in both the adult and developing mouse brain, while the e-Mouse Atlas of Gene Expression (EMAGE) [22] and EurExpress [44] focused on the developing mouse brain. Comparable atlases of gene expression in the human brain are far less abundant due to the challenges posed by difference in size between the human and mouse brain as well as the scarcity of post-mortem tissue. However, several studies have profiled the human brain transcriptome to analyze expression variation across the brain [45], expression developmental dynamics [46–48] and differential expression in the autistic brain [49], albeit in a limited number of coarse brain regions. The Allen Institute for Brain Science provides the most comprehensive maps of gene expression in the mouse and human brain in terms of the number of genes, the spatial-resolution, and the developmental stages covered [23]. Several atlases have been released which map gene expression in the adult and developing mouse brain [24, 25], the adult and developing human brain [26, 27], and the developing non-Human Primate (NHP) brain; Figure 2.1. Sunkin *et al*.. provides a complete review of the Allen Brain Atlas resources [50].

The availability of genome-wide spatially-mapped gene expression data provides a great opportunity to understand the complexity of the mammalian brain. It provides the necessary data to decode the molecular functions of different cell populations and brain nuclei. However, the diversity of cell types and their molecular signatures and the effect of mutations on the brain remain poorly understood. For example, de novo loss-of-function mutations in autistic children have been shown to converge on three distinct pathways: synapse, Wnt signaling and chromatin remodeling [51, 52]. Except for the synaptic role of autism-related genes, it is not clear how alternations in basic cell functions such as Wnt signaling and chromatin remodeling can result is the complex phenotype of autism spectrum disorders (ASD). A recent effort to map somatic mutations in cortical neurons using single-cell sequencing has shown that neurons have on average 1,500 transcription-associated mutations [19]. The significant association of these single-neuron mutations and genes with cortical expression indicates the vulnerability of genes active in human neurons to somatic mutations, even in normal individuals. Efforts to understand genotype-phenotype relationships in the brain face several challenges including the complexity of the underlying molecular mechanisms and the poor definition of clinically-based neurological disorders. In addition, the high dimensionality of the data yields most studies underpowered to detect any associations. This is especially true in the case of testing genetic associations with phenotype markers, such as imaging measurements [53]. A combination of efforts to map the genomic landscape of the brain, and data-driven approaches can add to our understanding of the underlying genetic etiology of neurological processes and how they are altered in neurological disorders.

Several review articles provide extensive insights into the gene expression maps of

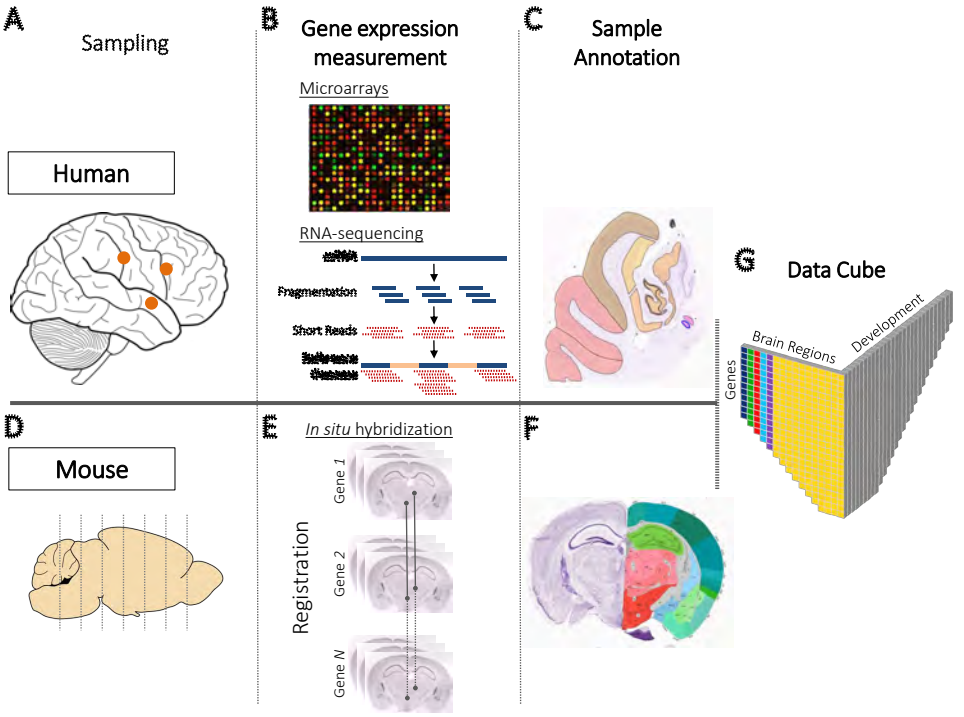


Figure 2.1: **Spatially-mapped gene expression in the mammalian brain.** To map gene expression across the human and mouse brains, the Allen Institute for Brain Sciences followed two different strategies. In the human brain, samples covering all brain regions are extracted (A) and gene expression is measured using either microarray or RNA-sequencing (B). Accompanying histology sections and MRI scans are acquired to localize samples. Manual delineation of anatomical regions on the histology sections allowed for accurate sample annotation (C). In the mouse brain, gene expression is measured in coronal and sagittal sections using in situ hybridization (D). Several slices covering the mouse brain are extracted per gene. Image registration methods are used to align the set of sections acquired for each gene to a common reference atlas (E). Anatomical regions are delineated on the reference atlas allowing for sample annotation (F). Data from the mouse and human atlases can be represented in a data matrix of three dimensions representing: genes, brain regions and developmental stages (in case of the developmental atlases) (G).

the brain. French and Pavlidis [54] provide a global overview of neuroinformatics including ontology, semantics, databases, connectivity, electrophysiology, and computational neuroscience. Jones *et al*. gave an overview on developing the mouse atlas, the challenges faced, the community reaction, limitations, and atlas usage examples, as well as the data mining tools provided by the Allen institute [55]. Pollock *et al*. provide a detailed review of the technology and tools which are currently advancing the field of molecular neuroanatomy [23]. Recently, Parikshak *et al*. illustrated the power of using network approaches to leverage our understanding of the genetic etiology of neurological disorders [11]. Yet, a global overview of the computational methodologies applied to brain transcriptome atlases to increase our understanding of neurological processes and disorders remains missing.

In this review, we provide an overview of the computational approaches used to expand our understanding of the relationship between gene expression on one hand and the anatomical and functional organization of the mammalian brain on the other hand. We focus our discussion on spatial and temporal brain transcriptomes mapped by the Allen Institute for Brain Sciences. Nevertheless, we also discuss how the methods can be extended to epigenomes and proteomes of the brain and other human tissues. We describe the different computational approaches taken to analyze the high-dimensional data and how they have contributed to our understanding of the functional role of genes in the brain, molecular neuroanatomy, and genetic etiology of neurological disorders. Finally, we discuss how these methods can help solve some of the data-specific challenges, and how the integration of several data types can further our understanding of the brain at different scales, ranging from molecular to behavioral.

2.2. COMPUTATIONAL ANALYSIS OF SPATIAL AND TEMPORAL GENE EXPRESSION DATA IN THE BRAIN

Spatio-temporal transcriptomes of the brain pose several challenges due to their high-dimensionality. In this section we identify the different types of approaches taken to analyze the spatially-mapped gene expression data. We show the strengths of each approach and demonstrate how it enriched neuroscience research. We divide the different methods into two categories. First, we describe a class of methods used to analyze the expression profile of gene(s) across different brain regions, cell types and developmental stages. Second we discuss methods focusing on the molecular organization and the genetic signature of the brain.

ANALYZING THE EXPRESSION PATTERNS OF GENE(S) IN THE BRAIN

Mapping gene expression across the brain is very helpful in determining the neural function of a gene of interest by associating it to a specific brain region and/or developmental stage or in identifying genetic markers of those brain regions and developmental stages. Brain transcriptome atlases, such as the Allen Brain Atlases, provides useful information about the expression of a gene under "normal" conditions. Such information can be used to further direct in depth studies about a specific gene in biologically/clinically relevant cohorts. With the increasing number of genes implicated in neurological diseases as well as the realization that complex phenotypes of the brain likely result from the combined activity of several genes, several studies analyze gene sets rather than individual candidate-genes. By studying the expression of a gene set rather than a single gene, neuroscientists are faced with a challenge on how to summarize this data in order to understand the relationship between genes and neuronal phenotypes.

GENE EXPRESSION VISUALIZATION

High throughput data visualization approaches can facilitate the exploration of complex patterns in multivariate high-dimensional gene expression data sets [56]. For example, heatmaps are commonly used to visualize gene expression levels across a set of samples using a two-dimensional false-color image (Figure 2.2F). However, techniques like heatmaps are not ideal to represent brain transcriptomes because they fail to capture

the multivariate nature of the data (genes, samples, and time-points) and to represent the inherent spatial and temporal relationships between different brain regions and developmental stages, respectively. In order to acquire high resolution gene expression maps, the Allen atlases of the developing and adult mouse brain rely of ISH images (Figure 2.2A). The Brain Explorer 3D viewer [57] is an interactive desktop application that allows the visualization of the 3D expression of one or more genes with the possibility to link them back to the high resolution ISH images [50] (Figure 2.2B). ISH images can be synchronized between different genes and also with the anatomical atlas of the mouse brain (Figure 2.2C), facilitating the analysis of a group of genes. For the adult and developing human atlases, the gene expression data (microarray or RNA-seq) is mainly visualized using heatmaps (Figure 2.2D). In the adult human atlas, the expression data can also be visualized on top of the magnetic resonance images (Figure 2.2E). The Brain Explorer 3D viewer [57] can be used to visualize gene expression from cortical samples using an inflated cortical surface, a surface-based representation of the cortex that allows better representation of the relative locations of laminar, columnar, and areal features (Figure 2.2F). In addition, gene expression can be mapped to an anatomical representation of the brain to facilitate interpretation (Figure 2.2G). French *et al*. developed a pipeline to map the expression of any gene from the Allen Human brain atlas to the cortical atlas built into the FreeSurfer software, which shall facilitate integration with medical imaging studies [58]. Similarly, Ng *et al*. developed a method to construct surface-based flatmaps of the mouse cortex that enables mapping of gene expression data from the Allen Mouse Brain Atlas [59].

SUMMARY STATISTICS AND VISUALIZATION-BASED METHODS

Early studies employing the Allen Brain Atlases used a variety of visualization and qualitative measurements to analyze the expression of gene sets associated with consummatory behavior in the mouse brain [60], changes in locomotor activity in the mouse brain [61], midbrain dopaminergic neurons [62], and dopamine neurotransmission [63]. Kondapalli *et al*. used a similar qualitative approach to analyze the expression of Na+/H+ exchangers (NHE6 and NHE9), which are linked to several neuropsychiatric disorders, in the adult and developing mouse brain atlases [64].

In order to provide better quantitative representations of the expression of gene sets, several studies relied on basic summary statistics, such as the mean, standard deviation and summation. Zaldivar *et al*. used summations to summarize the expression of cholinergic, dopaminergic, noradrenergic, and serotonergic receptors in the amygdala, and in neuromodulatory areas [65]. By plotting the average expression of genes harboring de novo loss-of-function mutations identified by means of exome sequencing across human brain development, Ben-David and Shifman identified two clusters with antagonistic expression patterns across development [66]. Dahlin *et al*. developed their custom score (expression factor) of gene expression in the mouse brain based on the ISH images of the Allen Mouse Brain Atlas [67]. They computed the mean and the standard deviation of the expression factor to assess the global expression and heterogeneity of solute carrier genes, respectively. To deal with the qualitative ISH-based expression data from the Allen Mouse Brain Atlas, Roth *et al*. used a non-parametric representation of the data (using ranks instead of raw expression values) to study the relationship between genes associated with grooming behavior in mice and 12 major brain structures [68].

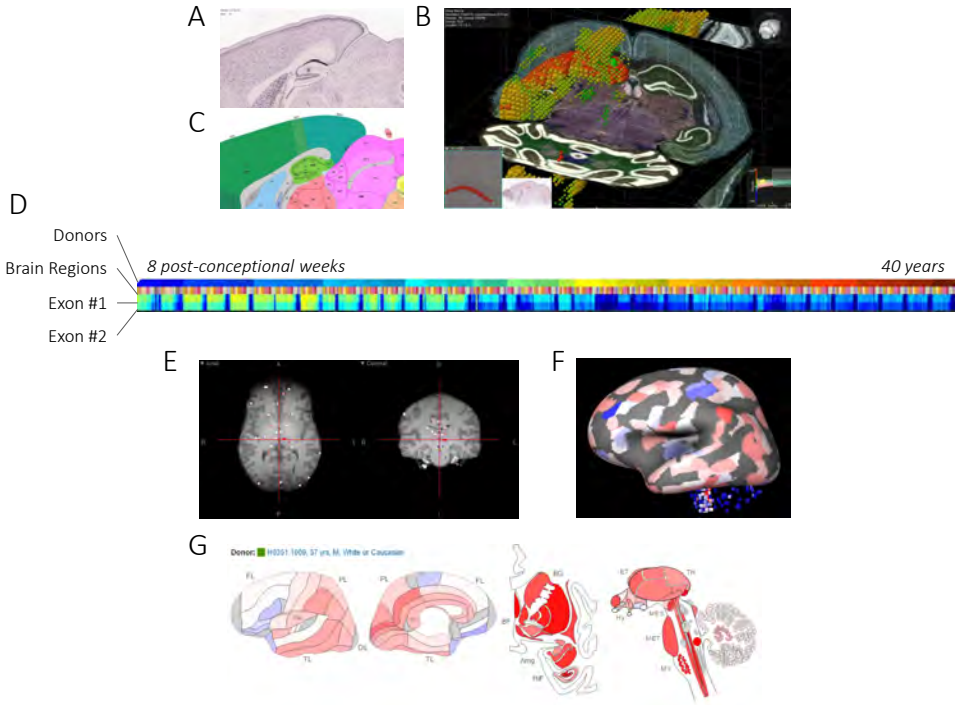


Figure 2.2: **Gene expression visualization.** Gene expression of spatially-mapped samples can be visualized using several approaches. (A) The mouse gene expression data of the gene Man1a can be investigated using the original ISH sections. (B) The BrainExplorer software allows visualization of the 3D expression volume with an overlay of the anatomical atlas and the ability to go back to the original high-resolution ISH section. (C) Simultaneously viewing the ISH section and the corresponding atlas section helps in localizing gene expression to brain regions. (D) Heat-maps are commonly used to visualize gene expression. Expression of the two exons of the NEUROD6 gene from the BrainSpan Atlas are visualized using a heat-map in which samples are ordered according to the age of the donor. (E) Samples from the Allen Human Brain Atlas are associated with coordinates of their location in the corresponding brain MRI. (F) Using the BrainExplorer, expression values of MECP2 can be mapped to an inflated white matter surface for better visualization of the cortex. (G) Alternatively, expression values can be mapped on an anatomical atlas of the human brain.

Most of the studies analyzing gene expression in the brain focused mainly on scores describing the expression of a gene or a gene set within each brain region of interest. Liu *et al*. [69] proposed a characterization of the stratified expression pattern of sonic hedgehog (*Shh*), a classical signal molecule required for pattern formation along the dorsal-ventral axis, and its receptor *Ptch1*. Using a combination of differential expression, transcription factor motif analysis and ChIP-seq, they identified the role of *Gata3*, *Fox2*, and their downstream targets in pattern formation in the early mouse brain. These results illustrate the power of characterizing complex expression patterns across the brain rather than the solely summarizing the expression of each gene within individual brain regions.

Box 1 | Gene Sets

Complex biological functions and disorders usually involve several rather than a single gene. Gene sets are groups of genes that share common biological functions that can be defined either based on prior knowledge (e.g. about biochemical pathways or diseases) or experimental data (e.g. transcription factor targets identified using CHIP-seq). Gene set databases organize existing knowledge about these groups of genes by arranging related genes in a set where each set is associated with a functional term, such as a pathway name or a transcription factor that regulates the genes. Gene set databases can be classified into five types of sets:

Gene Ontology (GO)

The Gene Ontology project [70] developed three hierarchically structured vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions. Genes annotated with the same GO term(s) constitute a gene set.

Biological Pathways

Biological pathways are networks of molecular interactions underlying biological processes. Pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [71] and REACTOME [72], catalog physical entities (proteins and other macromolecules, small molecules, complexes of these entities and post-translationally modified forms of them), their subcellular locations and the transformations they can undergo (biochemical reaction, association to form a complex and translocation from one cellular compartment to another).

Transcription

Transcription databases include information on transcription regulation of genes by transcription factors (TFs) binding to the DNA or post-transcriptional regulation by microRNA binding to the mRNA. Determining these physical interactions can be done either in silico using computational inference (motif enrichment analysis) or using experimental data such as (ChIP-seq and microRNA binding data). For the motif enrichment analysis, position weight matrices (PWMs) from databases TRANSFAC [73] and JASPER [74] can be used to scan the promoters of genes in the region around the transcription factor start site (TSS). ChIP-seq data, such as the large collection of experiments from the Encyclopedia of DNA Elements (ENCODE) project [35] and the Roadmap Epigenomics consortium [75], is used to identify genes targeted by the TFs. Similarly, microRNA targets can be extracted from databases such as TargetScan [76].

Cell-type markers

Cell type-specific transcriptional data provide a very rich source of cell type marker genes. Genes are identified as a cell type marker if they are up-regulated in one cell population compared to other cell populations. Several studies have used microarrays and RNA-seq to profile the transcriptome of several neuronal cell types [36, 37]. Recently, studies are using single-cell sequencing to precisely capture the transcriptome of individual neuronal cells [39, 77].

Disease

Genes can be grouped into sets based on their association to the same diseases. Public databases, such as OMIM [78] and DisGeNet [79], contain curated information from literature and public sources on gene-disease association. Another source to obtain disease-related gene sets is by identifying genes harboring variants identified using GWAS [80, 81], exome-sequencing [82], or whole-genome sequencing.

IDENTIFYING GENES WITH LOCALIZED EXPRESSION PATTERN

The complexity of the brain implies that genes are involved in more than one function and that their function is region- or cell-type-specific. Neuronal cell types have been classically defined using cell morphology, electrophysiological and connectivity properties. Similarly, classical neuroanatomy identifies regions based on their cyto-, myelo-, or chemo-architecture. Genomic transcriptome measurements provide an alternative route to define functional cell types and brain regions based on their genetic makeup.

Several studies have analyzed the ISH-based gene expression data of the Allen Mouse Brain Atlas in order to identify cell-type specific genes and genes with localized gene expression. Li *et al*. identified cell-type-specific genes using scale-invariant feature transform (SIFT) features of the ISH images [83]. They further classified genes, using a supervised learning approach (regularized learning), based on their expression in different brain cell-types. Similarly, Kirsch *et al*. [84] described an approach to identify genes with a localized expression pattern in a specific layer of the mouse cerebellum. They represented each ISH image (gene) by using a histogram of local binary patterns (LBP) at multiple-scales. Predicting the localization of genes to each of the four cerebellar layers is done using two-level classification. First they used a support vector machine (SVM) classifier to assign a cerebellar layer to each image and then used multiple-instance learning (MIL) to combine the resulting image classification into gene classification. At the brain regions level, David and Eddy developed ALLENMINER [85], a tool that searches the Allen Mouse Brain Atlas for genes with a specific expression pattern in a user-defined brain region. More application specific methods include the identification of genetic markers of the ventromedial hypothalamus [86] and the localization of human age-related gene expression changes in different neuronal cell types [87].

More recently, Ramsden *et al*. [88] studied the molecular components underlying the neural circuits encoding spatial positioning and orientation in the medial entorhinal cortex (MEC). They developed a computational pipeline for automated registration and analysis of ISH images of the Allen Mouse Brain Atlas at laminar resolution. They showed that while very few genes are uniquely expressed in the MEC, differential gene expression defines its borders with neighboring brain structures, and its laminar and dorsoventral organization. Their analysis identifies ion channel-, cell adhesion- and synapse- related genes as candidates for functional differentiation of MEC layers and for encoding of spatial information at different scales along the dorsoventral axis of the MEC. Finally, they reveal laminar organization of genes related to disease pathology and suggest that a high metabolic demand predisposes layer II to neurodegenerative pathology.

SPATIAL AND TEMPORAL GENE CO-EXPRESSION

Genes with similar expression patterns over a set of samples are said to be co-expressed and are more likely to be involved in the same biological processes (guilt by association) [89]. Applying the same approach to brain transcriptomes can identify co-expressed genes based on their spatial and/or temporal expression across the brain. This can serve as a powerful tool to characterize genes with respect to their context-specific functions.

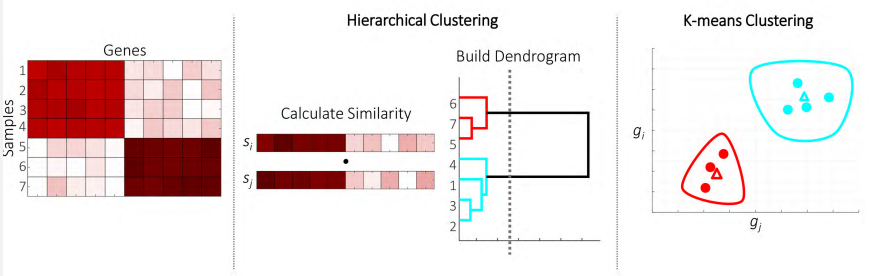
Box 2 | Dimensionality Reduction

The high dimensionality of transcriptomes, and other biological data (e.g. proteomes, epigenomes, etc.), provides a challenge for visualization as well as for selecting informative features for clustering and classification. Dimensionality-reduction approaches aim at finding a smaller number of features that can adequately represent the original high dimensional data in a lower dimensional space. The conventional principal component analysis (PCA) is the most commonly used dimensionality reduction method. Despite its utility, PCA can only capture linear rather than non-linear relationships, which are inherent in many biological applications. Several non-linear dimensionality reduction techniques have been proposed (e.g. Isomap [90]), see Lee and Verleysen [91] for an extensive review. The t-distributed stochastic neighbor embedding (t- SNE) method [32] has been widely used to visualize biological data in two dimensions by preserving both the global and local relationships between the data points in the high-dimensional space [92].

Several similarity/distance measurements have been used to characterize the similarity in spatial/temporal expression patterns between a pair of genes. Of these, correlation-based measures are mostly used to assess gene co-expression patterns across the brain. NeuroBlast [93] is a search tool developed by the Allen Institute for Brain Sciences to identify genes with a similar 3D spatial expression to that of a gene of interest in a given anatomical region, based on Spearman's correlation. Figure 2.3A shows an example of the obtained correlations of estrogen receptor alpha (*Esr1*) in the mouse hypothalamus. The ISH sections in Figure 2.3B shows that correlation can effectively be used to identify genes functional association with *Esr1*. For example, the top correlated gene to *Esr1* in the hypothalamus is insulin receptor substrate 4 (*Irs4*), a target gene of *Esr1* associated with sex specific behavior [94]. NeuroBlast was used to identify genes with a similar expression profile to Wnt3a, a ligand in the Wnt signaling pathway, in the developing mouse brain and identified eight Wnt signaling genes among the top correlated genes [25]. Using Spearman's correlation coefficient, French *et al*. [95] analyzed gene-pairs with positive and negative co-expression in the mouse brain. By focusing on genes with a strong negative correlation, they showed that variation in gene expression in the adult normal mouse brain can be explained as reflecting regional variation in glia to neuron ratios, and is correlated with degree of connectivity and location in the brain along the anterior-posterior axis. Tan *et al*. [96] extended the analysis to the adult human brain and identified conserved co-expression patterns between the mouse and the human brain. In order to characterize the role of SNCA, a gene harboring a causative mutation for Parkinson's disease, Liscovitch and French [97] analyzed the co-expression relationships of SNCA in the adult and developing human brain. They identified a negative spatial co-expression between SNCA and interferon-c (IFN-c) signaling genes in the normal brain and a positive co-expression in postmortem samples from Parkinson's patients, suggesting an immune-modulatory role of SNCA that may provide insight into neurodegeneration.

Box 3 | Clustering

Clustering is the unsupervised learning process of identifying distinct groups of objects (clusters) in a dataset [98]. There are two main types of clustering: hierarchical and partitional. Hierarchical clustering algorithms start by calculating all the pair-wise similarities between samples and then building a dendrogram by iteratively grouping the most similar sample pairs. By cutting the tree at an appropriate height, the samples are grouped into clusters. On the other hand, partitional clustering optimizes the number of simple models to fit the data. Examples of partitional clustering include k-means, Gaussian mixture models (GMMs), density-based clustering, and graph-based methods.



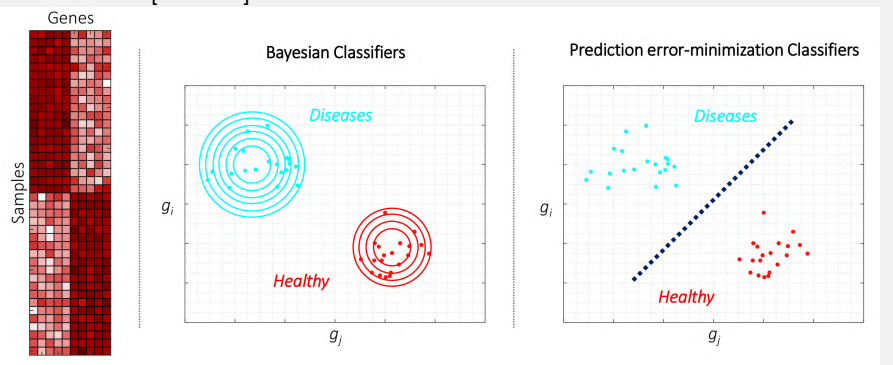
In order to cluster the samples hierarchically, all the pair-wise similarities between sample Si and Sj are calculated. Samples are then grouped iteratively based on the calculated similarities (grouping the most similar first). Once the full dendrogram is built, a cut-off (dashed line) is used to group samples into two groups. For k-means we set the number of clusters to two based on the data heatmap. K-means groups samples by minimizing the within-cluster sum of square distances between each point in the cluster to the cluster center.

Gene co-expression can serve as a very powerful tool for in silico prediction and prioritization of disease genes, by identifying genes with similar expression pattern to known disease genes. Piro *et al.* [99] described a candidate gene prioritization method using the Allen Mouse Brain Atlas. They showed that the spatial gene-expression patterns can be successfully exploited for the prediction of gene–phenotype associations by applying their method to the case of X-linked mental retardation. By extending their methods to the human brain atlas, they showed that spatially mapped gene expression data from the human brain can be employed to predict candidate genes for Febrile seizures (FEB) and genetic epilepsy with febrile seizures plus (GEFS+) [100]. Both examples illustrate the power of using computational approaches to prioritize disease genes before carrying out empirical analysis in the lab.

In measuring gene co-expression, correlation-based methods are not specific to spatially-mapped expression data and hence do not fully model the complexity of the brain transcriptomes. In order to identify gene-pairs with similar expression patterns in the adult mouse brain based on the ISH images, Liu *et al.* [101] compared three image similarity metrics: a naïve pixel-wise metric, an adjusted pixel-wise metric, and a histogram- row-column (HRC) metric. They showed that HRC performs better than voxel-based methods, indicating the superiority of methods that capture the local structure in spatially-mapped data. Miazaki and Costa [102] used Voronoi diagrams to measure the similarity of the density distribution between gene expressions in the adult

Box 4 | Classification

Classification is a supervised learning process of labeling unseen objects (test set) given a set of labeled objects (training set) [98]. Classification approaches can be divided into Bayesian methods and prediction error minimization methods. The former group is based on Bayesian decision theory and uses statistical inference to find the best class for a given object. Bayesian methods can be further divided into parametric classifiers (e.g nearest-mean classifier and Hidden Markov Models) and non-parametric classifiers (e.g. Parzen window or k-nearest neighbor classifier). Alternatively, classifiers can be designed to minimize a measure of the prediction error. Famous classifiers in this category include: regression classifiers (e.g. Lasso regression), support vector machines, decision trees and artificial neural networks. Neural networks (Deep Learning), have become very successful in solving problems in a wide range of applications, including bioinformatics [104–106].



A low dimensional embedding of the samples is generated using two features (genes). Baysian Classifiers assigns each sample to one of the two classes (Diseases or Healthy) based on statistical inference. Prediction error-minimization classifiers updates the classification boundary (dashed line) based on the prediction error and terminates when a certain criteria is met.

mouse brain. Inspired by computer vision algorithms, Liscovitch *et al.* [103] used the similarity of scale-invariant feature transform (SIFT) descriptors of the ISH images of the mouse brain to predict the gene ontology (GO) labels of genes.

GENE CO-EXPRESSION NETWORKS

As we have shown, the guilt by association paradigm has been successfully employed to identify pairs of spatially co-expressed genes sharing the same function, based on various similarity measures. To extend this notion beyond gene pairs, clustering and network-based approaches are used to identify molecular interaction networks of a group of genes that signal through similar pathways, share common regulatory elements, or are involved in the same biological process. Co-expression networks avoid the problem of relying on prior knowledge, such as protein-protein interactions and pathway information, which are valuable but incomplete. Co-expression networks are widely used to identify disrupted molecular mechanisms in cancer and aging [107–109].

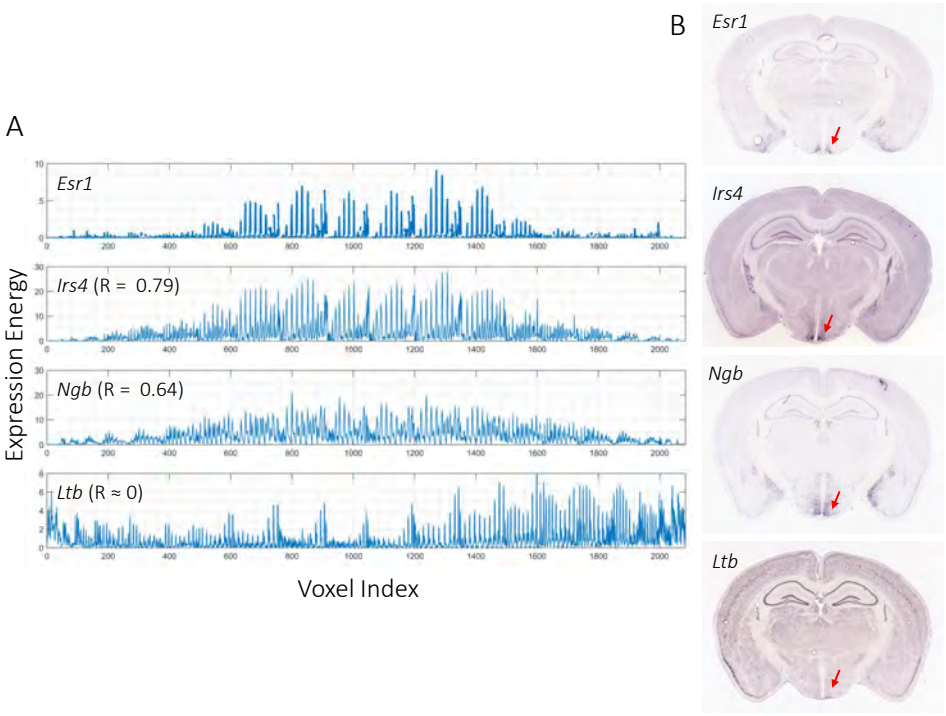


Figure 2.3: **Spatial gene co-expression in the mouse brain.** (A) Expression energy profiles of voxels in the hypothalamus region of the mouse brain using the same linear ordering. The estrogen receptor alpha (*Esr1*) gene shows high expression in the hypothalamus. The expression patterns of *Irs4* and *Ngb* are highly correlated with that of *Esr1* ($R = 0.79$ and $R = 0.64$, respectively). On the other hand, the expression pattern of *Ltb* is not correlated with that of *Esr1* ($R = 8.01 \times 10^{-4}$). Correlation is calculated using Pearson's correlation. (B) *Esr1* and its highly correlated genes (*Irs4* and *Ngb*) are highly expressed in the hypothalamus (red arrow) while *Ltb* is not.

Hierarchical clustering is a widely used unsupervised approach to identify groups of co-expressed genes across a set of samples. Using hierarchical clustering, Gofflot *et al.* [110] identified the functional networks of nuclear receptors based on their global expression across different regions of the mouse brain. By focusing on subsets of brain structures involved in specialized behavioral functions, such as feeding and memory, they elucidated links between nuclear receptors and these specialized brain functions that were initially undetected in a global analysis. Dahlin *et al.* [67] used hierarchical clustering to explore potential functional relatedness of the solute carrier genes and anatomic association with brain microstructures.

Another approach to unsupervised clustering is to use gene co-expression relationships to construct a co-expression network where nodes are genes and edges represent the similarity of the expression profile of those genes. Weighted gene co-expression network analysis (WGCNA) [111] is a commonly used method to construct modules of co-regulated genes based on the topological overlap between genes in a weighted co-

Box 5 | Co-expression Measurements

Gene co-expression is widely used for functional annotation, pathway analysis, and the reconstruction of gene regulatory networks. Co-expression measurements assess the similarity between a pair of gene expression profiles by detecting bivariate associations between them. These co-expression measurements can be summarized in five categories [114, 116–118]:

Correlation The most widely used co-expression measure is Pearson correlation, due to its straightforward conceptual interpretation and computational efficiency. However, Pearson correlation can only capture linear relationships between variables. Alternatively, Spearman correlation is a measure of non-linear monotonic associations. Other correlation-based methods include Renyi correlation, Kendall's rank correlation, and Bi-weight mid-correlation [119].

Partial correlation Partial correlation is used to measure direct relationships between a pair of variables, excluding indirect relationships. These conditional dependencies are used in Gaussian graphical models, and can be calculated using the precision matrix (the inverse of the covariance matrix). Note that partial correlations can only be calculated with more samples than variables unless a regularized estimate is used, such as the graphical lasso [120].

Mutual-Information Mutual information-based methods measure general statistical dependence between two variables rather than a specific type of bivariate association. Based on information theory, mutual information does not assume monotonic relationships and hence can capture non-linear dependencies. Recently, Maximal Information Coefficient has been proposed as non-parametric way of estimating MI.

Other measures Other geometric measures of co-expression are Euclidean distance, cosine similarity, and distance covariance. The Kullback-Leibler divergence and Hoeffding's D are probabilistic measures, just like mutual information. In Bayesian networks, relationships between genes are modelled as causal, directed links.

expression network. WGCNA has been widely used to identify transcription networks in the mammalian brain. Oldham *et al.* [112] demonstrated the first application of WGCNA to examine the conservation of co-expression networks between the human and chimpanzee brains. They found that module conservation in cerebral cortex is significantly weaker than module conservation in sub-cortical brain regions, which is in line with evolutionary hierarchies. WGCNA has been applied to identify modules of co-regulated genes in the developing human brain transcriptome [47], the developing rhesus monkey brain [113], the developing mouse brain [25], the prenatal human cortex [27] and the adult human brain [26], see Figure 2.3B. The identified modules provide a valuable insight into the molecular organization of the brain by identifying modules reflecting primary neural cell types and molecular functions. For example, modules constructed based on the prenatal human cortex correspond to cortical layers and age while no areal patterning was observed [27]. There are numerous technical considerations to take into account while constructing co-expression networks that go beyond the scope of this review [114, 115].

CO-EXPRESSION OF DISEASE RELATED GENES

Complex neuropsychiatric and neurological disorders involve dysregulation of multiple genes, each conferring small but incremental risk, which potentially converging in deregulated biological pathways or cellular functions. Using genome-wide association studies (GWAS), exome sequencing, and whole genome sequencing (WGS), hundreds of variants have been linked to complex neurological disorders, such as autism [52, 121–125], schizophrenia [126, 127], Migraine [128], and Alzheimer's [129, 130]. With the increasing numbers of samples included in these studies, the number of variants associated to each disease is set to increase [51]. Gene co-expression networks provide a framework to identify the underlying molecular mechanisms on which these variants converge. Ben-David and Shifman [66] analyzed co-expression networks of genes affected by common and rare variants in autism using WGCNA. Menashe *et al.* [131] used the cosine similarity of expression profiles to build a co-expression network of autism-related genes in the mouse brain. Both studies provide an important link between gene networks associated with autism and specific brain regions. However, for neurodevelopmental disorders such as autism and schizophrenia, it is more beneficial to study when and where autism genes are expressed during brain development. Gulsuner *et al.* [132] studied the transcriptional co-expression of genes harboring de novo mutations in schizophrenia patients using the BrainSpan atlas of the Developing Human Brain. Parikshak *et al.* [133] used WGCNA to identify modules of co-expressed genes during human brain development using the BrainSpan atlas. They identified modules with significant enrichment in autism-related genes (Figure 2.4). Willsey *et al.* [134] used the BrainSpan atlas to generate co-expression networks around nine genes harboring recurrent de novo loss-of-function mutations in autism probands. Mahfouz *et al.* [135] used a combination of differential and genome-wide co-expression analysis to identify shared pathways among autism-related genes.

Using gene co-expression networks to study relationships between disease-related genes is a valuable approach to understand disease mechanisms. In addition, using networks facilitates the integration of different types of interactions between genes, including but not limited to: co-expression, protein-protein interactions, and literature-based interactions. This can be very useful to our understanding of the etiologies of complex neurological diseases at different levels. In a recent study, Hormozdiari *et al.* [29] integrated gene co-expression based on the BrainSpan atlas and protein-protein interaction (PPI) networks to identify networks of genes related to autism and intellectual disability. For a review on using gene networks to investigate the molecular mechanisms underlying neurological disorders we refer to Gaiteri *et al.* [136] and Parikshak *et al.* [11].

ANALYZING GENETIC SIGNATURE OF BRAIN REGIONS

Spatially-mapped gene expression data allows the exploration of neuroanatomy from a molecular point of view. Individual genes with spatially differential expression have long been used to define the structural organization of the brain and to break it down into regions and sub-regions. Genes have also been used to identify different classes of neuronal cell types. Studying the "genetic signature" of different brain regions can be useful for a multitude of applications. Spatially-mapped gene expression data allows the analysis of the similarity between brain regions in terms of their expression profiles. Regions

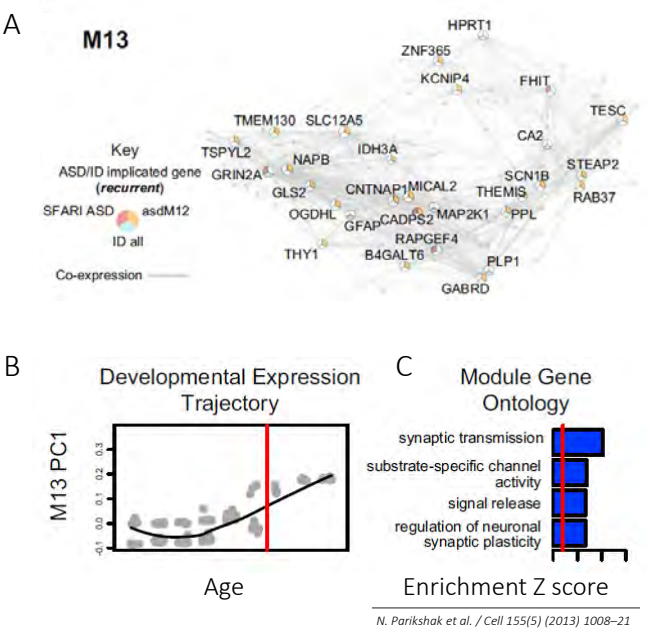


Figure 2.4: **Gene co-expression networks.** (A) Module M13 of co-expressed genes from Parikshak *et al.* [133][1]. The shown module is significantly enriched in autism-related genes. The shown network comprises the top 200 connected genes (highest correlation) and their top 1,000 connections in the subnetwork (also ordered on correlation). Genes are labeled if they are members of relevant gene sets. (B) The pattern of gene expression of genes in the shown module is summarized using the first principal component (eigengene). The red line indicates birth. (C) Gene Ontology terms enriched in the shown module. The blue bars indicate relative enrichment compared to all cortex-expressed genes in terms of Z-score. The red line indicates $Z = 2$.

sharing an expression profile are likely to be involved in the same neuronal functions or be part of the same neuronal circuit. Moreover, studying the expression profiles of functionally and anatomically connected structures provides valuable insights into the molecular basis of brain connectivity.

VOXEL-BASED SIMILARITY AND SPATIAL CLUSTERING

Each of the Allen Brain Atlases assigns a spatial location to each sample, allowing the exploration of the structural organization of the brain based on spatial similarity between different brain regions across the expression of thousands of genes. The Anatomic Gene Expression Atlas (AGEA) [137] is a web-based tool to calculate voxel-wise correlations based on gene expression in the adult and developing mouse brain atlases. To show the

[1]Reprinted from Cell, 155/5, Neelroop N. Parikshak, Rui Luo, Alice Zhang, Hyejung Won, Jennifer K. Lowe, Vijayendran Chandran, Steve Horvath, Daniel H. Geschwind, Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism, 1008-1021, Copyright (2016), with permission from Elsevier.

value of using gene expression patterns to study anatomical organization, Dong *et al.* [138] used AGEA to identify three distinct functional domains in the CA1 region of the mouse hippocampus. Hawrylycz *et al.* [139] used AGEA to show that a consistent expression based organization of areal patterning in the mouse cortex exists when clustered on a laminar basis.

Voxel correlation maps, such as those obtained by AGEA, can be used to cluster the mouse brain voxels into regions with similar gene expression profile. To analyze whether anatomically delineated regions, as defined classically, can also be distinguished based on their expression profile, Bohland *et al.* [140] clustered the adult mouse brain voxels based on the similarity of their expression profiles. Using k-means clustering, they showed that their parcellations are quantitatively similar to the classically-defined neuroanatomical atlas. These results show that the spatially-mapped gene expression data can be very valuable in identifying the molecular basis of brain organization.

In order to identify which genes are responsible for brain organization, Ko *et al.* [141] used a similar approach to cluster brain voxels based on their expression of gene markers of different cell types. Their results show that the neuroanatomical boundaries within a mouse brain can be defined by the clustering of only 170 neuron-specific genes. To identify the driving mechanism of spatial co-expression of genes in the brain, Grange *et al.* [142] modeled co-expression patterns based on the spatial distribution of underlying cell types. Their model can be used to estimate cell-type specific maps of the mouse brain and to identify brain regions based on their genetic signatures. The model proposed in [142] was used to estimate the similarity between the expression profiles of two cliques of co-expressed autism genes [131] and the spatial distribution of cell types [143].

GENE EXPRESSION AND BRAIN CONNECTIVITY

Another way to study brain organization and function is to study brain connectivity. Brain connectivity has been linked to many neurological disorders, such as ischemic stroke, autism and schizophrenia. The relationship between gene expression and neuronal connectivity has long been studied in model organisms, such as Caenorhabditis elegans, to identify genes involved in synaptogenesis and axon guidance [144–146].

Zaldivar *et al.* [65] used the Allen mouse brain atlas to study the expression patterns of neurotransmitters in the brain. Since the expression of a transmitter must be coupled with expression of appropriate receptors in the postsynaptic target, they have also analyzed the expression of receptors in target regions. This study shows that known neurobiological concepts can be seen back in the Allen brain atlas. In order to take it one step further, French and Pavlidis [147] and Wolf *et al.* [148] analyzed the relationship between gene expression similarity of brain regions and their connectivity. Both studies used the Allen mouse brain atlas to calculate the similarity in gene expression between different regions and the neural connectivity data of the rat brain from the Brain Architecture Management System (BAMS) [149]. Genes involved in brain development and neurodevelopmental disorders, such as autism, showed strong correlations with anatomical connectivity patterns.

With the recent availability of the Allen mouse connectivity atlas, it has become possible to study the relationship between gene expression and brain connectivity within the same species. Rather than assessing the correlation between the gene expression

similarity and connectivity, Ji *et al*. [150] and Fakhry *et al*. [151] set to predict connectivity based on gene expression patterns. Recently, Richardi *et al*. [152] analyzed the relationship between gene expression similarity and synchronized activity as measured by fMRI in the human brain. By analyzing the relationship between genes with consistent expression patterns across individuals and resting-state functional connectivity data from the Human Connectome Project, Hawrylycz *et al*. [153] suggested that functional circuits are linked to conserved gene expression patterns across the cortex. Another study by Mahfouz *et al*. [154] analyzed the similarity between gene expression patterns of brain regions during human development. Using a network-based approach they characterized the topology of the connectivity network of autism-related genes across development.

STUDYING BRAIN ORGANIZATION USING DIMENSIONALITY REDUCTION METHODS

An alternative approach to analyze the relationship between gene expression and neuroanatomy is dimensionality reduction (Box 2). Mapping high dimensional data in two dimensions allows the exploration of how gene expression patterns relate to brain organization. Ji [155] used t-distributed stochastic neighborhood embedding (t-SNE) to map the Allen developing mouse brain atlas and showed that t-SNE clearly outperforms PCA. Their results show that clustering voxels in the low dimensional space is more consistent with neuroanatomy than those in the original space. Mahfouz *et al*. [156] used a computationally-efficient implementation of t-SNE, named Barnes-Hut-SNE, to map the sagittal and coronal adult mouse atlas and the brain transcriptome of the 6 human donors (Figure 2.5). They quantitatively showed that BH-SNE maps are superior in their separation of neuroanatomical regions in comparison to PCA and MDS. Similarly, dimensionality-reduction approaches can be used to analyze the gene-gene relationships. A low dimensional embedding of genes in which distances represent similarity of the spatial and/or temporal expression profile of genes across the brain can be very informative.

Box 6 | Co-expression Networks

Gene co-expression networks provide a framework to uncover the molecular mechanisms underlying biological processes based on gene expression data. A co-expression network consists of nodes to represent genes and edges to encode the co-expression between two genes. A weighted network is a network in which the edges have continuous values to indicate the strength of co-expression. Networks with binary edges (an edge either exists or not) are termed binary networks. Construction of co-expression networks can be summarized in three main steps.

Network Construction The first step in building a co-expression network is to construct a similarity matrix, by quantifying the similarity between the expression profiles of each pair of genes (i.e. co-expression). Several methods to measure gene co-expression are discussed in Box 5. For non-regularized estimations of co-expression, all off-diagonal elements of this similarity matrix will be non-zero. We can take these similarities as edge weights in the network, but that will give a fully connected network (each gene is connected to each gene). An additional step can be to threshold the similarity matrix, either to prune edges, or to binarize (absent/present) the similarities to obtain an adjacency matrix. In the latter case, pairs of genes with co-expression

values above a threshold will be connected in a binary network. In the weighted gene co-expression network analysis (WGCNA) framework the similarity matrix undergoes a power transformation and a weight diffusion step, to optimize the topological properties and stability of the network [111].

Network Characterization The obtained networks can be analyzed in a number of ways. Topological measures characterize the structure of the network, and quantify the importance of genes in their network context. These measures have been extended to weighted networks [111], and can capture topology on different levels of scale [30]. Sets of networks can also be aligned and compared [157–159]. Network comparison can be used either to assess changes between different conditions, or to replicate a network in an independent dataset for validity assessment.

Module Identification To interpret a network, it can be divided into sub-networks, or gene modules. To do this, the network edges are often treated as similarities in a clustering approach (see Box 3). Alternatively, graph properties, such as topological overlap or modularity, can be used to divide a network into modules [160].

Module Characterization Finally, modules can be characterized using a wide range of approaches. The expression profile of genes within the same module can be summarized using the average or the first principle component (also called eigengene [112]). Alternatively, one can characterize a module according to its hub genes: genes with the largest number of connections within the module. Another option is to assess the association of a module to external data by testing statistical enrichment in various gene sets (see Box 1 for different types of gene sets). In addition, modules can be characterized based on changes between conditions (e.g. health and disease) in their summary statistics (average expression profile), their topological measures (inter-connectivity), or the number of differentially-expressed genes they include.

2.3. PERSPECTIVE ON THE FUTURE OF COMPUTATIONAL ANALYSIS OF BRAIN TRANSCRIPTOMES

CELL-TYPE SPECIFICITY

The identification of the molecular profile of the different cell types in the brain, their connectivity patterns, and their electrophysiological properties are crucial to our understanding of the functional organization of the brain. Despite the undoubtedly valuable information provided by the brain transcriptomes, these resources remain limited in their ability to quantify cell-type-specific expression of genes. New technologies targeting specific cell populations, such as viral, optogenetic and single-cell sequencing approaches, will allow us to better characterize cell types and their role in brain function. So far these techniques are limited in their scalability and computational methods still provide a feasible alternative approach. Using spatial clustering of gene expression patterns of cell type-specific genes in the adult mouse, Ko *et al.* [141] showed that astrocytes and oligodendrocytes differ between brain regions, but that these regional differences in expression are less pronounced than differences in neuronal composition. Similarly, Grange *et al.* [142] proposed a model to estimate cell-type specific maps of the mouse brain. Kuhn *et al.* [161] developed a method to analyze brain samples of varying cellular composition. Their method detected myelin-related abnormalities in brain

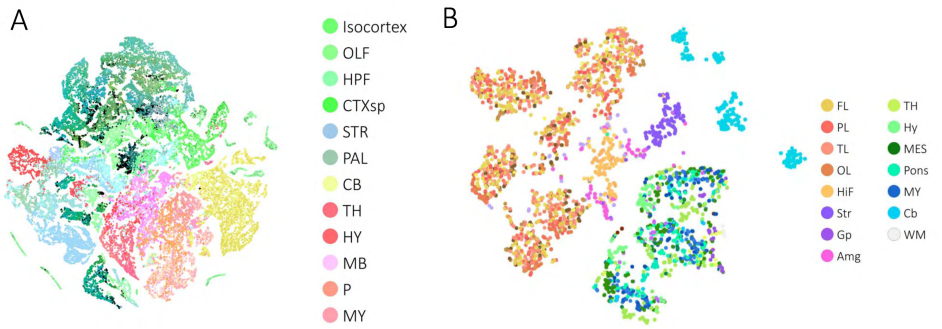


Figure 2.5: **Dimensionality reduction of brain transcriptomes.** Samples from brain transcriptomes can be embedded in a low dimensional space by means of dimensionality reduction methods. (A) 2D embedding of 60,000 voxels from the Allen Mouse Brain Atlas. (B) 2D embedding of 3,700 samples from the 6 donors in the Allen Human Brain Atlas. Both embeddings were generated using Barnes-Hut t-SNE. In both maps, colors correspond to anatomical regions of the mouse and human brain. Data from Mahfouz *et al*. [156].

samples from Huntington's disease patients, which was not detected using standard differential expression. These examples illustrate the power of computational models in untangling the complex composition of the different cell types in the brain.

With the recent advances in single-cell mRNA sequencing, it has become feasible to measure the expression of thousands of genes and their variability between different cell types [162]. Single-cell sequencing has indicated that neurons from small cortical regions come from different clones with distinct somatic mutations [19]. Understanding how these different clones of neurons contribute to the aggregated gene expression from a specific brain region will be of great interest to understand the role of mutations in neurological disorders. The vast amount of data generated by these projects yields computational methods that can identify distinct groups of cells with a common functional role highly valuable [163, 164].

SPATIAL RESOLUTION MATTERS

There are several limitations associated with the current spatial and temporal brain transcriptomes. Despite their unprecedented spatial and temporal resolution, human brain transcriptomes are still of low resolution with 1000 samples per brain. This relatively low resolution presents a fundamental limitation specially when integration with imaging-based data (e.g. MRI or PET) is considered. The ISH-based mouse transcriptomes offer a much higher resolution. Although the original ISH data provides a near-cellular resolution (1µm), the genome-wide data registered to the common 3D space offers a much lower resolution (200µm). Several studies used re-registration of a limited set of the high-resolution ISH images from the Allen Mouse Brain atlas to acquire genome-wide data at a higher resolution. The aforementioned study by Ko *et al*. [141] found more transcriptionally distinct brain regions than a previous study [140], mainly due to the usage of cell-type specific genes. However, Ko *et al*. have also realigned the ISH images of the mouse brain atlas and performed their analysis on a higher-resolution grid (100µm). Ramsden *et al*. [88] used non-linear registration to realign the ISH data of the mouse.

By analyzing genome-wide data at a resolution of 10μm, they were able to identify genes whose expression pattern delineates the borders and layers of the medial entorhinal cortex.

There is still need for more generic approaches to map spatially-mapped gene expression data (from ISH experiments) generated at different labs to the standard 3D space of the Allen Reference Atlas. Tools such as BrainAligner [165] are available for analyzing Drosophila melanogaster neural expression patterns. The availability of similar tools for the mouse and human brain could enormously enhance our understanding of disease molecular mechanisms by allowing researchers to map their own data to the same space.

BEYOND PROTEIN-CODING MRNA

Most of the atlases profiling the mammalian brain transcriptome and its relationship to brain development and function have mainly focused on profiling the expression of protein-coding mRNA. These atlases mostly provided limited or no information about other RNA species such as non-coding RNA (ncRNA) and microRNA (miRNA) despite their recognized role in brain development and neurological disorders [166, 167]. Long ncRNAs show regionally enriched expression patterns, such as those observed for protein-coding mRNAs [168], further supporting their functional role in the brain. By profiling the developmental transcriptome of the neocortex using deep sequencing, Fertuzinhos *et al*. [169] characterized the dynamics of mRNA, miRNA, and ncRNA across the different layers of the mouse cortex. The BrainSpan atlas provides the most comprehensive map of miRNA expression in the developing human brain. Ziats and Rennert [170] used the BrainSpan miRNA data to define a pattern of increased inter-regional expression differences of miRNA through development potentially driving regional specialization. Moreover, targets of differentially expressed miRNA were mostly related to transcriptional regulation and neurodevelopmental disorders, highlighting the importance of studying miRNA as potential biomarkers. Additional measurement of ncRNAs and miRNAs as well as a detailed analysis of their role in gene regulatory networks can help our understanding of their relationship to genes related to neurodevelopmental disorders.

INTEGRATING BRAIN TRANSCRIPTOME ATLASES WITH OTHER NEURO-OMICS DATA

Advances in high-throughput molecular profiling have facilitated acquiring various omics data sets spanning a wide spectrum of cellular processes. For instance, the rapid developments in next-generation sequencing (NGS) technology allowed for genome-wide measurement of genomic, transcriptomic, and epigenomic data of brain tissues. While transcriptomes provide detailed information of the abundance of RNA, epigenomic features, such as histone modifications, methylation and chromatin interactions, describe the underlying mechanisms of distinct cell-specific transcriptomes. Moreover, most disease-related variants are in the non-coding regulatory regions of the genome, yielding epigenomic studies crucial to uncover a larger proportion of the genetic contribution to complex traits than that explained by coding variants only. Increasingly, studies are gathering data across different platforms

from a wide range of tissues and cell types to uncover mechanisms underlying complex phenotypes and disease. The Encyclopedia of DNA Elements (ENCODE) [171] and the Roadmap Epigenome project [75] have profiled the epigenome of several tissues and cell-types, while the Genotype Tissue Expression project (GTEx) [45] is generating genotype and gene expression data from many human tissues. In contrast, The Cancer Genome Atlas project (TCGA) [172] and the International Cancer Genome Consortium (ICGC) [173] provide comprehensive genomic and transcriptomic and epigenomic data from multiple cancer types. However, most of these projects have profiled samples from cancer cell lines or normal cells from non-brain tissues due to limitations specific to the brain, such as the requirement of large amount of genomic material and the high heterogeneity of cell types within the same sample [174]. Currently, the isolation of more homogeneous samples from the brain as well as developments in single-cell analysis are greatly advancing the field of neuro-epigenomics [174, 175]. For example, efforts have been made to map the brain methylome [176] and to identify cis-regulatory elements across brain regions [177]. The PsychENCODE consortium [178] is an ongoing project to profile the neurobiological epigenetic landscape of the healthy and diseased developing and adult human brains. Systems genomics approaches which integrates different genome-wide data types can minimize false positive discoveries as well as unravel the complete molecular mechanism underlying the phenotype or disease of interest. Several approaches have been developed to integrate multi-omics data [179], clearly illustrating the added value of collecting multiple omics measurements from a large number of samples.

INTEGRATING BRAIN TRANSCRIPTOME ATLASES WITH IMAGING MASS SPECTROSCOPY

Over the past few years imaging mass spectrometry (IMS) [180] has emerged as a powerful technique to capture the spatial distribution of large biomolecules such as proteins, peptides and lipids in biological samples. Similar to ISH, imaging mass spectroscopy hold great potential in studying the chemical organization of complex samples from the brain [181]. Methods have been developed to align IMS-based sections of the mouse brain to histology-based sections from the Allen Mouse Brain Atlas to anatomically localize biomolecules within the brain [182, 183]. But recently, these methods have been extended to link protein expression to the expression of the encoding genes as well as their co-expressed genes based on the Allen Mouse Brain Atlas [184]. There is a great potential for applications based on the integration of ISH-based gene expression and IMS-based protein expression measurements to help our understanding of translational mechanisms in the brain. Yet, more complex modeling of the two data types is needed. Methods developed to integrate spatially-mapped gene and protein expression data can also be used to study spatial localization within the cell using data from the Human Protein Atlas [185].

INTEGRATING BRAIN TRANSCRIPTOME ATLASES WITH IMAGING DATA (IMAGING-GENETICS)

In an attempt to better understand gene-disease associations, researchers are searching for genes that affect intermediate disease biomarkers. Brain imaging studies can be

used to reveal genetic effects on brain structure, function and circuitry, providing valuable mechanistic insights. Imaging genetics have emerged as a field concerned with finding associations between genetic variants (typically SNPs) and imaging-based measurements [186]. Due to the millions of statistical tests that need to be performed, stringent statistical thresholds are required to limit the false discovery rate [53]. Recently, the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium [18] analyzed SNPs' associations with the volume of subcortical structures in 30,000 individuals, providing the first large scale analysis of the genetic causes of human brain variability. Several methods have been developed to limit the number of statistical tests performed in genome-wide, brain-wide analysis by either exploiting the dependency between brain voxels and/or testing for associations with genes or pathways instead of individual variants [187]. In addition, efforts have been made to jointly model imaging and genetic observations from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data (adni.loni.ucla.edu), using multivariate statistical methods [188, 189]. These methods remain computationally very expensive, limiting the number of variables analyzed. Brain transcriptomes can play an important role in imaging genetics by providing region-specific information about gene expression that can be used to prioritize genes and variants for testing. For example, incorporating spatial gene co-expression of amyloid-related candidate genes from the Allen Human Brain Atlas as prior knowledge to their statistical model significantly improved the prediction of associations between SNPs in the APOE gene and amyloid deposition measures among cortical regions [190]. Rizzo *et al*. [191] tested the predictive power of mRNA transcription maps extracted from the Allen Human Brain Atlas and in vivo protein distributions acquired using positron emission tomography (PET). By analyzing genes involved in two neurotransmission systems with different regulatory mechanisms, their results show the possibility to predict in vivo protein distributions using mRNA transcription maps when translational mechanisms rather than posttranscriptional regulation determine expression. There is need for more advanced methods to link genomic measurements which is usually collected from blood samples to intermediate disease phenotypes observed in brain images.

UNEXPLORED COMPUTATIONAL AVENUES

The multiple dimensions of the brain transcriptomes (genes, regions, and time) provides a framework to explore spatio-temporal regulation of gene expression during development. Clustering the data along one dimension only yields global patterns of similarity, while in a complex system such as the brain it is always more useful to identify more localized patterns of correlation. For example, the effect of steroid hormones on the brain is highly region-specific, depending on the availability of target genes and co-regulators affecting the steroid receptors at the site of action. Analyzing the region-specific co-expression relationships of steroid receptors and their coactivators can be used to predict steroid responsiveness and selective activation of particular circuits with synthetic ligands [192].

Biclustering is a type of technique to simultaneously identify a subset of genes associated with a subset of conditions (this can be brain regions and/or time points), allowing the identification of local spatial or temporal patterns of co-expression. Biclustering was

shown to be particularly effective in analyzing time-series expression data [193]. Similarly, applying bi-clustering to expression data from the Allen Mouse Brain Atlas resulted in more GO-enriched clusters than those obtained by independently clustering genes or regions [194]. Ji and Zhang [195] described a co-clustering method based on graph approximation to explore the spatiotemporal regulation of gene expression during the mouse brain development. Yet, they apply biclustering to each developmental stage independently and do not consider the time-varying nature of the developing mouse brain data, due to the lack of correspondence between the voxels across different stages. In order to fully exploit the multi-dimensionality of the developing brain transcriptomes, triclustering methods provide an interesting approach to identify groups of genes that show spatial and temporal co-expression [196]. Recently, Jung *et al*. [197] used three-component analysis to identify genes associated with aging by analyzing longitudinal gene expression, methylation and histone modification data of human skin fibroblasts. Their three-component analysis is an integrative approach to jointly model temporal changes in different data types. An extension of their methods to incorporate spatial information available in brain transcriptomes can lead to a complete approach of modeling spatial and temporal changes of different omics data from the brain.

Graphical models (e.g. conditional random fields) are commonly used for data segmentation using local features, especially in computer vision applications. The Roadmap Epigenome project has used a Hidden Markov Model to classify the human genome into chromatin states based on epigenetic markers [75]. These models can be used to model the spatial and/or temporal relationships between genes in brain transcriptome atlases.

A greater challenge lies in identifying causal relationships rather than associations in gene-gene interactions and the brain is no exception. Systems biology approaches provide an interesting avenue to explore causal relationships between genes by means of quantitative modeling. The resulting mathematical models enable formal analysis and simulation of complex biological processes [198]. However, inferring causal relationships between the different variables requires a vast amount of data, limiting the application to a small number of genes [199]. Hwang *et al*. presented a system approach to analyze genes differentially expressed in the mouse brain across time in Prion disease [200]. An extension of such a model to include spatial information on gene expression can help refine the model as well as associate disease-related changes to specific brain areas.

Chapter 3

Visualizing the Spatial Gene Expression Organization in the Brain through Non-Linear Similarity Embeddings

Ahmed Mahfouz*
Martijn van de Giessen*
Laurens van der Maaten
Sjoerd MH Huisman
Marcel JT Reinders
Michael J Hawrylycz
Boudewijn PF Lelieveldt

This Chapter is published as: *Methods* (2015) 73:79-89, doi: 10.1016/j.ymeth.2014.10.004.
*Equal contribution.

THE Allen Brain Atlases enable the study of spatially resolved, genome-wide gene expression patterns across the mammalian brain. Several explorative studies have applied linear dimensionality reduction methods such as Principal Component Analysis (PCA) and classical Multi-Dimensional Scaling (cMDS) to gain insight into the spatial organization of these expression patterns. In this paper, we describe a non-linear embedding technique called Barnes-Hut Stochastic Neighbor Embedding (BH-SNE) that emphasizes the local similarity structure of high-dimensional data points. By applying BH-SNE to the gene expression data from the Allen Brain Atlases, we demonstrate the consistency of the 2D, non-linear embedding of the sagittal and coronal mouse brain atlases, and across 6 human brains. In addition, we quantitatively show that BH-SNE maps are superior in their separation of neuroanatomical regions in comparison to PCA and cMDS. Finally, we assess the effect of higher-order principal components on the global structure of the BH-SNE similarity maps. Based on our observations, we conclude that BH-SNE maps with or without prior dimensionality reduction (based on PCA) provide comprehensive and intuitive insights in both the local and global spatial transcriptome structure of the human and mouse Allen Brain Atlases.

3.1. INTRODUCTION

The mammalian brain is a complex system governing all high-level cognitive tasks. The complexity of this system is reflected in the large number of cell types, organized into hundreds of distinct structures [201]. A major challenge facing the neuroscience community is to collect, integrate and analyze data across different levels and scales to produce new insights about the brain's anatomical and functional organization [93]. At the molecular level, each brain structure has a specific cellular composition with a distinct gene expression signature that dictates its functional role [137]. Therefore, to understand the basic anatomical and functional organization of the brain in relation to gene functions, it is crucial to study the spatial localization of genome-wide gene expressions in the brain.

Given the high cellular diversity in the brain, mapping genes at a sufficient spatial resolution is essential to analyze the transcriptome architecture of the brain. Several studies have previously mapped the expression of genes across the mammalian brain, but they have all been limited either in terms of the number of genes analyzed and/or the number of brain structures assessed [20–22]. The Allen Institute for Brain Sciences provides comprehensive genome-wide maps of gene expression across the mouse and human brain, providing a unique opportunity to study the transcriptome architecture of the mammalian brain. In the Mouse Brain Atlas [24] the expression of ~20,000 genes at a cellular resolution using in situ hybridization (ISH) is mapped on an anatomical atlas of the mouse brain. Comparably, the Human Brain Atlas [26] employed microarrays to produce a genome-wide map of gene expression distribution across the entire human brain. These two resources allow the unprecedented study of how the transcriptome architecture of different brain regions instructs their functional role.

The high diversity of spatially-mapped gene expression patterns in the brain, ranging from globally-expressed genes to highly-specialized regional markers, poses great challenges for computational approaches. Univariate approaches involving the analysis of the expression profiles of few genes of interest using prior knowledge of their site of ac-

tion in the brain are not suitable to capture the full complexity of the data. In order to capture the complex patterns of expression of thousands of genes across the entire brain (thousands of samples), multivariate approaches should be employed to accommodate the high-dimensionality of the data. However, visualizing high-dimensional data for intuitive interpretation is challenging.

Several studies have used Principal Component Analysis (PCA) or classical Multidimensional Scaling (cMDS) to reduce the dimensionality of the voxel level genome-wide gene expression data of the mouse brain [26, 140, 141]. These low-dimensional maps are then used either to enable visual exploration of the gene expression patterns or as an input to a clustering algorithm where the resulting clusters are compared to the classical neuroanatomy. Classical methods such as PCA and cMDS focus on appropriately modeling large pairwise distances between gene expression profiles [202]. The focus on modelling large pairwise distances comes at the price of substantial errors in modelling small pairwise distances. However, it is exactly this local similarity structure that is essential in clustering and visual exploration: the goal of clustering is to find groups of nearby data points and, similarly, the goal of visual exploration is to determine which parts of the data are similar to a reference data point [203]. Therefore, we advocate to employ embedding techniques that focus on preserving local similarity structure, as is done by techniques such as t-distributed stochastic neighbor embedding (t-SNE) [32]. Since its introduction in 2008, t-SNE has been proven to outperform linear dimensionality reduction methods, but also non-linear embedding methods such as ISOMAP [90], in several research fields including machine-learning benchmark datasets and hyperspectral remote sensing data [204].

Recently, t-SNE has been employed to analyze high dimensional proteomic and genomic data. Shekhar *et al*. [205] used t-SNE to differentiate between cellular phenotypes of the immune system based on mass cytometry data. Ji [155] used t-SNE to analyze the relationship between gene expressions and neuroanatomy in the developing mouse brain showing that t-SNE is able to capture the local similarities in the high-dimensional space. Fonville *et al*. [206] have shown that t-SNE outperforms PCA and self-organizing maps when used for modeling of mass spectrometry imaging data, where each pixel represents a molecular mass spectrum. All the previously mentioned applications demonstrate the high potential of t-SNE in the visual analysis of high-dimensional molecular data.

The goal of this work is to explore the effectiveness and limitations of t-SNE for spatial mapping of gene expression patterns in both the mouse and the human Allen Brain Atlases. By applying Barnes-Hut-SNE (BH-SNE) [207], a recently developed optimization algorithm for t-SNE, we show the consistency of the low dimensional embedding across the 6 human brains as well as between the sagittal and coronal experiments of the mouse brain. In addition, we quantitatively show the superiority of BH-SNE over PCA and cMDS in separating neuroanatomical regions in the low-dimensional 2D embeddings. Finally, we assess the effect of higher-order principal components on the local and global structure of the spatial transcriptome similarity maps.

3.2. MATERIAL AND METHODS

MOUSE BRAIN GENE EXPRESSION

The Allen Mouse Brain Atlas [24, 208] provides genome-wide cellular-resolution in situ hybridization (ISH) gene expression data for approximately 20,000 genes of the 8-week old adult C57BL/6J male mouse brain. For each gene, sagittal ISH sections were sampled at 25μm intervals across the entire brain and the high-resolution 2D image series from each experiment were reconstructed in 3D and registered to the Nissl stain-based reference atlas (Allen Reference Atlas). The data were then aggregated into isotropic voxels defined by a uniform 200μm grid in the reference space by averaging the expression levels and densities of all pixels (in the high-resolution ISH sections) within each voxel. The ontology of the reference atlas is used to label individual voxels with their anatomical nomenclature. In addition, coronal sections are available for a set of approximately 4,000 genes that showed marked regional expression patterns in the sagittal plane [137]. More information about the ISH sections alignment and registration to the Allen Reference Atlas can be found in [209].

We retrieved all expression energy volumes from [208] using the Allen Brain Atlas application programming interface (API). Expression energy is a measurement combining the expression level (the integrated amount of signal within each voxel) and the expression density (the amount of expressing cells within each voxel) [57].

We focused our analysis on a subset of high confidence genes for which coronal and sagittal experiments are available, as in [140]. For each gene, we computed the Spearman's rank correlation between the corresponding coronal and sagittal experiments and selected genes in the top-three quartiles of correlation (3,241 genes). The coronal and sagittal experiments corresponding to those 3,241 genes were retained for further analysis (Supplementary Table 1). For genes with more than one sagittal experiment, the maximum correlation value was used. A mask was applied to exclude all non-brain voxels, resulting in a 61,164 × 3,241 (voxels × genes) matrix for the coronal experiments and a 27,365 × 3,241 matrix for the sagittal experiments.

HUMAN BRAIN GENE EXPRESSION

The Allen Human Brain Atlas [26, 210] includes RNA microarray data collected from the postmortem brains of six donors, with no known neuropsychiatric or neuropathological history; see Table 3.1 for detailed information about the donors. Magnetic resonance (MR) T1-weighted (T1W), T2-weighted (T2W) and Diffusion Tensor (DT) images were collected in-cranio, prior to dissection for anatomic visualization of each brain.

Approximately 1,000 samples were dissected using manual macrodissection for large regions and laser captured microdissection for smaller regions from two donor brains (H0351.2001 and H0351.2002), representing all structures across the whole brain. For the other four donor brains, approximately 500 samples were taken from one hemisphere only. Each sample is associated with a 3D (x, y, z) coordinate on its corresponding donor's MRI volume. Moreover, the MNI coordinates of each sample is reported (registration to the MNI reference space was done using FreeSurfer software). The dataset contained expression profiles of 29,191 genes represented by 58,692 probes, with 93% of known genes represented by at least 2 probes. The data was already normalized across samples and

across different brains using the procedure explained in [211]. Probes with no Entrez ID or gene symbol were excluded and the expression profiles of all the probes representing one gene were averaged, resulting in 20,737 genes (Supplementary Table 2).

Table 3.1: **Human Donors Information.**

Donor ID	Number of Samples	Sex	Age (years)	Race/Ethnicity
H0351.2001	946	Male	24	African American
H0351.2002	893	Male	39	African American
H0351.1009	363	Male	57	Caucasian
H0351.1012	529	Male	31	Caucasian
H0351.1015	470	Female	49	Hispanic
H0351.1016	501	Male	55	Caucasian

CELL TYPE MARKERS

Lists of cell-type specific genes were extracted from a previously published work by Cahoy *et al*. [36] who profiled gene expression patterns in purified populations of neurons, astrocytes, and oligodendrocytes using microarrays. We selected genes enriched by at least 10-folds in one cell type, compared to the two other cell types, out of the 3,241 high-confidence genes included in our analysis, resulting in 195 neuron-specific, 60 astrocyte-specific, and 43 oligodendrocyte-specific genes in the mouse data (Supplementary Table 3). Using the same 10-fold enrichment threshold for the human data resulted in 247 neuron-specific, 151 astrocyte-specific, and 92 oligodendrocyte-specific human orthologous genes (Supplementary Table 4).

NON-LINEAR DIMENSIONALITY REDUCTION

Three different multivariate data analysis methods were used to visualize the high dimensional expression data, namely: Barnes-Hut-SNE (see Theory), principle component analysis (PCA) and classical multi-dimensional scaling (cMDS). For the mouse data, the expression profile of each gene, i.e. column of the voxel × gene expression matrix, was Z-score normalized across all voxels. The human brain expression data was already Z-score normalized [211]. PCA was applied to the expression matrices (voxel × gene for the mouse data and sample × gene for the human data). The human data was also analyzed with cMDS. The first two components of PCA and the first two dimensions of cMDS were used to visualize the data in each case. The goodness-of-fit criterion for cMDS was the stress, normalized by the sum of squares of the inter-point distances [212]. Distances within cMDS between two samples s_1 and s_2 were computed as genetic distances: $d(s_1, s_2) = (1 - \rho(s_1, s_2)^2)^{\frac{1}{2}}$, where $\rho(s_1, s_2)$ denotes the correlation between gene expression levels.

For the non-linear dimensionality reduction using Barnes-Hut-SNE (BH-SNE), we used the full data dimensionality and mapped it to a 2D BH-SNE plot. In addition, we assessed the effect of prior dimensionality reduction using PCA in order to reduce noise in the final maps. The data was first mapped either to the first 2, 3, 5, 10, or 20 components and then embedded into 2D BH-SNE maps.

REGIONAL GENE EXPRESSION VISUALIZATION

For both the mouse and the human data, the mapped data points (voxels or samples) were colored in the low-dimensional 2D map according to their associated reference atlas ontology colors, as obtained from the mouse and human atlases [24, 26] (Supplementary Table 5 and 6). This ontology was colorized so that each brain structure has a unique color and anatomically related structures (e.g., substructures of the hypothalamus) are coded with similar colors.

To visually analyze the ability of the different methods (BH-SNE, PCA, and MDS) to segment different regions of the mouse brain, the data points (voxels) were colored by spanning an "L*a*b*" color map [213], that maps the a* and b* colormap axes to the horizontal and vertical axes of the PCA, cMDS and BH-SNE maps. The L*a*b* color space was selected because it spans all perceivable colors and the a* and b* axes span a two-dimensional space with all perceivable colors at a constant perceived "lightness" L* that is perceptually linear. L* was fixed at 50 for all plots providing a good color contrast in the 2D maps.

Using the MNI152 coordinates associated with each of the human brain samples, we mapped each sample back to the Automated Anatomical Labeling (AAL) human brain atlas [214]. Direct visualization of these samples is, however, hampered by the spatial sparsity of the data, i.e. there are very few samples per anatomical regions. Therefore, we colored each voxel in the brain where there is no sample available (no gene expression data) according to the closest sampled voxel based on the Euclidean distance between the unsampled voxels and the sampled voxel.

EVALUATION OF THE MAPPED DATA

To evaluate the capability of each of the dimensionality reduction methods to separate different brain regions, we analyzed the separation between different brain-region clusters in the low-dimensional 2D maps produced by these methods. The separation between two brain structures can be characterized by the similarity of the distributions of the data points belonging to each region in the low-dimensional space. In this work we use the Jensen-Shannon divergence [215] to compute this similarity. Briefly, for each brain structure, a 2D histogram is computed by calculating the density of the data points belonging to that structure in the 2D map. This is achieved by overlaying a 40×40 rectangular grid that covers all the samples in the low dimensional space. The divergence between the two histograms (distributions) of two brain structures P and Q is then calculated as: $JSD(P \parallel Q) = \frac{1}{2} KL(P \parallel M) + \frac{1}{2} KL(Q \parallel M)$, where $KL(P \parallel Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}}$ is the Kullback-Leibler divergence between P and Q, and $M = \frac{1}{2}(P+Q)$. By using the same grid size (the same number of bins), divergences between brain regions are comparable between the different dimensionality reduction methods regardless of their scaling factor. For the mouse data, divergence scores were calculated between: cortex (Isocortex), olfactory areas (OLF), hippocampal formation (HPF), cortical subplate (CTXsp), striatum (STR), palladium (PAL), cerebellum (CB), thalamus (TH), hypothalamus (HY), midbrain (MB), pons (P), and medulla (MY). In the human atlas, divergences were calculated between: frontal lobe (FL), parietal lobe (PL), temporal lobe (TL), occipital lobe (OL), hippocampal formation (HiF), striatum (Str), Globus pallidus (Gp), amygdala (Amg), thalamus (TH), hypothalamus (Hy),

mesencephalon (MES), Pons, myelencephalon (MY), cerebellum (Cb), and white matter (WM).

3.3. THEORY

t-Distributed Stochastic Neighbor Embedding (t-SNE) [32] constructs a two-dimensional scatter plot in which each point represents a gene expression profile. In such a t-SNE map, nearby points correspond to similar profiles, whereas distant points correspond to dissimilar profiles. The map is constructed by (1) measuring similarities between gene expression profiles and (2) moving points around in the map in such a way as to minimize some difference measure between similarities of points in the map and the corresponding gene-expression profile similarities.

Mathematically, t-SNE operates by converting the gene expression profiles into a probability distribution over pairs of profiles in such a way, that similar pairs have a high probability of being picked. The distribution is defined as a standard Gaussian kernel (with a particular choice for σ^2) that is normalized to sum to one. Next, t-SNE constructs a map in which each point corresponds to an expression profile by: (1) defining a similar distribution over the pairs of points in the map, and (2) minimizing the divergence between the two distributions with respect to the coordinates of the points in the map using gradient descent. Mathematically, for a given sample pair t-SNE defines pairwise similarities between gene expression profiles $x \in \mathbb{R}^D$ (with D the number of genes) as:

$$p_{ij} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)}{\sum_{k \neq l} \exp\left(\frac{-\|x_k - x_l\|^2}{2\sigma^2}\right)} \tag{3.1}$$

where p_{ij} is the similarity between the expression profiles x_i and x_j in the high dimensional space $\mathbb{R}^D$. Likewise, the similarity between the corresponding low-dimensional models of these samples $y \in \mathbb{R}^d$ (with d the dimensionality of the t-SNE map) is defined as:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{3.2}$$

where q_{ij} is the similarity between the expression profiles y_i and y_j in the low dimensional space $\mathbb{R}^d$. In the definition of map similarity, a heavy-tailed Student-t distribution is used to measure similarity in the map to account for the large difference in volume between the high-dimensional gene expression space and the low-dimensional map. The low-dimensional map $y_1, \ldots, y_N$, with N the number of gene expression profiles, is learned by minimizing the Kullback-Leibler divergence between both distributions:

$$KL(P \parallel Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{3.3}$$

The asymmetry of the Kullback-Leibler divergence encourages modeling large $P-values$ (similar expression profile) by large $Q-values$ (nearby points). As a result, in contrast to techniques like PCA, t-SNE focuses on appropriately modeling the local gene expression profile structure in the map.

The gradient that is used for learning a t-SNE map can be interpreted as an N-body system [207]: each point in the map exerts a force onto all other points, and the gradient for each point is the resultant force on that point (i.e. the sum of all incoming forces). Specifically, the gradient on point y_i comprises springs between y_i and all other points, where the force in a spring depends on the difference between the corresponding $P-values$ and $Q-values$. The gradient computes the resultant force on map point y_i in this spring system:

$$\frac{\partial KL}{\partial y_i} = 4 \sum_{i \neq j} \left(p_{ij} - q_{ij}\right) \left(1 + \|y_i - y_j\|^2\right)^{-1} \left(y_i - y_j\right) \tag{3.4}$$

The interpretation of t-SNE as an N-body simulation facilitates the use of approximation techniques that were originally developed in astronomy to simulate large galaxies of stars, such as the Barnes-Hut approximation [216] or fast multipole approximations [217]. We focus on the Barnes-Hut approximation [207][1], which exploits the fact that a group of nearby map points exerts very similar forces on another point that is relatively far away. Therefore, the (resultant) force exerted on the latter point can be approximated by the force between the center-of-mass of the group of points and the point under consideration, multiplied by the number of points in the group. In practice, this is implemented by storing all points in a quadtree (for 2D maps) or octtree (for 3D maps) and performing a depth-first tree search on this tree, pruning away nodes for which the aforementioned approximation can be used. The resulting algorithm has an average-case complexity of $O(NlogN)$. For more details, we refer to [207].

3.4. RESULTS AND DISCUSSION

GENETIC SIMILARITY WITHIN THE MOUSE BRAIN

We used BH-SNE to embed the mouse coronal expression data in a 2D space, see Figure 3.1A. In order to reduce the noise in the data, we first reduced the data dimensionality by mapping the data on the first 10 principal components and then used the reduced data as an input to BH-SNE. The BH-SNE mapped data show that anatomical regions are in many cases in disjoint and visually distinct clusters. By comparison, Figure 3.1D plots the reduced data when only PCA is being used, showing that then samples of the same anatomical region are close but no clear clustering and regional separation is visible. This difference between BH-SNE and PCA is also reflected in Figure 3.1B and E, where the borders between anatomical regions seem sharper with coloring based on BH-SNE. In Figure 3.1B for example the hippocampal formation (in red) and cerebral nuclei (in light blue) can be easily distinguished. More strikingly, the transversal views in Figure 3.1B shows that within the deeper brain structures there is a clear difference between medial gene expression (green) and posterior gene expression (pink). The first 2 principal components do not pick up on this variation and suggest a strong similarity (Fig-

[1] A variant of t-SNE based on fast multipole methods is presented in [218].

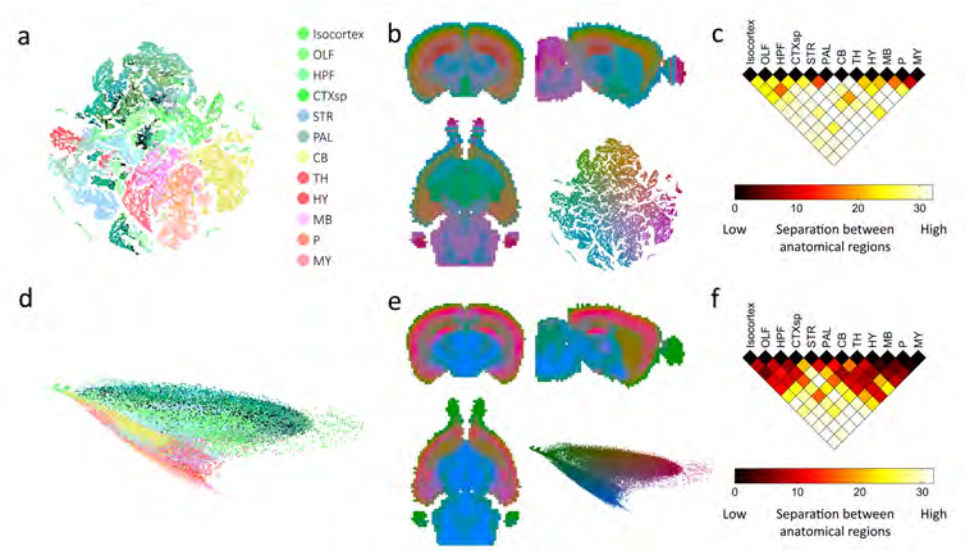


Figure 3.1: **Coronal mouse brain transcriptome similarities.** (a) BH-SNE map of the mouse coronal data initialized with 10 principal components and colored by anatomical region labels from the Allen Reference Atlas: cortex (Isocortex), olfactory areas (OLF), hippocampal formation (HPF), cortical subplate (CTXsp), striatum (STR), palladium (PAL), cerebellum (CB), thalamus (TH), hypothalamus (HY), midbrain (MB), pons (P), and medulla (MY). (b) The mouse coronal data mapped back to the 3D volume of the mouse atlas (3 views) and colored by the L*a*b* colormap of the BH-SNE mapping at a constant L* value. (c) Divergence plot for BH-SNE showing the similarity between pairs of neuroanatomical regions. A higher divergence value (lighter colors) indicates better separation between a pair of neuroanatomical regions in the 2D BH-SNE map. (d) The first two PCA components of the mouse coronal data colored by anatomical region labels from the Allen reference Atlas. (e) The mouse coronal data mapped back to the 3D volume of the mouse atlas (3 views) and colored by the L*a*b* colormap of the PCA mapping. (f) Divergence plot for PCA showing the similarity between pairs of neuroanatomical regions. A higher divergence value (lighter colors) indicates better separation between a pair of neuroanatomical regions in the 2D PCA map.

ure 3.1E). To quantify these observations, Figure 3.1F shows the Jensen-Shannon divergence between classical anatomical regions in both the first two principal components and the two-dimensional BH-SNE. A lower divergence value indicates that the map suggests higher similarity between the corresponding brain regions in the 2D map. In general, BH-SNE yields much higher divergence values between pairs of anatomical regions compared to PCA. Particularly, the cortex (Isocortex) and the hippocampus (HPF) are not separable in the PCA map, that is why they both retain the same color (pink) in Figure 3.1E. This high similarity between the cortex and hippocampus is also reflected with a low divergence value in Figure 3.1F. On the other hand, BH-SNE can clearly separate the cortex and the hippocampus giving them different colors in Figure 3.1B and resulting in a high divergence value in Figure 3.1C.

Adding more PCA components in the BH-SNE preprocessing step increases the information available to the BH-SNE algorithm and thereby influences the separation of neuroanatomical regions in the embedded map. To study the effect of PCA preprocessing step on the BH-SNE mapping we varied the number of principal components. Fig-

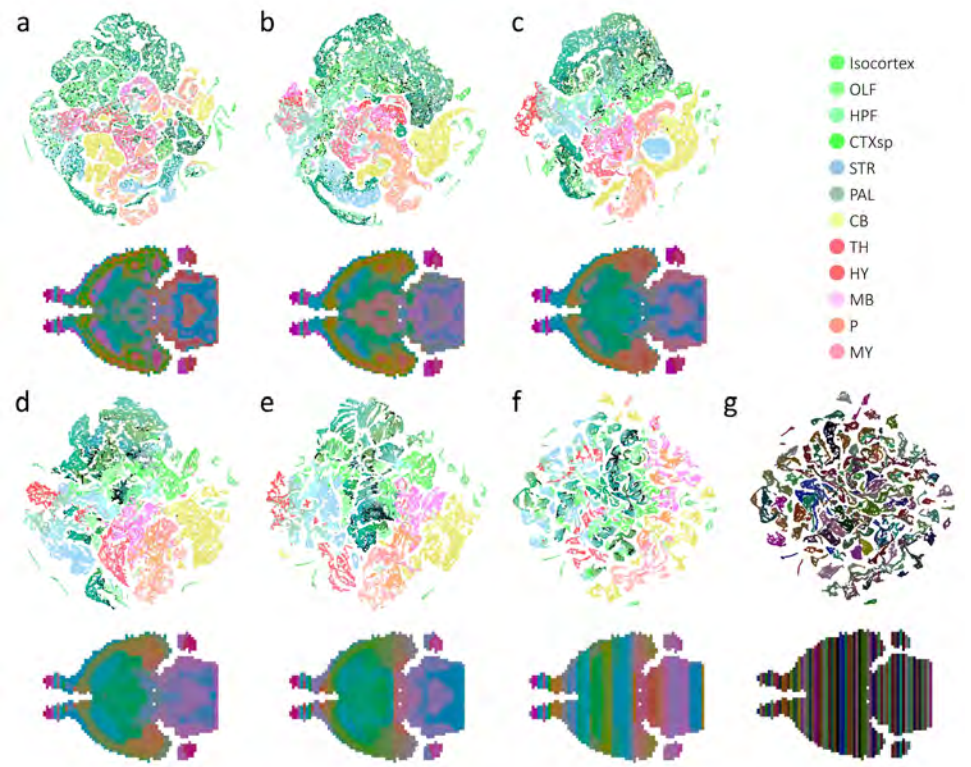


Figure 3.2: **BH-SNE maps the mouse coronal data using different initializations.** BH-SNA mappings of the mouse coronal data using (a) 2, (b) 3, (c) 5, (d) 10 and (e) 20 principal components to reduce the dimensionality of the data before applying BH-SNE. (f) BH-SNE mapping of the mouse coronal data without any prior dimensionality reduction. BH-SNE maps are colored by anatomical region labels from the Allen reference Atlas. Axial sections of the mouse 3D atlas space are colored with the corresponding L*a*b colors of each voxel in the BH-SNE maps. (g) BH-SNE mapping of the mouse coronal data without any prior dimensionality reduction, colored by the corresponding coronal plane.

ure 3.2 shows the BH-SNE plots and the L*a*b* mappings to the mouse brain (transversal sections) using a range of components to reduce the dimensionality of the data before applying BH-SNE. When using a small number of components (5 or less), insufficient information is retained to achieve good separation between different brain regions. This is visible from the mixing of samples from the same anatomical region in the BH-SNE maps and in the axial brain slices by ragged edges (Figure 3.2A and B, e.g., yellow, blue and green clusters). By increasing the number of components to 5 or 10, BH-SNE produces much better results as seen by clear separation between voxels belonging to different brain structures (Figure 3.2C and D). When we further increased the number of components (20 or more), the clusters of voxels belonging to one brain structure started to break into smaller sub-clusters. This is particularly clear in Figure 3.2F (BH-SNE without prior dimension reduction). These sub-clusters seem to be formed by voxels belonging to different coronal planes, visible when the points in the BH-SNE map are projected back to

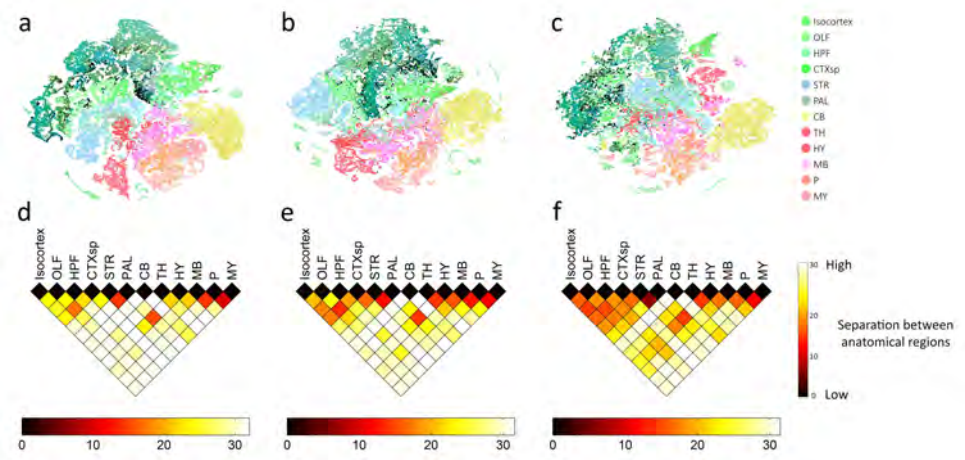


Figure 3.3: **BH-SNE embeddings of cell-type specific genes in the mouse coronal data.** (a) neuron-specific, (b) astrocyte-specific, and (c) oligodendrocyte-specific gene sets based BH-SNE embeddings, colored by anatomical region labels from the Allen reference Atlas. (d)-(f) Divergence plots of the BH-SNE embeddings using (d) neuron-specific, (e) astrocyte-specific, and (f) oligodendrocyte-specific gene sets. A higher divergence value (lighter colors) indicates better separation between a pair of neuroanatomical regions in the 2D BH-SNE map.

the mouse brain (Figure 3.2F, lower panel). When the BH-SNE map is colored according to the coronal plane from which the data was extracted, the observed sub-clusters could indeed be attributed to the different coronal planes (Figure 3.2G). Since the gene expression data was generated using ISH if coronal section of the mouse brain, we can reason that by including more components BH-SNE starts to pick up inter-slice differences. Neighboring voxels within the same brain structure and within the same slice have a more similar expression profile than neighboring voxels within the same brain structure but in different coronal slices. This leads to the fragmentation of brain region clusters into smaller, within-slice clusters, and hence the visible color gradient over the coronal brain slices (Figure 3.2F).

Ko *et al.* [141] demonstrated that k-means clustering on cell-type specific genes reveals that neuron-specific genes show the most neuroanatomically similar pattern across the mouse brain. To explore if cell type composition gives rise to the expression differences that separate anatomical regions, we performed BH-SNE on cell type marking genes only. Figure 3.3 shows the BH-SNE mappings for three disjoint sets of genes. The neuron-specific genes yield the highest separation between neuroanatomical regions (Figure 3.3A). In order to quantify the ability of different gene sets to partition neuroantomy we use the Jensen-Shannon divergence plot. The neuron-specific gene set leads to stronger separation of the cortex (Isocortex), olfactory area (OLF), hippocampus (HPF), and the cortical subpalate (CTXsp) as well as between the thalamus (TH) and the hypothalamus (HY), indicated by higher divergence values (lighter color) in Figure 3.3D–F. Overall oligodendrocyte-specific genes show the weakest separation between classical anatomical regions, although all gene sets give clear distinctions between cortex and non-cortex areas. This confirms the observations

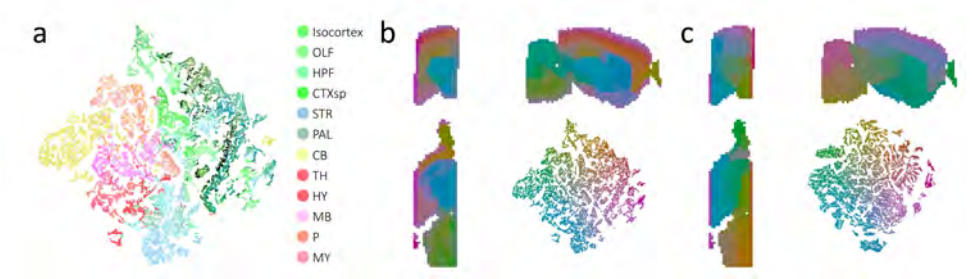


Figure 3.4: **BH-SNE maps of the mouse sagittal data.** (a) BH-SNE map of the mouse sagittal data using 10 principal components for the initial dimensionality reduction and colored by anatomical region labels from the Allen reference Atlas. (b) Mapping of the BH-SNE embedding back to the 3D volume of the mouse atlas (3 views), colored by the L*a*b* colormap of the BH-SNE embedding. (c) Mapping of the BH-SNE embedding using the high dimensional sagittal data without prior dimensionality reduction.

reported in [141] on k-means clustering based on expression of cell-type specific genes.

The BH-SNE analysis was also performed on the sagittal data, which spans a smaller volume of interest than the coronal data (27,365 voxels compared to 61,164). This map (in Figure 3.4A and B) is highly comparable to that of the coronal experiments, with especially clear distinctions between cerebral nuclei (medium blue) and cerebral cortex (orange). In this case, selecting fewer principal components for the BH-SNE initialization does not lead to slicing effects as observed in the coronal samples, but it does emphasise within-slice similarities which are visible by a smoother color transitions in Figure 3.4C.

GENOME-WIDE GENE EXPRESSION SIMILARITY WITHIN THE HUMAN BRAIN

In the human brain atlas platform paper, Hawrylycz *et al*. [26] used PCA and cMDS mappings to show that the transcriptional relationships between cortical samples mimic the spatial topography of the cortex. To visualize the anatomical organization of the high-dimensional expression data through the entire adult human brain, we mapped the data to a 2D map using BH-SNE without prior dimensionality reduction. By mapping the expression data of each of the six brains separately, we observed that BH-SNE is able to map the samples with clear clustering of samples belonging to the same anatomical region, see Figure 3.5. Particularly, the cortex (red, yellow and brown samples) and the cerebellum (light blue) are clearly separated from all other brain regions across the six brains, indicating that both regions have distinct expression profiles from the rest of the brain. Moreover, the thalamus (light green), hippocampus (ochre) and the caudate nucleus and putamen (purple) are consistently close to each other in the low dimensional space, indicating that the expression profiles of these regions are more similar to each other than to other regions of the brain. Remarkably, the caudate nucleus and putamen are clustered, while the third organ of the basal ganglia, the globus pallidus, has a separate cluster. The relationships between samples in the low-dimensional space are very consistent across the six brains, suggesting a global organization of the human brain transcriptome across individuals. It is worth noting that BH-SNE optimizes the pair-wise distances between pairs of data points, but not the absolute location of each

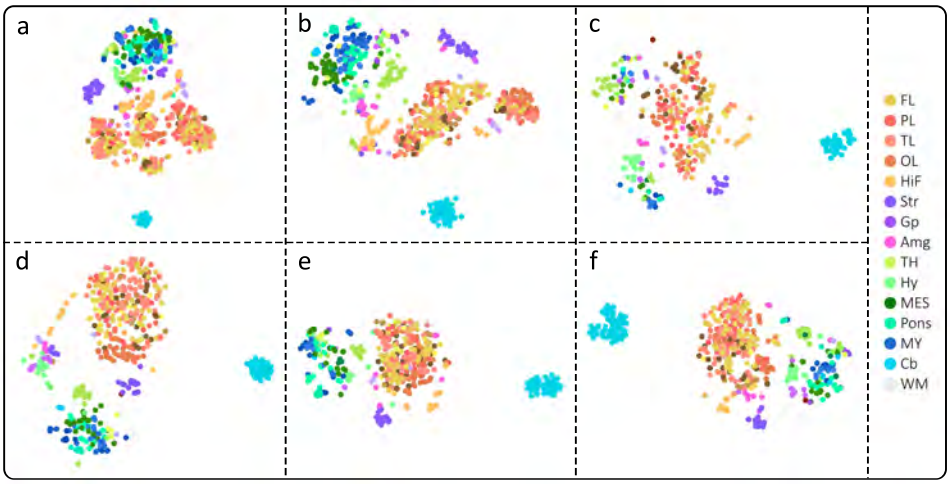


Figure 3.5: **Genome-wide gene expression similarity within six human brains.** BH-SNE maps of the expression data for each of the six human brain donors individually without prior dimensionality reduction by PCA. (1) Donor H0351.2001. (b) Donor H0351.2002. (c) Donor H0351.2009. (d) Donor H0351.2012. (e) Donor H0351.2015. (f) Donor H0351.2016. BH-SNE maps are colored by anatomical region labels from the Allen reference Atlas: frontal lobe (FL), parietal lobe (PL), temporal lobe (TL), occipital lobe (OL), hippocampal formation (HiF), striatum (Str), Globus pallidus (Gp), amygdala (Amg), thalamus (TH), hypothalamus (Hy), mesencephalon (MES), Pons, myelencephalon (MY), cerebellum (Cb), and white matter (WM). Note the higher density in (a) and (b) is due to the larger number of samples.

data point in the 2D map. Therefore, when analyzing the maps in Figure 3.5, one should consider the relative distances between samples from different neuroanatomical regions rather than the absolute geometric location within the map.

To compare the similarity of the expression profiles of different brain regions to the classical neuroanatomy, we compared the L*a*b* colors of the BH-SNE mapping to the original structural labels on the MNI152 atlas for Donor H0351.2001 and Donor H0351.2002, from whom both hemispheres were sampled in the ABA. In Figure 3.6, the differences between cortical, cerebellar, and brain stem samples are clearly visible. However, we could not identify differences between the frontal, medial, and anterior regions of the cortex in the t-SNE map. On the other hand, regions surrounding the ventricles clearly differ from the adjacent brain regions and there are no clear differences within the brainstem. These maps reveal the global symmetry between hemispheres in regional gene expression that was also reported in [26].

To gain insight into the effect of cell-type specific genes on the mapping, we inspected embeddings of human brain samples when creating the maps using only the expression profiles of cell-type specific genes. Similar to our observation from the mouse data, neuron-specific genes encode better separability between anatomical regions (Figure 3.7A) with a BH-SNE map very similar to the BH-SNE map obtained using the entire set of genes. Furthermore, astrocyte-specific genes resulted in a more separable map compared to oligodendrocyte-specific genes (Figure 3.7B and C). In all three mappings, there is a strong overlap between the frontal lobe (FL), temporal lobe (TL), occipital lobe

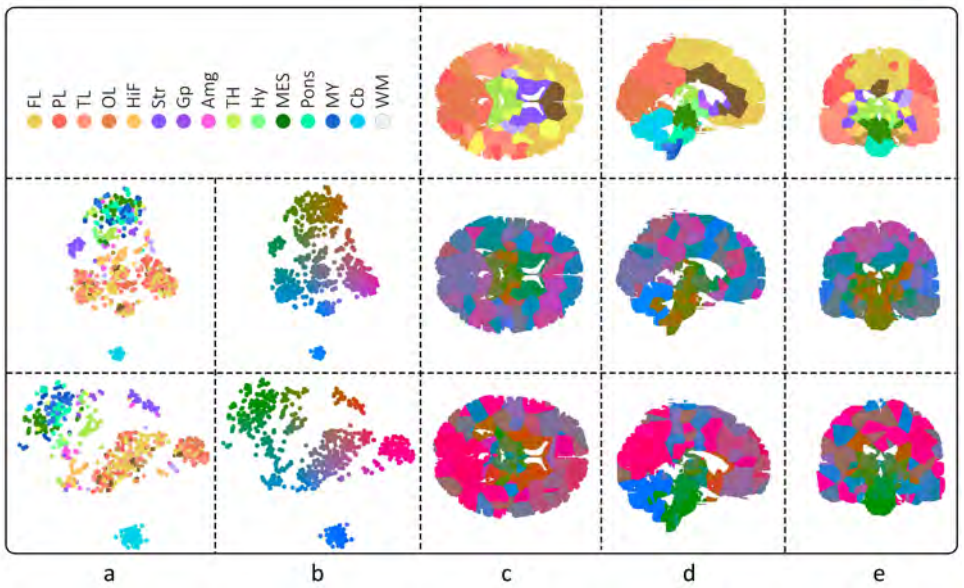


Figure 3.6: **Anatomical view on the genome-wide gene expression similarity within two human brains.** BH-SNE maps of the expression data without prior dimensionality reduction by PCA, (a) colored by anatomical region labels from the Allen reference Atlas (see legend) and (b) colored by the L*a*b* colormap. (c) Transverse, (d) sagittal, and (e) coronal views of the brain colored according to the L*a*b* colors of the samples in the BH-SNE maps. Top Row: Anatomical labels from the ABA projected on the full brain. Middle Row: the brain of H0351.2001. Bottom Row: the brain of H0351.2002. Note the clear separation of the cerebellum and brain stem in (b) as well as that the global preservation of the symmetry in regional gene expression between hemispheres in (c,e).

(OL), and parietal lobe (PL) (Figure 3.7D–F), which was also visible in the brain maps in Figure 3.6. It is also worth noting that the overall differences between the cell-type specific maps are much smaller in the human data that the corresponding maps in the mouse data (Figure 3.3).

We then pooled all the samples from the six brains and mapped them to the low-dimensional 2D space using BH-SNE (with prior dimensionality reduction to 10 principal components), PCA, and cMDS. The BH-SNE map of the concatenated data resembled those of the individual donors to a large extent, with samples from the cortex and the cerebellum clearly separated from samples in the rest of the brain, see Figure 3.8. Again, BH-SNE (Figure 3.8A) retains a better separation between samples belonging to different anatomical regions as compared to PCA (Figure 3.8B) and cMDS (Figure 3.8C). For PCA, the overlap between the cerebellum (light blue) and the cortex (red, yellow and brown) can be resolved when the 3rd component is taken into account. Within the cerebellum, none of the three methods could separate frontal (FL), temporal (TL), occipital (OL) and parietal lobes (PL), consistent with our findings from individual brains (Figure 3.7). This further supports the superiority of BH-SNE to retain variations in higher components in the 2D space. Separation between donors becomes apparent in BH-SNE (Figure 3.8D), which is much less apparent in PCA and cMDS. In order to quantify the

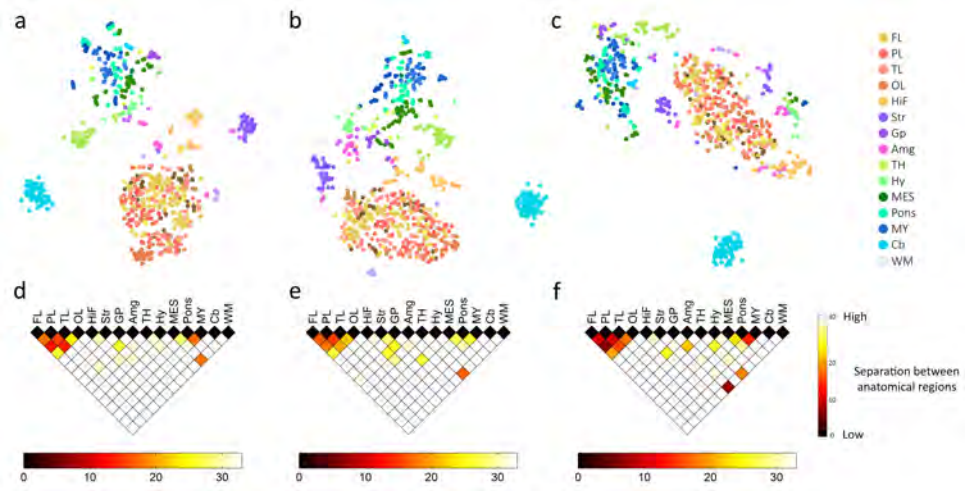


Figure 3.7: **BH-SNE mappings of the human brain based on cell-type specific genes.** BH-SNE embeddings of samples from donor H0351.2002 using (a) neuron-specific, (b) astrocyte-specific, and (c) oligodendrocyte-specific gene sets, colored by anatomical region labels from the Allen Reference Atlas. (d)-(f) Divergence plots of the BH-SNE embeddings using (d) neuron-specific, (e) astrocyte-specific, and (f) oligodendrocyte-specific gene sets. A higher divergence value (lighter colors) indicates better separation between a pair of neuroanatomical regions in the 2D BH-SNE map.

separation between different anatomical structures in the low dimensional space, we computed the Jensen-Shannon divergences between the regions; see Figure 3.8G–I. The divergence plots show clearly that the BH-SNE map has higher divergence values, hence retaining a better separation between all neuroanatomical regions followed by PCA and subsequently cMDS. The sensitivity of BH-SNE is also demonstrated by its ability to distinguish samples from different donors. Figure 3.8A and D show different clusters per brain region and per donor. In the PCA and cMDS mappings, samples from different donors are fully mixed.

In Figure 3.2F, we have shown that by retaining more PCA components, one can separate the mouse expression data based on inter-slice differences. To analyze the effect of retaining more PCA components prior to the BH-SNE embedding in the human brain, we gradually increased the number of principal components used to initialize the BH-SNE mapping. Figure 3.9 (top row) shows that by increasing the number of principal components, i.e. increasing the data dimensionality, before applying BH-SNE, samples belonging to the same anatomical structure, but to different donors, start to deviate from each other. When colored according to the source brain, Figure 3.9 (bottom row), deviations in the BH-SNE maps appear to reflect differences between brains only when more components are used in the prior dimensionality reduction. At 5 components (Figure 3.9C) we start to observe a separation of the samples from H0351.2001 brain regions (red) and H0351.2002 brain regions (yellow) from the other samples, especially in the cortical and cerebellar regions. The other four brains (H0351.2009, H0351.2012, H0351.2015, and H0351.2016) become clearly separated when much higher components

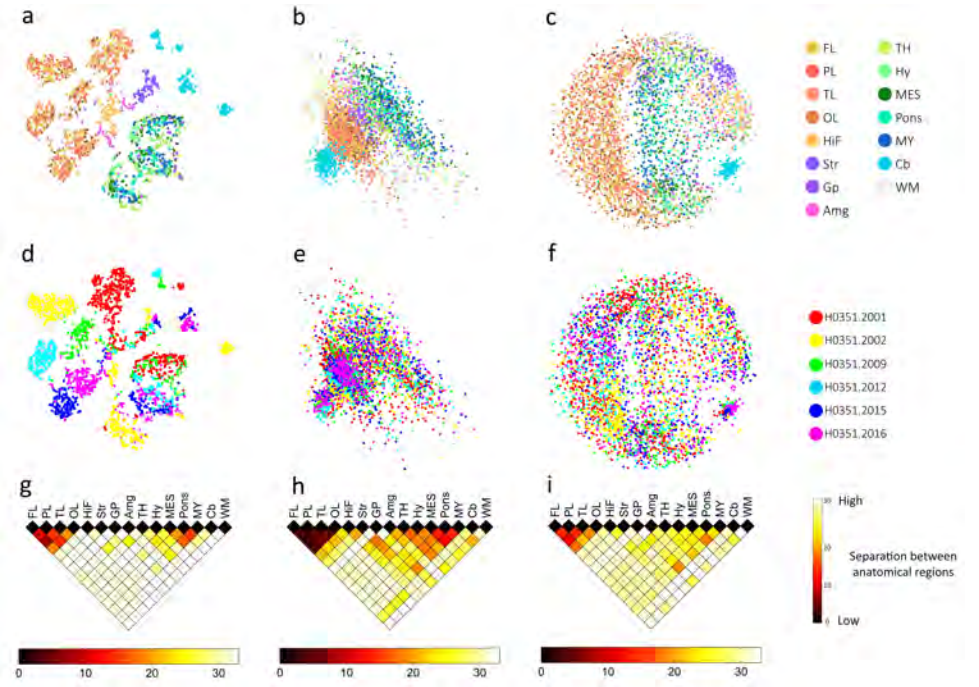


Figure 3.8: **Linear and non-linear embeddings of the human brain transcriptome.** (a) BH-SNE using 10 principal components for prior dimensionality reduction, (b) PCA, and (c) cMDS embeddings of the aggregated gene expression data of the six human brains. Maps a-c are colored by anatomical region labels from the Allen reference Atlas, maps d-f are colored by donor . (g – i) Jensen-Shannon divergence plots between neuroanatomical regions for BH-SNE, PCA and cMDS, respectively.

are included (20 components, Figure 3.9E). However, BH-SNE is still able to maintain the separation between different anatomical structures even when higher components are added, but at the costs that each brain region in each donor then forms its own cluster. The clustering in anatomy related clusters for a low number of retained components shows that the variations in normalized gene expression levels between brain regions are dominant over the variations between donors.

3.5. CONCLUSIONS

We have explored the effectiveness of using t-Distributed Stochastic Neighbor Embedding (t-SNE) to assess the spatial organization of genome-wide expression data across the mammalian brain. We have used Barnes-Hut-SNE (BH-SNE), a recently developed, computationally efficient t-SNE optimization algorithm, to map the large volumes of data in the mouse and human Allen Brain Atlases. Our results show that the mapped gene-expression data is highly consistent between the coronal and sagittal mouse atlases as well as between the six human brain datasets, with the cortex and cerebellum always being the most distinct from other brain regions. Additionally, the BH-SNE maps of the human brain show clear expression symmetry between hemispheres. The separa-

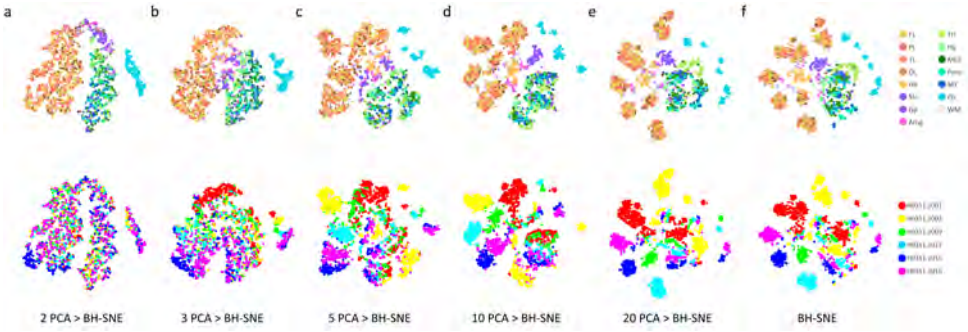


Figure 3.9: **BH-SNE maps of the human data when using different PCA initializations.** BH-SNE embeddings of the aggregated data from six human brains when using (a) 2, (b) 3, (c) 5, (d) 10 and (e) 20 principal components before applying BH-SNE. (f) BH-SNE mapping of the human data without any prior dimensionality reduction. (Top row) BH-SNE maps colored by anatomical region labels from the Allen reference Atlas. (Bottom row) BH-SNE maps colored by donor brain.

tion of neuroanatomical regions in the BH-SNE embedding is better than the separation in the PCA and MDS embeddings, further supporting the need for non-linear embedding methods to capture the complex organization of the Allen Brain Atlas data. We have employed the Jensen-Shannon divergence to quantify the ability of different gene sets and different embedding methods to map brain samples in 2D while preserving the known neuroanatomy. Additionally, we studied the effect of keeping more PCA components prior to the BH-SNE mapping. Due to its emphasis on local structures, BH-SNE is sensitive to having more PCA components as input, even when they may encode for non-anatomical information such as inter-slice differences in the coronal mouse data and different donor brains in the human data. These results suggest that to map high-dimensional spatial transcriptome data to a two dimensional space, a combination of a linear PCA mapping followed by a non-linear BH-SNE mapping gives the best tradeoff between preserving local and global structure in one 2D map.

3.6. SUPPLEMENTARY MATERIAL

The online version of this article contains supplementary material[2].

[2] http://www.sciencedirect.com/science/article/pii/S1046202314003211

CHAPTER 4

COMPREHENSIVE ISOFORM ANALYSIS CHARACTERIZES DYSTROPHIN FUNCTION IN HUMAN BRAIN DEVELOPMENT

Ahmed Mahfouz*
Nathalie Doorenweerd*
Maaike van Putten
Rajaram Kaliyaperumal
Peter AC 't Hoen
Jos GM Hendriksen
Annemieke Aartsma-Rus
Erik H Niks
Jan JGM Verschuuren
Marcel JT Reinders
Hermien E Kan
Boudewijn PF Lelieveldt

In preparation.
*Equal contribution.

DYSTROPHINOPATHIES are muscular dystrophies with a high incidence of learning and behavioural problems and comorbidity with neurodevelopmental disorders. However, the pathophysiology of central nervous system involvement remains elusive. We provide a detailed analysis of the expression profiles of the dystrophin isoforms in the human brain across development. Contrary to expectation, we found that the purkinje dystrophin isoform was virtually absent from the human brain, which we validated using additional data of promoter activity and epigenomic markers as well as ex vivo experiments. Furthermore, a co-expression analysis suggests a strong association between dystrophin transcripts and genes implicated in neurodevelopmental disorders, providing an underlying genetic basis to the co-morbidity of these disorders in dystropinopathy patients.

4.1. INTRODUCTION

Duchenne (DMD) and Becker (BMD) muscular dystrophies are X-linked genetic neuromuscular disorders characterized by severe and progressive muscle weakness. Mutations in the *DMD* gene result in absent/non-functional muscle dystrophin protein in DMD and shortened/partially functional protein in BMD.

In addition to skeletal muscle pathology, DMD is characterized by cognitive and behavioural problems with 30% of boys with DMD showing cognitive impairment (IQ below 70) [219] and 40% having reading deficits similar to those observed in patients with phonological dyslexia [220–222]. Moreover, there is a higher incidence of attention-deficit/hyperactivity disorder (ADHD) (32%), anxiety disorder (27%), autism spectrum disorders (ASD) (15%), epilepsy (6.3%), and obsessive-compulsive disorder (OCD) (4.8%) in patients with DMD [223–225]. In BMD patients, frequencies of learning difficulties or comorbidity with neurodevelopmental disorders have not been systematically reviewed. However, one report of 24 patients does indicate spelling (32%), arithmetic (26%), and reading (21%) difficulties, as well as behavioral problems and occurrence of epilepsy despite absent deviations from full scale IQ (FSIQ) distributions [226, 227]. A case report of four patients may suggest the possibility of BMD presenting with central nervous system (CNS) symptoms in the absence of muscle weakness [228]. Caution is warranted, however, when projecting these percentages to the general BMD population as selection bias cannot be excluded.

The DMD gene contains at least seven independent, tissue-specific promoters and two polyA-addition sites, producing several isoforms that are named to reflect their length and splice-variants (Figure 4.1). The localization and function of the full-length muscle isoform Dp427m is well characterized both in humans and animal models. It is a crucial component of the dystrophin-glycoprotein complex (DGC), which bridges the inner cytoskeleton and the extracellular matrix providing structural stability to muscle fibers [229, 230]. However, information on brain dystrophin is almost solely derived from animal models and cell culture studies, with only a few case studies in man [231]. It is believed that the cortical isoform Dp427c is predominantly expressed in neurons of the cortex and the CA regions of the hippocampus [232, 233]. The Purkinje isoform Dp427p has two variants which are reported to be expressed in cerebellar Purkinje cells [234]. The shorter Dp260 and Dp116 isoforms are expressed primarily in the retina and the peripheral nerve, respectively [235, 236]. There is very limited information on the

sites of expression of the Dp140 isoform and its splice variants. A study of one 3.5 month old fetus and one 60 year old brain suggested that the Dp140 isoform is predominantly expressed during fetal life stages [237]. Finally, the Dp71 isoform is ubiquitously expressed, with higher levels in the CNS [238, 239].

The risk of cognitive impairment in DMD has been associated to the location of mutations within the DMD gene which results in the absence of specific dystrophin isoforms. FSIQ scores correlate with the number of isoforms missing. Patients missing all isoforms due to mutations in the distal part of the gene have the lowest scores, whereas patients missing only the full-length isoform have the highest scores [240]. Moreover, patients lacking Dp140 isoforms performed worse on all neuropsychological tests (general cognitive abilities, verbal memory, attention and executive functions) compared to those with preserved Dp140 [241]. The relationship between the isoforms that are affected and the cognitive profile is further supported by the higher incidence of neurodevelopmental disorders in patients missing Dp140 compared to patients missing only Dp427 [223, 242]. This group distinction was already detectable below the age of four, as assessed with developmental quotients [243]. Finally, imaging studies has shown reduced grey matter volume and altered white matter microstructure compared to age-matched healthy controls, which also was more profound in patients also missing Dp140 [244].

Despite this mounting evidence on the association between the absence of shorter dystrophin isoforms and higher incidence of learning and behavioral disabilities, the etiology of the CNS pathology in DMD and BMD remains elusive. In this study, we provide detailed analysis of the spatial and temporal expression patterns of the dystrophin isoforms in the pathology-free adult and developing human brain. Using co-expression analysis, we characterize the functional role of the dystrophin isoforms as well as their relationships to other neurological disorders across brain development.

4.2. RESULTS

DIFFERENTIAL DYSTROPHIN ISOFORM EXPRESSION DURING DEVELOPMENT

We used the BrainSpan atlas of the developing human brain transcriptome [27] to assess the dystrophin isoforms expression throughout development. The BrainSpan atlas provides RNA-sequencing expression profiling of 16 brain structures from 42 donor brains spanning early pre-natal development (8 weeks post-conception) to adulthood (40 years of age). In order to assess the expression of the different dystrophin isoforms, we used the expression of the unique first exons of Dp427p Dp427c, Dp427m, Dp260, Dp140, Dp116 and the shared first exon of Dp71 and Dp40 (Figure 4.2). We grouped the donors into 10 developmental stages (Supplementary Table 1). Figure 4.2 shows the expression of all exons within the DMD gene, across different brain regions and through development. The expression of Dp427c, and Dp427m is low during fetal development, shows a slight increase around the age of two, and is low throughout middle adulthood. This pattern is consistent across the different brain regions, though more prominent in the cerebral cortex when assessing Dp427 exon two (Figure 4.2B-D).

In contrast to previous reports [234, 245, 246], the Purkinje isoform Dp427p was virtually absent in the brain throughout development, with expression levels even lower

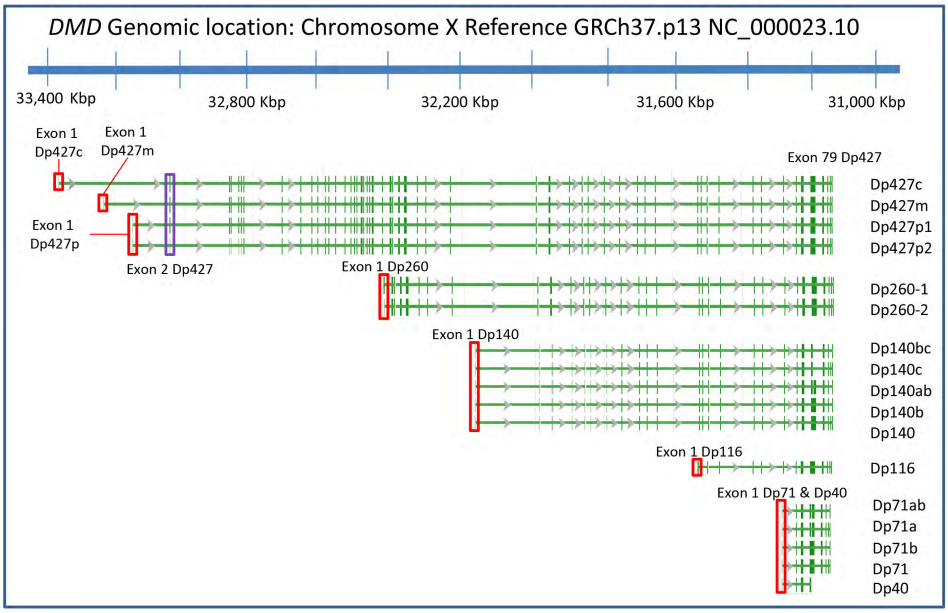


Figure 4.1: **Human dystrophin isoforms.** Dystrophin transcripts located on the X chromosome (GRCh37.p13, RefSeq Release 74: NG_012232.1). The vertical green dashes indicate individual exons. The full length dystrophins (Dp427) have 79 exons, with isoforms starting at unique first exons. For some isoforms multiple splice variants have been identified (indicated on the right hand side). The shorter isoforms (relative to the full length isoforms Dp427) have unique first exons (i.e. not included in any other isoform), with the exception of Dp71 and Dp40 which share a first exon but use alternative polyadenylation sites. The red boxes indicate the position of the promoter region of each isoform. The second exon of Dp427c,m,p was used to represent the full length dystrophin as a group (indicated by a purple box). This figure was generated using the NCBI Sequence Viewer.

than muscle dystrophin Dp427m. To verify that Dp427p is indeed expressed in mouse, but not in human brain, we analyzed the expression of Dp427p in cerebellum and cerebral cortex samples from control adult human brain (provided by the Netherlands Brain Bank) using quantitative polymerase chain reaction (qPCR). Indeed, we did not observe Dp427p expression in the human cerebral cortex and also not in the cerebellum, where the Purkinje neurons are located (Figure 4.3A). Yet in line with previous studies [245, 246], Dp427p was expressed in the mouse cerebellum and not in the mouse cerebral cortex. This sharp contrast in Dp427p expression in the cerebellum suggests a different role for Dp427p in human than in mouse.

As expected, the samples representing retinal Dp260 and peripheral nerve Dp116 have virtually no expression in the brain (their unique first exons are not expressed). By contrast, Dp140 is clearly expressed in the fetal brain, with high expression in the early to mid-fetal stages, but very low expression from the late fetal stage onwards. Nevertheless, Dp140 is still expressed in the cerebellum and cerebral cortex at middle adulthood, which has never been reported before (Figure 4.2E). To verify that Dp140 is indeed expressed in the adult cerebral cortex and cerebellum, we analyzed the expression of

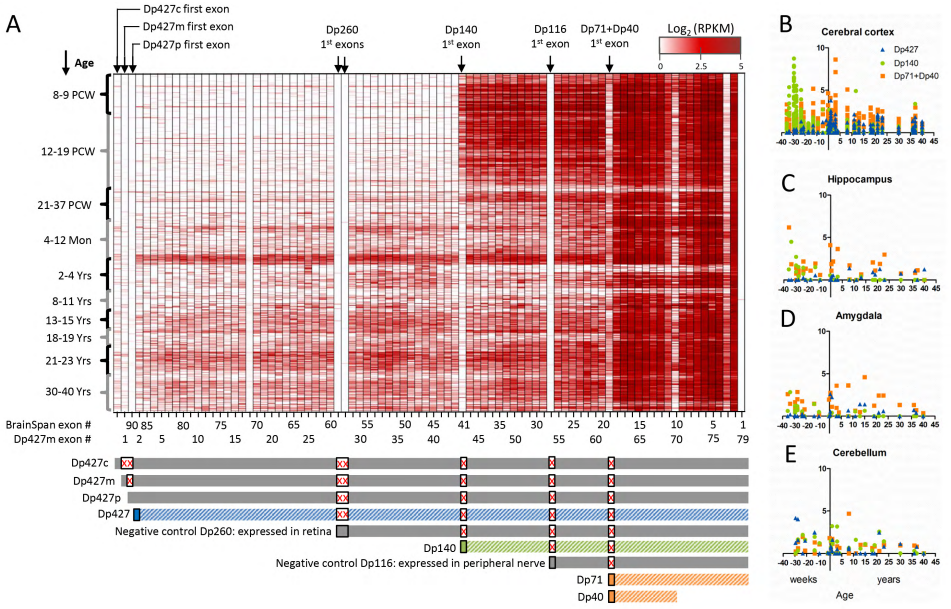


Figure 4.2: **Dystrophin isoforms expression across brain development.** (A) Dystrophin exons expression throughout brain development. The isoform unique first exons are indicated on top of the heatmap. The developmental stages are indicated on the left in post-conceptual weeks and months or years after birth. The BrainSpan atlas exon number is indicated below the heatmap together with the Dp427m exon numbering for reference. Bars below the heatmap indicate the different isoform groups. The grey bars corresponding to Dp427c,m,p are grouped together using exon 2 in further analysis (dark blue). The grey bars corresponding to Dp260 and Dp116 are expressed in the retina and peripheral nerves and are excluded from further analysis. The first exon of Dp140 (dark green) and Dp71+Dp40 (dark orange) is used for further analysis. Boxes with a red 'X' indicate exons that are not part of the transcript. Expression values are presented as $log_2(RPKM)$. Brain region specific expression across development is shown for the cerebral cortex (B), hippocampus (C), amygdala (D) and cerebellum (E) of Dp427 (second exon), Dp140 (first exon) and Dp71+Dp40 (first exon).

Dp140 as described above for Dp427p. Results confirm expression on Dp140 in the adult human cerebral cortex, as well as much higher expression in the adult human cerebellum (Figure 4.3B). The Dp71+Dp40 expression is high during fetal stages and remains high after birth and later in life showing little regional specificity, in line with earlier reports indicating ubiquitous expression [231, 247]. This is further supported by qPCR results showing comparable expression levels of Dp71 between the cortex and cerebellum (Figure 4.3B).

DMD EXPRESSION IN THE ADULT HUMAN BRAIN IS HIGH IN THE HIPPOCAMPUS AND AMYGDALA BUT LOW IN THE CEREBELLUM RELATIVE TO THE BRAIN AVERAGE EXPRESSION

To analyze the spatial distribution of DMD gene expression across the adult brain, we used the Allen Human Brain Atlas (AHBA) [26], which has a much higher spatial resolution than the BrainSpan atlas but lacks the temporal dimension. The AHBA provides mi-

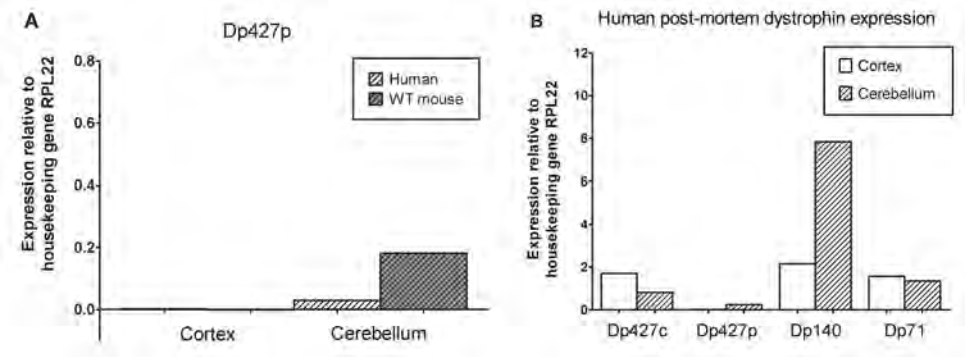


Figure 4.3: **Ex vivo dystrophin isoforms expression in the mouse and human cerebral cortex and cerebellum.** (A) The expression of the Purkinje isoform Dp427p was measured using qPCR in samples from the cortex and cerebellum of human and wild type mouse brains. (B) Expression levels of Dp427c, Dp427p, Dp140, and Dp71 in the human cortex and cerebellum. Expression levels are shown relative to the housekeeping gene *RPL22*.

croarray gene expression data from hundreds of samples extracted from six adult human brains, allowing detailed analysis of the regional expression of genes across the human brain. However, as oligo-dT primers were used for sample preparation, which capture the distal part of the gene, it is not possible to distinguish between different isoforms, nor is exon specific data available.

Relative to the average expression levels across the six donors, the highest expression levels of *DMD* were found in the hippocampus and amygdala (Figure 4.4). Within the hippocampus, expression was highest in the CA4 region, and lowest in the CA2 region (Figure 4.4A, Supplementary Table 2). The expression of *DMD* in the amygdala was highest in the basolateral complex, the input side of the amygdala that receives information from the prefrontal cortex, which is implicated in complex behaviour. Relatively low *DMD* expression was found on the output side with the central nucleus which connects with the brainstem and pons. Of the basolateral complex, highest *DMD* expression was found in the lateral nucleus which receives information from the neocortex, thalamus and hippocampus.

Animal studies have thus far consistently shown high dystrophin expression in the cerebellum [248, 249]. Surprisingly, the lowest levels of *DMD* expression in the human brain were found in the cerebellum and the pons (Figure 4.4B). Within the cerebellum, *DMD* expression was lowest in the globose (GL), fastigial (Fas) and dentate nuclei (DN) which receive inhibitory (GABAergic) input from Purkinje cells and excitatory (glutamatergic) inputs from mossy fibres and climbing fibre pathways. Second lowest expression was located in the regions associated with working memory, in the biventral lobule (Bl). The regions implicated in timing and coordination as well as attention through the prefrontal cortex, in the tonsilla (TO) and semilunal lobule (SL) were third lowest.

TRANSCRIPTION START SITES IN THE *DMD* GENE

Gene expression is regulated by multiple factors that integrate at transcription start sites (TSSs) to control the transcription of target genes in a cell-specific manner [250]. To

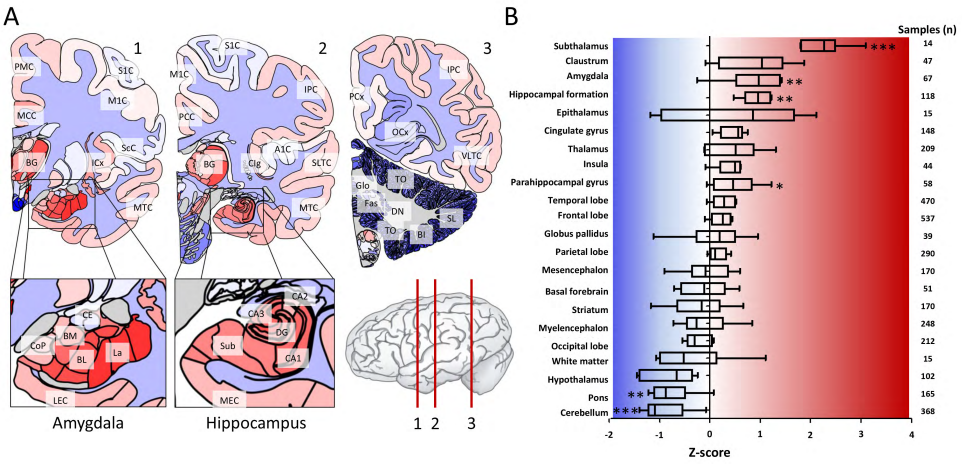


Figure 4.4: ***DMD* gene expression across the adult human brain.** *DMD* expression in the adult human brain at high spatial resolution averaged from six adult donors (five males and one female; mean age 42 years). Data is shown relative to the average expression across the whole brain (z-score normalization). (A) The spatial distribution is highlighted in three cross-section of the brain showing the high sub-structural expression in the amygdala and hippocampus in contrast to the low expression throughout the cerebellum. (B) The brain was subdivided into 22 non-overlapping anatomical regions. For each region, the average expression in each of the donors was calculated separately (after z-score normalization) and all six average values are shown in a boxplot. The number of samples from which these samples were derived are indicated to the right. Significantly higher expression was found in the subthalamus, amygdala, parahippocampus and hippocampal formation. Significantly lower expression was found in the cerebellum and pons (Mann-Whitnney U-test; $*P < 0.05, **P < 0.01, ***P < 0.001$). Full structure names for the indicated acronyms can be found in the Materials and Methods.

better characterize the activity of the dystrophin transcripts across different tissues and cell-types, we analyzed TSS usage within the *DMD* gene across different brain regions. Genome-wide TSS usage has been detected across many human cell-types in the FANTOM5 consortium data set (using cap analysis of gene expression; CAGE) [251, 252] and by the Roadmap Epigenomic Consortium using chromatin markers specific to TSSs [75].

Using the FANTOM5 data, we mapped the usage of TSSs from tissue samples of the amygdala, hippocampus, cerebellum and cerebral cortex based on the TSS expression in the adult human brain (Figure 4.5). In total, there were 25 TSSs within a window of 1kb of the first exons of Dp427c, Dp427m, Dp427p, Dp260, Dp140, Dp116, and Dp70+Dp41 (Supplementary Table 3). Consistent with our findings from the BrainSpan data and the qPCR experiment (Figure 4.2 and Figure 4.3), the TSSs of the Purkinje isoform were not expressed in any of the samples analyzed. Similarly, we did not observe expression of the TSSs of Dp260 and Dp116. In addition, the expression of the TSSs of Dp427c was highest in the amygdala and hippocampus, in line with the observations from the AHBA analysis (Figure 4.4). The short isoforms Dp71+Dp40 were consistently expressed across the brain with lower expression in the cerebellum, in line with results from the BrainSpan analysis (Figure 4.2A). The TSSs of Dp140 were expressed throughout the adult brain with higher expression in the cerebellum compared to the rest of the brain. The higher expression in the cerebellum is in line with the BrainSpan data (Figure 4.2E). In contrast

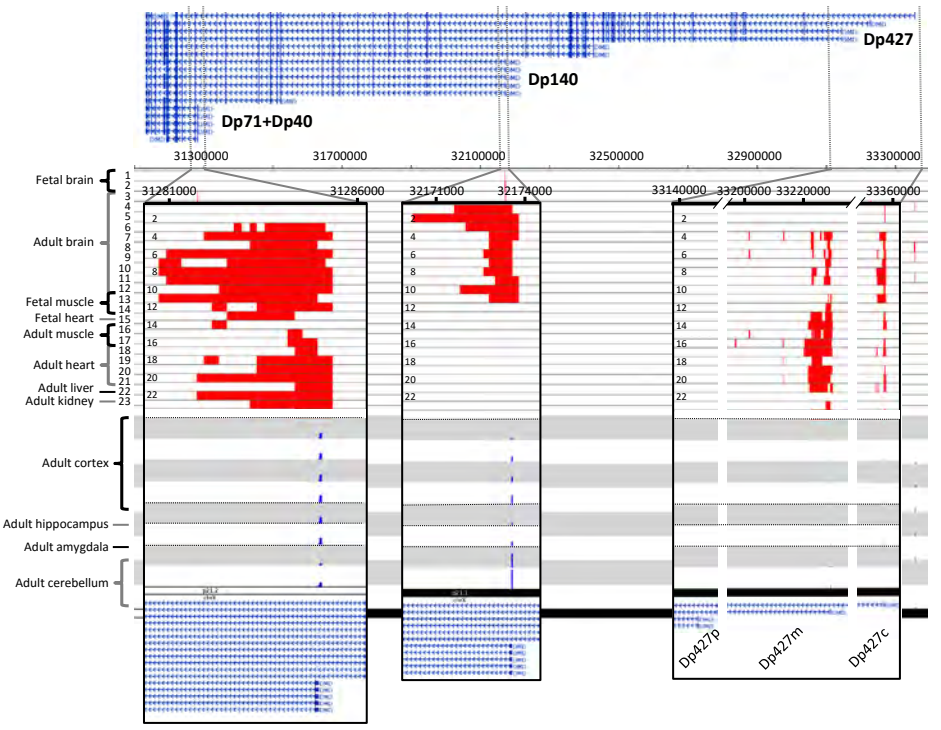


Figure 4.5: ***DMD*transcription start sites.** Genome browser view of the different TSSs within the *DMD* gene. Active TSS state based on histone markers within the *DMD* gene are shown for 23 samples including fetal and adult brain, muscle, heart, liver and kidney. See Materials and Methods for detailed sample information. Red bars indicate an active TSS state as defined by the Roadmap Epigenomic Consortium [75]. The bottom eight tracks show the TSS activity (blue bars) within the first exons of the different isoforms captured by CAGE sequencing from the FANTOM5 project [251, 252]. All the active TSSs have been highlighted by zooming in on the first exons of the different isoform groups, from left to right: Dp71+Dp40, Dp140, and Dp427. Note the absence of any TSS activity or epigenetic markers for Dp427p. Data is aligned to the human reference genome (GRCh37) and RefSeq transcripts are shown at the top. Data is plotted using the WashU Epigenome Browser [253].

to the low expression levels observed after birth in the BrainSpan data, the expression levels of Dp140 TSSs were high in the cerebral cortex, hippocampus and amygdala of the adult brain. However, this matches our results from the qPCR analysis of the adult cerebral cortex and cerebellum (Figure 4.3B). Finally, the expression of the TSSs of the muscle isoform (Dp427m) was low across the brain except in the cerebellum.

To further investigate TSS usage within the DMD gene, we used the data from the Roadmap Epigenomic Consortium [75] to identify TSSs based on their chromatin signatures. We analyzed the chromatin signatures (i.e. chromatin states) across the DMD gene, focusing on *active TSS*, in adult and fetal brain samples as well as samples from the muscle, heart, liver, aorta, and kidney (see Materials and Methods for a full list of samples). In general, brain samples showed high TSS activity for the Dp140 and Dp71+Dp40 isoform groups (Figure 4.5B), while the muscle and heart samples showed high TSS ac-

tivity for the Dp427 isoforms group. The fetal brain samples showed active TSS at the first exon of the Dp140 isoform but no active TSS for the Dp71+Dp40 isoform, supporting the expression patterns shown in Figure 4.2B. The Dp140 isoform contained active TSS markers in the Neurospheres Cortex Derived, Angular Gyrus, Germinal Matrix and Mid Frontal Lobe samples and no active TSS in the Substania Nigra, Anterior Caudate, Cingulate Gyrus, and Inferior Temporal Lobe samples.

DYSTROPHIN ISOFORMS ARE SIGNIFICANTLY CO-EXPRESSED WITH GENES IMPLICATED IN NEURODEVELOPMENTAL DISORDERS

To get more insight into the functional role of dystrophin throughout human brain development and its association to other neurodevelopmental disorders, we analyzed the spatial and temporal co-expression relationships of the *DMD* gene and the different dystrophin isoforms. Co-expression analysis is a well-established approach to infer functional associations of genes using high-throughput expression data based on the 'guilt by association' principle [89]. First, we ranked all genes based on the correlation of their expression pattern to the *DMD* gene in the AHBA and to the three dystrophin isoform groups (Dp427, Dp140 and Dp71+Dp40) in the BrainSpan atlas, resulting in four ranked gene lists (Supplementary Table 4). Next, we tested whether genes related to five disorders with high incidence in DMD patients (ASD, intellectual disability (ID), ADHD, OCD, and dyslexia; Supplementary Table 5) are overrepresented among genes which are strongly co-expressed with *DMD* and the three isoforms (Figure 4.6).

Genes associated with ASD and ID were significantly co-expressed with dystrophin expression patterns for both the full-length and smaller isoforms, especially Dp140 (FDR-corrected $P < 5.66 \times 10^{-4}$; one-sided Mann-Whitney U-Test; Figure 4.6). In addition, ADD- and OCD-related genes were significantly co-expressed with Dp427 (FDR-corrected $P < 4.3 \times 10^{-4}$; one-sided Mann-Whitney U-Test), and dyslexia-related genes with the adult dystrophin expression, as well as Dp427 and Dp71+Dp40 expression in the developing brain (FDR-corrected $P < 2.98 \times 10^{-3}$; one-sided Mann-Whitney U-Test).

We mapped the top 25 genes based on their co-expression with Dp427, Dp140 and Dp71+Dp40 in the developing human brain to a co-expression network, together with their disease associations from DisGeNET [79] (Figure 4.7). The Dp140 network shows a higher co-expression between genes compared to the networks of Dp427 and Dp71+Dp40. The overlaid disease annotations show strong co-expression between dystrophin isoforms and other relevant diseases such as epilepsy, mental retardation, obesity, nervous system malformation, neurodevelopmental disorders and cardiovascular problems. These co-expression relationships point toward a functional association between *DMD* and genes related to these disorders.

To get an insight into the functional role of dystrophin throughout brain development, we assessed genes with strong co-expression to the three different dystrophin isoform groups for enrichment in gene ontology (GO) terms (Figure 4.7, Supplementary Table 6). GO-terms associated to Dp427 mainly relate to signal transduction by regulating membrane transport of ions, cations, synaptic transmission or membrane potential regulation. Genes co-expressed with Dp140 were enriched in GO-terms related to early neurodevelopment via regulation of neuron differentiation and neuron projec-

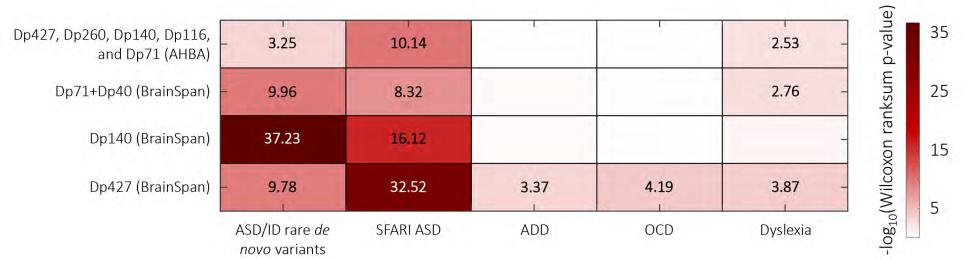


Figure 4.6: **Genes co-expressed with dystrophin isoforms are enriched in disease-related genes** Genes co-expressed with *DMD* gene across the adult human brain as well as the dystrophin isoforms across brain development (rows) are analyzed for enrichment in genes harboring rare de novo variants in ASD and ID probands, a curated set of ASD risk genes (SFARI ASD), ADD-, OCD-, and dyslexia-related genes. Heatmap colors correspond to $-log_{10}(FDR-correctedPvalue)$. All enrichment values for the lists enriched at $P < 0.05$ (one-sided Mann-Whitney U-test; FDR-corrected) are shown.

tion morphogenesis as well as chromatin modification. Finally, genes co-expressed with Dp71+Dp40 are enriched in terms related to signal transduction again with transmembrane receptor binding, but in relation to growth factors.

4.3. DISCUSSION

Despite the evidence supporting brain abnormalities in DMD patients, the mechanisms underlying the CNS involvement in this disorder is largely unknown. We provide a comprehensive study of the expression of dystrophin isoforms in the healthy human brain across anatomical regions and developmental stages. The detailed analysis of the expression patterns of the dystrophin isoforms and their co-expression relationships provides a better understanding of the role of dystrophin role in human brain function and the association between dystrophin and brain abnormalities .

The full length isoforms Dp427c and Dp427m show very low yet detectable expression throughout human development, confirming earlier reports [233, 254]. However, the Purkinje isoform (Dp427p) showed almost no expression in the developing human brain data which is in contrast an earlier study by Gorecki *et al.* [255] in which they established the Purkinje specificity of Dp427p by showing its expression in the mouse cerebellum . A later report by Holder *et al.* [234] showed expression of Dp427p in one adult cortical sample and no expression signal in a 20-week old fetal brain sample using PCR. To further confirm our findings from the BrainSpan atlas, we validated the absence of Dp427p expression from adult human cerebral cortex and cerebellar samples using qPCR and by interrogating TSS usage information from the FANTOM5 and Roadmap Epigenomic projects. As previously reported, we could detect Dp427p expression in mouse brain by qPCR. These results illustrate discrepancies in dystrophin isoforms expression between human and mouse brains and highlight the importance of comprehensive maps of expression in mouse and human brains for better translation between animal experiments to human conditions in which the *DMD* gene is implicated.

The virtual lack of Dp260 and Dp116 expression confirms earlier hypotheses on the exclusive expression of these isoforms in the retina and peripheral nerves, respectively,

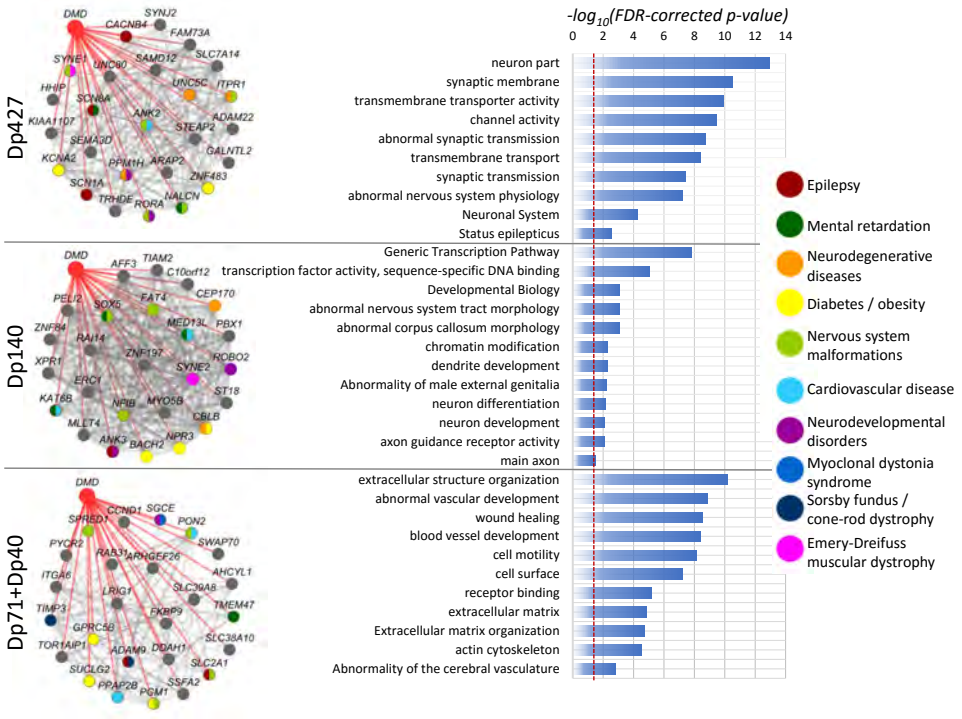


Figure 4.7: **Dystrophin isoforms co-expression networks and their associated GO-terms.** Co-expression networks showing the top 25 positive correlating genes to the dystrophin isoforms Dp427, Dp140 and Dp71+Dp40 across development. Edges between nodes represent correlation strength (the thicker the line the more correlated the genes). Correlations to dystrophin isoforms are shown in red (not weighted). Genes are color-coded according to their disease associations from DisGeNet. The bar plot shows the top terms enriched within the top 200 correlated genes. The vertical red line indicates the significance level ($FDR-corrected Pvalue < 0.05$).

despite a recent report indicating that Dp260 may be expressed in brain as well [256]. Based on the BrainSpan data, Dp140 is expressed mainly during the fetal stages of life across the brain and at middle adulthood in few samples from the cerebellum and cortex. These results are in line with earlier results from fetal and adult human brain samples, where western-blot analysis showed Dp140 in the fetal brain but not in the adult brain [237, 247, 248]. Interestingly, the AHBA data indicates that *DMD* expression is low in the cerebellum relative to the rest of the brain. Based on our validation we observed high expression of Dp140 in the adult human cerebellum and much lower expression in the cortex, in contrast to the earlier observation from AHBA. Further analysis of the expression of the Dp140 TSS from the FANTOM5 data indicates that in deed Dp140 is expressed in the adult human cerebellum.

The Dp71 and Dp40 isoforms showed consistent expression throughout brain development. Although we could not differentiate between Dp71 and Dp40 isoforms because they share a first exon, the high expression level of the last exon (belonging to Dp71 and

not Dp40) suggests some specificity for the high signal to Dp71.

Cerebellar dysfunction has been suggested to underlie deficits in reading and verbal working memory, an important component of the DMD cognitive deficit [257]. However, our finding of low *DMD* expression in the healthy adult human cerebellum does not support this notion. Our results rather emphasize the amygdala, involved in emotion regulation, and the hippocampus, involved in memory, based on their high expression of *DMD* and the supporting evidence of memory and emotion deficits in DMD from animal and neuropsychological studies in humans [223, 242, 258, 259]. Moreover, we show a relatively even distribution of *DMD* expression throughout the cortex, involved in higher-order cognitive functioning. This is well in line with magnetic resonance imaging (MRI) studies demonstrating reduced grey matter volume, altered white matter microstructure and reduced cerebral blood flow in DMD patients compared to healthy age-matched controls [244]. This correspondence between the imaging and gene expression analysis implies that the alterations observed in MRI are due to the lack of one or more dystrophin isoforms in brain rather than secondary effects of muscle weakness, limited mobility, and corticosteroid treatment or cardiac medication.

Co-expression analysis indicates a strong association between the dystrophin isoforms and genes implicated in ASD and ID, suggesting a common genetic mechanism which might explain the high incidence of these disorders in DMD and BMD patients. Functional enrichment analysis of strongly co-expressed genes advocates a functional distinction between Dp427 and Dp71+Dp40 isoforms, involved signal transduction, and Dp140 which is related to early neurodevelopment. The association of Dp427 and signal transduction further supports the proposed dystrophin-glycoprotein complex-like structure positioned in post-synaptic densities of GABA-ergic neurons in the brain [231]. The association of Dp71+Dp40 signal transduction plus transmembrane receptor binding in relation to growth factors implies that the structural alterations thus far observed in the brains of patients missing Dp427 and Dp140 [244] may be further aggravated in patients missing Dp71 and Dp40. Future studies on cerebral structural integrity in the absence of Dp71 and Dp40 can test this hypothesis.

Genes with strong co-expression to dystrophin isoforms across brain development may provide novel insights into the molecular mechanisms in DMD. For instance, Dp427 and Dp140 are strongly co-expressed with genes associated Emery-Dreifuss muscular dystrophy (*SYNE1* and *SYNE2*, respectively). Cerebral thromboembolism is one of the risk factors in Emery-Dreifuss muscular dystrophy and a link to *DMD* might explain the reduced cerebral blood flow seen in DMD patients. Similarly, Dp427 is strongly co-expressed with *NALCN*, the gene mutated in infantile hypotonia with psychomotor retardation and characteristic facies (IHPRF; OMIM: 615419) [260, 261]. Both DMD and IHPRF patients suffer from speech delay in early in development, suggesting a common mechanistic pathway involving *DMD* and *NALCN* [262]. Even more interestingly, Dp140 is strongly co-expressed with *FAT4*, the gene responsible for the Van Maldergem Syndrome 2 (VMLDS2; OMIM: 615546) which is characterized by mental retardation, deafness and skeletal and limb malformation and involves neuronal migrational abnormalities [263]. This association together with the higher incidence of mental retardation in DMD patients missing Dp140 [223, 241] as well as the expression of Dp140 in early developmental stages strongly suggests a role for Dp140 in neuronal migration. It

has long been thought that Dp260, the dystrophin isoform expressed in the retina, is responsible for the high incidence of color blindness and abnormal retinal neurotransmission [256, 264]. However, the strong co-expression of Dp71+Dp40 with *TIMP3* and *ADAM9*, which are associated with Sorsby fundus dystrophy [265] and cone-rod dystrophy 9 [264], respectively, points toward a brain role in retinal abnormalities in DMD patients.

Based on the expression profiles and the co-expression relationships of the dystrophin isoforms across brain regions and developmental stages, we have found a strong association between dystrophin and neurodevelopmental processes and disorders. Our results indicate a necessity to profile the expression of the dystrophin isoforms in DMD brains in order to elucidate the transcriptional mechanisms underlying the behavioral and learning problems in DMD. This can greatly facilitate risk assessments of comorbid disorders and guide screening for early detection and targeted treatment.

4.4. MATERIALS AND METHODS

BRAINSPAN DEVELOPING HUMAN BRAIN TRANSCRIPTOME

RNA-sequencing-derived exon-level expression data of the different isoforms of *DMD* was downloaded from the BrainSpan atlas of the developing human brain transcriptome [27](`http://brainspan.org`). RNA sequencing (RNA-seq) data generated from 524 tissue samples collected from 42 post-mortem brains collected from neurologically unremarkable individuals spanning early pre-natal development (8 post-conception weeks, PCW) to late adulthood (40 years of age). Samples were extracted using macro dissection from 8–16 regions per brain. Details of tissue acquisition and data processing can be found at (`http://brainspan.org`). Gene annotation of the RNA-seq data was derived from Gencode version 10 (GRCh37 – Ensembl 65; `http://www.gencodegenes.org/releases/10.html`). The expression level of the exons was measured in RPKM (reads per kilobase of exon model per million mapped reads).

The *DMD* gene (Chromosome X: 31,115,794-33,357,558) data included 94 exons. There are only 79 exons in the well-known muscle dystrophin Dp427m (NM_004006.2). We therefore mapped the exon locations to the Dp427m exon annotation (Figure 4.2). In this process, we found values specific to the isoform first exons (which are not exons in Dp427m) and we found genomic coordinates that mapped to an isoform called Dp427l. This isoform is no longer included in the latest release of the human genome (GRCh38), due to lack of evidence [266]. As such, we excluded these exon coordinates from further analyses.

ADULT HUMAN BRAIN EXPRESSION DATA

Spatial gene expression data from six adult human brains was obtained from the Allen Human Brain Atlas database (AHBA) [26] (`http://human.brain-map.org/`). Samples were collected from postmortem brains from 5 males and one female between 24 and 57 years of age (mean age 42), with no known psychopathologies, by either manual macrodissection (cortical and some subcortical structures) or by laser-based microdissection (subcortical and brainstem areas). For each brain, RNA was extracted from 363

to 946 different samples per brain (3,702 samples in total) and measured on custom Agilent microarrays containing the 4 × 44K Agilent Whole Human Genome probes as well as an additional 16,000 custom probes. For genes with 2 probes, the one with the highest variance was selected. For genes with at least 3 probes, the connectivity of each probe was calculated (sum of the Pearson correlations to all other probes, measured per brain and then averaged) and the one with the highest connectivity was selected. Expression data of the 19,991 genes was z-score normalized per brain. The expression of the DMD gene was measured using six probes, of which A_24_P185854 (NM_004023.1) has the highest connectivity and hence was used for further analysis. This probe is located at the distal part of the gene and captures the Dp71, Dp116, Dp140, Dp260, and Dp427 isoforms (all except Dp40). For visualization, z-score values of the *DMD* gene expression were mapped to anatomical atlas images acquired from the Allen Human Brain Atlas [26]. The following acronyms are indicated in Figure 4.4A; BG: Basal ganglia, BL: Basolateral nucleus, BM: Basomedial nucleus, BL: Biventral lobule, CA1: CA1 region of the hippocampus, CA2: CA2 region of the hippocampus, CA3: CA3 region of the hippocampus, Cig: Caudal granular insular cortex, CE: Central nuclear group, DG: Dentate gyrus, DN: Dentate nucleus, Fas: Fastigial nucleus, Glo: Globose, Icx: Insular neocortex, LEC: Lateral entorhinal cortex, La: Lateral nucleus, MEC: Medial lateral entorhinal cortex, MCC: Midcingulate gyrus, MTC: Midlateral temporal cortex, Ocx: Occipital neocortex, PCx: Parietal neocortex, PCC: Posterior cingulate cortex, CoP: Posterior cortical nucleus, IPC: Posteroventral parietal cortex, IPC: Posteroventral parietal cortex, PMC: Premotor cortex, A1C: Primary auditory cortex, M1C: Primary motor cortex, S1C: Primary somatosensory cortex, SL: semilunal lobule, SsC: Subcentral cortex, Sub: Subicular cortex, SLTC: Superolateral temporalcortex, TO: Tonsilla, VLTC: Ventrolateral temporal neocortex.

4

CO-EXPRESSION ANALYSIS

To characterize the functional association of the *DMD* gene in the adult human brain, we calculated the spatial correlation (Pearson's) between each gene in the AHBA (19,991 genes) and the *DMD* gene using all samples concatenated from the six donors (3,702 samples). Genes were ranked based on the correlations in a descending order. To assess the functional association of the different dystrophin isoforms across development, we calculated the spatial-temporal correlation (Pearson's) between each exon in the BrainSpan dataset (241,690 exons) and the exons that are specific to each of the three dystrophin isoforms groups: exon 2 for full length Dp427, isoform specific exons located in intron 44 for Dp140 and intron 62 for Dp71+Dp40, with respect to Dp427m nomenclature. These isoforms were selected because virtually all DMD patients have mutations affecting Dp427c, Dp427m and Dp427p. A proportion of patients additionally cannot produce Dp140. And a small number of patients cannot produce any isoforms, including the shortest Dp71 and Dp40. For each isoform group, we ranked all exons in a descending order based on correlation. To get a ranked gene list, each gene was assigned the rank of its most correlated exon. For each gene set, functional enrichment analysis was performed on the top 1% (most positively correlated) using ToppGene [267]. We returned all terms enriched at an FDR-corrected $q-value < 0.05$ from the categories: GO Molecular Function, GO Biological Process, GO Cellular Component, Human Phenotype, Mouse Phenotype, and Pathway.

DISEASE GENE SETS OVER-REPRESENTATION

Enrichment analysis of disease-related gene sets was performed using a two-sided Wilcoxon rank sum test (Mann-Whitney U-test). For each list of all genes ranked based on their co-expression with dystrophin expression we used the rank sum test to assess the significance of the ranks of each disease gene set. To control the false discovery rate, we corrected for multiple testing using the Benjamini-Hochberg method [268]. In case of the Adult Human Brain we tested the set of 19,991 genes, ranked based on their correlation to the DMD gene across all samples (Pearson's correlation). Similarly, for the BrainSpan developing human brain transcriptome we tested three sets of 21,164 genes ranked based on their co-expression with the exons corresponding to the three dystrophin isoforms: Dp71+Dp40, Dp140 and Dp427.

We tested for the enrichment of five disease-related gene sets. The ASD-ID list contained 827 genes harboring de novo mutations from four ASD [122–125] and two ID [269, 270] exome sequencing studies. The ASD-ID was retrieved from [29]. The SFARI ASD list contained 706 genes associated to ASD using manual curation of published scientific literature from the Simons Foundation Autism Research Initiative (SFARI) AutDB database [271]. The list includes candidate genes implicated by common variant association, candidate gene studies, genes within ASD-associated CNV, and genes implicated in syndromic forms of ASD. Lists of genes related to ADD, OCD and dyslexia were retrieved from DisGeNet v3.0, a database that integrates human gene-disease associations from various expert curated sources and text-mining of literature [79].

FANTOM5 DATA

We used the FANTOM5 samples ontology and the linked data version of FANTOM5 data, which was exposed as nanopublications [251, 252]. We queried the FANTOM5 data to get all transcription start sites which are overlapping with the first exons of Dp427c, Dp427m, Dp427p, Dp260, Dp140, Dp116, and Dp70+Dp41. We selected only samples belonging to the cerebral cortex, hippocampus, amygdala and cerebellum brain regions. Further, we removed samples pooled from multiple donors since they spanned a wide age range, which could dilute the expression of a TSS varied through development. Our analysis resulted in 25 TSSs across 8 samples (Supplementary Table 6).

EPIGENETIC DATA

Data from the Roadmap Epigenomics Consortium [75] was visualized using the WashU EpiGenome Browser v40.0.0 [253]. We visualized only the track corresponding to Active transcription start site (TSS) chromatin state. Details of the 23 selected samples are in Supplementary Table 7.

VALIDATION USING EX VIVO QPCR

Frozen tissue samples from a 51 years old male non-demented control brain of the anterior orbital gyrus and cerebellum were obtained (post-mortem delay: 07:45 hr; pH 7.05, stored in cryovial at -80^{o}). For total RNA isolation, tissue was disrupted in tubes with MagNA Lyser Green Beads (Roche Diagnostics) with TriPure Isolation Reagent (Roche Diagnostics). Isolation was performed with chloroform, and RNA was precipitated with isopropanol. The NucleoSpin RNA II kit including DNase digestion (Bioke, Leiden, The

Netherlands) was used for RNA purification. RNA was used for cDNA synthesis with random hexamer primers. Expression of Dp427c, Dp427p, Dp140 and Dp71 (primer sequences available on request) was determined by SYBR Green-based real-time quantitative PCR (95°C for 10s, 60°C for 30s, and 72°C for 20s, 45 cycles followed by melting curve analysis) on the Roche LightCycler 480 (Roche Diagnostics). Housekeeping gene *RPL22* was used as a reference gene. Primer efficiencies were determined and analysis was performed with LinREgPCR66.

Chapter 5

Shared Pathways among Autism Candidate Genes determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome

Ahmed Mahfouz*
Mark N Ziats*
Owen M Rennert
Boudewijn PF Lelieveldt
Marcel JT Reinders

This Chapter is published as: *J Mol Neurosci* (2015) 57(4):580-594, doi: 10.1007/s12031-015-0641-3.
*Equal contribution.

AUTISM spectrum disorder (ASD) is a neurodevelopmental syndrome known to have a significant but complex genetic etiology. Hundreds of diverse genes have been implicated in ASD; yet understanding how many genes, each with disparate function, can all be linked to a single clinical phenotype remains unclear. We hypothesized that understanding functional relationships between autism candidate genes during normal human brain development may provide convergent mechanistic insight into the genetic heterogeneity of ASD. We analyzed the co-expression relationships of 455 genes previously implicated in autism using the BrainSpan human transcriptome database, across sixteen anatomical brain regions spanning prenatal life through adulthood. We discovered modules of ASD candidate genes with biologically relevant temporal co-expression dynamics, which were enriched for functional ontologies related to synaptogenesis, apoptosis, and GABA-ergic neurons. Furthermore, we also constructed co-expression networks from the entire transcriptome and found that ASD candidate genes were enriched in modules related to mitochondrial function, protein translation, and ubiquitination. Hub genes central to these ASD-enriched modules were further identified, and their functions supported these ontological findings. Overall, our multi-dimensional co-expression analysis of ASD candidate genes in the normal developing human brain suggests the heterogeneous set of ASD candidates share transcriptional networks related to synapse formation and elimination, protein turnover, and mitochondrial function.

5.1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental syndrome characterized clinically by impairments in verbal and non-verbal communication, deficits in social interaction, and repetitive and/or restrictive patterns of behavior [272]. Despite an estimated prevalence of 1 in 88 newborns [273], and an exponential increase in recent efforts to elucidate autism neurobiology, a clear understanding of the molecular mechanisms underlying the development of ASD remains elusive. However, recent studies have firmly established a substantial role for genetic etiologies in the development of ASD. Evidence for a strong heritable risk of ASD was initially described in twin and sibling epidemiological studies of autism [274–278], and has since been firmly established through multiple genetic approaches [51, 279, 280]. For instance, genome-wide association studies (GWAS) [281–283], copy number variation (CNV) analysis [284–289], and whole-exome sequencing projects [52, 122–125, 290] have implicated hundreds of genes in ASD. Yet understanding how this diverse set of genes relates to the underlying molecular mechanisms and subsequent neuropathology of ASD is still unclear.

Mechanistic understanding of how ASD candidate genes relate to the neurobiology of autism is a difficult task, since genes encode multiple highly complex functions at different stages of development and across different regions of the brain. Moreover, the set of genes implicated in ASD is highly heterogeneous, and many of their functions are completely unknown. Furthermore, understanding how disruption in different genes with disparate functions still results in a common clinical phenotype makes developing common targeted biomarkers and treatments for ASD challenging. Therefore, in addition to attempts to identify genes that are causative for ASD, it is important to understand how ASD candidate genes may relate to each other during human neurodevelopment in

order to identify potential shared molecular pathways.

One validated approach to integrate heterogeneous gene sets, in order to uncover shared molecular mechanisms, is through the analysis of gene co-expression patterns, which invokes the guilt-by-association heuristic that is pervasive in genomics research [89, 291]. Several studies have demonstrated that genes with similar brain co-expression patterns are likely to function together in common cellular pathways [46, 292]. These transcriptional co-expression relationships are particularly relevant to neurodevelopment, as the precise regulation of gene expression across brain regions at different ages instructs the exquisite specialization and connectivity within the brain. Since neurodevelopmental disorders such as autism are believed to result from functional aberrations within brain regions and/or disruption of inter-regional connectivity between regions [279], investigating the gene expression profiles of autism candidate genes across brain regions and throughout normal human neurodevelopment may provide insight into the complex functional genomics of this neurodevelopmental disorder.

A global survey of ASD gene co-expression patterns across normal human neurodevelopment could therefore facilitate our translation of ASD candidate genes to ASD candidate pathways, but this has not yet been undertaken. A recent study that assessed autism gene co-expression patterns in two adult human brains is an important step toward this goal [293], but as autism is a neurodevelopmental disorder it is imperative to understand the relationship of autism candidate genes in a developmental context. Conversely, other studies have explored the expression profiles of individual ASD candidates in human brain development [47], but lack an assessment of the relationships among these ASD candidates and how they relate to global transcriptional pathways important in brain development.

Transcriptome-based studies of the developing human brain have previously been limited in the sample size, number of brain structures analyzed, and developmental time points assessed, hampering the ability to evaluate the genetic contributors to neurodevelopmental disease comprehensively [294–298]. However, the recent availability of broad developmental surveys of gene expression, which cover many brain regions over multiple developmental stages, can greatly facilitate such analysis [47]. The BrainSpan transcriptional atlas of the developing human brain is a repository of RNA-seq expression profiling of 16 brain structures spanning early pre-natal development (8 weeks post-conception) to adulthood (40 years of age). This publicly available atlas presents a unique opportunity to understand the spatial and temporal specificity of ASD candidate genes.

A few studies have recently assessed for co-expression relationships between subsets of autism-related genes and/or certain developmental windows using human brain gene expression relationships. For instance, Parikshak *et al*. analyzed the co-expression of autism and intellectual disability risk genes in the neocortex and among cortical laminae from samples representing early development using Weighted Gene Co-Expression Network Analysis (WCGNA). They demonstrated that ASD risk genes were enriched in modules expressed in superficial cortical layers and glutamatergic projection neuron and functionally related to transcription and synaptic development [133]. Willsey *et al*. studied co-expression networks derived from nine genes harboring recurrent de novo loss-of function mutations in autism patients, and showed principally that the autism

risk gene expression is most prominent in layer 5/6 cortical projection neurons during mid-fetal gestation [134]. Finally, using a different computational approach, Hormozdiari *et al*. integrated co-expression networks and protein-protein interaction networks of autism and intellectual disability risk genes identified in a recent cohort of 116 patients, and also showed that the autism genes enrich into networks related to transcription and synaptogenesis [29]. Despite the importance of these results and their largely overlapping findings, no study has yet assessed very broad sets of autism risk genes across all brain regions and development time points to gain insight into potentially shared molecular pathways or affected brain regions among the incredibly heterogeneous autism genetic subtypes.

Here we present an analysis of the spatial-temporal co-expression of ASD candidate genes across the normal developing human brain using the BrainSpan atlas. We developed a biologically driven computational approach to deduce functional relationships among this diverse set of genes. We first discovered modules of ASD candidates with biologically relevant temporal co-expression dynamics. These modules were related to the processes of synaptogenesis, apoptosis, and the neurotransmitter γ-aminobutyric acid (GABA). Then, we created a transcriptome-wide co-expression network from all genes expressed in the brain, to discover significant 'Molecular Interaction Modules,' and demonstrated that ASD candidate genes are enriched only in modules related to the processes of synaptogenesis, mitochondrial function, protein translation, and ubiquitination. Lastly, we identified hub genes within the ASD-enriched Molecular Interaction Modules, whose functions supported our ontological results, and which may serve as additional ASD candidate genes. Our analysis of this multi-dimensional expression data suggests pathways previously independently implicated in autism are related to each other through shared neurodevelopmental transcriptional networks.

5.2. RESULTS

SPATIO-TEMPORAL GENE CO-EXPRESSION ANALYSIS OF ASD CANDIDATE GENES

In order to identify functional relationships between ASD candidate genes, we investigated patterns of gene co-expression change across developmental stages between each pair of genes from the ASD list. First, the correlation between each pair of ASD genes was calculated separately within each developmental stage based on the Spearman's rank correlation between the two genes across all brain regions. For each gene-pair, this resulted in a correlation value for each of the seven developmental stages, representing the brain-wide transcriptional similarity between the genes at each developmental stage (Figure 5.1C and D). Gene-pairs were retained only if they had an absolute correlation value greater than 0.8 in at least one developmental stage. We have used the Spearman's Rank Correlation as it focuses more on the similarity in the change of gene expression; as opposed to similarity in the absolute values of gene expression (See the Supplementary Information for more details).

Second, the surviving gene-pairs were hierarchically clustered into distinct modules based on the similarity of their correlation profiles over time (using the Euclidean distance between the profiles and a complete linkage to merge clusters). Finally, the

correlation pattern for each module was summarized by averaging all the gene-pair correlation patterns included in the respective module. It is worth noting that the patterns within the modules represent changes in co-expression across development (which should not be confused with actual expression levels of genes).

ASD GENE MODULES DISPLAY DISTINCT TEMPORAL DYNAMICS AROUND BIRTH

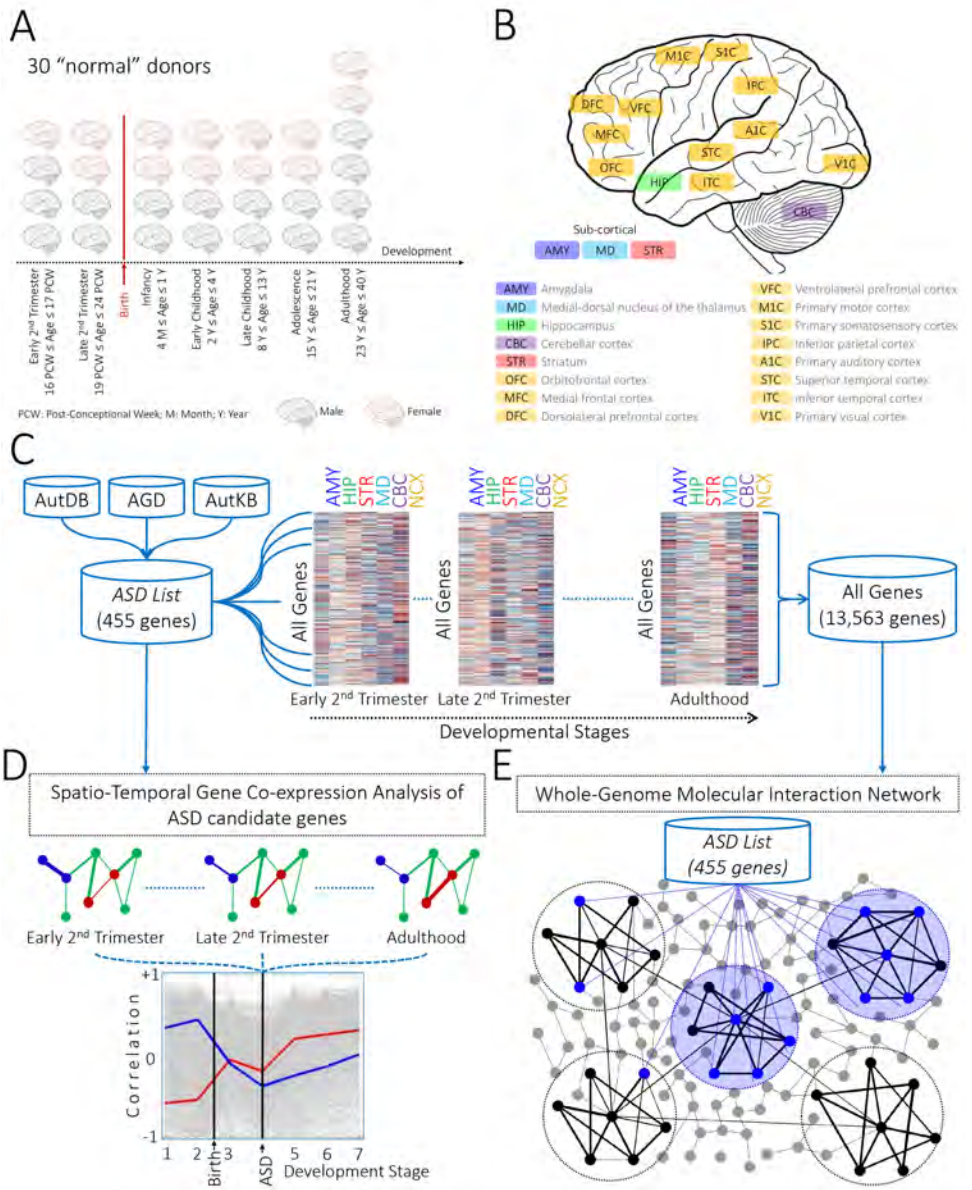
Figure 5.2A shows the hierarchical clustering of the retained ASD gene-pairs. In total there were 103,285 pair-wise correlations between the 455 ASD candidate genes in the ASD list, of which 1,168 remained after applying the stringent threshold of an absolute correlation greater than 0.8. The surviving gene-pairs clustered into three distinct modules. Two of these modules, the "Green" module and the "Blue" module, displayed distinct correlation patterns relative to pre- versus post-natal development. The Green module (Figure 5.2B) consisted of gene-pairs that lose correlation in the middle stages of development (infancy and childhood); that is, each pair of genes within the Green module has highly correlated spatial expression profiles at prenatal developmental stages but this correlation is lost at birth. In contrast, the Blue module (Figure 5.2C) consisted of gene-pairs that gain correlation during development. These genes do not show correlation at prenatal stages but progressively increase correlation throughout postnatal development. The "Red" module did not show any coordinated pattern of expression over developmental time (Figure S1). Genes forming gene-pairs in each of the three modules are listed in Table S1.

To further characterize these modules, we used the gene ontology (GO) enrichment analysis tool DAVID v6.7 [299] to discover whether genes in these modules relate to specific molecular mechanisms, cellular pathways or disease annotation terms. The top significantly enriched terms (Benjamini-Hochberg corrected $P < 0.01$) are summarized in Figure 5.2D. All the three modules were enriched for annotation terms related to neuron projection, synapse, synaptic transmission and behavior. The three modules were also enriched for disease terms including mental retardation and epilepsy. The Green and Blue modules were significantly enriched for neuron differentiation, cell morphogenesis, and learning/memory. The Green module was specifically enriched in functional terms related to regulation of apoptosis and regulation of cell death, while the Blue module was specifically enriched in terms related to ion channel, neurotransmitter receptor activity and GABA receptor activity. Table S2 includes the full list of enriched gene-annotation terms for these two modules.

None of the GO terms that were significantly enriched in the three ASD modules showed any significant enrichment in modules from 10 randomly created sets (see Table S3). We also assessed how many gene-pairs remained after thresholding them on co-expression (absolute correlation > 0.8 at any developmental stage) in 10,000 random gene sets of 455 genes. The results are summarized in Figure S2, where we show that the number of gene-pairs remaining after thresholding the ASD list (1,168 gene-pairs) is highly significant ($P < 10^{-4}$).

MODULES OF ASD CANDIDATE GENES ARE ENRICHED IN NEURONS

We then assessed if these modules were enriched in specific brain cell types. Lists of cell-type specific genes were obtained from previously published work [36]. These lists


A
30 "normal" donors
Development
Early 2nd Trimester 16 PCW ≤ Age ≤ 17 PCW
Late 2nd Trimester 19 PCW ≤ Age ≤ 24 PCW
Birth
Infancy 4 M ≤ Age ≤ 1 Y
Early Childhood 2 Y ≤ Age ≤ 4 Y
Late Childhood 8 Y ≤ Age ≤ 13 Y
Adolescence 15 Y ≤ Age ≤ 21 Y
Adulthood 23 Y ≤ Age ≤ 40 Y
PCW: Post-Conceptional Week; M: Month; Y: Year
Male
Female
B
M1C
S1C
IPC
DFC
VFC
MFC
A1C
STC
OFC
V1C
HIP
ITC
CBC
Sub-cortical
AMY
MD
STR
AMY Amygdala
MD Medial-dorsal nucleus of the thalamus
HIP Hippocampus
CBC Cerebellar cortex
STR Striatum
OFC Orbitofrontal cortex
MFC Medial frontal cortex
DFC Dorsolateral prefrontal cortex
VFC Ventrolateral prefrontal cortex
M1C Primary motor cortex
S1C Primary somatosensory cortex
IPC Inferior parietal cortex
A1C Primary auditory cortex
STC Superior temporal cortex
ITC Inferior temporal cortex
V1C Primary visual cortex
C
AutDB
AGD
AutKB
ASD List (455 genes)
AMY HIP STR MD CBC NCX
All Genes
All Genes (13,563 genes)
Early 2nd Trimester
Late 2nd Trimester
Adulthood
Developmental Stages
D
Spatio-Temporal Gene Co-expression Analysis of ASD candidate genes
Early 2nd Trimester
Late 2nd Trimester
Adulthood
Correlation
+1
0
-1
1 2 3 5 6 7
Birth
ASD
Development Stage
E
Whole-Genome Molecular Interaction Network
ASD List (455 genes)

included 1,465 neuron-, 1,529 oligodendrocyte-, and 1,829 astrocyte-specific genes (Table S4). ASD candidate gene modules were assessed for enrichment of these cell types using the hypergeometric probability test (see Methods). Both the Green and Blue modules were significantly enriched in neurons, whereas the Red module demonstrated no significant enrichment, as shown in Figure 5.2E.

ENRICHMENT OF ASD CANDIDATE GENES IN TRANSCRIPTOME-WIDE MOLECULAR INTERACTION MODULES

Given the marked genetic heterogeneity of ASD and the large number of genes involved, it is also important to understand the role of ASD candidate genes in normal brain development within the context of the whole transcriptome, as sub-networks of the entire brain transcriptome may be perturbed by the ASD candidates. An analysis of these sub-networks could reveal ASD-related pathways that would be missed by analyzing the ASD candidates alone, as it is unlikely that all ASD candidate genes have been identified to date [125]. Moreover, this top-down approach allows the identification of other genes that might also relate to ASD. Therefore, we performed a transcriptome-wide co-expression network analysis to identify functionally related gene modules throughout the normal developing brain transcriptome (Molecular Interaction Modules). Then, we assessed whether these modules were specific to distinct brain regions or developmental stages, and if they were related to specific pathways, cellular processes, or disease annotation terms. Finally, we determined if ASD candidate genes were enriched in any of the resultant Molecular Interaction Modules.

NO EVIDENCE FOR REGION-SPECIFIC MODULES

The transcriptome-wide co-expression network was constructed from all genes expressed in the brain (13,563 genes), based on their expression profile across all samples (480 samples, i.e. all brain structures and developmental stages). Genes were hierarchically clustered based on Spearman's rank correlation and complete linkage

Figure 5.1 *(preceding page)*: **Analyzing ASD candidate genes in the BrainSpan Atlas.** (A) Temporal description (i.e. age points) of the number and sex of the assessed brains. The data were grouped into seven developmental stages based on age. Black-colored brains indicate male donors and red-colored brains indicate female donors. (B) A representation of the 16 structures sampled in the BrainSpan Atlas. (C) Each heat-map shows the expression of all genes across six representative brain regions (AMY, HIP, STR, MD, CBC and NCX) in three representative developmental stages. The ASD list was created by combining lists of ASD candidate genes from three sources (AutDB, AGD, and AutKB-484). (D) A co-expression network of ASD candidate genes was generated for each developmental stage by correlating the expression vectors across brain regions. The correlation between each gene-pair was tracked over the developmental stages. The blue gene-pair represents two genes that are moderately correlated at early developmental stages, but gain correlation through development. Stronger correlation is represented by a thicker edge between the two nodes. By contrast, the red gene-pair represents two genes that lose correlation over development. The lower panel shows the correlation patterns of all gene-pairs in the network (grey) across development. Correlation patterns of the blue and red pairs are shown in respective colors. Birth and the average age of ASD diagnosis are indicated. (E) The transcriptome-wide Molecular Interaction Network was constructed based on the pairwise correlation between each pair of genes expressed in the BrainSpan atlas (13,563 genes). Each node in the network represents a gene while the weighted edges represent correlations between genes based on their expression across all samples. Nodes were clustered into modules (dashed circles). Genes from the ASD list are highlighted within each module (blue nodes). Blue circles indicate modules that are significantly enriched in genes from the ASD list.

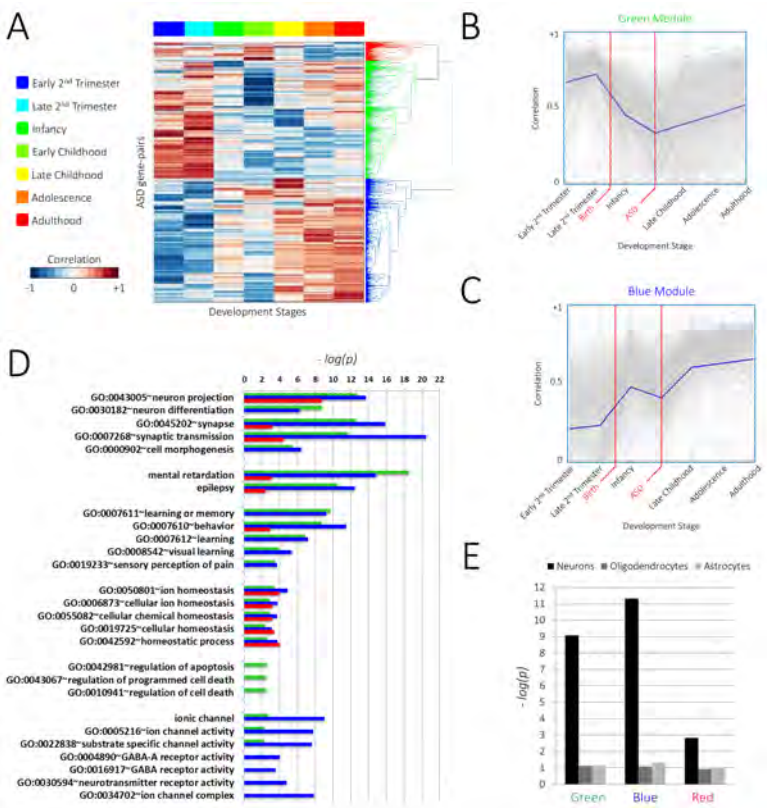


Figure 5.2: **Spatio-temporal Gene Co-expression Analysis of ASD candidate genes.** (A) Heat-map of the temporal correlation patterns of ASD gene-pairs (rows) through different developmental stages (columns). The dendrogram to the right shows the clustering of ASD gene-pairs into three modules (Red, Green and Blue). (B) The average correlation pattern of gene-pairs in the Green module shows loss of correlation at childhood. Vertical lines indicate birth and average age of ASD diagnosis. (C) The average correlation pattern of gene-pairs in the Blue module shows progressive gain of correlation across development. (D) Gene Ontology terms enriched in each of the three modules (represented in $-log_{10}(P)$, Benjamini-Hochberg corrected). Bars are colored according to the module's name. (E) Enrichment scores for each of the ASD modules in neurons, astrocytes and oligodendrocytes (represented in $-log_{10}(P)$, FDR-corrected).

between pairs of genes. The resulting network consisted of 32 modules of varying size (from 36 to 1,386 genes), as shown in Figure 5.3A. Visual analysis of the heat-map and average expression patterns of member genes from each of the 32 modules demonstrated that none were specific to particular anatomical regions. This observation is consistent with the results from a similar dataset of human brain development assessed by microarray [293]. We did not observe any pre/post natal specific expression patters in any of the 32 modules (Figure S3). The genes comprising each of the 32 modules are listed in Table S5.

MODULES ENRICHED FOR ASD GENES RELATE TO SYNAPTOGENESIS, PROTEIN TURNOVER, AND MITOCHONDRIA

The resulting transcriptome-wide co-expression modules were then assessed for enrichment of genes belonging to the ASD list using the hypergeometric probability test. Four modules—Magenta, Brown, Orange, and Purple—were significantly enriched for ASD candidate genes (FDR-corrected $P < 0.001$), as shown in Figure 5.3B. The Magenta module (Figure 5.4A) contained highly co-expressed genes during early childhood. The Brown module (Figure 5.4B) included genes with weak co-expression during childhood and differential spatial co-expression at late developmental stages. The Orange Module (Figure 5.4C) contained genes with progressively increasing co-expression during development. Finally, the Purple module (Figure 5.4D) included genes with varied co-expression during development and high differential spatial co-expression in adolescence and adulthood.

Then, these ASD-enriched modules were tested for enrichment of gene ontology terms, as shown in Figure 5.4 (see Table S6 for full list). The Magenta and Orange modules were significantly enriched for mitochondrial processes. Additional GO terms that were significantly enriched in the modules included ribosome and protein translation, transit peptide, ubiquitination, and alternative splicing. Significant enrichment for synapse was also found in the Brown module and the Purple module. Enrichment of ASD candidate genes into transcriptome-wide synapse modules further supports our previous finding of ASD modules (Green and Blue modules), above, which were also related to synaptogenesis. Neurological disease terms were also significant in the ASD-enriched modules: epilepsy (Brown module), Parkinson's (Magenta and Orange modules), Alzheimer's (Magenta and Orange modules) and Huntington's (Magenta and Orange modules).

ASD-ENRICHED MOLECULAR INTERACTION MODULES ARE MAINLY NEURONAL

Each module was also tested for enrichment of specific neural cell populations (i.e. neurons, oligodendrocytes, and astrocytes), as described earlier. Three out of the four ASD-enriched modules were enriched for neurons (Magenta, Brown and Purple modules), as shown in Figure 5.5. The Orange module, which was related to mitochondrial functioning, was highly enriched in astrocytes but not neurons. This finding is of relevance, as multiple recent studies have implicated glia, and specifically astrocytes, in the brain pathology of autistic subjects [300, 301].

ASD-ENRICHED MOLECULAR INTERACTION MODULE HUB GENES PROVIDE MOLECULAR TARGETS

An alternative approach to annotate the function of each ASD-enriched module is to analyze the genes with the strongest correlations within each module. It has been shown that within an interaction network, genes with the most connections to other genes, termed hub genes, are informative for the network as a whole, and are potential high yield therapeutic targets [302]. The strongest correlations within a module were explored using Cytoscape v2.8 [303]. First, each ASD-enriched module (Magenta, Brown, Orange and Purple) was imported as a graph with genes acting as nodes and pair-wise correlations between genes representing edges between the nodes. Figure 5.6 shows a subset of the connected nodes within each graph.

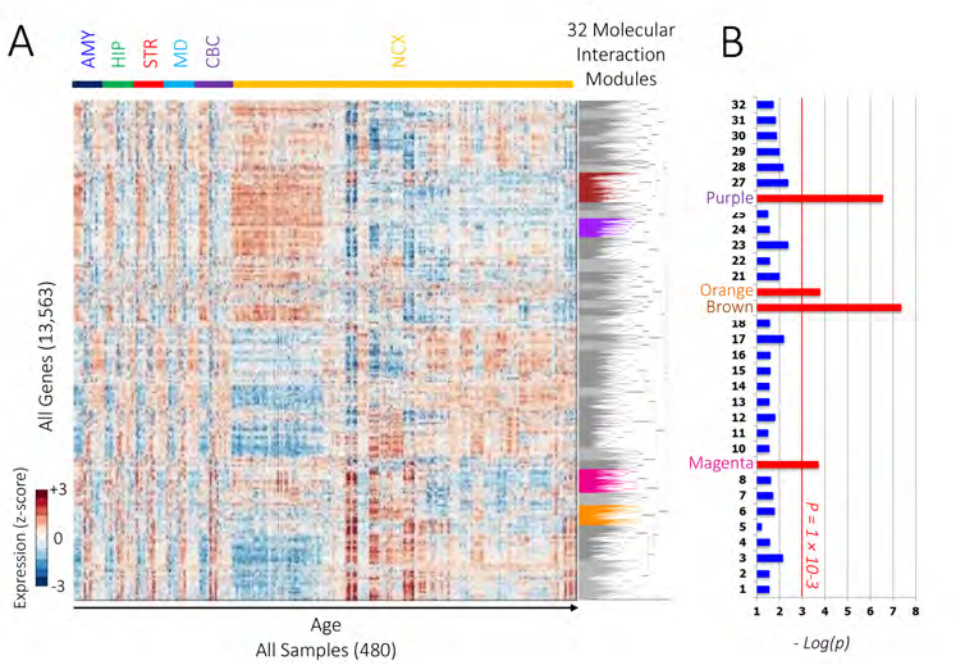


Figure 5.3: **Transcriptome-wide Molecular Interaction Network.** (A) A heat-map of the expression of 13,563 genes (rows) across all 480 samples (columns). Samples are ordered first by brain region (color-code at the top) and then by age. The dendrogram to the right shows the clustering of all the genes into 32 modules. Modules with significant enrichment ($P < 10-3$) of genes from the ASD list are colored while other modules are shown in gray. (B) Enrichment of ASD candidate genes in each of the modules showing high significance in the Magenta, Brown, Orange and Purple modules (represented by $-log_{10}(P)$, FDR-corrected).

The 10 most highly connected nodes (genes) within each graph were extracted and their putative functions determined by manual curation of the literature. Among these most highly connected hub genes, a number were of note. The most striking observation was that most of the highly connected hub genes in the Magenta and Brown modules are known to function in the processes of chromatin remodeling, transcription, or translation (*HMGN3, EIF3K, ZFAND6, DNAJC1, C6orf130, ERCC1, LCMBT2, MBTPS2, KIAA1191, C14orf138, GDA,* and *NCOA7*). This result is in line with the gene ontology enrichment for these modules (Figure 5.4). A number of other central hub genes are involved in intracellular signaling pathways (*PROCA1, TBC1D22B, PPP2R2D,* and *HACE1*), and a few are known to function as membrane ion channels (*PRRT1, KCTD4, SLC26A1,* and *KCNA4*). In addition, a number of hub genes function in apoptosis or myeloid/microglia cell processes (such as: *RNF11, CD200,* and *FAF1*). These hub gene functions largely recapitulate the ontologies of their respective networks, supporting our enrichment results and highlighting potential critical regulatory molecules of these networks.

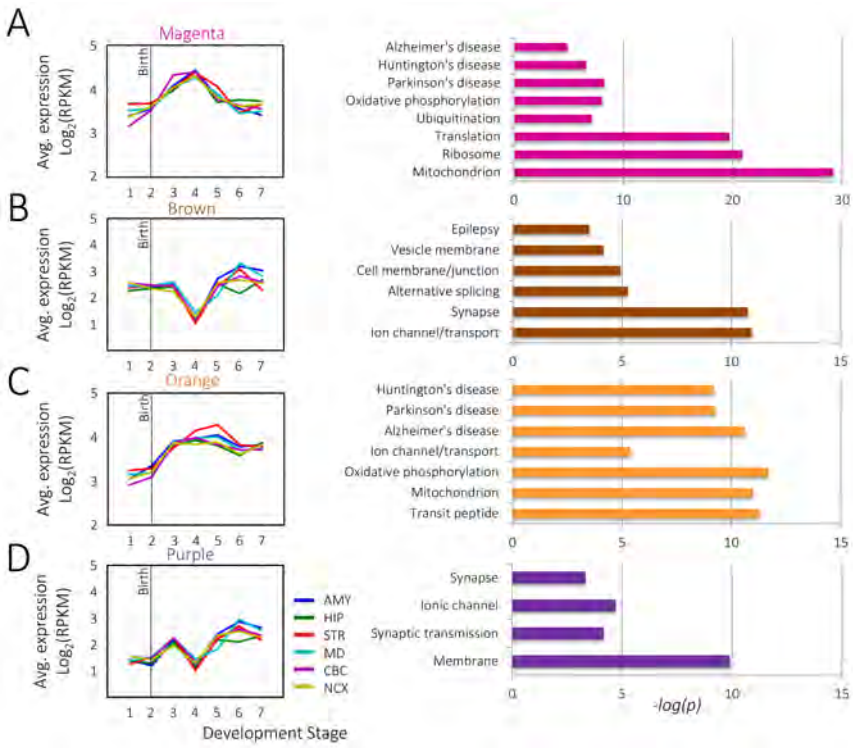


Figure 5.4: **ASD Modules.** (A) Left: Average expression pattern of the Magenta module genes across different brain regions (different plot colors). Right: Top GO terms enriched in the Magenta module. (B) Left: Average expression pattern of the Brown module genes. Right: Top GO terms enriched in the Brown module. (C) Left: Average expression pattern of the Orange Module genes. Right: Top GO terms enriched in the Orange module. (D) Left: Average expression pattern of the Purple module genes. Right: Top GO terms enriched in the Purple module. All enrichment values are represented by $-log_{10}(P)$, Benjamini-Hochberg corrected.

5.3. DISCUSSION

In order to gain insight into the molecular pathogenesis of ASD, we present a biologically-driven computational approach to analyze a heterogeneous set of genes previously independently implicated in ASD, to understand if they may relate to each other through shared functional genomic mechanisms. The main goal of this work is to understand if ASD candidate genes relate to common cellular/molecular pathways when considered in the context of transcription during normal human brain development. Identifying such pathways has profound implications for understanding the pathophysiology of ASD, especially since the majority of ASD patients do not have an identifiable genetic mutation [304]. Yet those patients are still likely to have alterations in the same pathways that are affected as those ASD patients with genetic mutations, although the alterations may be caused by environmental, epigenetic, or other non-genetic factors.

We intentionally analyzed a very broad collection of genes associated with ASD, in

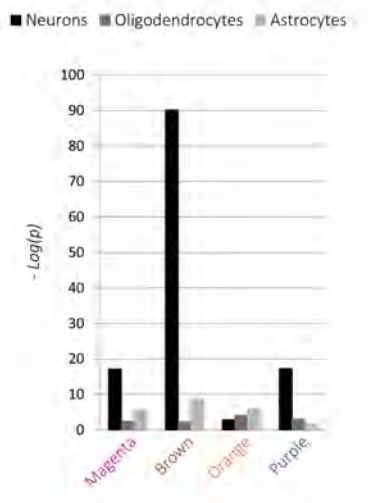


Figure 5.5: **Enrichment of the ASD Modules in Cell-type Specific Genes.** Enrichment of ASD-enriched modules in: neurons, oligodendrocytes and astrocytes (represented in $-log_{10}(P)$, FDR-corrected).

an attempt to understand if there are cellular or molecular pathways that may represent common mechanisms across all patients. Despite the fact that some of the genes in our ASD list are essentially causative for ASD (for instance, single gene mutation syndromes such as Fragile X), while others are not as strongly associated, we have weighted all genes equally to avoid bias toward more severely-affected patient cases. Future work could attempt to weight genes differently within the co-expression networks to study different genetic subtypes of autism.

We discovered subsets of ASD candidate gene modules that displayed biologically-relevant co-expression dynamics, which were enriched for the processes of synaptogenesis, apoptosis, and GABA-ergic signaling. In addition, we assessed for functional genomic relationships between ASD candidate genes and the entire developing human brain transcriptome. This analysis revealed that ASD candidate genes are enriched within transcriptome-wide modules related to synaptogenesis, mitochondrial function, alternative splicing, protein translation, and ubiquitination. By identifying gene modules that have similar expression patterns in the brain (regardless of time period), we were able to infer that they are likely functioning in similar pathways. This allowed us to infer which cellular and molecular mechanisms are likely to be disrupted in autism. We also demonstrated the cell-type specific enrichment of these modules being mostly neurons, further supporting the biological relevance of our computational approach, as the broad ASD phenotype is generally consider to ultimately result from neuronal/synaptic abnormalities [305]. Although several brain regions have been highlighted in neuroimaging and connectivity studies of autistic brains (namely cortical regions and the cerebellum) [306, 307], interestingly, none of the transcriptome-wide modules were specific to particular anatomical regions, which supports previous reports of the BrainSpan dataset via microarray [47]. Finally, by

assessing genes with the highest connectivity within the transcriptome-wide Molecular Interaction Modules that were enriched for ASD candidates, we identified hub genes that may represent critical regulatory molecules in these networks, and their functions further supported our enrichment findings.

The number of strongly connected gene-pairs from the ASD list were found to be highly significant ($P = 10^{-4}$), indicating that – based on their significantly strong co-expression across development – those ASD-associated genes are likely to be functionally related. We discovered three subsets of ASD–associated genes with distinct co-expression profiles around birth, even though the co-expression network for each developmental stage was calculated separately to avoid any bias towards pre/post natal expression changes. All three of these modules were significantly enriched for the processes of synaptogenesis and behavior, in addition to the disease annotations of mental retardation and epilepsy. Two of the modules (the Green and Blue modules) were also significantly associated with cell morphogenesis, neuron differentiation, and learning. Moreover, the Green module, which had highly correlated spatial expression at prenatal developmental stages with a dramatic loss of correlation at birth, was uniquely enriched for the process of apoptosis. Conversely, the Blue module displayed an opposite co-expression trajectory —poor correlation in expression prior to birth, but strong co-expression beginning in infancy and increasing through adulthood— and was uniquely related to GABA-ergic signaling and ion channels. The distinct, biologically relevant expression patterns of these two modules around birth, a developmental period with the greatest shifts in gene expression [47], suggests a key role of these networks in brain development and autism.

ASD-associated genes were highly co-expressed later in development in some of the identified modules (childhood and adulthood), whereas autism symptoms are generally apparent by the age of two. Our results suggest that a heterogenous set of genes which were independently associated to ASD converge into few functional pathways late in normal development. However, our findings do not preclude the possibility that the pathways implicated by these modules are involved in ASD pathogenesis, as our analysis was on co-expression patterns, not absolute gene expression levels. It is possible that the genes in these modules are still expressed in early neurodevelopment, but that they are most strongly co-expressed with other genes in the same module later in life. Consequently, disruption of the integrity of these genes (through inherited mutations, de novo mutations, mis-expression, etc.) early in development is likely to disrupt the functions of those modules later in life.

The functional ontologies of these networks are all pathways previously implicated ASD. Disrupted synaptogenesis has been one of the most replicated findings in ASD research [308], and autism is largely considered to be a disorder that results from a convergence of factors into synaptic dysfunction [305]. Our finding of multiple ASD gene co-expression networks enriched for the function of synaptogenesis is in line with these previous studies. Additionally, our analysis shows these same transcriptional networks are related to the processes of GABA-ergic signaling and apoptosis, which have been independently associated with ASD through various approaches. GABA-ergic neurons are the main inhibitory cell of the brain, and much research has suggested that an imbalance in the ratio of inhibitory to excitatory neurons may underlie autism at the cellular

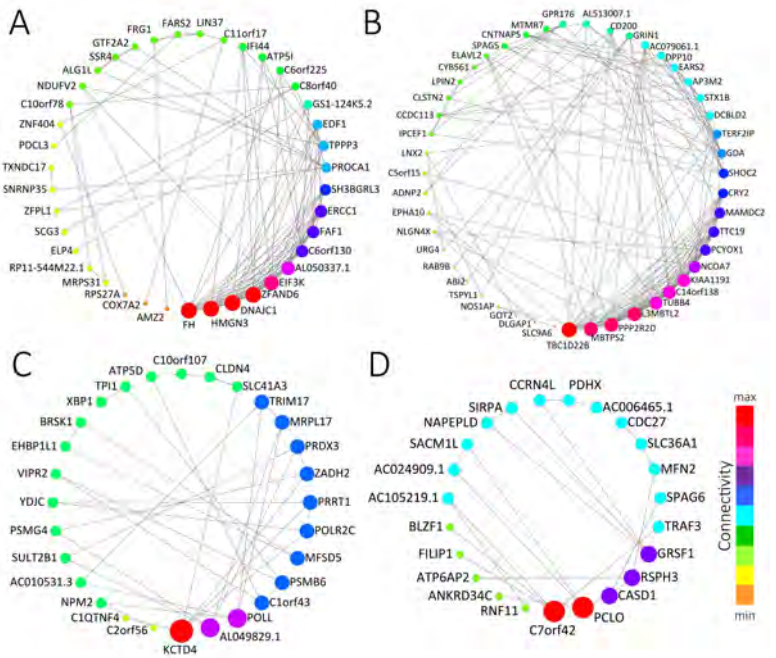


Figure 5.6: **Hub Genes of ASD Modules.** Each of the four ASD-enriched modules is presented with the Degree Sorted Circle layout of Cytoscape, with the nodes' size and color reflect the level of connectivity within the network. The bigger the node, the more connections it has. For clarity, edges with correlation values smaller than 0.9 are removed. (A) Top connected genes of the Magenta module. (B) Top connected genes of the Brown module. (C) Top connected genes of the Orange module. (D) Top connected genes of the Purple module.

circuit level [309]. Furthermore, a number of clinical trials are currently ongoing to test GABA-ergic modulators for the treatment of ASD [310]. Likewise, apoptosis —and more specifically the pruning of overabundant neural connections in early development— has recently been shown to be a critical process in the developing mammalian brain [311], and a number of studies have suggested this process may be aberrant in ASD [312, 313]. A delicate balance between formation of needed synaptic connections and pruning of overabundant connectivity (and their excitatory/inhibitor ratio) is a main component of early experience-dependent brain development, and both human and animal studies have previously shown deficiencies in these processes in ASD [307]. Our results suggest these processes may relate to each other and to ASD candidate genes through shared transcriptional networks.

ASD candidate gene modules with distinct temporal co-expression profiles around birth, which are highly related to synaptogenesis, support the notion that the pathogenesis of ASD is strongly related to this process. Additionally, the demonstration that the same transcriptional networks are also related to GABA-ergic signaling and apoptosis—both also suggested to be aberrant in autism—suggests that these disparate pathways may relate to each other through underlying shared transcriptional networks, providing a potential mechanism for functional convergence of ASD candidate genes into

common pathways underlying autism.

By incorporating the ASD candidate genes into the context of the entire brain transcriptome, our results suggest that the disruption of synaptogenesis in autism may also relate to underlying basic cellular processes —alternative splicing, protein translation, and ubiquitination— which have previously been implicated in ASD [314–317]. Defects in protein translation in particular have recently been shown to be a prominent feature in multiple animal models of ASD [318–320].

Two transcriptome-wide modules that were enriched for ASD candidate genes were both related to mitochondrial function, and one was specifically enriched in glia but not neurons. A large body of evidence has associated mitochondria dysfunction with rare syndromic forms of autism [321] and recent evidence suggests that altered mitochondrial gene expression may contribute to non-syndromic autism as well [322, 323]. Furthermore, these modules were also related to Huntington's and Alzheimer's disease, both known to have mitochondrial defects associated with their pathogenesis [324]. While the ASD-only gene modules in the first part of this study did not implicate mitochondrial function, significant enrichment of ASD genes in two different transcriptome-wide networks related to mitochondria suggests that additional ASD genes related to mitochondria may remain to be discovered, and our hub gene analysis provides potential high confidence candidates.

While the phenotype of autism may ultimately result from dysfunctional synaptogenesis, it is possible that such fundamental cellular processes as protein translation, ubiquitination, alternative splicing, and mitochondrial function may underlie the synaptic dysfunction. Furthermore, this may help explain the incredibly variable clinical spectrum of autism, and account for the increased prevalence of other complex medical problems in both the brain and other systems that ASD patients experience [5]. Moreover, a recent meta-analysis of de novo mutations in autism demonstrated enrichment for genes related to transcriptional regulation, and showed they have similar neurodevelopmental expression patterns to the Green and Blue modules of ASD candidates we identified [66]. Multiple recent whole-exome sequencing studies of individuals and family trios have continued to support the role of transcription-related and synaptogenesis-related genes in ASD [52, 290, 325]. Furthermore, a similar network analysis approach that assessed specifically for enrichment of de novo variants implicated in ASD and intellectual disability found similar shared transcriptional networks [29]. By integrating co-expression and protein-protein interaction networks they demonstrated that ASD-related genes converge into two modules related to basic intracellular processes including transcriptional regulation and synaptogenesis, and that the former process was more operant in prenatal time periods and the later in post-natal development [29]. These results are in line with earlier findings using either co-expression networks only [133, 134] or protein-protein interaction networks only [326]. Our results, despite assessing a much broader set of ASD candidate genes, are largely in agreement with these recent results. Whether and how defects in these basic cellular mechanisms result in altered synaptogenesis, are a reaction to altered synaptogenesis, or are mutually- exclusive from synaptogenesis is unclear. However, our results in addition to these previous studies suggest that a complex interplay between these processes and synaptogenesis are related to each other through overlapping co-expression networks.

A number of studies have assessed for changes in gene expression in post-mortem autistic brain directly (for a review see [327] and [328]. These studies have repeatedly shown that the autistic transcriptome is abnormally expressed compared to control brains across many different brain regions. The genes that are mis-expressed in autistic brains have been consistently demonstrated to be involved in pathways related to the synapse [49, 329], immune response/apoptosis [49, 329, 330], neurotransmitter receptors [331], RNA splicing [49, 170, 329], and mitochondrial function [323, 332]. These findings in autistic brain complement our results by showing that the ASD co-expression modules we discovered in the normal developing brain are functioning in the same pathways that are consistently disrupted in autistic brains.

Finally, the identified hub genes of ASD-enriched modules recapitulate the gene ontology analysis of these modules, strengthening the observation that basic cellular functions related to genome processing and mitochondrial function may represent a nexus in the genomic pathology of ASD. In addition, a number of hub genes relate to myeloid cells and apoptosis. There is a growing body of evidence implicating cytokine signaling, microglia-mediated synaptic pruning, and other immune-related processes in ASD [313], and this finding suggests the autism candidate genes may indirectly relate to processes that interact with these pathways through the transcriptional machinery. Furthermore, this supports our finding that the Green module of autism candidate genes relates to apoptotic pathways. However, because comprehensive lists of microglia-specific marker genes are not available, we were unable to assess for enrichment of ASD candidate genes into this cell type in this study. By highlighting individual genes that are most central in the identified molecular interaction networks, the hub gene analysis may provide potential additional high-yield ASD candidates for their respective transcriptional networks.

In summary, we have profiled the transcriptional co-expression networks of autism candidate genes throughout normal human brain development to identify modules of ASD candidate genes with biologically-relevant expression patterns. We have shown that these ASD modules are enriched for synaptogenesis, apoptosis, and GABA-ergic signaling, suggesting that pathways previously independently implicated in autism are related to each other through shared neurodevelopmental transcriptional networks. In addition, we expanded the analysis of ASD candidates to consider their relationship with the entire brain transcriptome. We demonstrated that ASD-enriched transcriptome-wide Molecular Interaction Modules are related to mitochondrial function, splicing, and protein turnover, which suggests further ASD candidates related to these functions may remain to be discovered.

Our comprehensive analysis of the global co-expression relationships between ASD candidates demonstrates that the various pathways implicated in autism separately may relate to one another when considered in a broader functional genomics framework. Furthermore, our Molecular Interaction Module analysis represents a valuable strategy to identify and prioritize other potential ASD candidate genes. Moreover, this approach can be used to assess genes from other complex neurodevelopmental and psychiatric disorders like schizophrenia, to uncover potential overlapping transcriptional pathways in the developing human brain among other gene sets.

5.4. MATERIALS AND METHODS

DEVELOPING HUMAN BRAIN TRANSCRIPTOME DATA

We downloaded the BrainSpan transcriptional atlas from `http://www.brainspan.org`. Details of tissue acquisition and data processing can be found in the BrainSpan documentation. The atlas contains next-generation RNA sequencing (RNA-seq) data generated from 579 tissue samples. These samples were collected from 41 developing and adult post-mortem brains of neurologically unremarkable donors spanning early prenatal development (8 post-conception weeks, PCW) to late adulthood (40 years of age). Some donor brains in the BrainSpan Atlas have missing data from certain brain regions. We excluded donors that had more than six regions missing. For donors with six or less missing regions, we imputed the data for the missing brain regions using a nearest neighbor approach. A full mathematical description of this is provided in the Supplementary Information. The resulting dataset contained 30 donor brains. From these donor brains, only the sixteen brain regions that were present in all 30 donor datasets were analyzed. This filtration resulted in a final dataset derived from 30 donor brains across 16 brain regions, or 480 brain samples in total.

The 30 donor brains used in our analysis were further grouped into seven developmental stages according to the BrainSpan classification system (Figure 5.1A). The transcriptomes of the cerebellar cortex (CBC), medial-dorsal nucleus of the thalamus (MD), striatum (STR), amygdala (AMY), hippocampus (HIP), and 11 areas of the neocortex (NCX) were assessed (Figure 5.1B).

The resultant dataset contained RNA-seq expression values aligned to composite gene models, and given in units of reads per kilobase of exon model per million mapped reads (RPKM) [333]. Genes whose RPKM values were likely to represent noise rather than actual sequenced reads were discarded by removing any gene that did not have at least one expression value greater than or equal to five RPKM in any of the 480 tissue samples. The remaining set consisted of 13,563 genes expressed in the 30 donor samples assessed. The expression data was then normalized across all samples using quantile normalization. Finally, the data was log_2-transformed for further analysis.

ASD GENE LIST

A comprehensive yet high confidence list of common ASD susceptibility genes (herein named "ASD list") was created by combining (taking the union) lists from three main ASD genes databases: AutDB [271], Autism Genetics Database (AGD) [334], and AutKB-484 [335] (a subset of AutKB determined by the Xu *et al*. through ranking and scoring algorithm to be the most high confidence ASD candidates). These databases each independently collected genes that have previously been associated with autism through a number of different experimental studies using various methods (namely GWAS, single-gene deletion syndromes that have autism as a component, genome-wide expression profiling, and genome-wide sequencing/CNV/linkage studies). ASD genes that were not present in the 13,563 genes we considered from the BrainSpan Atlas (for instance, mitochondrially-encoded genes) were discarded. The final ASD list consisted of 455 ASD susceptibility genes (Table S7).

CO-EXPRESSION OF ASD CANDIDATE GENES

We calculated the Spearman's rank correlation between each pair of ASD candidate genes within each of the seven developmental stages separately. For each donor, the correlation between each gene-pair is calculated across all 16 brain regions. The correlation between a gene-pair in each developmental stage is the average of their correlation across all donors within the developmental stage. We focused our analysis on gene-pairs with an absolute correlation value greater than 0.8 in at least one developmental stage (1,168 out of 103,285 gene-pairs). We used hierarchical clustering to cluster gene-pairs using the Euclidean distance between the profiles and a complete linkage to merge clusters. Based on the heatmap of gene-pair correlations across development, we cut the dendrogram to produce three clusters. The correlation pattern for each module was summarized by averaging all the gene-pair correlation patterns included in the respective module.

TRANSCRIPTOME-WIDE CO-EXPRESSION NETWORK

We constructed a transcriptome-wide co-expression from all genes expressed in the brain (13,563 genes), based on the similarity of their expression profile across all samples (480 samples). We used hierarchical clustering to cluster gene-pairs using Spearman's rank correlation between the profiles and a complete linkage to merge clusters. We cut the dendrogam to produce 32 modules of varying size (from 36 to 1,386 genes).

GENE SET ENRICHMENT AND GENE ONTOLOGY ENRICHMENT ANALYSIS

Enrichment of transcriptome-wide Molecular Interaction Modules in ASD candidate genes and cell-type specific genes was assessed using the hypergeometric probability density function (hygepdf) in MATLAB R2011a (The MathWorks, Inc.). The resulting P values were corrected for multiple testing by controling the false discovery rate (FDR) using the Storey method [336]. All results reported are the $-log_{10}$ of FDR-corrected P values, and only $P < 0.001$ were considered significant. Gene list were assessed for shared biological pathways by testing for enrichment of gene ontology terms (GO) using DAVID Bioinformatics Resources v6.7 [299]. The complete list of expressed genes in this study's dataset (13,563 genes) was used as the background. Only gene ontology terms with a Benjamini-Hochberg corrected $P < 0.01$ are presented as significant.

5.5. SUPPLEMENTARY MATERIAL

The online version of this article contains supplementary material[1].

[1] http://link.springer.com/article/10.1007/s12031-015-0641-3

Chapter 6

Genome-wide co-expression of steroid receptors in the mouse brain: identifying signaling pathways and functionally coordinated regions

Ahmed Mahfouz
Boudewijn PF Lelieveldt
Aldo Grefhorst
Lisa TCM van Weert
Isabel M Mol
Hetty CM Sips
José K van den Heuvel
Nicole A Datson
Jenny A Visser
Marcel JT Reinders
Onno Meijer

This Chapter is published as: *Proc Natl Acad Sci* (2016) 113: 2738–2743, doi:10.1073/pnas.1520376113

STEROID receptors are pleiotropic transcription factors that coordinate adaptation to different physiological states. An important target organ is the brain, but even though their effects are well studied in specific regions, brain-wide steroid receptor targets and mediators remain largely unknown due to the complexity of the brain. Here, we tested the idea that novel aspects of steroid action can be identified through spatial correlation of steroid receptors with genome-wide mRNA expression across different regions in the mouse brain. First, we observed significant co-expression of six nuclear receptors (NRs) [androgen receptor (*Ar*), estrogen receptor alpha (*Esr1*), estrogen receptor beta (*Esr2*), glucocorticoid receptor (*Gr*), mineralocorticoid receptor (*Mr*), and progesterone receptor (*Pgr*)] with sets of steroid target genes that were identified in single brain regions. These co-expression relationships were also present in distinct other brain regions, suggestive of as yet unidentified coordinate regulation of brain regions by, for example, glucocorticoids and estrogens. Second, co-expression of a set of 62 known NR coregulators and the six steroid receptors in 12 nonoverlapping mouse brain regions revealed selective downstream pathways, such as *Pak6* as a mediator for the effects of *Ar* and *Gr* on dopaminergic transmission. Third, *Magel2* and *Irs4* were identified and validated as strongly responsive targets to the estrogen diethylstilbestrol in the mouse hypothalamus. The brain- and genome-wide correlations of mRNA expression levels of six steroid receptors that we provide constitute a rich resource for further predictions and understanding of brain modulation by steroid hormones.

6.1. INTRODUCTION

Steroid receptors are part of the superfamily of nuclear receptors (NRs), which act as transcription factors regulating expression of numerous biologically important target genes [337]. Their transcriptional activity is induced by steroid hormones, which respond to changed demands in terms of reproductive status, mineral balance, or stressful physical and psychological challenges. A crucial site of action is the brain, where these hormones have strong modulatory effects on physiological regulation, cognitive function, mood and behavior. They do so by changing cellular responsiveness to a variety of neurotransmitters and peptides, and by inducing morphological changes [338, 339].

Understanding the effects of steroid hormones on the brain faces the challenge to identify in as many as 900 different brain nuclei [24] both the highly cell specific target genes that mediate the hormone effects [340, 341], as well as the signaling factors that mediate or influence steroid receptor signaling. The latter include proteins affecting prereceptor metabolism, interacting transcription factors [342], and downstream nuclear receptor co-regulator proteins [337]. Even if steroid hormones effects are well-studied in specific regions [337, 343], overall the brain steroid receptor targets and mediators remain largely unknown.

In situ hybridization (ISH) has been used to identify the functional roles of the 49 NR genes in adult mouse brain based on the clustering of the NR expression patterns in anatomical and regulatory networks [110]. In this study, we substantially extended this approach to identify targets and signaling partners of the steroid receptors, and relationships between different regions of the mouse brain, based on genome wide co-expression with steroid receptors. The Allen Brain Atlas (ABA) [24] is the most comprehensive repository of ISH-based gene expression in the adult mouse brain. We used the

ABA to identify genes that have three dimensional (3D) spatial gene expression profiles similar to steroid receptors.

To validate the functional relevance of this approach, we analyzed the co-expression relationship of the glucocorticoid receptor (*Gr*) and estrogen receptor alpha (*Esr1*) and their known transcriptional targets in specific brain regions. We then exploited these associations to derive new hypotheses about the functional role of receptors in brain regions with no previously known effects of steroids. Furthermore, we studied the region-specific co-expression of nuclear receptors and their downstream mediators (co-regulators) to identify specific partners mediating the hormonal effects on dopaminergic transmission. Finally, to illustrate the potential of using spatial co-expression to predict region-specific steroid receptor targets in the brain, we identified and validated genes which responded to changes in estrogen in the mouse hypothalamus.

6.2. RESULTS

SPATIAL EXPRESSION REVEALS KNOWN SITES OF ACTION OF STEROID RECEPTORS IN THE MOUSE BRAIN

We first analyzed the mRNA expression of six nuclear steroid receptors (Estrogen Receptor alpha, *Esr1*; and beta, *Esr2*; Androgen Receptor, *Ar*; Progesterone Receptor, *Pgr*; Glucocorticoid Receptor, *Gr*; and Mineralocorticoid Receptor, *Mr*) across the brain using the 3D spatial gene expression data from the ABA [24]. We generated a general overview of the expression of each receptor across 12 non-overlapping brain structures covering the entire brain: Isocortex; olfactory areas (OLF), hippocampal formation (HPF), cortical subplate (CTXsp), striatum (STR), pallidum (PAL), cerebellum (CB), thalamus (TH), hypothalamus (HY), midbrain (MB); pons (P), and medulla (MY) (Figure 6.1A). The expression profiles generally correspond to the known distribution and sites of actions of different receptors [110], and provide comprehensive information at the higher aggregation level of brain regions described here. For example, *Esr1* is highly expressed in the hypothalamus, olfactory and the cortical sub-palate. Within the hypothalamus, *Esr1* shows high expression in the arcuate nucleus (ARH), and medial preoptic nucleus (MPO) (Figure 6.1B). *Gr* is highly expressed in the CA1 and dentate gyrus (DG) areas of the hippocampus, cortex and the thalamus, while *Mr* is predominantly expressed in the hippocampus (Figure 6.1A). These expression patterns are well in line with the known sites of action of the different receptors across the brain [344, 345].

GENES SPATIALLY CO-EXPRESSED WITH STEROID RECEPTORS INDICATE REGIONAL FUNCTIONAL SPECIFICITY

To go beyond the expression profiles of steroid receptors as reported in the literature, we identified genes with similar expression profiles to each of the receptors. Based on the principle of 'guilt by association', these co-expressed genes are likely to be enriched in receptor target genes and receptor signaling partners such as co-regulators. For each steroid receptor, we ranked genes based on their spatial co-expression across the whole brain as well as in each of the aforementioned 12 brain structures separately, resulting in 13 ranked lists per receptor (Dataset S1)[1]. For each steroid receptor, strongly co-

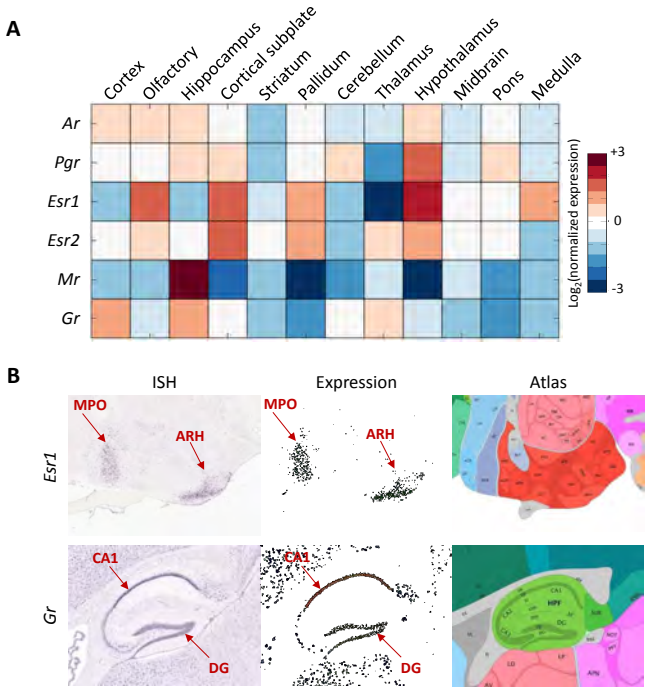


Figure 6.1: **Expression of steroid receptors in the mouse brain.** (A) Expression of six steroid receptors (Androgen Receptor, *Ar*; Glucocorticoid Receptor, *Gr*; Mineralocorticoid Receptor, *Mr*; Progesterone Receptor, *Pgr*; Estrogen Receptor alpha, *Esr1*; and beta, *Esr2*) across the 12 brain regions. Reported values are the average expression energy per region normalized to the average expression across the whole brain and then log_2-transformed. (B) Example sagittal sections from the Allen Brain Atlas showing the ISH (left), expression mask (middle), and the corresponding atlas section (right) of *Esr1* in the hypothalamus (top) and the expression of *Gr* in the hippocampus. Red arrows indicated the medial preoptic area (MPO) and arcuate hypothalamic nucleus (ARH), Cornu Ammonis subdivision 1 (CA1), and the dentate gyrus (DG).

expressed genes within a brain region are likely related to the localized functional role of the receptor. For example, out of the top 10 genes co-expressed with *Esr1* across the whole brain, 4 were previously shown to be regulated by *Esr1* and/or estrogens in various tissues (*Gpr101*, *Calcr*, *Ngb*, and *Gpx3*) [346–349]. These genes were also co-expressed with *Esr1* in hypothalamus, in line with their functional relationship to *Esr1* in mediating reproductive and metabolic processes. However, whole brain correlation of these genes with *Esr1* were also driven by thalamus, midbrain and pallidum, demonstrating less obvious relationships between *Esr1* and these target genes. Strikingly, among the top 10 genes co-expressed with *Gr* across the whole brain none are strongly co-expressed with *Gr* in the hypothalamus, indicating that *Gr* signaling in hypothalamus is rather distinct from that in cortex, striatum, thalamus and midbrain.

[1] Datasets S1–S6 are available at:
`data.3tu.nl/Q:11repository/uuid:ecc3b182-d312-4216-9053-a824d0e04d5e`.

In addition, we analyzed the functional enrichment of genes co-expressed with *Gr* and *Esr1* in the 12 brain regions (Table S1). *Esr1*-co-expressed genes were enriched for neuropeptide regulation in the hypothalamus as well as cerebellum. A number of *Gr* associated genes in hypothalamus were related to glia and oligodendrocyte development, supporting the known effects of *Gr* on these processes in the hypothalamus [350].

GLUCOCORTICOID-RESPONSIVE GENES ARE HIGHLY EXPRESSED WITH *Gr* IN HIP, P, MB AND WHOLE BRAIN

To test the validity of our hypothesis that co-expressed genes constitute candidate targets of steroid receptors, we assessed the extent of co-expression between *Gr* and known GR target genes. Since *Gr* has an important role in mediating transcription of genes involved in coping with stress within the hippocampus [338], we analyzed the co-expression of glucocorticoid (GC)-responsive genes (i.e. likely GR targets) with *Gr* in the whole brain, the hippocampus and its substructures the dentate gyrus (DG) and the different subregions of the cornu ammonis (CA) (Dataset S2; Figure 6.2A). The set of GC-responsive genes we considered originates from experiments where male rats were exposed to glucocorticoid treatment in chronic restraint stress (CRS) condition as well as in a control situation [351]. These experiments resulted in three sets of genes differentially expressed in dentate gyrus neurons: 1) GC-responsive genes in CRS rats, 2) GC-responsive genes in control rats, and 3) genes that show differential expression in GC treatment for both conditions (common GC-responsive genes).

As expected, GC-responsive genes are significantly co-expressed with *Gr* in the DG (where they were identified), but interestingly also in the whole brain and in the CA3 region (FDR-corrected $P < 1.8 \times 10^{-3}$; Mann-Whitney U-Test). The significant co-expression of GC-responsive genes in the CA3 area indicates that those cells in CA3 that do express *Gr* [344] may be functionally linked to DG granule cell in terms of their response to GCs. Of note, only those genes that responded to GC treatment in stressed and control rats (common GC-responsive genes) showed a significant co-expression with *Gr* in DG, CA1, and (very substantially) CA3 regions of the hippocampus. The data reveal that only the subset of invariant, context-independent GR target genes is related to constitutive co-expression with *Gr*, even if the correlation data come from 'control' conditions

The co-expression of the GC-responsive gene sets with *Gr* was not significant for areas such as the hypothalamus and the cortex. We initially considered these negative control regions, given that the target genes were identified in micro-dissected DG granule neurons [351], and the presumed high degree of cell-specificity. However, the co-expression of *Gr* with GC-responsive genes in CA3 prompted us to test whether this co-expression also occurs in other brain areas. Figure 6.2B shows that the set of common GC-responsive genes is not only co-expressed with *Gr* in the hippocampus (DG, CA1, and CA3) but also in the cortical subplate, pallidum and midbrain. These associations indicate a potential, as yet unknown relationship between these three brain areas in terms of endocrine regulation, in accordance with the notion that the cellular responses to glucocorticoids can be similar in distributed parts of brain networks [352]. Taken together, these results show that GR targets are co-expressed with *Gr* in the DG, the region where responsiveness was measured, as well as point to other brain regions that might share

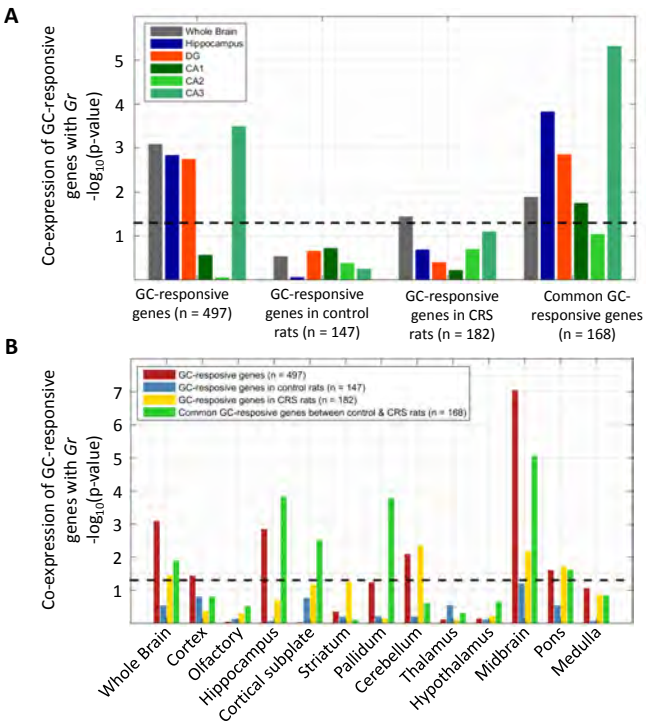


Figure 6.2: **Co-expression of glucocorticoid (GC)-responsive genes and *Gr* in the hippocampus.** Co-expression of four glucocorticoid (GC)-responsive gene sets with *Gr* in: whole brain, hippocampus, dentate gyrus, Cornu Ammonis subfields CA1, CA2 and CA3. (B) Co-expression of four glucocorticoid-responsive gene sets with *Gr* across the whole brain as well as the 12 major brain structures. All bars indicate the $-log_{10}(P)$ of the Wilcoxon ranksum test and the dashed line indicates the significance level at $P = 0.05$.

the same regulation mechanism.

SEXUALLY DIMORPHIC GENES ARE HIGHLY CO-EXPRESSED WITH *Esr1* IN THE HYPOTHALAMUS

To illustrate the generalizability of our approach to other receptors and brain regions, we followed the same approach to analyze the co-expression of *Esr1* and its putative targets. Xu *et al*. [94] showed that a set of 16 genes, including *Esr1*, has sexual dimorphic expression in the adult mouse hypothalamus. In addition, they showed that these 16 genes are sensitive to gonadal steroids (also in the male mouse brain) and that some are necessary for effects of estrogens on sexually dimorphic behavior [94], making this a valuable set of *Esr1* targets in the hypothalamus.

Table S2 shows the correlation values for each of the 15 sexually dimorphic genes with *Esr1* in whole brain, as well as in the hypothalamus, based on data from the ABA. The set of 15 genes is significantly correlated to *Esr1* based on whole brain analysis (FDR-corrected $P = 8.69 \times 10^{-14}$; Mann-Whitney U-Test), as well as the hypothalamic expression pattern (FDR-corrected $P = 3.85 \times 10^{-10}$; Mann-Whitney U-Test). In order to test

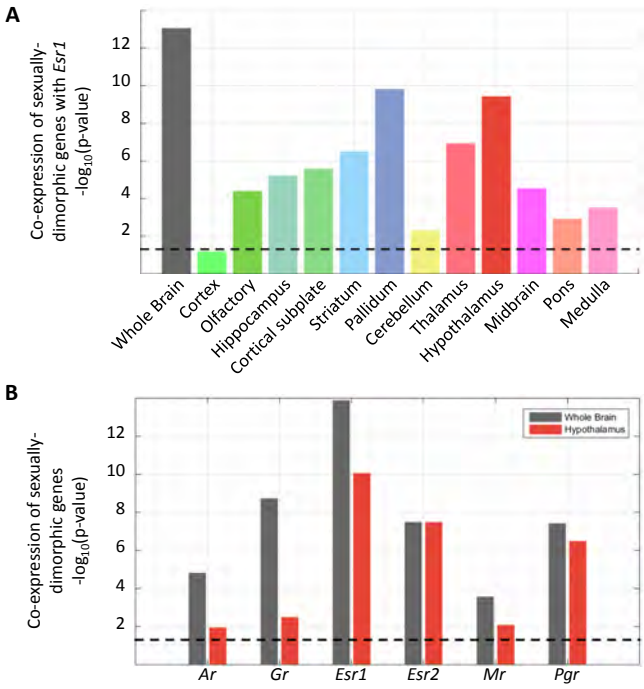


Figure 6.3: **Co-expression of sexually dimorphic genes and *Esr1* in the hypothalamus.** (A) Co-expression of 15 sexually dimorphic genes with *Esr1* across the mouse brain. (B) Co-expression of the 15 sexually dimorphic genes with the 6 steroid receptors across the whole brain as well as the hypothalamus. All bars indicate the $-log_{10}(P)$ of the Wilcoxon ranksum test and the dashed line indicates the significance level at $P = 0.05$.

whether the correlation between the 15 genes and *Esr1* is hypothalamus-specific, we repeated the analysis for all 12 brain structures. Figure 6.3A shows that sexually dimorphic genes are mostly correlated to *Esr1* in the hypothalamus (HY), pallidum (PAL), thalamus (TH), and the striatum (STR) ($P < 10^{-6}$). Similar to the results obtained for GR target genes, we observed high co-expression outside the main region of action (e.g. in the pallidum) suggesting that these brain regions share aspects of their transcriptional response to estrogen receptor activation. Furthermore, sex steroid receptors (*Esr1*, *Esr2*, and *Pgr*) showed higher co-expression levels with the sexually dimorphic genes with respect to the stress-steroid related *Mr* and *Gr* in the hypothalamus (Figure 6.3B). The strongest co-expression was with *Esr1*, indicating that the hypothalamic sexual dimorphism genes are mainly - but probably not exclusively - related to *Esr1*. Taken together, these results show that spatial co-expression can pinpoint context-specific actions of steroid receptors (in this case *Gr* and *Esr1*) and yields region-specific co-expressed genes, a very rich resource to generate hypothesis about steroid receptor targets.

REGION SPECIFIC CO-REGULATOR ANALYSIS POINTS TO DOPAMINERGIC TRANSMISSION VIA *Pak6*

So far, we have analyzed the potential of genes co-expressed with receptors to include region-specific targets. However, since correlation only indicates association rather than causation, co-expressed genes can also include co-regulators of steroid receptors. Previous studies have shown the signaling pathways of steroid receptors to differ across brain regions in a gene-specific manner [337, 353]. In order to identify putative region-dependent co-regulators of steroid receptors, we analyzed the co-expression relationships of each steroid receptor and a set of 62 nuclear receptor co-regulators as present on a peptide array [354] (complete data in Dataset S3). Figure 6.4A shows that the expression of co-regulators varies greatly across the different brain regions. For example, while *Ncoa1* is expressed in a fairly homogeneous manner, conforming with earlier results [353], *Ncoa4* is substantially enriched in the caudal brain regions.

The co-expressions of co-regulators with the *Ar, Gr* and *Mr* differ greatly across different brain regions, indicating selective co-regulation (Figure 6.4B–D). For example, the *Ar*/*Gr* coactivators *Pias2* [355] and *Ncoa4* [356] are highly co-expressed with *Gr* in the midbrain and hypothalamus, respectively (Figure 6.4C). However, both coactivators are not co-expressed with *Ar* within the same regions even though the relative abundance of *Ar* in the midbrain and the hypothalamus is higher than *Gr* (Figure 6.1A). *Mr* is predominantly expressed in the hippocampus where it is highly co-expressed with *Ncoa1*, *Txnrd1*, *Tref1*, *Ncor1*, *Wipi1* and *Ncor2* (Figure 6.4D). While *Ncoa1* is a known MR co-regulator [357], little is known about the effect of the other co-regulators on MR function in the hippocampus and they might be good candidates for further functional analysis.

Because there still is substantial heterogeneity across the 12 brain regions that we initially analyzed, we narrowed down our analysis to well-established target regions of steroid hormone action. We analyzed the co-expression of the 62 co-regulators with the steroid receptors in dopaminergic regions in the Ventral Tegmental Area (VTA) and Substantia Nigra (SN), known targets of steroid actions (Figures S1 and S2) [358, 359]. We found three significantly co-expressed co-regulators with *Ar* in VTA/SN: *Pnrc2*, *Pak6* and *Trerf1* (Dataset S4 and Figure 6.4E), suggesting that these may be involved in mediating AR effects on dopaminergic transmission. Furthermore, only *Pak6* was strongly co-expressed with *Gr* in the dopaminergic regions ($P < 0.01$). Thus, AR and GR may share some but not all co-regulators, much like the fact that AR binding sites may overlap in part with GR binding sites [360]. These results indicate that not only can we use genome-wide spatial co-expression to analyze the relationship between the receptors and their targets, but also to identify region-specific co-regulators.

PREDICTIVE VALUE OF CO-EXPRESSION FOR HORMONE RESPONSIVENESS: *Magel2* IS LIKELY A TARGET OF ESR1

Finally, we set out to test the predictive value of high co-expression with a steroid receptor to identify transcriptional targets. We measured the response of genes that are highly co-expressed with *Esr1* in the hypothalamus to estrogen diethylstilbesterol (DES) in castrated male mice using quantitative polymerase chain reaction (qPCR) (Materials and Methods). In the male brain, testosterone can be metabolized to estrogen or act directly via the androgen receptor. To avoid interpretation difficulties, we decided to di-

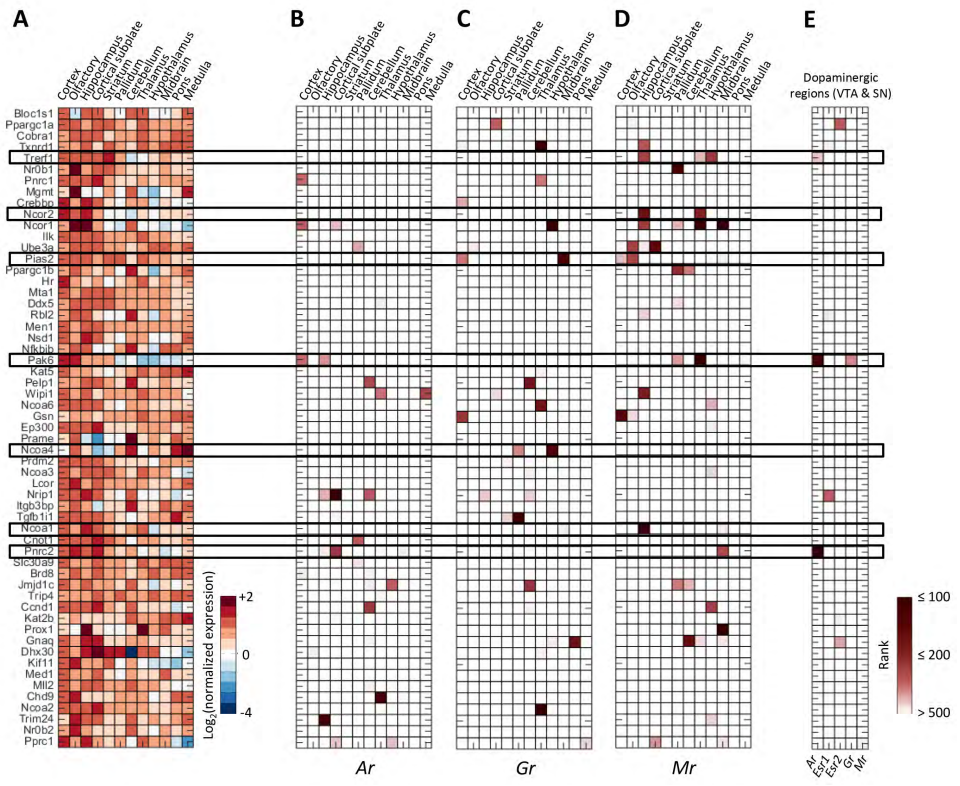


Figure 6.4: **Co-expression of co-regulators and steroid receptors.** (A) Expression of 62 co-regulators in 12 brain regions. Reported values are the average expression energy per region normalized to the average expression across the whole brain and then log_2-transformed. Co-expression ranks of the 62 co-regulators with (B) *Ar*, (C) *Gr*, and (D) *Mr*. Dark red corresponds to high rank (i.e. strong co-expression). (E) Rank sum of the co-expression rank of each co-regulator with five steroid receptors (*Ar*, *Esr1*, *Esr2*, *Gr*, and *Mr*) in the dopaminergic regions (Ventral Tegmental Area; VTA and Substantia Nigra; SN).

rectly activate brain estrogen receptors with the selective ligand DES. We selected the top 10 most strongly co-expressed genes with *Esr1* in the hypothalamus. As a negative control we used the set of genes that are not co-expressed with *Esr1* in the hypothalamus. Figure 6.5A shows examples, from the ISH experiments of the ABA, of *Irs4* and *Magel2*, two of the strongly co-expressed genes selected for validation. Since *Esr1* is not homogenously expressed across the hypothalamus (Figure 6.1B), we analyzed the responsiveness of the set of top 10 genes to DES in the anterior (MPO) and posterior (ARH) parts of the hypothalamus separately. Fold-change upregulation was modest, which may be due to non-responsiveness, a modest transcriptional response of brain targets, or to dilution of the signal in the hypothalamic homogenates (Table S3).

To further confirm co-localization, we performed quantitative double in situ hybridization (dISH) for *Esr1* and the six mRNAs (*Irs4*, *Magel2*, *Adck4*, *Unc5*, *Ngb*, and *Gdpd2*) that showed more than 1.3 fold enrichment in qPCR. *Esr1* mRNA was

consistently down-regulated more than 2-fold upon DES treatment, validating the treatment (Figure S3). *Irs4* and *Magel2* mRNA were both significantly upregulated by DES treatment in MPO (1.9 and 2.4-fold, respectively) while only *Magel2* was upregulated in ARH (2.6-fold) (Figure 6.5B and D). A 1.3-fold induction of *Ngb* mRNA in ARH did not reach statistical significance, while *Gdpd2*, *Unc5d*, and *Adck4* mRNA levels showed no trend of regulation after DES treatment (Figure S4).

The data indicate that additional criteria are necessary for reliable target prediction. As *Irs4* and *Magel2* are among the top genes expressed in the hypothalamus (ranked 1 and 11, respectively) compared to a ranking of 141 for *Adck4* and 284 for *Unc5d*, these may include a combination of expression, co-expression filters and other criteria.

IDENTIFYING GR-RELATED CORTICOSTERONE TARGETS IN THE HIPPOCAMPUS

Using gene expression measurements (qPCR and double ISH) we validated the responsiveness of *Irs4* and *Magel2* as predicted ESR1 targets to DES treatment. Despite its importance especially in detecting co-localization, gene expression remains an indirect measurement of interaction. Therefore we set out to directly detect genomic binding of steroid receptors using chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq). Previously, we used ChIP-seq to identify genomic binding sites of GR in the rat hippocampus [361]. Reanalyzing this data, we identified 694 corticosterone target genes with GR binding sites out of which 16 were within the top 200 genes co-expressed with *Gr* in the hippocampus (16/200; $P = 9.97 \times 10^{-5}$; one-sided Fisher's Exact Test), Table S4. Figure 6.6 shows examples of the GR binding sites we identified in genes strongly co-expressed with *Gr*. We did not observe any significant enrichment of corticosterone target genes in the 200 genes with the lowest correlation to *Gr* in the hippocampus (5/200; $P = 0.62$; one-sided Fisher's Exact Test) nor in the set of 200 genes with the highest correlation to *Esr1* in the hippocampus (1/200; $P = 1$; one-sided Fisher's Exact Test).

6.3. DISCUSSION

Since nuclear steroid receptors act as transcription factors, they may a priori be expected to co-express with their target genes and signaling partners. In the brain, the effects of steroid receptors are region-specific and by analyzing their spatial co-expression relationships across different brain regions, we can define potential targets and partners, as well as parallels between brain areas. The complexity and large variability in gene expression across the brain has forced many studies to analyze either brain-wide expression of a small set of genes or genome-wide expression in a few regions. The availability of high-resolution ISH-based expression maps of the mouse brain in the ABA allows the identification of all genes with a similar expression pattern across many brain regions that might indicate functional similarity between the gene products [138]. We provide a comprehensive description of the co-expression of genes with six receptors of gonadal and adrenal steroid hormones in the male mouse brain. Our results demonstrate that genes that are spatially co-expressed with receptors in a region-specific manner can enhance our understanding of brain modulation by steroid hormones.

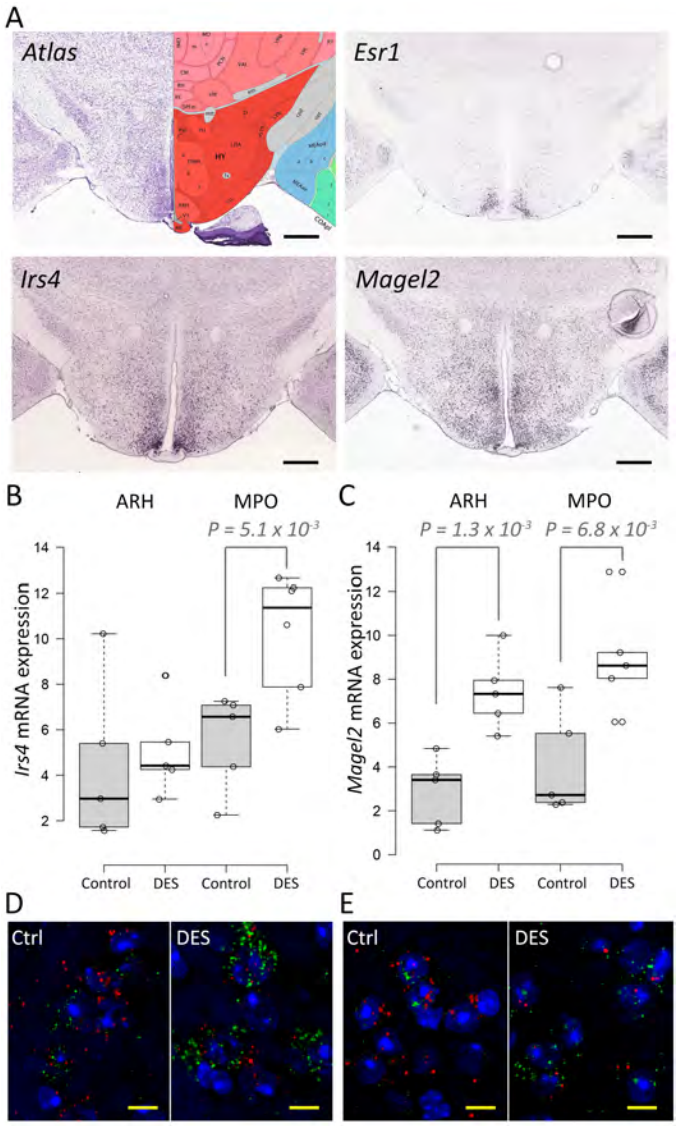


Figure 6.5: **Highly co-expressed genes are potential steroid targets.** (A) Coronal ISH sections showing the expression of: *Esr1*, *Irs4*, and *Magel2* (Scale bars, 600 µm). Data taken from the Allen Brain Atlas. Response of (B) *Irs4* and (C) *Magel2* to DES treatment in castrated mice in the anterior (medial preoptic area; MPO) and posterior hypothalamus (arcuate hypothalamic nucleus; ARH) using double ISH. (D) Double ISH of *Esr1* (Red) and *Irs4* (Green) in the anterior and posterior hypothalamus. (E) Double ISH of *Esr1* (Red) and *Magel2* (Green) in the anterior and posterior hypothalamus. (Scale bars, 10 µm. Magnification, 100×). mRNA expression in ISH was quantified as the percentage of the image surface with positive signal. Reported P values are calculated with a one-sided two-sample t-test with significant level at $P < 0.05$.

Using genome-wide spatial co-expression analysis, we observed strong co-expression of known GR transcriptional targets in the hippocampus and known ESR1 transcriptional targets in the hypothalamus. These observations support our hypothesis that genes showing strong co-expression with a steroid receptor are enriched in transcriptional targets and/or co-regulators of that receptor. In addition, the unanticipated co-expression of genes with these receptors outside their known sites of action may extend our understanding of the coordinated steroid response of the brain. For example, the high co-expression between *Gr* and its GC-responsive target genes (originally derived from the DG) in CA3, midbrain and pallidum is in line with a network that has been referred to as the neurocircuitry of stress [362]. Likewise, dendritic complexity of neurons and excitability are modulated by glucocorticoids and stress across different brain regions simultaneously [352]. Such similar responses of distinct brain regions suggest similar cellular machinery and thus similarly correlated gene expression with the responsible receptor.

For the genes that are expressed in a sex-specific manner, we confirmed their co-expression with *Esr1*, *Esr2*, and *Pgr* which is in accordance with their regulation by gonadal steroids [94]. Lack of co-expression with *Ar* may reflect the fact that many testosterone effects on hypothalamus are mediated by estrogen receptors after aromatization of testosterone into estradiol. It is as yet unclear whether the significant co-expression with *Pgr* reflects simply co-expression of *Esr1* and *Pgr*, or also points to progesterone regulation of these genes. Regardless, we extended the co-expression between sexually dimorphic genes to extra-hypothalamic sites, pointing to a parallel regulation in at least the pallidum, a region that includes the bed nucleus of the stria terminalis where regulation by (non-specified) gonadal hormones has been observed [94].

Our analysis of the co-expression of co-regulators and steroid receptors identified known relationships, such as the high co-expression between *Ncoa1* and *Mr* in the hippocampus [353]. More importantly, this brain-wide analysis provides an overview of potentially unknown relationships between steroid receptors and co-regulators. By focusing on dopaminergic regions (VTA and SN), we identified strongly co-expression of *Pak6* with *Ar* as well as *Gr*. Of interest, *Pak6* is a known AR co-regulator [363] and *Pak6* knockout mice show several locomotion and behavioral deficits which are likely related to disturbed dopaminergic transmission [364]. Thus, this example of *Pak6* co-expression underscores the feasibility of our methodology to find potential partners of nuclear steroid receptors. Of note, steroid receptor-coactivator interactions may be induced with a certain degree of specificity by selective modulator types of steroid receptor ligands [357, 365]. The co-expression of steroid receptors with their coactivators may not only predict steroid responsiveness, but also point to selective activation of particular circuits with synthetic ligands [357].

To test whether spatial co-expression can be used to predict transcriptional targets of steroid receptors in the brain, we used qPCR and double ISH to assess if genes strongly co-expressed with *Esr1* in the hypothalamus include any ESR1-targets. Among the tested genes, we identified two estrogen-regulated genes, one previously known (*Irs4*) [94] and a new target (*Magel2*). Loss of *Magel2* leads to impaired reproduction, providing an immediate link to estrogen regulation [366]. This gene is deleted in Prader-Willi syndrome, that is associated with hypogonadotropic hypogonadism, obesity and hyperpha-

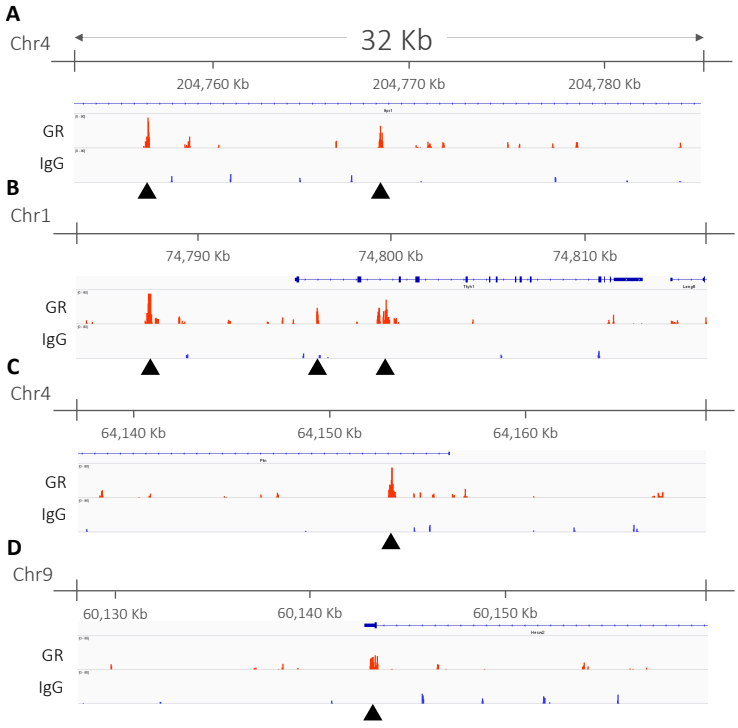


Figure 6.6: **Validating GR targets in hippocampus using ChIP-Seq.** Examples of GR binding sites to genes strongly co-expressed with Gr in the hippocampus: *Itpr1* (A), *Ttyh1* (B), *Ptn* (C), and *Hecw2* (D). For each gene, a genomic region of 32 kb centered around the identified peak is shown using the Integrative Genomics Viewer (IGV) [371]. Black arrows indicate the intergenic peaks identified within the gene.

gia [367]. Likewise, *Irs4* has a role in hypothalamic leptin signaling and regulation of metabolism [368]. Therefore, hypothalamic estrogen responsiveness of *Magel2* and *Irs4* may be related to estrogen-effects on metabolism [369]. The presently modest predictive power may be improved by incorporating the effect size (i.e. the absolute expression of a gene), given the values for true positives *Irs4* and *Magel2*. Also the presence of conserved steroid response elements on the DNA could be a useful additional filter [370].

The enrichment of known targets and co-regulators of a certain nuclear receptor within the same brain regions where the nuclear receptor is expressed confirms the validity of our analysis. Our approach is even strengthened by the notion that the receptor and its targets and/or co-regulators are significantly co-expressed despite the genome-wide and brain-wide qualitative approach of measuring mRNA levels using ISH. However, there are some intrinsic limitations to the analysis. First, while the quality of ISH is overall high, it is insufficient for some genes. Of the three datasets covering expression of *Gr*, only one was of sufficient quality. Also, *Ncoa1* which codes for an important co-regulator for *Esr1* and *Gr* [337, 353] is expressed at low levels and not significantly associated with the two receptors. Consequently, there is the risk for increased false negative results associated with a genome-wide approach using this data. Second, the ABA maps

the expression of all genes under the same normal conditions. This dataset, while unique in its brain-wide and genome-wide coverage, does not include variations between individuals as well as context-specific expression patterns (Materials and Methods).

Our approach relies on Pearson's correlation as a measure of similarity between 3D expression patterns of genes, summarized to 200µm isotropic voxels. While using the expression volumes instead of the original ISH slices simplifies computations and reduces noise effects, the lower resolution yields the analysis of small brain nuclei unreliable. For example, the very small number of voxels representing dorsal raphe nucleus in the 3D atlas hampered analysis of the serotonergic dorsal raphe nucleus. By using correlation as a measure of co-expression we detect both direct and indirect statistical associations between genes rather than causal relationships, yielding functional validation using expression measurements (qPCR and/or double ISH) and ChIP analysis crucial to confirm predicted associations as well as causality.

Concluding, we have shown that nuclear steroid hormone receptors co-express with genes not known to associate, and in brain regions where steroids were not known to be active. These findings point towards brain region specific signaling machinery of the steroid receptors.

6.4. MATERIALS AND METHODS

ALLEN MOUSE BRAIN ATLAS

The Allen Brain Atlas of the Mouse Brain [24] (`http://mouse.brain-map.org`) is a spatially-mapped in situ hybridization gene expression atlas of the 8 week old adult C57BL/6J male mouse brain. The genome-wide atlas contains expression data for ~20,000 genes. For each gene, ISH brain sections were sampled at $25\mu m$ intervals across the entire brain. The high (in-plane) resolution primary data from each experiment were reconstructed in 3D and registered to the Nissl stain-based reference atlas (Allen Mouse Reference Atlas; ARA), created specifically for this project. For each gene, the data were then aggregated into isotropic voxels defined by a uniform $200\mu m$ grid in the reference space. Resulting data consists of a spatially aligned $67 \times 41 \times 58$ (rostral-caudal, dorsal-ventral, left-right) volume for each gene. The ontology of the ARA is used to label individual voxels with their anatomical nomenclature. Some genes were assayed more than once, using a different probe or plane of sectioning (sagittal or coronal). Generally, ~20,000 genes were assayed using sagittal-sectioning experiments, while the coronal-sectioning experiments were carried out for ~4,000 genes.

The Allen Mouse Brain Atlas provides expression of genes under normal conditions. Brain sections were collected from thousands of animals and hence do not represent a single individual. In brief, each brain sectioned either in sagittal or coronal planes was used to generate 8 series (each series contain 5 slides, each slide contains four sections) [24]. Each of these series was hybridized to a single gene with each physical brain used to survey several independent genes [209]. For many genes in the dataset several experiments were conducted, resulting in multiple measurements of those genes from different animals. Moreover, for genes assayed using both sagittal and coronal sectioning experiments, sections are collected from different animals. A visual analysis of the expression pattern of the *Man1a* gene which is measured using 19 different experiments

(18 sagittal and 1 coronal) shows a high consistency of expression patterns, although these come from different animals.

DATA PREPROCESSING

We downloaded the 26,069 expression energy volumes corresponding to all experiments (21,722 sagittal and 4,347 coronal) through the Application Programming Interface (API) of the ABA on the 12th of February 2013. Expression energy $E(S)$ is a measurement combining the expression level $I(v)$ (the integrated amount of signal within each voxel) and the expression density (the amount of expressing cells within each voxel). The average expression energy of gene g in region S is calculated as:

$$E_g(S) = \frac{\sum_{v \in S} M(v) \times I(V)}{|S|} \tag{6.1}$$

where v is a voxel in region S, $|S|$ is the total number of voxels representing S, $M(v)$ is a binary expression mask with $1'$s and $0'$s representing expressing and non-expressing voxel, respectively.

In sagittal-sectioning experiments, data was generated from the left hemisphere of the brain only while in coronal-sectioning experiments; data was generated from both hemispheres. Voxels with more than 20% missing data (no gene expression value) were removed from further analysis, resulting in 27,365 voxels in the sagittal datasets and 61,164 voxels in the coronal datasets.

SPATIAL CO-EXPRESSION

We used Pearson's correlation coefficient as a measure of similarity between 3D spatial expression profiles. Given a steroid receptor of interest (seed gene), we calculate the Pearson's correlation between the spatial expression profile of that seed gene and every other gene in the ABA based on the expression values within any structure of interest (e.g. for the whole brain correlations were calculated based on the expression across 27,365 voxels in the sagittal datasets and 61,164 voxels in the coronal dataset).

Together with the co-expression calculations we also calculated the average expression energy of gene g in structure S, $E_g(S)$, as well as the normalized average expression $K_g(S)$:

$$K_g(S) = \frac{E_g(S)}{E_g(Brain)} \tag{6.2}$$

where $E_g(Brain)$ is the average expression of gene g in the whole mouse brain.

ENRICHMENT ANALYSIS

To characterize the functional associations of nuclear receptors in each of the 12 brain regions, we performed functional enrichment analysis on the top 200 spatially co-expressed genes. Functional enrichment analysis was performed using Enrichr [372]. For each region we report the top ten enriched Gene Ontology (GO) Biological Process and Molecular Function.

GENE SET ANALYSIS

In order to assess the co-expression between a set of genes and a steroid receptor of interest, a Mann-Whitney U-test is used. The test assesses the null hypothesis that the correlations of the set of targets or mediators on one hand and the correlations of all other genes on the other hand are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians. The returned P values from different experiments are corrected for multiple testing by controlling the false discovery rate using Benjamini-Hochberg method [268] (FDR-corected).

SELECTING TARGETS FOR VALIDATION

In order to select genes for validation, we generated a list of genes ranked by the strength of their co-expression with *Esr1* in the hypothalamus. We restricted our analysis to the list of genes with a coronal experiment in the ABA (4,345 genes), Dataset S5. In order to improve our predictions of co-expressed genes, we filtered out genes with an normalized average expression $< T_{HY}$ in the hypothalamus. In our experiments, we used $T_{HY} = 0.5$ since the average expression of our receptor of interest, *Esr1*, in the hypothalamus was 0.44. After filtering, we selected the top ten co-expressed genes as *Esr1*-related genes for validation using quantitative polymerase chain reaction (qPCR). In contrast, we selected the 10 genes showing the weakest co-expression with *Esr1* (correlation ≈ 0) as a background set.

SUM OF RANKS ANALYSIS

In order to assess a set of co-expressed genes for a seed gene in a set of functionally-related brain structures, such as the dopaminergic system composed of the ventral tegmental area (VTA) and substantia nigra SN (Figure S1), we used a ranksum analysis. We ranked all genes based on their correlation to the seed gene within the structure of interest. Since the number of samples (voxels) used in the correlation calculation varies between different brain structures (i.e. brain structures have different numbers of voxels), comparisons of co-expression across different structures are carried out based on the rank of the gene in a specific list rather than comparing correlation values. Given a set of structures S, we calculate the sum of ranks RS such that: $RS^S_{i,j} = \sum_{s \in S} R^s_{i,j}$, where $R^s_{i,j}$ is the rank of the correlation between gene i and gene j in structure s. The rank of the correlation is calculated as the rank of the correlation value between seed gene i and target gene j among the list of correlations of all genes with the seed gene i.

We assessed the significance of the sum of ranks value based on permutations. Given a set of n functionally-related brain structures, we randomly draw n random integers from a discrete uniform distribution ranging from 1 to the total number of genes (26,022 in case of all genes and 4,345 in case of genes with coronal only experiment). We repeated the experiment 10,000 times and calculated the sum of the randomly drawn numbers to obtain a probability distribution function (PDF) of obtaining a certain sum of ranks.

VALIDATION USING QPCR AND DOUBLE IN SITU HYBRIDIZATION

C57Bl/6J mice were obtained from Charles River Laboratories (Maastricht, The Netherlands) at the indicated age and were kept one week under standard housing conditions before they were enrolled in an experimental set-up. Nine-week-old male mice underwent gonadectomy or a sham operation under isoflurane anesthesia. Gonadectomy involved a small incision in the skin after which the testes were removed. After one week of recovery, these mice received daily subcutaneous injections with $100\mu g/kg$ diethylstilbesterol (DES) (Steraloids Inc., Newport, RI) dissolved in olive oil or the olive oil vehicle alone for one week before they were terminated by cardiac puncture under isoflurane anesthesia. Brains were rapidly dissected and frozen on powdered dry ice, and stored at −80 C. All animal experiments were performed with the Approval of the Animal Ethics Committee at Erasmus MC, Rotterdam, The Netherlands. To collect mRNA, frozen brains were cut sagitally over the midline and $60\mu m$ sections containing hypothalamus from one hemisphere were collected on uncoated glass slides (Menzel-Gläser, Braunschweig, Germany). Hypothalamic tissue was punched out using appropriate Harris Uni-core punching needles (Tedpella, Redding, CA, USA) and pooled per anterior (+0.26 to $-1.22mm$ relative to Bregma) or posterior (-1.22 to $-2.7mm$ relative to Bregma) division. RNA isolation and cDNA synthesis have been performed as described in [373]. Quantitative polymerase chain reaction (qPCR) was performed on a IQ5 PCR platform (Bio-Rad) as described in [374], using 36b4 as housekeeping gene. Primer sequences are listed in Dataset S6.

For Non-isotopic double label semi-quantitative in situ hybidrization we used the Panomics View-RNA method (Affymetrix, Santa Clara, CA, USA). Probe sets were designed by, and are available from the manufacturer. $12\mu m$ thick cryosections on Superforst plus microscope slides (Menzel Gläser, Braunschweig, Germany) were postfixed in 4% (vol/vol) formaldehyde (Sigma-Aldrich). Pre-incubation was performed following manufacturer's instructions (`https://www.panomics.com/products/rna-in-situ-analysis/viewrna-ish-tissue-assay/how-it-works`). Probes were hybridized for 4 hours in a Startspin thermobrite stove (Iris sample processing, Westwood, MA, USA). Linear amplification and visualization steps were performed following manufacturer's instructions. Slides were lightly counterstained with Mayer's hematoxylin, and DAPI (1 minute incubation at $3\mu g/ml$), and embedded in Innovex mounting medium (Innovex Biosciences, USA).

VALIDATION USING CHIP-SEQ

We remapped the ChIP-seq data from Polman *et al.* [361] to the rattus norvegicus genome version 5 (rn5) using Burrow-Wheeler Aligner [375] on default settings. GR peaks were called using Model-based Analysis of ChIP-Seq (MACS) [376] – version 2.14, with the IgG antibody binding dataset as the background using the following settings: P value cut-off = 0.05; model fold = $[10,40]$; λ = 1000/10000; effective genome size = 2.5×109. In total, we identified 694 genes with intergenic GR binding peaks. Data were visualised by uploading bigwig files to Integrative Genomics Viewer (IGV) [371].

6.5. SUPPLEMENTARY MATERIAL

The online version of this article contains supplementary material[2].

[2] www.pnas.org/lookup/suppl/doi:10.1073/pnas.1520376113/-/DCSupplemental

Chapter 7

Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex

Sepideh Babaei*
Ahmed Mahfouz*
Marc Hulsman
Boudewijn PF Lelieveldt
Jeroen de Ridder
Marcel JT Reinders

This Chapter is published as: *PLoS Comput Biol* (2015) 11(5):e1004221, doi:10.1371/journal.pcbi.1004221.
*Equal contribution.

THE three dimensional conformation of the genome in the cell nucleus influences important biological processes such as gene expression regulation. Recent studies have shown a strong correlation between chromatin interactions and gene co-expression. However, predicting gene co-expression from frequent long-range chromatin interactions remains challenging. We address this by characterizing the topology of the cortical chromatin interaction network using scale-aware topological measures. We demonstrate that based on these characterizations it is possible to accurately predict spatial co-expression between genes in the mouse cortex. Consistent with previous findings, we find that the chromatin interaction profile of a gene-pair is a good predictor of their spatial co-expression. However, the accuracy of the prediction can be substantially improved when chromatin interactions are described using scale-aware topological measures of the multi-resolution chromatin interaction network. We conclude that, for co-expression prediction, it is necessary to take into account different levels of chromatin interactions ranging from direct interaction between genes (i.e. small-scale) to chromatin compartment interactions (i.e. large-scale).

7.1. INTRODUCTION

The three dimensional (3D) conformation of chromosomes in the cell nucleus plays an important role in determining which genes are expressed in a cell [377–382]. In particular, it has been shown that genes are often regulated by elements that are located far away in terms of the linear genome sequence [383, 384]. In fact, transcribed genes tend to spatially associate with their regulatory elements which results in 3D clustering of co-regulated genes [383, 385]. Moreover, there is increasing evidence that transcription occurs at specific nuclear sites, sometimes called transcription factories [383, 386].

Chromosome conformation capture techniques, such as 3C, 4C, 5C, and Hi-C, allow direct measurement of chromatin interactions and thereby the study of the role of these interactions in gene regulation [387–389]. Using 4C, for instance, it was demonstrated that the 3D structure of the yeast genome correlates with gene co-expression [379]. Dong *et al*. used Hi-C data from two human cell lines to demonstrate that chromatin interactions associate with co-expression [378]. Both studies, however, have shown that it is difficult to explain the relationship between co-expression and the 3D structure of the genome by considering direct chromatin interactions only. Thus, while a clear relation between chromatin interaction and co-expression exists [378–380], this relation may be better understood if more comprehensive characterizations of long-range chromatin interactions, i.e. those involving also indirect interactions, are taken into account [390].

A more comprehensive characterization of long-range chromatin interactions can be obtained by considering the chromatin conformation data as a network [391, 392]. In such network, termed Chromatin Interaction Network (CIN), a genomic locus is represented by a node while links between the nodes denote chromatin interactions. Investigation of the CIN topology may reveal properties of the 3D genome organization that are important for understanding its function, such as co-expression of genes.

Characterizing the topology in biological networks has been extensively explored, for instance to gain insight into the functional relationships encoded in such networks [393, 394]. Standard network topological measures, such as shortest path, betweenness centrality and clustering coefficient, have been used to capture either the topology

around a single node or the global topology of the whole network [395, 396]. As a result, these measures of network topology operate at a fixed zoom-level. Recently, scale-aware topological measures have been shown to superiorly predict gene function and interactions by characterizing the topology of protein interaction networks at different scales [30, 394]. In this work, we explore the use of scale-aware topological measures (STMs), proposed in [30], to describe the CIN topology. Analyzing the CIN topology enables us to study the relation between long-range chromatin interactions and co-expression.

The CIN constructed in this study is based on Hi-C measurements from the mouse cortical cells [384]. In the brain, genes with a common expression pattern across the brain may have a common role in influencing the function of the brain region in which they are co-expressed [101]. In order to study spatial co-expression in the mouse brain, and mammals in general, it is necessary to map the expression at sufficient resolution to decode the high complexity [131]. The Allen Mouse Brain Atlas (ABA) [24], a genome-wide map of gene expression across the brain, provides sampled cellular-resolution in situ hybridization sections at a $25\mu m$ interval across the entire brain. We use this high-resolution dataset to obtain spatial co-expression relationships between genes at the cellular level (Figure 7.1), i.e. two genes will be co-expressed if they are expressed in the same set of cells across the brain.

To test the hypothesis that co-expression in the cortex is encoded in the CIN, we employ a supervised learning procedure. More specifically, we aim to predict the spatial co-expression between gene-pairs based on a set of features that describe the topology of the connection between the two genes in the CIN. We show that the resolution at which the chromatin interactions are captured affects the prediction of co-expression from genomic organization. In particular, our results reveal that the accuracy of the prediction is increased when measures from different Hi-C resolutions are integrated. Finally, we clearly demonstrate the importance of using descriptions of the CIN topology at different scales, ranging from specific interactions between transcription start sites of genes (small-scale) through interactions between whole genes (medium-scale) and interaction between chromatin compartments (large-scale).

7.2. RESULTS

INTRA-CHROMOSOMAL HI-C DATA

We collected the intra-chromosomal Hi-C data from Shen *et al.* [384]. They obtained Hi-C measurements in the mouse cortex following the methods proposed in Lieberman-Aiden *et al.* [388]. About 20-30 million cortex cells from 8-week old male C57Bl/6 mice were used to generate Hi-C contact matrices [384]. The resulting Hi-C matrices contain pair-wise chromatin contact frequencies between pairs of $40kb$ genomic segments (i.e. bins). Experimental biases, such as GC content of trimmed ligation junctions and distance between restriction sites, were eliminated by an integrated probabilistic background model as described by Yaffe *et al.* [397]. Hi-C technology measures only steady-state chromosome conformations across a population of cells. So, the resulting genome-wide interactions are averaged across the cells and are not exactly the same in any given cell [384, 398]. Yet, the variability of chromatin interactions is mostly confined to local interactions, while long-range interactions are relatively well conserved and stable [399].

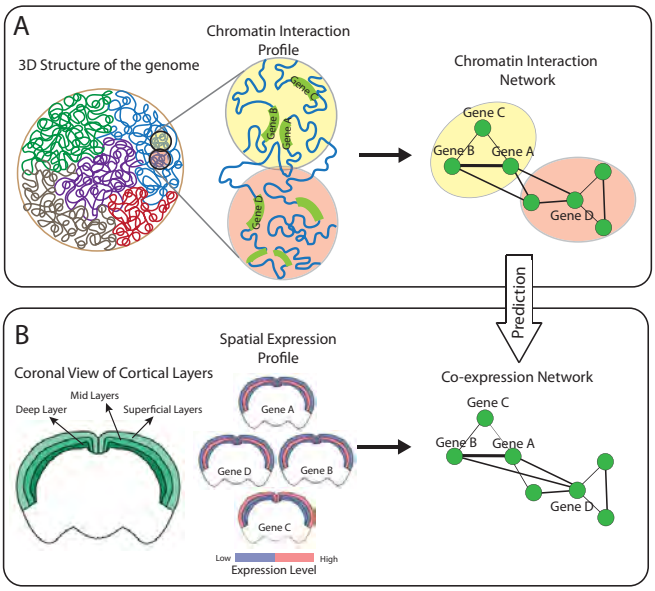


Figure 7.1: **Association between chromatin interaction and co-expression of gene-pairs.** (A) Co-regulated genes are co-localized in 3D structure of the genome through frequent chromatin interactions. Chromatin interactions can be at different levels from direct interaction between genes (interaction between Gene A and Gene B) to chromatin compartment interactions (interaction between Gene D and Gene B). Chromatin interactions between gene-pairs can be characterized by a network, termed Chromatin Interaction Network (CIN). (B) Co-expression between gene-pairs based on their spatial expression pattern across the mouse cortex. Gene A, Gene B and Gene D are expressed in the mid layers of the mouse cortex and are hence highly co-expressed. Gene C, on the other hand, is expressed in the superficial cortical layers and therefore is not co-expressed with the other three genes. The chromatin interaction profile of a gene-pair, encoded by the topological structure of the CIN, can be used to predict the co-expression status as captured by the co-expression network.

This demonstrates that different cell types share a common global architecture of their chromosomes which can be well described by the chromatin contact matrix.

Two regions that are close-by in the linear genome are expected to have higher chromatin interaction frequency, irrespective of the actual 3D organization of the genome (S1_Fig). To account for this, several studies have defined normalized Hi-C contact matrices assuming that the Hi-C interactions are normally distributed [388, 400] or independent [401]. Alternatively, we used a non-parametric rank based normalization method [402] to describe the Hi-C score distributions for a certain distance, which we found to be more powerful for detecting variations across the genomic distance.

MULTI-RESOLUTION HI-C DATA

Since we are interested in predicting co-expression patterns of genes, each bin-based Hi-C matrix is converted to a gene-based Hi-C matrix based on the Hi-C interaction between the corresponding bins in which the genes reside (see Materials and Methods).

While assigning Hi-C interactions between genes, the bin size of the Hi-C data controls the genomic neighborhood considered around genes. In order to capture interactions between genes at variable linear genomic distances we varied the resolution of the Hi-C data matrices, before constructing gene-based matrices. This was achieved by considering different bin sizes between $40kb$ (high-resolution) and $1Mb$ (low-resolution). The lower resolution matrices were obtained by summing the contact frequencies of consecutive bins in the higher resolution matrices.

CHROMATIN INTERACTION NETWORK (CIN)

To determine the Hi-C interactions between each gene-pair we take the Hi-C interaction between the corresponding bins in which the genes reside. However, some genes might span multiple bins, depending on gene size and bin size. In this case, we determine the Hi-C interaction for a gene-pair (x, y) by one of two approaches. In the first approach, referred to as MAX-mapping, we define a link as the maximum Hi-C value among all possible interactions, i.e. $\hat{h}_{xy} = max_{i \in x, j \in y}(\hat{h}_{ij})$. In the second approach, referred to as TSS-mapping, we define a link as the Hi-C score between the bin-pair which contains the transcription start sites (TSS) of the two genes, i.e. $\hat{h}_{xy} = \hat{h}_{ij}$; where: $TSS(x) \in i$ and $TSS(y) \in j$. We applied a threshold to convert the weighted gene-based Hi-C matrix to an un-weighted matrix by retaining only interactions that exceed the 90^{th}-percentile of all Hi-C score across all chromosomes at the corresponding bin size.

We constructed one CIN per chromosome per resolution because the employed Hi-C data contains only intra-chromosomal interactions. For each CIN $H^R_{chr} = (G, I_H)$, G represents the set of nodes corresponding to genes and I_H represents the set of links corresponding to Hi-C interactions between genes that exceed the 90^{th}-percentile of all Hi-C scores across all chromosomes at a resolution R.

CIN TOPOLOGY

There are several topological measures which capture graph structure for nodes and/or links in a network [393, 395]. In this work, we calculated five graph-topological measures of the chromatin interaction network: shortest path length, Jaccard index, degree (and closeness) centrality, betweenness centrality, and clustering coefficient (Table 7.1). Since our goal is to predict co-expression between gene-pairs, all features used by the classifier should be link-based. Therefore, we converted all the node-based topological measures (degree-closeness centrality, betweenness centrality and clustering coefficient) to link-based measures by taking the average and the difference between the values of the gene-based measure for each gene-pair. For example, for a gene-pair (x, y), the clustering coefficient of the link between x and y is described by $\{\left|(cc(x) - cc(y)\right|, \frac{1}{2}(cc(x) + cc(y))\}$. As a result, each link in the interaction network is represented by eight link-based topological features.

In addition to the standard topological measures, we used the scale-aware topological measures (STMs) described by Hulsman *et al.* [30] to capture the network characteristics across different scales. STMs are based on diffusion kernels [403], a network smoothing process in which the diffusion strength β parameter determines the scale at which the network is considered [404]. By varying the scale at which we consider the CIN, different types of interactions are taken into account. For example, specific interac-

Table 7.1: **Topological Measures.**

Measure	Description	Scale-aware version
Shortest Path	The minimum number of vertices connecting node x and y, $s(x,y)$	$s^\beta(x,y) = -\log(K^\beta_{x,y})$
Jaccard Index	The proportion of shared nodes between x and y relative to the total number of nodes connected to x or y, $J(x,y) = \frac{n(x)\cap n(y)}{n(x)\cup n(y)}$	$J^\beta(x,y) = \frac{\sum_i min(K^\beta_{x,i}, K^\beta_{i,y})}{\sum_i max(K^\beta_{x,i}, K^\beta_{i,y})}$
Degree & closeness Centrality	The degree centrality reflects the connectivity of a node in terms of the number of edges connected to it, $deg(x)$ and closeness centrality reflects the farness of a node x, by summing the shortest path distances to all other nodes, $c(x) = \frac{1}{\sum_{i\setminus x} s(x,i)}$	$c^\beta(x) = 1 - K^\beta_{x,x}$
Betweenness Centrality	The number of shortest paths that pass through a node, $b(x) = \sum_{i,j\setminus x} \frac{q_{ij}(x)}{q_{ij}}$ where q_{ij} is the number of shortest paths between nodes i and j, and $q_{ij}(x)$ the number of those paths that pass through x	$b^\beta(z) = \frac{1}{N^2}\sum_{x,y} \left(s^\beta(x,y) - (s^\beta(x,z) + s^\beta(z,y))\right)$
Clustering Coefficient	The number of edges between its direct neighbors including itself, divided by the maximum number of possible edges, $cc(x) = \frac{2\lvert e_x\rvert}{deg(x)(deg(x)-1)}$	$cc^\beta(x) = \sum_{i\setminus x} K^\beta_{x,i} J^\beta(x,i)$

N is the set of all nodes in the network, and n is the number of nodes. (x,y) is a link between nodes x and y, $(x,y \in N)$. $a(x,y)$ is the connection status between x and y: $a(x,y) = 1$ when link (x,y) exists; $a(x,y) = 0$ otherwise. Scale-aware versions are base on diffusion kernel where $K^\beta = e^{\beta(A-D)}$, A is the adjacency matrix and D is the degree matrix of the network. The diffusion level β determines the scale. $K^\beta(x,y)$ is the diffusion strength between node x and y.

tions between transcription start sites of genes are more pronounced at the small-scale while interactions between chromatin compartments are more pronounced at the large-scale.

CO-EXPRESSION NETWORK

The Allen Mouse Brain Atlas (ABA) [24]; (http://mouse.brain-map.org/) provides a genome-wide cellular-resolution, in situ hybridization (ISH)-based, gene expression map of the 8−week old adult $C57BL/6J$ male mouse brain. A spatial co-expression map was constructed based on the similarity of the spatial expression profiles of each pair of genes across the cortex (see Materials and Methods).

The employed Hi-C data contains only intra-chromosomal interactions. Therefore,

one co-expression network was constructed per chromosome and is denoted by $E_{chr} = (G, I_E)$, where G indicates a set of nodes representing genes and I_E indicates set of links representing intra-chromosomal co-expressions between gene-pairs. The largest and smallest networks E_2 and E_{18} (S2_Fig) consisted of 338 and 119 genes (i.e. nodes), respectively. To focus our predictions on reliable interactions, we included only strongly co-expressed genes and gene-pairs without strong co-expression (see Materials and Methods).

HIGHLY CO-EXPRESSED GENES ARE SPATIALLY CO-LOCALIZED

To examine whether gene-pairs with high spatial co-expression frequently interact in the 3D conformation of chromosomes, we defined two sets of gene-pairs: strongly co-expressed genes and gene-pairs without strong co-expression (see Methods). We used a Wilcoxon rank-sum test to determine if strongly co-expressed gene-pairs have stronger Hi-C interactions, and hence are closer to each other in the 3D conformation of the chromosome, compared to gene-pairs without strong co-expression.

Figure 7.2A (and S3_Fig) shows that co-expressed genes are significantly co-localized in the nucleus in most of the chromosomes and most CIN-resolutions (Wilcoxon rank-sum test; $P < 0.0002$, Bonferroni corrected for 260 tests: 20 chromosomes × 13 resolutions). Strikingly, we observe that the resolution for which the strongest co-localization is attained is different for different chromosomes (Figure 7.2B). This observation underscores the importance of a multi-resolution approach to characterize chromatin interactions which apparently can occur between loci in the direct vicinity of genes as well as between broader regions (domains) in which these genes reside.

CHROMATIN INTERACTION PROFILES AS CO-EXPRESSION PREDICTORS

To determine whether strong co-expression can be predicted from chromatin interactions, we calculated the correlation between the Hi-C matrix and the co-expression matrix for each chromosome at different resolutions. S4_Fig shows that the correlation is very low across different chromosomes and Hi-C resolutions (−0.4 to +0.1). Additionally, training a classifier on the presence or absence of links in the CIN results in a poor classification performance (0.55 median AUC across chromosomes at $40kb$ resolution). S5_Fig shows that only 2% (average across all chromosomes) of all gene-pairs are co-expressed and connected in the CIN of each chromosome. This observation further highlights the importance of indirect chromatin interactions in explaining co-expression. Taken together, these results indicate that chromatin interaction and co-expression do not have an injective (one-to-one) relation. The relation between chromatin interaction and co-expression would be better described by a more comprehensive characterization of long-range interactions, i.e. indirect interactions.

A compelling example is given in Figure 7.3A. In Chromosome 16, *Synj1* and *Dyrk1a* genes are co-expressed (dashed red line) while their corresponding genomic loci do not frequently interact, i.e. there is no link (solid blue line) between them in the CIN at $200kb$ resolution. A classifier only taking direct chromatin interactions into account will mistakenly predict that the two genes are not co-expressed. However, both *Synj1* and *Dyrk1a* genes have strong chromatin interactions with *Pam16*, *Fstl1*, *Hmox2*, *Sidt1* and

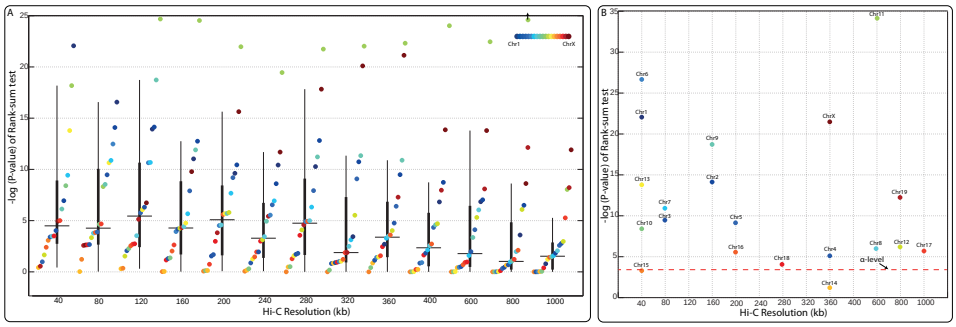


Figure 7.2: **Co-expressed genes are co-localized in 3D structure of the genome.** (A) Assessment of the enrichment of Hi-C interactions between strongly co-expressed gene-pairs compared to gene-pairs with no co-expression across different Hi-C resolutions. The y-axis indicates $-log_{10}(P)$ of the one-tailed Wilcoxon rank-sum test used for the enrichment analysis. Hi-C interactions were mapped to genes using the MAX-mapping method. In each box, the horizontal line represents the median. The thick vertical line represents the interval of $q_1 = 25^{th}$ and $q_3 = 75^{th}$ percentiles. The thin vertical line represents the interval of $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$.

(B) Overview of the Hi-C resolution at which Hi-C interactions are most significantly associated with co-expressed gene-pairs for each chromosome.

their strong co-expression can be correctly predicted if these indirect interactions are considered. For this particular example, the indirect interactions between the two genes can be characterized by the Jaccard index which captures to what extent the two genes have common direct neighbors.

Another example is the interaction between *Kcnc4* and *Tspan5* in Chromosome 3 (Figure 7.3B). *Kcnc4* and *Tspan5* directly interact in the $200kb$-CIN (solid blue line) but they are not strongly co-expressed (no dashed red line). Nevertheless, this direct chromatin interaction may explain the strong co-expression between gene-pairs in the CIN neighborhood that lack a direct chromatin interaction themselves. For example, *Wdr47* and *Lphn2* are co-expressed although they are not directly connected in the CIN (no solid blue line) but their co-expression can be explained by the chromatin interaction path through the *Lhfp*, *Kcnc4* and *Tspan5* genes. Similarly, the co-expression of *Wdr47* and *Rap1gds1* can be explained by the chromatin interaction path through *Lhfp* and *Kcnc4*. For this example, the importance of the Hi-C link between *Kcnc4* and *Tspan5* to describe strong co-expression between their neighboring genes in the CIN can be captured using the betweenness centrality of both genes. Both examples illustrate that strong co-expression between gene-pairs can be better explained by their chromatin interaction profile, defined as the path connecting two genes in the context of the CIN.

TOPOLOGICAL DESCRIPTIONS OF MULTI-RESOLUTION INTERACTION NETWORKS INCREASE THE PREDICTION PERFORMANCE

For each CIN of a certain resolution, we calculated the standard graph-topological measures and trained a random neural network (RNN) classifier using the resulting topological features (see Materials and Methods). The classification results are summarized in

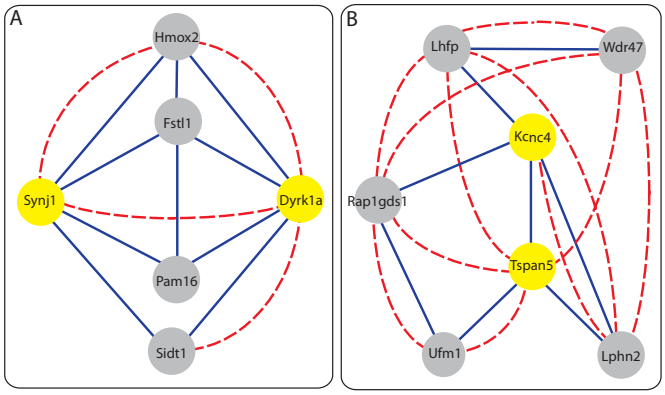


Figure 7.3: **Chromatin interactions of gene-pairs in the CIN at** $200kb$ **resolution.** (A) *Synj1-Dykr1a* (yellow nodes) in Chromosome 16 are co-expressed (dashed red link) but their corresponding genomic loci do not interact frequently (no blue link). Both genes have strong chromatin interactions with 4 other genes (grey nodes) resulting a high Jaccard index between them. (B) *Kcnc4-Tapan5* (yellow nodes) in Chromosome 3 directly interact (solid blue line) but they are not strongly co-expressed (no dashed red line). This direct chromatin interaction explains the strong co-expression between other gene-pairs in their neighbourhood, such as *Wdr47-Lphn2* and *Wdr47-Rap1gds1*, which are not directly connected in the CIN themselves (no solid blue line). The betweenness centrality measure of the link between *Kcnc4-Tapan5* can describe the strong co-expression between their neighbouring genes. Chromatin interaction and co-expression are shown by solid blue and dashed red links, respectively.

Figure 7.4 (Box 1-4 and 7). The figure shows that an increased - yet moderate - classification performance is obtained when standard topological measures of the CIN at a single resolution (median AUC of 0.72 for $200kb$ and 0.73 for $40kb$, Figure 7.4 (Box 1, 2)) are used as features (compared to 0.55 AUC when using only direct interactions).

To evaluate the effect of Hi-C resolution on co-expression prediction, we applied the RNN classifier to a concatenated set of standard topological measures obtained from CINs at different Hi-C resolutions ($40, 80, 120, 160$, and $200kb$), i.e. the topological descriptions of each resolution are concatenated in one feature representation. At a high Hi-C resolution ($40kb$) we mainly focus on chromatin interactions between pairs of genes. On the other hand, at a low Hi-C resolution ($200kb$) we consider interactions between larger genomic domains. Our multi-resolution approach increased the power of the interaction data to predict co-expression (Figure 7.4, Box 3, 4, 7), supporting our earlier observation that gene regulation occurs at different regional scales of chromatin interaction, such as the gene-level or the level of broad regions. So far, the best prediction performance is obtained by concatenating standard topological measures of CINs built using both TSS- and MAX-mapping methods (0.77 median AUC, Figure 7.4, Box 7).

STMS IMPROVE THE PREDICTION PERFORMANCE

To examine the effect of indirect chromatin interactions on the prediction of co-expression, we described the CIN topology at multiple topological scales using STMs (see Methods). We calculated STMs of the CIN at each Hi-C resolution separately and

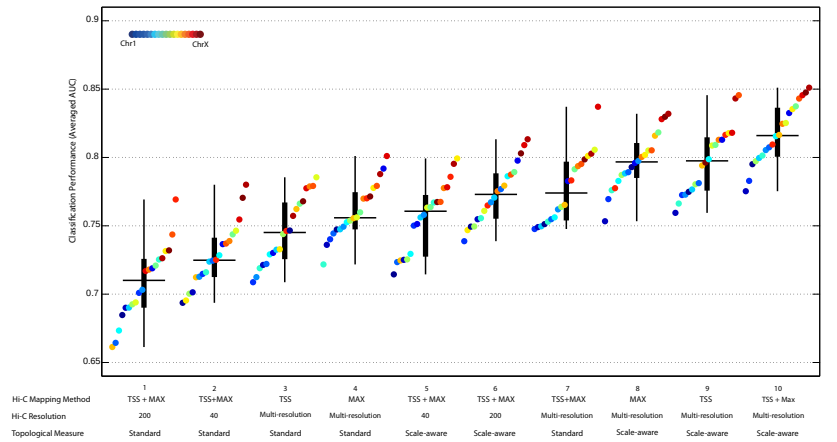


Figure 7.4: **Classification performance for the co-expression prediction based on intra-chromosomal chromatin interaction networks.** Each box encompasses the classifier performance in terms of AUC for all mouse chromosomes. Boxes are sorted based on their medians. The method that was used for computing the input feature set is given under each box. TSS or MAX refers to the mapping method for assigning Hi-C interaction between gene-pairs when the CIN is built. TSS+MAX refers to concatenated feature set of topological measures of CINs built using both TSS- and MAX-mapping methods. Multi-resolution refers to concatenated feature set of topological measures obtained from CINs at Hi-C resolutions of $40, 80, 120, 160$, and $200kb$. In each box, the horizontal line represents the median. The thick vertical line represents the interval of $q_1 = 25^{th}$ and $q_3 = 75^{th}$ percentiles. The thin vertical line represents the interval of $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$. All the values shown in the figure are also available in S6_Table

then concatenated all STMs, resulting in 800 features; 8 STMs at 10 scales applied to 10 CINs; 5 different resolutions and two mapping methods (see Methods for more details). We then followed the same procedure as before and trained a RNN classifier on this combined feature set.

Figure 7.4 (Box 5-6 and 8-10) summarizes the results obtained when using STMs rather than the standard topological measures. The performance obtained using STMs calculated at a single resolution CIN (Figure 7.4, Box 5,6) is comparable to the performance obtained by concatenating standard topological measures from multi-resolution networks (Figure 7.4, Box 7). However, by combining features from STMs applied to multi-resolution CINs, the power to predict co-expression improves significantly (Wilcoxon rank-sum test; $P < 0.00001$) (0.82 AUC, Figure 7.4, Box 10). The best performances are obtained for Chromosome 16 (0.86 AUC) and Chromosome X (0.85 AUC). The observed performance improvement demonstrates that it is important to use a scale-aware topological description of the CIN to capture the complex 3D organizational features of the genome that determine gene co-expression.

In order to analyze the effect of considering only strongly co-expressed genes on the classification performance, we assessed the performance when all co-expression links are included. In this analysis, a gene-pair is labeled co-expressed (i.e. positive class) or not co-expressed (i.e. negative class) if their correlation is above or below the median (i.e. 50^{th}-percentile) of all correlations across all chromosomes, respectively. The resulting AUCs across all chromosomes show that STMs performs better than standard mea-

sures to distinguish between co-expressed and non co-expressed gene-pairs (S7_Fig). As expected, the classification performance is lower with respect to the case where we excluded weakly co-expressed gene-pairs (i.e. gene-pairs that have a co-expression that is in between the 50^{th} and 90^{th}-percentile of all correlations across all chromosomes) (Figure 7.4). Most likely this is caused by a noisy class assignment for weakly correlated gene-pairs which confuses the classifier during training.

We also performed the classification procedure by including Hi-C scores above the median of all Hi-C scores across all chromosomes. The resulting AUCs across all chromosomes show that STMs perform better than standard measures to distinguish between co-expressed and non-co-expressed gene-pairs (S7_Fig). The classification performance is, however, less than the AUC when we defined strong Hi-C interactions as Hi-C scores above the 90^{th}-percentile of all Hi-C scores across all chromosomes (Figure 7.4).

To compare the rank-based normalization method [402] with the average-based method proposed by Lieberman *et al.* [388], we trained the classifier on the standard and scale-aware topological measures of the CIN that was built using the average-based normalized Hi-C matrices. The performance of these classifiers is lower than when constructing the CIN on using the rank-based normalized Hi-C data (S8_Fig), underscoring the usefulness of the rank-based normalization for predicting co-expression from chromatin interaction data. Nevertheless, STMs perform better than standard measures for both normalization methods, indicating that the classifier is not biased towards the normalization method.

To investigate the effect of chromatin interactions between non-genic and genic regions on the co-expression prediction we built a bin-based CIN (instead of a gene-based CIN). In the bin-based CIN, nodes represent non-overlapping bins with size of $200kb$ and links represent Hi-C interactions between bins that exceed the 90^{th}-percentile of all Hi-C scores across all chromosomes at a $200kb$ resolution. We calculated standard and scale-aware topological measures (8 measures) for all links in the bin-based CIN. The classifier was trained on topological measures of the portion of links that connect two gene-loci. In this strategy, the interaction profile between two gene-loci is characterized by chromatin interactions of all genomic regions across the scales. The resulting AUCs across all chromosomes show that STMs performs better than standard measures to distinguish between co-expressed and non-co-expressed gene-pairs (S9_Fig). It is interesting to observe that the classification performance is approximately similar to that obtained when gene-based CINs were used. This suggests that the STMs can capture all the necessary information from the genic Hi-C links.

CIN TOPOLOGY DIFFERS PER CHROMOSOME

To investigate the variation in topological properties of the CIN of different chromosomes, we performed a leave-one-chromosome-out experiment. If the CINs of all 20 mouse chromosomes share the same topological properties, then it would be possible for a classifier trained on all but one chromosome to accurately predict the co-expression labels of the left-out chromosome. To test this hypothesis, we trained the RNN classifier on the STMs (800 features) extracted from 19 chromosomes and then tested the performance on the left-out chromosome. We repeated the procedure 20 times and each time,

a different chromosome was left out of training and used for testing. The maximum AUC obtained was 0.54, which indicates that the CIN of each chromosome has a unique topology, to which the high-scale STM feature values are sensitive.

The variation in topological properties of CINs across chromosomes is also observed when we trained an RNN classifier on individual topological measures. The classification performance using individual standard measures (S10_Fig) and individual STMs (S11_Fig) is highly variable across chromosomes, which explains the poor performance obtained in the leave-one-chromosome-out experiment. For instance, the clustering coefficient STM is a good descriptor of the CIN of Chromosome 3 at medium-resolution and low-scale, while it is a good descriptor of the CIN of Chromosome 10 at high-resolution across the scales (S11_Fig).

TOPOLOGICAL SIGNATURES OF CINS

To analyze the topological properties that are most predictive we trained the classifier on individual topological measures. The classification performance using individual standard measures (S10_Fig) and individual STMs (S11_Fig) shows that none of the topological measures has dominant power to predict co-expression. Therefore, the classifier requires more than a single topological descriptor to describe chromatin interaction profile between two gene-loci. In order to determine the set of STMs that characterizes the CIN of each chromosome the best, we performed forward feature selection in combination with the RNN classifier. To facilitate this computationally, we reduced the number of nodes in the hidden layer to 100 and applied 5-fold cross validation. To ease interpretation, we used the STMs derived from multi-resolution CINs using the MAX-mapping method only (400 STMs, 8 measures × 5 resolutions × 10 scales). S12_Fig shows that the classification performance achieved using feature selection (0.8 AUC) is higher than the performance achieved using all features (0.72 AUC). For most chromosomes, the top 5 selected features in all 5 folds are clustering coefficient (at small-scale, $\beta < 0.5$), closeness centrality (at medium-scale ($0.5 < \beta < 3$) and Jaccard index (at large-scale, $\beta > 3$) STMs (S13_Table).

The clustering coefficient measures to what extent a gene is embedded in a well-connected component of the CIN. Selecting the small-scale clustering coefficient implies that co-expressed genes are usually embedded in a locally well-connected component in the CIN (e.g. chromatin compartment). The Jaccard index determines the fraction of common interacting genes between gene-pairs in the CIN. At a large scale it takes more indirect neighboring nodes (e.g. genes located in different chromatin compartments) into account. The closeness centrality reflects the farness of a gene by summing the shortest path distances to all other genes and at a medium scale it thus takes somewhat longer paths into account. Both Jaccard index and closeness centrality explain that common indirect interacting genes (e.g. interaction between chromatin compartments) are important to describe the co-expression pattern of a pair of genes.

Additionally, we observed that all scale-levels (small, medium and large) were selected reflecting the importance of characterizing CINs at different scales. The selection of various scale-levels could be explained by the hierarchical structure of the chromatin folding in the cell nucleus ranging from looping between the promoter regions of genes to larger chromatin compartments [387, 391]. This is corroborated in the work by

Sandhu *et al*. [391] who have shown that genomic regions are organized into a hierarchical chromatin interaction network.

STMS EFFECTIVELY CHARACTERIZE THE CIN OF CHROMOSOME 16 TO PREDICT CO-EXPRESSION

We analyzed the top selected STMs of the 200*kb*-CIN of Chromosome 16, for which the highest prediction performance is achieved, to gain insight into the topological measures and scales that best describe the network. The best classification performance (AUC = 0.84) is obtained using 206 of the 400 STMs (S13_Table) which are selected by forward feature selection. We mapped these 206 features to a 2D space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [32, 207] (see Materials and Methods).

The 2D map of all gene pairs in Chromosome 16 (Figure 7.5) shows that there are few distinct clusters of co-expressed and not co-expressed gene-pairs, i.e. clustering of red and blue dots in Figure 7.5B respectively. However, it is difficult to discriminate between the majority of gene-pairs (big cluster in the middle of Figure 7.5B), further supporting our observation of complex organization of chromatin interactions. Coloring the t-SNE with two of the top selected features, the clustering coefficient at small-scale (Figure 7.5A) and the Jaccard index at the medium-scale (Figure 7.5C), shows that gene pairs are characterized by different values of those two features, indicating their importance for the classification performance.

Since the clustering coefficient at small-scale is one of the top selected features for Chromosome 16, we used the t-SNE map to select a co-expressed gene-pair with a high clustering coefficient at a small-scale (Figure 7.5A). We constructed a sub-network of the selected gene-pair by retrieving all the Hi-C and co-expression interactions surrounding the gene-pair (Figure 7.5D). *B3galt5* and *Carhsp1* are co-expressed (dashed red link in Figure 7.5D and red dot in Figure 7.5B) although there is no direct Hi-C interaction between them (no blue link). However, it is possible to predict their co-expression because they are both part of a very well connected cluster, which is captured by a high average clustering coefficient at small scale.

Similarly, we select a co-expressed gene-pair with a high Jaccard index at the medium-scale, another top selected STM of the CIN of Chromosome 16 (Figure 7.5C). The sub-network including the selected gene-pair *Masp1* and *Abat* (Figure 7.5E), shows that they are co-expressed although no direct Hi-C interaction exists between them (no blue link in Figure 7.5E). The two genes also do not share many direct neighbors. At a medium scale, however, the Jaccard STM takes indirect neighbors into account, resulting in a high Jaccard index based on the Hi-C links between the neighbors of *Masp1* and *Abat*.

7.3. DISCUSSION

We proposed a network-based approach to better understand the 3D structure of the genome based on scale-aware topological measures of the chromatin interaction network. Previous studies have shown a strong correlation between co-expression and chromatin interaction, for example in model organisms (e.g. yeast) [379] or cell lines (human gm06990 and K562 cells) [378]. Our results demonstrate that the co-expression

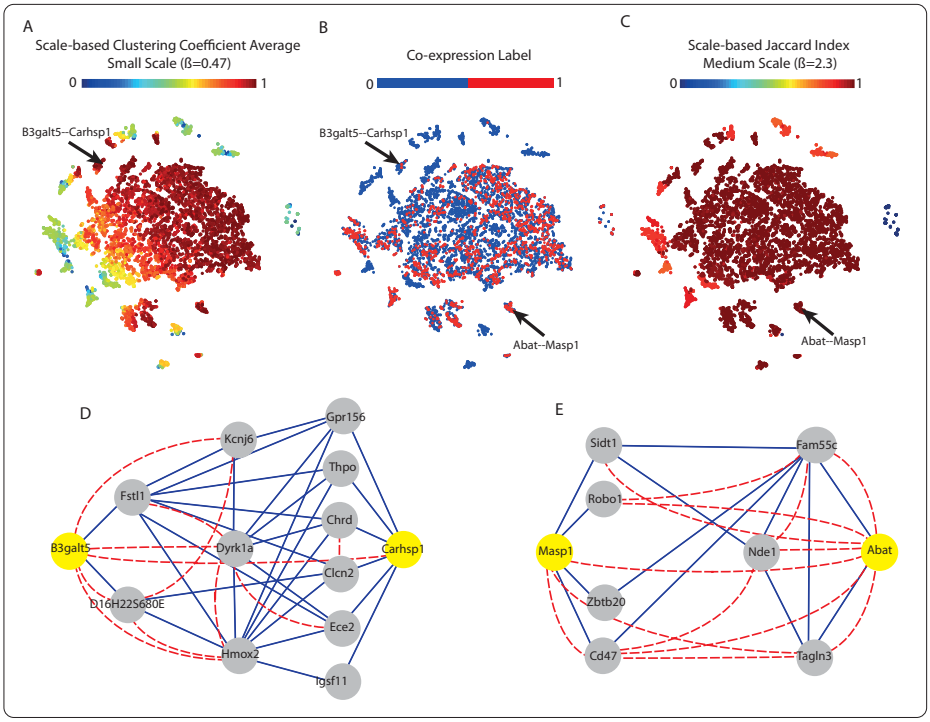


Figure 7.5: **Topological signature of Chromosome 16 CIN.** t-SNE maps of the 200kb-CIN of Chromosome 16. Each point in the map corresponds to a link between a gene-pair which is colored according to (A) Clustering coefficient at small-scale, (B) Co-expression label, and (C) Jaccard index at the medium-scale. (D) and (E) show sub-networks of the CIN surrounding selected gene-pairs (indicated in the 2D maps): (D) *B3galt5-Carhsp1* (yellow nodes) with the high clustering coefficient average; and (E) *Masp1* and *Abat* (yellow nodes) with the high Jaccard index. Chromatin interaction and co-expression are shown by solid blue and dashed red links, respectively.

relationship between a pair of genes in the mouse cortex could be accurately predicted from their chromatin interaction profile, extending previous observations in [378, 379]. Furthermore, the predictive power of our model depends greatly on the resolution at which the interactions are observed as well as the scale at which the topological properties on the interaction network are calculated. By integrating scale-aware topological measures at multiple Hi-C resolutions, we were able to predict spatial co-expression between gene-pairs with an AUC performance of 0.82. To our knowledge, this is the first attempt to predict co-expression based on genome-wide chromatin interactions.

The results also showed a general trend of the prediction performance (Figure 7.4) suggesting that STMs across multiple Hi-C resolutions are necessary to accurately capture the 3D structural features in the genome that determine spatial co-expression between genes in the mouse cortex. While the multi-resolution approach captures direct chromatin interactions between genes at variable linear genomic distances, standard

topological measures extracted from a single-resolution CIN fail to represent the complex 3D structure of genome. By using STMs [30] to describe each single-resolution CIN, we were able to capture both direct and indirect interactions between genes, and hence correctly predict their co-expression status.

The 2D t-SNE maps of the CINs using 80 standard topological measures (S14_Fig) and 800 STMs (S15_Fig) reveal a complex organization of chromatin interactions, indicating that the discrimination between co-expression labels (blue and red points in S14_Fig and S15_Fig) is a difficult task. These observations may also explain the poor classification performance obtained using a simple classifier such as nearest mean (NM). The RNN classifier, however, is able to capture the complex chromatin interaction profile of a gene-pair and their co-expression status.

Comparing the t-SNE map of standard topological measures and STMs of Chromosome 16's CIN shows that STMs are indeed more powerful in discriminating co-expression labels (S16_Fig). For example, the t-SNE map of standard topological measures shows that most of the interactions in the CIN of Chromosome 16 are characterized by a low Jaccard index value and consequently, the contribution of the Jaccard index to the classification performance is very low (S16_Fig). The scale-aware Jaccard index, however, captures indirect neighbors between a gene-pair which improves the classification performance.

Furthermore, we showed that each STM characterizes the CIN differently across scales and resolutions. For instance, the t-SNE map of STMs shows that the chromatin interaction profiles between gene-pairs in a well-connected component, indicated by a high clustering coefficient, are better captured at low resolution, whereas other well-connected components are better characterized at the high-resolution (different color pattern in S16_Fig). Additionally, some interactions are well discriminated using the clustering coefficient (a node-based STM) while other interactions are better discriminated using the Jaccard index (a link-based STM) (S16_Fig). This highlights the importance of both link- and node-based STMs in characterizing the topology of connectivity and neighborhood, respectively, of gene-pairs in the CIN to predict co-expression.

Our observations are in line with the two complementary models of how regulatory elements, such as enhancers and insulators, act to regulate the expression of distant genes [405]. The looping model assumes that loops along the genome are formed to bring distal regulatory sequences in direct contact with the promoters of target genes. Alternatively, genes undergoing transcription might co-localize in the nucleus in transcription factories, and enhancers facilitate the movement of genes into or out of these factories. Our finding that a multi-resolution scale-aware encoding of the CIN topology better predicts co-expression indeed shows that chromatin interactions occur at different levels, ranging from direct interactions between the transcription start sites of genes (small-scale) through interactions between genes (medium-scale) up to interaction between chromatin compartments (large-scale).

The topology of different chromosomes might be radically different, due to both chromosome length and different fractions of chromatin types. High-scale STM values are in particular sensitive to such a change in topology, and are likely to be one of the causes for the differences in performance. Indeed, a classifier, such as the one proposed

here, might also be used to characterize chromatin conformation.

In the current study, we used only intra-chromosomal interactions. Nevertheless, our proposed methods could principally be applied to inter-chromosomal interactions given that the data is normalized properly across chromosomes [397, 406]. Furthermore, the method is not tissue- or organism-specific and can be generalized to predict any functional relationships (not only co-expression) between genomic loci (bins or genes) based on the characterization of the CIN.

The brain is a very complex structure with large variability in gene expression patterns across different regions. Using the high-resolution maps of the ABA, this variability could be used to identify distinct groups of genes with a similar expression pattern indicating their functional similarity [138, 142]. For example, several studies analyzed the relationship between spatial-co-expression and connectivity in the mouse brain [147, 148, 150, 151]. Menashe *et al*. [131] used a spatial co-expression network of the mouse brain to identify common neuro-functional properties of autism-related genes. We expect that within the brain, and especially the cortex, many genes vary and that their biologically meaningful spatial correlation patterns are reflected by long-range chromatin interactions.

With the recent association of dozens of mutations in chromatin regulators to neuropsychiatric disorders [407], our method provides a promising approach to investigate the effect of those regulators on the cortical regulatory network. A good characterization of interactions in the CIN and their relationship to co-expression can add to our understanding of the genetic etiology of these diseases.

7.4. MATERIALS AND METHODS

RANK-BASED NORMALIZATION OF HI-C CONTACT MATRICES

In order to eliminate genomic distance bias in a Hi-C matrix, each Hi-C contact value is replaced by its relative rank compared to Hi-C contacts between bins with a similar genomic distance, measured in base-pairs [402]. The normalized Hi-C score $\hat{c}_{ij}$ is defined as the rank of c_{ij} in the vector C^d, where c_{ij} is the Hi-C contact between bin i and j with genomic distance of d base pairs (bp). The vector C^d is the mth super-diagonal of the Hi-C contact matrix with $m = \frac{d}{binsize}$ which contains Hi-C scores between all bin pairs that have the same genomic distance d. Ranks are adjusted for ties by using the average rank whenever values in C^d are tied.

Note that by increasing the genomic distance, the length of C^d decreases. Therefore, $C^d s$ are extended to have an equal length L. The extension is done by adding elements from n neighboring super-diagonals around mth super-diagonal to reach the constant length L. As we move further from the main diagonal, the number of elements on the mth super-diagonal becomes very small. Therefore, a substantial number of elements from neighboring super-diagonals are included. This is acceptable since the distributions of C^d are more similar for large d, and can thus be pooled. We set L equal for all chromosomes to determine a genome-wide threshold of strong Hi-C scores between gene-loci. So, the normalized Hi-C scores (i.e. ranks) are set to be in the same range across all chromosomes. We set L to be equal to twice the number of bins on Chromosome 1, the largest chromosome in the mouse genome.

SCALE-AWARE TOPOLOGICAL MEASURES

STMs were acquired by calculating the five topological measures described in Table 7.1 on a diffused network, across a range of scales (β). We empirically choose 10 values for beta in range of [0,10] according to:

$$\beta = \frac{2^{6b}-1}{2^6-1} \times (10-0.0001) + 0.0001 \tag{7.1}$$

with $b = 0.0, ..., 1.0$ in 10 steps resulting $\beta : [0.0001, 0.09, 0.24, 0.47, 0.8, 1.4, 2.3, 3.8, 6.2, 10]$. As a result, for the scale-aware classification, 80 features (8 measures × 10 scales) were extracted from the chromatin interaction network.

SPATIALLY-MAPPED GENE EXPRESSION DATA

We downloaded all the expression energy volumes of the 4,345 genes with coronal experiments from (`http://mouse.brain-map.org/`) [24], using the ABA Application Programming Interface (API). Expression energy is a measurement combining the expression level, defined as the integrated amount of signal within each voxel, and the expression density, defined as the amount of expressing cells within each voxel. We selected all voxels belonging to the cortex, defined as *Isocortex* in the ABA, and all the RefSeq genes, resulting in an expression matrix of 15,410 rows (voxels) and 4,230 columns (genes). We used Spearman's Rank correlation as a measure of similarity between the spatial expression profiles of each pair of genes, resulting in a 4,230 × 4,230 spatial co-expression matrix. Gene entries from the spatial co-expression matrix were mapped to their genomic locations to determine the Hi-C contact frequency between gene-pairs based on the mouse reference genome (mm9: NCBI m37, $GCA_0$00001635.18).

We considered a gene-pair to be strongly co-expressed (i.e. positive label) if their correlation exceeds the 90^{th}-percentile of all correlations across all chromosomes. Conversely, gene-pairs are considered to be without strong co-expression (i.e. negative label) when their correlation falls below the median of all correlations across all chromosomes.

SUPERVISED LEARNING PROCEDURE

We used a random neural network (RNN) classifier from the PRTools toolbox [408] (Matlab 2012b) to predict the co-expression label of gene pairs using the topological measures of the link connecting them in the CIN as features. RNN is a feed-forward neural network with one hidden layer. We set the number of hidden nodes to 800, the maximum number of input features (8 STMs at 10 scales applied to 10 CINs; 5 different resolutions and two mapping methods).

The performance of the classifier was determined using 10-fold cross validation and reported in terms of the area under the ROC (receiver operating characteristic) curve (AUC). The ROC curve represents the true positive rate (sensitivity) as a function of the false positive rate (1 - specificity) for different discrimination thresholds of the classifier (S17_Fig). An AUC of 1 represents a perfect classification and 0.5 represent a random classification.

T-SNE MAP

t-Distributed Stochastic Neighbor Embedding (t-SNE) [32, 207] was used to map the links of each chromosome's CIN to a 2D space by reducing the dimensionality of the $N \times M$ data, where N is the number of gene-pairs in each chromosome and M is the number of topological features. In the resulting map, each Hi-C link is represented by a point in the 2D space where the distance between points reflect the similarity between their corresponding topological profiles. We applied t-SNE with perplexity of 30 and initial dimensionality reduction using 50 principal components.

7.5. SUPPLEMENTARY MATERIAL

The online version of this article contains supplementary material[1].

[1] http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004221#sec022

Chapter 8
Discussion

ATLASES of spatial and temporal gene expression in the mammalian brain are essential to understand how genetic variation affects structural and functional organization of the brain. Nevertheless, the growing number of genetic variants implicated in neurobiological processes and the high dimensionality of these atlases poses several computational challenges. This thesis described a set of computational methods developed to address some of these challenges. In the following sections, we discuss our conclusions and potential extensions to this work. We briefly discuss our perspective on the future of brain transcriptomes in terms of the availability of additional data sources and the need for data integration methods to enhance our understanding of the brain.

8.1. DIMENSIONALITY REDUCTION

In Chapter 3, we have analyzed the relationship between gene expression and neuroanatomy using nonlinear dimensionality reduction. We have shown that t-SNE provides a better representation of both local and global relationships between anatomical regions in the mouse and human brain transcriptomes, compared to PCA and classical MDS. The resulting low dimensional embeddings were consistent between the sagittal and coronal mouse brain atlases and across the six human brains. We showed that such low dimensional maps can be used to assess the contribution of cell-type markers towards the structural organization of the brain.

We can use the same approach to analyze similarities between the spatial and temporal expression of genes by creating a low dimensional gene map using t-SNE. A simultaneous view of both maps (gene and sample maps) can provide valuable information about the spatial or temporal localization of groups of co-expressed genes. This can enhance our understanding of the genetic influence on brain connectivity. An interactive platform which allows visualization of the dual t-SNE map with real-time data brushing and map recalculation will be crucial. Such a platform will allow online hypothesis testing on the contribution of genes to the anatomical organization of the brain, as well as identifying regions driving co-expression relationships between genes. However, this requires a computationally efficient implementation of t-SNE that allows online mapping of large amounts of data.

One of the greatest challenges is how to integrate multiple types of data to study the relationships between samples representing different brain regions and time points. For example, for a subset of the samples in the BrainSpan Atlas of the developing human brain, gene expression, methylation and microRNA measurements are available. 2D representations generated using dimensionality reduction methods can be used to explore similarities between samples based on each of these data types separately. However, integrating different types of genomic data as features before dimensionality reduction can be challenging and methods that jointly model such data are highly needed.

8.2. SPATIAL AND TEMPORAL CO-EXPRESSION

In this thesis, we used co-expression analysis to identify common biological processes between groups of genes. Using the BrainSpan Atlas and the Allen Human Brain Atlas we analyzed the spatial and temporal expression of the *DMD* gene and its isoforms in the human brain (Chapter 4). This analysis provides the first comprehensive overview of the

DMD expression in the human brain in contrast to earlier studies including small sample sizes (i.e. few subjects, few anatomical regions, and few time points). Our analysis of the co-expression relationships of the dystrophin isoforms indicates a strong association with genes implicated in neurodevelopmental disorders such as autism and intellectual disability. These results might explain the high incidence of learning and behavioural problems in DMD patients.

Similarly, we employed co-expression analysis to identify signaling pathways of steroid receptors of nuclear steroid receptors in different regions of the mouse brain (Chapter 6). We showed that known targets and co-regulators of steroid receptors are highly co-expressed with the receptor in brain regions where they were identified. For example, genes that are sensitive to sex steroids (such as estrogen) are strongly co-expressed with the estrogen receptor alpha gene in the hypothalamus, the brain region responsible for sexual behavior in animals. In addition, we also observed strong co-expression of genes with steroid receptors outside their known sites of action. This unanticipated co-expression may extend our understanding of the coordinated steroid response of the brain.

In order to build a predictor of brain region-specific targets of steroid receptor, or transcription factors in general, additional features, other than co-expression, can be incorporated. Our experiments showed that the effect size (i.e. the absolute expression of a gene) is a good indicator of responsiveness. In addition, the presence of (conserved) binding motif in the promoter region could be a useful additional predictor [370, 409]. Epigenetic features such as methylation and histone modifications from the ENCODE project can also be incorporated. However, the ENCODE data does not cover all neuronal cell-types and brain regions and the computationally-predicted binding motifs are not cell-type specific.

8.3. CO-EXPRESSION NETWORKS OF DISEASE-RELATED GENES

In Chapters 4 and 6, we were interested in the co-expression of a certain gene of interest (*DMD* or nuclear steroid receptors) with other genes, based on the spatial and temporal expression patterns, in order to identify (neural) functional relationships. Using the BrainSpan Atlas of gene expression across different brain regions and developmental stages of the normal human brain, we studied the co-expression relationships of genes associated with autism spectrum disorders (Chapter 5). A co-expression network analysis of autism-related genes identified three groups with distinct co-expression profiles across development and distinct functional enrichment. Moreover, we constructed a genome-wide co-expression network of the developing human brain transcriptome and we found that autism-related genes were enriched in modules related to mitochondrial function, protein translation, and ubiquitination.

In comparison to methods that address one gene at a time, such as differential gene expression analysis, networks can model the relationships of each gene in the context of its molecular system. By modeling these relationships, network analysis usually results in more coherent sets of genes which facilitates biological interpretation of the results. In addition to similarity of molecular profiles, networks can be used to model other

types of interactions between genes; i.e. edges in a network can represent physical interactions (e.g. protein-protein interaction networks) or computational predictions (e.g. motif enrichment analysis to infer transcription factor binding). Protein-protein interaction (PPI) data is not complete [410], biased towards published literature [411], and not tissue- and species-specific. On the other hand, gene co-expression networks may contain spurious gene–gene correlations due to small study sizes and low signal-to-noise ratio [89]. A major advantage of networks is that they facilitate the integration of multiple data sources as well as information on the functional relationships between genes. A simultaneous integration of PPIs and gene co-expression based on the BrainSpan Atlas identified modules containing functionally related genes enriched for deleterious mutations in ASD and ID [29]. These results emphasize the power of using networks to integrate data. Despite the significant role of networks to improve our understanding of the genetic mechanisms underlying neuropsychiatric and neurodevelopmental disorders [11], the increasing availability of public resources of molecular profiling data requires efficient methods to model complex interactions at multiple molecular levels [28]. There is need for network methods to model multiple types of interactions between genes in order to provide a global overview of the different molecular changes (transcriptomic, epigenomic, and proteomic) associated with neurological processes.

While networks provide an attractive approach to identify common molecular mechanisms between the hundreds of genetic loci implicated in neurological disorders, there is great interest in identifying disease risk genes. Exome- and whole-genome sequencing studies of autism families showed that de novo loss-of-function (LoF) mutations occurred twofold more often in children with ASD compared to their unaffected siblings [51]. Genes harboring recurrent de novo LoF mutations in multiple independent samples have been implicated in ASD risk. Several methods have been developed to prioritize mutations based on their prior probabilities of conferring risk of disease [412] and their deleteriousness (CADD [413]). Despite the power of these methods to prioritize disease risk genes, they don't incorporate relationships between genes which are efficiently encoded in networks. The Detecting Association With Networks (DAWN) method [414] combines TADA scores [415] and co-expression networks and uses a hidden Markov random field to identify ASD risk genes based on TADA scores. Developing more methods to incorporate disease risk scores and network information will be crucial. Machine learning methods provide a promising approach to predict neurological disease risk genes by incorporating different types of features including gene network information.

8.4. CHROMATIN INTERACTIONS AND GENE CO-EXPRESSION

In Chapter 7, we showed that spatial gene co-expression in the mouse cortex can be predicted from long-range chromatin interactions based on Hi-C data. We showed the usefulness of encoding chromatin interactions as a network and using topological measurements to describe it. Moreover, our results illustrate the power of using a multi-scale, multi-resolution scheme to capture different ranges of chromatin interactions; i.e. from direct interaction between genes (i.e. small-scale) to chromatin compartment interactions (i.e. large-scale).

Dozens of mutations in chromatin modifiers, such as *CHD8*, have been implicated in ASD [52, 407]. ASD risk genes converge into co-expression modules through brain

development [29, 133–135]. These observations yield the study of chromatin interactions and gene co-expression in the cortex crucial to our understanding of how ASD variants affects transcription early during development. For instance, we can identify co-expression relationships which are uniquely predictable from chromatin interactions in the cortex compared to other brain regions. This can enrich our understanding of how genomic variants targeting such a basic cellular function can result in a complex neurological disorder such as ASD by altering transcription in the cortex in contrast to other brain regions.

In our work, we have used chromatin interactions to predict gene co-expression based on the hypothesis that the 3D chromatin structure of the genome has a functional regulatory role. Alternatively, we can test the hypothesis whether we can predict chromatin interactions from gene co-expression. If possible, this can give new insight into the role of transcriptional dysregulation in neurological disorders. In addition, acquiring gene expression data is much easier than Hi-C data. A good predictor of chromatin interactions from gene expression data can provide a simpler computational overview of the 3D structure of the genome. Such a model can be used to guide targeted measurement of chromatin interactions (e.g. 4C) of specific genomic regions based on their effect on transcription regulation.

8.5. PERSPECTIVE ON THE FUTURE OF BRAIN TRANSCRIPTOMES

In Chapter 2, we have discussed our perspective on how to enrich our understanding of the brain by means of computational analysis of brain transcriptome atlases. Here, we provide a brief summarization.

While brain transcriptome atlases provide detailed information about gene expression across brain regions and developmental stages, epigenomic and proteomic measurements from the brain can provide invaluable information about regulatory and translational alternations. There is an increasing availability of epigenetic data from different neuronal cell types, brain regions, and time points and from large consortia (ENCODE project [171], Roadmap Epigenomics Mapping Consortium [75], and the ongoing efforts of PsychENCODE consortium [178]). Similarly, imaging mass spectroscopy allows capturing the spatial distribution of large biomolecules, such as proteins, to study the chemical organization of the brain. In addition, single-cell sequencing is a rapidly developing field allowing genetic, epigenetic, and transcriptional measurements from homogenous cell populations. Creative computational methods are desperately needed to integrate data across these different molecular levels [28, 416]. In addition, methods that can integrate cell-type-specific data with tissue-specific data can provide an insight into the dynamics of cell populations. Finally, computational approaches which can integrate tissue- and cell-type-specific gene expression data with data that is not tissue- or species-specific but yet valuable, such as PPIs, will be crucial to benefit from the large body of data generated pre single-cell era.

Imaging is important to diagnose and trace neurological disease progression in a highly inaccessible organ like the human brain. On the other hand, we have discussed in

details the impact of studying the molecular mechanisms underlying neurological disorders to help develop disease markers and identify treatment strategies. Combined imaging and genomic screening can elucidate genetic variations that influence brain structure, function and circuitry. However, the large number of statistical tests performed to detect associations pose a great challenge for small-sized studies and variants with small effect. Information about where and when a gene is expressed in the brain can be leveraged to prioritize genes and variants for testing, allowing imaging-genetic studies of smaller cohorts. For instance, integrating this information as a prior in existing graphical models of imaging and genetic data [190, 417, 418] can be beneficial.

The spatial and temporal resolution of the current brain transcriptomes poses several challenges to methods aiming to integrate these atlases with imaging and in-house gene expression measurements. For instance, the spatial resolution of the human brain transcriptome (~ 1000 samples per brain) is very low in contrast to imaging-based data (e.g. MRI), yielding data integration very challenging. On the other hand, the mouse brain transcriptome has a very high resolution since the ISH images can have a near-cellular resolution (~ $1\mu m$). However, the genome-wide data registered to the common 3D space offers a much lower resolution (~ $200\mu m$) in order to avoid the high computational cost associated with high-resolution genome- and brain-wide data. Hybrid methods which can analyze brain transcriptomes at multiple resolutions can be crucial. Such methods can employ a discovery approach to analyze the low resolution data and subsequently a more rigorous analysis of the high resolution data after converging to a specific brain region, developmental stage, or a small set of genes.

8.6. CONCLUDING REMARKS

In this thesis we have described several computational approaches to analyze brain transcriptome atlases in order to understand the genetic etiology of brain organization. Several methods have been used, including dimensionality reduction and gene co-expression networks, to associate groups of genes that share a common function to a specific anatomical region or developmental stage. Multi-scale network analysis has been used to integrate gene expression and the 3D chromatin structure of the genome in the mouse cortex. These methods have enriched our understanding of the underlying genetic etiology of DMD and ASD, but are generally applicable to other neurological disorders. Furthermore, we have shown that a genome-wide analysis of the co-expression of steroid receptors in the brain can be used to identify region-specific targets and co-regulators, which can be very valuable for selective drug targeting. This work illustrates the value of brain transcriptome atlases as well as the complex structure of the data. The increasing availability of data covering multiple molecular levels (transcriptomic, epigenomic, proteomic) as well as multiple organizational levels of the brain (single-cell, cell types, circuits, and networks) yields computational methods that can handle multi-scale data very crucial.

Over the past decade, brain transcriptome atlases have greatly facilitated our understanding of the functional elements in the brain. At the same time, the number of genetic variants implicated in neurological disorders as well as the amount of in vivo brain imaging data is rapidly increasing. These developments have been widely driven by significant advances in gene expression measurement, sequencing and imaging technologies

allowing high-throughput genomic and imaging measurements from large cohorts. The current challenge lies in interconnecting these various data sources for a better understanding of how cell-type specific mechanisms spanning several molecular levels contribute to different levels of cellular organization in the brain. A better understanding of brain region- and function- specific genetic mechanisms, can facilitate drug targets identification in a spatial and temporal specific manner.

BIBLIOGRAPHY

[1] J. P. Allen. *The Art of Medicine in Ancient Egypt*. Metropolitan Museum of Art Series. Metropolitan Museum of Art, New York (2005).

[2] G. E. Adelman. *Encyclopedia of Neuroscience*. Birkhauser Verlag AG (1987).

[3] L. M. Zimmerman, I. Veith. *Great Ideas in the History of Surgery*. Norman Surgery, No 7. Norman Pub. (1993).

[4] M. F. Bear, B. W. Connors, M. A. Paradiso. *Neuroscience*. Neuroscience: Exploring the Brain. Lippincott Williams & Wilkins (2007).

[5] S. Levy, D. Mandell, R. Schultz. Autism. *Lancet* 374(9701):1627–1638 (2009).

[6] W. M. Cowan, D. H. Harter, E. R. Kandel. The emergence of modern neuroscience: some implications for neurology and psychiatry. *Annual review of neuroscience* 23:343–391 (2000).

[7] E. Kandel, J. Schwartz, T. Jessell. *Principles of Neural Science, Fourth Edition*. McGraw-Hill Companies,Incorporated (2000).

[8] The Nobel Prize in Physiology or Medicine 1906.

[9] S. M. Sunkin. Towards the integration of spatially and temporally resolved murine gene expression databases. *Trends in genetics : TIG* 22(4):211–7 (2006).

[10] R. R. Kitchen, J. S. Rozowsky, M. B. Gerstein, A. C. Nairn. Decoding neuroproteomics: integrating the genome, translatome and functional anatomy. *Nature Neuroscience* 17(11):1491–1499 (2014).

[11] N. N. Parikshak, M. J. Gandal, D. H. Geschwind. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics* (July) (2015).

[12] P. a. Bandettini. Twenty years of functional MRI: The science and the stories. *NeuroImage* 62(2):575–588 (2012).

[13] S. Jbabdi, S. N. Sotiropoulos, S. N. Haber, D. C. V. Essen, T. E. Behrens. Measuring macroscopic brain connections in vivo. *Nature Neuroscience* 18(11):1546–1555 (2015).

[14] R. A. Poldrack, M. J. Farah. Progress and challenges in probing the human brain. *Nature* 526(7573):371–379 (2015).

[15] M. Wilhelm, et al. Mass-spectrometry-based draft of the human proteome. *Nature* 509(7502):582–587 (2014).

[16] M.-S. Kim, et al. A draft map of the human proteome. *Nature* 509(7502):575–81 (2014).

[17] K. Sharma, et al. Cell type– and brain region–resolved mouse brain proteome. *Nature Neuroscience* 18(12) (2015).

[18] D. P. Hibar, et al. Common genetic variants influence human subcortical brain structures. *Nature* 8(7546):224–229 (2015).

[19] M. A. Lodato, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350(6256):94 – 98 (2015).

[20] S. Gong, et al. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425(6961):917–925 (2003).

[21] N. Heintz. Gene expression nervous system atlas (GENSAT). *Nature neuroscience* 7(5):483 (2004).

[22] L. Richardson, et al. EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Research* 42(Database issue):D703–9 (2014).

[23] J. D. Pollock, D. Y. Wu, J. S. Satterlee. Molecular neuroanatomy: A generation of progress. *Trends in Neurosciences* 37(2):106–123 (2014).

[24] E. S. Lein, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124):168–76 (2007).

[25] C. L. Thompson, et al. A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron* 83(2):1–15 (2014).

[26] M. J. Hawrylycz, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489(7416):391–9 (2012).

[27] J. a. Miller, et al. Transcriptional landscape of the prenatal human brain. *Nature* 508(7495):199–206 (2014).

[28] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, D. Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16(2):85–97 (2015).

[29] F. Hormozdiari, O. Penn, E. Borenstein, E. E. Eichler. The discovery of integrated gene networks for autism and related disorders. *Genome research* 142–154 (2015).

[30] M. Hulsman, C. Dimitrakopoulos, J. De Ridder. Scale-space measures for graph topology link protein network architecture to function. *Bioinformatics* 30:237–245 (2014).

[31] S. Babaei, et al. Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex. *PLOS Computational Biology* 11(5):e1004221 (2015).

[32] L. V. D. Maaten, G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605 (2008).

[33] D. C. Van Essen, K. Ugurbil. The future of the human connectome (2012).

[34] M. I. Ekstrand, et al. Molecular profiling of neurons based on connectivity. *Cell* 157(5):1230–1242 (2014).

[35] J. G. Bernstein, P. a. Garrity, E. S. Boyden. Optogenetics and thermogenetics: Technologies for controlling the activity of targeted cells within intact neural circuits. *Current Opinion in Neurobiology* 22(1):61–71 (2012).

[36] J. D. Cahoy, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28(1):264–78 (2008).

[37] Y. Zhang, et al. An RNA-Sequencing Transcriptome and Splicing Database of Glia , Neurons , and Vascular Cells of the Cerebral Cortex. *The Journal of neuroscience* 34(36):1–19 (2014).

[38] J. Khattra, et al. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Research* 17(1):108–116 (2007).

[39] S. Darmanis, et al. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* 112(23):201507125 (2015).

[40] S. K. Kim, et al. A gene expression map for Caenorhabditis elegans. *Science (New York, NY)* 293(5537):2087–2092 (2001).

[41] W. C. Spencer, et al. A spatial and temporal map of C. elegans gene expression. *Genome Research21* 21(2):325–341 (2011).

[42] N. Milyaev, et al. The virtual fly brain browser and query interface. *Bioinformatics* 28(3):411–415 (2012).

[43] A. Visel, C. Thaller, G. Eichele. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic acids research* 32(Database issue):D552–D556 (2004).

[44] G. Diez-Roux, et al. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS biology* 9(1):e1000582 (2011).

[45] The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 45(6):580–5 (2013).

[46] M. C. Oldham, et al. Functional organization of the transcriptome in human brain. *Nature neuroscience* 11(11):1271–1282 (2008).

[47] H. J. Kang, et al. Spatio-temporal transcriptome of the human brain. *Nature* 478(7370):483–9 (2011).

[48] C. Colantuoni, et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* 478(7370):519–23 (2011).

[49] I. Voineagu, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474(7351):380–4 (2011).

[50] S. M. Sunkin, et al. Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research* 41(D1) (2013).

[51] N. Krumm, B. J. O'Roak, J. Shendure, E. E. Eichler. A de novo convergence of autism genetics and molecular neuroscience. *Trends in neurosciences* 37(2):95–105 (2014).

[52] S. De Rubeis, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 3 (2014).

[53] S. E. Medland, N. Jahanshad, B. M. Neale, P. M. Thompson. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nature neuroscience* 17(6):791–800 (2014).

[54] L. French, P. Pavlidis. Informatics in neuroscience. *Briefings in bioinformatics* 8(6):446–56 (2007).

[55] A. R. Jones, C. C. Overly, S. M. Sunkin. The Allen Brain Atlas: 5 years and beyond. *Nature reviews Neuroscience* 10(11):821–828 (2009).

[56] G. a. Pavlopoulos, et al. Visualizing genome and systems biology : technologies , tools , implementation techniques and trends , past , present and future. *GigaScience* (2015).

[57] C. Lau, et al. Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC bioinformatics* 9:153 (2008).

[58] L. French, T. Paus. A FreeSurfer view of the cortical transcriptome generated from the Allen Human Brain Atlas. *Frontiers in neuroscience* 9(September):1–5 (2015).

[59] L. Ng, et al. Surface-based mapping of gene expression and probabilistic expression maps in the mouse cortex. *Methods (San Diego, Calif)* 50(2):55–62 (2010).

[60] P. K. Olszewski, J. Cedernaes, F. Olsson, A. S. Levine, H. B. Schiöth. Analysis of the network of feeding neuroregulators using the Allen Brain Atlas. *Neuroscience and Biobehavioral Reviews* 32(5):945–956 (2008).

[61] P. Mignogna, D. Viggiano. Brain distribution of genes related to changes in locomotor activity. *Physiology & behavior* 99(5):618–626 (2010).

[62] K. N. Alavian, H. H. Simon. Linkage of cDNA expression profiles of mesencephalic dopaminergic neurons to a genome-wide in situ hybridization database. *Molecular neurodegeneration* 4:6 (2009).

[63] A. Björklund, S. B. Dunnett. Dopamine neuron systems in the brain: an update. *Trends in Neurosciences* 30(5):194–202 (2007).

[64] K. C. Kondapalli, H. Prasad, R. Rao. An inside job: how endosomal Na(+)/H(+) exchangers link to autism and neurological disease. *Frontiers in cellular neuroscience* 8(June):172 (2014).

[65] A. Zaldivar, J. L. Krichmar. Interactions between the neuromodulatory systems and the amygdala: exploratory survey using the Allen Mouse Brain Atlas. *Brain structure & function* 218(6):1513–30 (2013).

[66] E. Ben-David, S. Shifman. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Molecular psychiatry* 1–2 (2012).

[67] A. Dahlin, J. Royall, J. G. Hohmann, J. Wang. Expression profiling of the solute carrier gene family in the mouse brain. *The Journal of pharmacology and experimental therapeutics* 329(2):558–570 (2009).

[68] A. Roth, et al. Potential translational targets revealed by linking mouse grooming behavioral phenotypes to gene expression using public databases. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 40(1):312–325 (2013).

[69] J. Liu, et al. Reconstruction of the Gene Regulatory Network Involved in the Sonic Hedgehog Pathway with a Potential Role in Early Development of the Mouse Brain. *PLoS Computational Biology* 10(10):e1003884 (2014).

[70] M. Ashburner, et al. Gene ontology: tool for the unification of biology. *Nature genetics* 25(1):25–29 (2000).

[71] H. Ogata, et al. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 27(1):27–30 (1999).

[72] D. Croft, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research* 42(D1):D472–D477 (2014).

[73] V. Matys, et al. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucl Acids Res* 34(suppl_1):D108–110 (2006).

[74] E. Portales-Casamar, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* 38(Database):D105–D110 (2010).

[75] R. E. Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–330 (2015).

[76] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell* 115(7):787–798 (2003).

[77] A. Zeisel, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347(6226):1138–1142 (2015).

[78] Online Mendelian Inheritance in Man, OMIM (2015).

[79] J. Pinero, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015(0):bav028–bav028 (2015).

[80] D. Welter, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* 42(D1):D1001–D1006 (2014).

[81] J. Simón-Sánchez, A. Singleton. Genome-wide association studies in neurological disorders. *The Lancet Neurology* 7(11):1067–72 (2008).

[82] Exome Variant Server (2015).

[83] R. Li, W. Zhang, S. Ji. Automated identification of cell-type-specific genes in the mouse brain by image computing of expression patterns. *BMC bioinformatics* 15:209 (2014).

[84] L. Kirsch, N. Liscovitch, G. Chechik. Localizing genes to cerebellar layers by classifying ISH images. *PLoS computational biology* 8(12):e1002790 (2012).

[85] F. P. Davis, S. R. Eddy. A tool for identification of genes expressed in patterns of interest using the Allen Brain Atlas. *Bioinformatics* 25(13):1647–1654 (2009).

[86] D. M. Kurrasch, et al. The neonatal ventromedial hypothalamus transcriptome reveals novel markers with spatially distinct patterning. *The Journal of neuroscience* 27(50):13624–13634 (2007).

[87] P. M. Loerch, et al. Evolution of the aging brain transcriptome and synaptic regulation. *PloS one* 3(10):e3329 (2008).

[88] H. L. Ramsden, G. Sürmeli, S. G. McDonagh, M. F. Nolan. Laminar and Dorsoventral Molecular Organization of the Medial Entorhinal Cortex Revealed by Large-scale Anatomical Analysis of Gene Expression. *PLOS Computational Biology* 11(1):e1004032 (2015).

[89] J. M. Stuart, E. Segal, D. Koller, S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, NY)* 302(5643):249–55 (2003).

[90] J. B. Tenenbaum, V. de Silva, J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, NY)* 290:2319–2323 (2000).

[91] J. A. Lee, M. Verleysen. Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing* 67(1-4 SUPPL.):29–53 (2005).

[92] A. Saadatpour, S. Lai, G. Guo, G.-C. Yuan. Single-Cell Analysis in Cancer Genomics. *Trends in Genetics* 31(10):576–586 (2015).

[93] M. J. Hawrylycz, et al. Multi-scale correlation structure of gene expression in the brain. *Neural networks : the official journal of the International Neural Network Society* 24(9):933–42 (2011).

[94] X. Xu, et al. Modular genetic control of sexually dimorphic behaviors. *Cell* 148(3):596–607 (2012).

[95] L. French, P. P. C. Tan, P. Pavlidis. Large-Scale Analysis of Gene Expression and Connectivity in the Rodent Brain: Insights through Data Integration. *Frontiers in neuroinformatics* 5(July):12 (2011).

[96] P. P. C. Tan, L. French, P. Pavlidis. Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. *Frontiers in Neuroscience* 7(5):1–12 (2013).

[97] N. Liscovitch, L. French. Differential Co-Expression between α-Synuclein and IFN-γ Signaling Genes across Development and in Parkinson's Disease. *PloS one* 9(12):e115029 (2014).

[98] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification* (2000).

[99] R. M. Piro, et al. Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR. *Bioinformatics (Oxford, England)* 26(18):i618–24 (2010).

[100] R. M. Piro, I. Molineris, U. Ala, F. Di Cunto. Evaluation of candidate genes from orphan FEB and GEFS+ loci by analysis of human brain gene expression atlases. *PloS one* 6(8):e23149 (2011).

[101] Z. Liu, et al. Study of gene function based on spatial co-expression in a high-resolution mouse brain atlas. *BMC systems biology* 1:19 (2007).

[102] M. Miazaki, L. D. F. Costa. Study of cerebral gene expression densities using Voronoi analysis. *Journal of Neuroscience Methods* 203(1):212–219 (2012).

[103] N. Liscovitch, U. Shalit, G. Chechik. FuncISH: learning a functional representation of neural ISH images. *Bioinformatics (Oxford, England)* 29(13):i36–43 (2013).

[104] B. E. Engelhardt, C. D. Brown. Diving deeper to predict noncoding sequence function. *Nature Methods* 12(10):925–926 (2015).

[105] B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 33(November 2014):1–9 (2015).

[106] H. Y. Xiong, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347(6218):1254806–(2014).

[107] Y. Yang, et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications* 5:1–9 (2014).

[108] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology* 3(140):1–10 (2007).

[109] E. B. van den Akker, et al. Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging. *Aging Cell* 13(2):216–225 (2014).

[110] F. Gofflot, et al. Systematic gene expression mapping clusters nuclear receptors according to their function in the brain. *Cell* 131(2):405–18 (2007).

[111] B. Zhang, S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4:Article17 (2005).

[112] M. C. Oldham, S. Horvath, D. H. Geschwind. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences* 103(47):17973–8 (2006).

[113] J. a. Miller, et al. Conserved molecular signatures of neurogenesis in the hippocampal subgranular zone of rodents and primates. *Development (Cambridge, England)* 140(22):4633–44 (2013).

[114] J. D. J. Allen, Y. Xie, M. Chen, L. Girard, G. Xiao. Comparing statistical methods for constructing large scale gene networks. *PloS one* 7(1):e29348 (2012).

[115] S. Ballouz, W. Verleyen, J. Gillis. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* 31(February):2123–2130 (2015).

[116] L. Song, P. Langfelder, S. Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* 13(Mi):328 (2012).

[117] Y. X. R. Wang, M. S. Waterman, H. Huang. Gene coexpression measures in large heterogeneous samples using count statistics. *Proceedings of the National Academy of Sciences* 111(46):16371–6 (2014).

[118] S. Kumari, et al. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. *PLoS ONE* 7(11):e50411 (2012).

[119] R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*, vol. 47 (2005).

[120] J. Friedman, T. Hastie, R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441 (2008).

[121] S. Dong, et al. De Novo Insertions and Deletions of Predominantly Paternal Origin Are Associated with Autism Spectrum Disorder. *Cell Reports* 9(1):16–23 (2014).

[122] I. Iossifov, et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* 74(2):285–99 (2012).

[123] B. M. Neale, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485(7397):242–5 (2012).

[124] B. J. O'Roak, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485(7397):246–50 (2012).

[125] S. J. Sanders, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237–41 (2012).

[126] M. Fromer, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506(7487):179–84 (2014).

[127] S. Ripke, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511:421–427 (2014).

[128] T. Freilinger, et al. Genome-wide association analysis identifies susceptibility loci for migraine without aura. *Nature genetics* 44(7):777–82 (2012).

[129] K. Bettens, K. Sleegers, C. Van Broeckhoven. Genetic insights in Alzheimer's disease. *The Lancet Neurology* 12(1):92–104 (2013).

[130] B. Zhang, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 153(3):707–720 (2013).

[131] I. Menashe, P. Grange, E. C. Larsen, S. Banerjee-Basu, P. P. Mitra. Co-expression profiling of autism genes in the mouse brain. *PLoS computational biology* 9(7):e1003128 (2013).

[132] S. Gulsuner, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 154(3):518–29 (2013).

[133] N. N. Parikshak, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155(5):1008–21 (2013).

[134] a. J. Willsey, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155(5):997–1007 (2013).

[135] A. Mahfouz, M. N. Ziats, O. M. Rennert, B. P. F. Lelieveldt, M. J. T. Reinders. Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome. *Journal of Molecular Neuroscience* 57(4):580–594 (2015).

[136] C. Gaiteri, Y. Ding, B. French, G. C. Tseng, E. Sibille. 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 (2014).

[137] L. Ng, et al. An anatomic gene expression atlas of the adult mouse brain. *Nature neuroscience* 12(3):356–62 (2009).

[138] H.-W. Dong, L. W. Swanson, L. Chen, M. S. Fanselow, A. W. Toga. Genomic-anatomic evidence for distinct functional domains in hippocampal field CA1. *Proceedings of the National Academy of Sciences* 106(28):11794–9 (2009).

[139] M. J. Hawrylycz, et al. Areal and laminar differentiation in the mouse neocortex using large scale gene expression data. *Methods (San Diego, Calif)* 50(2):113–21 (2010).

[140] J. W. Bohland, et al. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods (San Diego, Calif)* 50(2):105–12 (2010).

[141] Y. Ko, et al. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proceedings of the National Academy of Sciences* 110(8):3095–100 (2013).

[142] P. Grange, et al. Cell-type-based model explaining coexpression patterns of genes in the brain. *Proceedings of the National Academy of Sciences* 111(14):5397–402 (2014).

[143] P. Grange, I. Menashe, M. J. Hawrylycz. Cell-type-specific neuroanatomy of cliques of autism-related genes in the Cell-type-specific neuroanatomy of cliques of autism-related genes in the mouse brain. *Frontiers in Computational Neuroscience* (2015).

[144] A. Kaufman, G. Dror, I. Meilijson, E. Ruppin. Gene expression of Caenorhabditis elegans neurons carries information on their synaptic connectivity. *PLoS computational biology* 2(12):e167 (2006).

[145] V. Varadan, D. M. Miller, D. Anastassiou. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics (Oxford, England)* 22(14):e497–506 (2006).

[146] L. Baruch, S. Itzkovitz, M. Golan-Mashiach, E. Shapiro, E. Segal. Using expression profiles of Caenorhabditis elegans neurons to identify genes that mediate synaptic connectivity. *PLoS computational biology* 4(7):e1000120 (2008).

[147] L. French, P. Pavlidis. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS computational biology* 7(1):e1001049 (2011).

[148] L. Wolf, C. Goldberg, N. Manor, R. Sharan, E. Ruppin. Gene expression in the rodent brain is associated with its regional connectivity. *PLoS computational biology* 7(5):e1002040 (2011).

[149] M. Bota, L. W. Swanson. Collating and Curating Neuroanatomical Nomenclatures: Principles and Use of the Brain Architecture Knowledge Management System (BAMS). *Frontiers in neuroinformatics* 4(March):3 (2010).

[150] S. Ji, A. Fakhry, H. Deng. Integrative analysis of the connectivity and gene expression atlases in the mouse brain. *NeuroImage* 84:245–53 (2014).

[151] A. Fakhry, S. Ji. High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods (San Diego, Calif)* 73C:71–78 (2014).

[152] J. Richiardi, A. Altmann, R. Jonas. Correlated gene expression supports synchronous activity in brain networks. Tech. Rep. 6240 (2015).

[153] M. J. Hawrylycz, et al. Canonical genetic signatures of the adult human brain. *Nature Neuroscience* 18(12) (2015).

[154] A. Mahfouz, M. N. Ziats, O. M. Rennert, B. P. F. Lelieveldt, M. J. T. Reinders. Genomic connectivity networks based on the BrainSpan atlas of the developing human brain. In *SPIE Medical Imaging*, 90344G—-90344G. International Society for Optics and Photonics (2014).

[155] S. Ji. Computational genetic neuroanatomy of the developing mouse brain: dimensionality reduction, visualization, and clustering. *BMC bioinformatics* 14:222 (2013).

[156] A. Mahfouz, et al. Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods* 73:79–89 (2015).

[157] V. Fionda. Biological network analysis and comparison: mining new biological knowledge. *Central European Journal of Computer Science* 1(2):185–193 (2011).

[158] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics* 23(2):e177–e183 (2007).

[159] M. Hayashida, T. Akutsu. Comparing biological networks via graph compression. *BMC systems biology* 4 Suppl 2(Suppl 2):S13 (2010).

[160] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10008(10):6 (2008).

[161] A. Kuhn, D. Thu, H. J. Waldvogel, R. L. M. Faull, R. Luthi-Carter. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature methods* 8(11):945–7 (2011).

[162] E. Shapiro, T. Biezuner, S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews Genetics* 14(9):618–30 (2013).

[163] J.-B. Pettit, et al. Identifying Cell Types from Spatially Referenced Single-Cell Expression Datasets. *PLoS computational biology* 10(9):e1003824 (2014).

[164] D. Grün, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* (2015).

[165] H. Peng, et al. BrainAligner: 3D registration atlases of Drosophila brains. *Nature methods* 8(6):493–500 (2011).

[166] J. Ponjavic, P. L. Oliver, G. Lunter, C. P. Ponting. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS genetics* 5(8):e1000617 (2009).

[167] I. a. Qureshi, M. F. Mehler. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nature Reviews Neuroscience* 13(8):528–541 (2012).

[168] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, J. S. Mattick. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences* 105(2):716–21 (2008).

[169] S. Fertuzinhos, et al. Laminar and temporal expression dynamics of coding and noncoding RNAs in the mouse neocortex. *Cell reports* 6(5):938–50 (2014).

[170] M. N. Ziats, O. M. Rennert. Identification of differentially expressed microRNAs across the developing human brain. *Molecular psychiatry* (May):1–5 (2013).

[171] B. E. Bernstein, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74 (2012).

[172] J. N. Weinstein, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* 45(10):1113–20 (2013).

[173] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker. International network of cancer genome projects. *Nature* 464(7291):993–8 (2010).

[174] J. Shin, G.-l. Ming, H. Song. Decoding neural transcriptomes and epigenomes via high-throughput sequencing. *Nature neuroscience* 17(11):1463–1475 (2014).

[175] I. Maze, et al. Analytical tools and current challenges in the modern era of neuroepigenomics. *Nature neuroscience* 17(11) (2014).

[176] R. S. Illingworth, et al. Inter-individual variability contrasts with regional homogeneity in the human brain DNA methylome. *Nucleic Acids Research* 43(2):732–744 (2015).

[177] M. W. Vermunt, et al. Large-Scale Identification of Coregulated Enhancer Networks in the Adult Human Brain. *Cell Reports* 9(2):767–779 (2014).

[178] S. Akbarian, et al. The PsychENCODE project. *Nature Neuroscience* 18(12):1707–1712 (2015).

[179] T. U. P. Consortium. The UK10K project identifies rare variants in health and disease. *Nature* 526(7571):82–90 (2015).

[180] R. M. Caprioli, T. B. Farmer, J. Gile. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical chemistry* 69(23):4751–4760 (1997).

[181] J. Hanrieder, N. T. Phan, M. E. Kurczy, a. G. Ewing. Imaging mass spectrometry in neuroscience. *ACS Chem Neurosci* 4(5):666–679 (2013).

[182] W. M. Abdelmoula, et al. Automatic registration of imaging mass spectrometry data to the Allen Brain Atlas transcriptome. *Analytical chemistry* 9034:90343M (2014).

[183] R. J. Carreira, et al. Large-Scale Mass Spectrometry Imaging Investigation of Consequences of Cortical Spreading Depression in a Transgenic Mouse Model of Migraine. *Journal of The American Society for Mass Spectrometry* 853–861 (2015).

[184] K. Škrášková, et al. Precise Anatomic Localization of Accumulated Lipids in Mfp2 Deficient Murine Brains Through Automated Registration of SIMS Images to the Allen Brain Atlas. *Journal of The American Society for Mass Spectrometry* 948–957 (2015).

[185] M. Uhlen, et al. Tissue-based map of the human proteome. *Science* 347(6220):1260419–1260419 (2015).

[186] D. P. Hibar, et al. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage* 56(4):1875–91 (2011).

[187] D. P. Hibar, O. Kohannim, J. L. Stein, M.-C. Chiang, P. M. Thompson. Multilocus genetic analysis of brain images. *Frontiers in genetics* 2(October):73 (2011).

[188] H. Wang, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics (Oxford, England)* 28(12):i127–36 (2012).

[189] N. K. Batmanghelich, A. V. Dalca, M. R. Sabuncu, P. Golland. Joint modeling of imaging and genetics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7917 LNCS:766–777 (2013).

[190] J. Yan, et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics (Oxford, England)* 30(17):i564–71 (2014).

[191] G. Rizzo, et al. The predictive power of brain mRNA mappings for in vivo protein density: a positron emission tomography correlation study. *Journal of cerebral blood flow and metabolism* 34(5):827–35 (2014).

[192] I. Zalachoras, R. Houtman, O. C. Meijer. Understanding stress-effects in the brain via transcriptional signal transduction pathways. *Neuroscience* 242:97–109 (2013).

[193] J. Goncalves, S. Madeira. LateBiclustering: Efficient Heuristic Algorithm for Time-Lagged Bicluster Identification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* XX(XX):1–1 (2014).

[194] M. Jagalur, C. Pal, E. Learned-Miller, R. T. Zoeller, D. Kulp. Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC bioinformatics* 8 Suppl 10:S5 (2007).

[195] S. Ji, W. Zhang, R. Li. A probabilistic latent semantic analysis model for coclustering the mouse brain atlas. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 10(6):1460–8 (2013).

[196] A. B. Tchagang, et al. Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm. *BMC Bioinformatics* 13(1):54 (2012).

[197] M. Jung, et al. Longitudinal epigenetic and gene expression profiles analyzed by three-component analysis reveal down-regulation of genes involved in protein translation in human aging. *Nucleic Acids Research* 43(15):1–14 (2015).

[198] W. Kolch, M. Halasz, M. Granovskaya, B. N. Kholodenko. The dynamic control of signal transduction networks in cancer cells. *Nature Reviews Cancer* 15(9):515–527 (2015).

[199] C. Lausted, et al. Systems Approach to Neurodegenerative Disease Biomarker Discovery. *Annual Review of Pharmacology and Toxicology* 54(1):457–481 (2014).

[200] D. Hwang, et al. A systems approach to prion disease. *Mol Syst Biol* 5(252):252 (2009).

[201] L. W. Swanson. *Brain Architecture: Understanding the Basic Plan.* OUP USA (2012).

[202] C. K. I. Williams. On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning* 46:11–19 (2002).

[203] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research* 11:451–490 (2010).

[204] D. Lunga, S. Prasad, M. M. Crawford, O. Ersoy. Manifold-Learning-Based Feature Extraction for Classification of Hyperspectral Data: A Review of Advances in Manifold Learning. *IEEE Signal Processing Magazine* 31(1):55–66 (2014).

[205] K. Shekhar, P. Brodin, M. M. Davis, A. K. Chakraborty. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proceedings of the National Academy of Sciences* 111(1):202–7 (2014).

[206] J. M. Fonville, et al. Hyperspectral visualization of mass spectrometry imaging data. *Analytical chemistry* 85(3):1415–23 (2013).

[207] L. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* 15:3221–3245 (2014).

[208] Allen Mouse Brain Atlas (2014).

[209] L. Ng, et al. Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 4(3):382–93 (2007).

[210] Allen Human Brain Atlas (2014).

[211] Allen Human Brain Atlas Technical White Paper: Microarray Data Normalization (2013).

[212] T. Cox, M. Cox. *Multidimensional Scaling, Second Edition*, vol. 88 (2000).

[213] G. Sharma. *Digital Color Imaging: Handbook*. The Electrical Engineering and Applied Signal Processing Series. CRC PressINC (2003).

[214] C. J. Holmes, et al. Enhancement of MR images using registration for signal averaging. *Journal of computer assisted tomography* 22(2):324–333 (1998).

[215] A. R. Webb. Statistical Pattern Recognition Statistical Pattern Recognition Second Edition. *Library* 9:0–470 (2002).

[216] J. Barnes, P. Hut. A hierarchical O(N log N) force-calculation algorithm (1986).

[217] L. Greengard, V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics* 73:325–348 (1987).

[218] M. Vladymyrov, M. Carreira-Perpinán, M. Carreira-Perpinan. Linear-time Training of Nonlinear Low-Dimensional Embeddings. In *17th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, 968–977 (2014).

[219] S. Cotton, N. J. Voudouris, K. M. Greenwood. Intelligence and Duchenne muscular dystrophy: full-scale, verbal, and performance intelligence quotients. *Developmental medicine and child neurology* 43(7):497–501 (2001).

[220] C. Billard, et al. Cognitive functions in Duchenne muscular dystrophy: a reappraisal and comparison with spinal muscular atrophy. *Neuromuscul Disord* 2(5-6):371–378 (1992).

[221] C. Dorman, A. D. Hurley, J. D. Avignon. Language and Learning Disorders of older Boys with Duchenne Muscular Dystrophy. *Developmental Medicine & Child Neurology* 30(3):316–327 (1988).

[222] a. Castles, M. Coltheart. Varieties of developmental dyslexia. *Cognition* 47(2):149–180 (1993).

[223] R. Banihani, et al. Cognitive and Neurobehavioral Profile in Boys With Duchenne Muscular Dystrophy. *Journal of Child Neurology* (2015).

[224] M. Pane, et al. Duchenne muscular dystrophy and epilepsy. *Neuromuscular disorders : NMD* 23(4):313–315 (2013).

[225] J. G. M. Hendriksen, J. S. H. Vles. Neuropsychiatric disorders in males with duchenne muscular dystrophy: frequency rate of attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder, and obsessive–compulsive disorder. *Journal of child neurology* 23(5):477–481 (2008).

[226] H. K. Young, et al. Cognitive and psychological profile of males with Becker muscular dystrophy. *Journal of child neurology* 23(2):155–162 (2008).

[227] F. Goodwin, F. Muntoni, V. Dubowitz. Epilepsy in Duchenne and Becker muscular dystrophies. *European journal of paediatric neurology* 1(4):115–119 (1997).

[228] K. N. North, et al. Cognitive dysfunction as the major presenting feature of Becker's muscular dystrophy. *Neurology* 46(2):461–5 (1996).

[229] H. L. Sweeney, E. R. Barton. The dystrophin-associated glycoprotein complex: what parts can you do without? *Proceedings of the National Academy of Sciences* 97(25):13464–13466 (2000).

[230] D. Townsend. Finding the sweet spot: assembly and glycosylation of the dystrophin-associated glycoprotein complex. *Anatomical record (Hoboken, NJ : 2007)* 297(9):1694–705 (2014).

[231] A. Waite, S. C. Brown, D. J. Blake. The dystrophin-glycoprotein complex in brain development and disease. *Trends in Neurosciences* 35(8):487–496 (2012).

[232] U. Nudel, et al. Duchenne muscular dystrophy gene product is not identical in muscle and brain. (1989).

[233] H. G. Lidov, T. J. Byers, S. C. Watkins, L. M. Kunkel. Localization of dystrophin to postsynaptic regions of central nervous system cortical neurons. (1990).

[234] E. Holder, M. Maeda, R. D. Bies. Expression and regulation of the dystrophin Purkinje promoter in human skeletal muscle, heart, and brain. *Human genetics* 97(2):232–9 (1996).

[235] V. N. D'Souza, et al. A novel dystrophin isoform is required for normal retinal electrophysiology. *Human molecular genetics* 4(5):837–842 (1995).

[236] T. J. Byers, H. G. Lidov, L. M. Kunkel. An alternative dystrophin transcript specific to peripheral nerve. *Nat Genet* 4(1):77–81 (1993).

[237] G. Morris, C. Simmons, N. Man. Apo-Dystrophins (DP140 and DP71) and Dystrophin-Splicing Isoforms in Developing Brain. *Biochemical and Biophysical Research Communications* 215(1):361–367 (1995).

[238] D. Lederfein, et al. A 71-kilodalton protein is a major product of the Duchenne muscular dystrophy gene in brain and other non-muscle tissues. *Proceedings of the National Academy of Sciences of the United States of America* 89(12):5346–5350 (1992).

[239] R. C. Austin, P. L. Howard, V. N. D'Souza, H. J. Klamut, P. N. Ray. Cloning and characterization of alternatively spliced isoforms of Dp71. *HumMolGenet* 4(9):1475–1483 (1995).

[240] P. J. Taylor, et al. Dystrophin gene mutation location and the risk of cognitive impairment in Duchenne muscular dystrophy. *PloS one* 5(1):e8803 (2010).

[241] T. Chamova, et al. Association between loss of dp140 and cognitive impairment in duchenne and becker dystrophies. *Balkan journal of medical genetics : BJMG* 16(1):21–30 (2013).

[242] V. Ricotti, et al. Neurodevelopmental, emotional, and behavioural problems in Duchenne muscular dystrophy in relation to underlying dystrophin gene mutations. *Developmental medicine and child neurology* (2015).

[243] M. Pane, et al. Early neurodevelopmental assessment in Duchenne muscular dystrophy. *Neuromuscular Disorders* 23(6):451–455 (2013).

[244] N. Doorenweerd, et al. Reduced cerebral gray matter and altered white matter in boys with Duchenne muscular dystrophy. *Annals of Neurology* 76:403–411 (2014).

[245] S. L. L. Kueh, S. I. Head, J. W. Morley. GABA(A) receptor expression and inhibitory post-synaptic currents in cerebellar Purkinje cells in dystrophin-deficient mdx mice. *Clinical and experimental pharmacology & physiology* 35(2):207–10 (2008).

[246] W. M. Snow, J. E. Anderson, M. Fry. Regional and genotypic differences in intrinsic electrophysiological properties of cerebellar Purkinje neurons from wild-type and dystrophin-deficient mdx mice. *Neurobiology of Learning and Memory* 107:19–31 (2014).

[247] H. G. Lidov, S. Selig, L. M. Kunkel. Dp140: a novel 140 kDa CNS transcript from the dystrophin locus. *Human molecular genetics* 4(3):329–335 (1995).

[248] H. G. W. Lidov, T. J. Byers, L. M. Kunkel. The distribution of dystrophin in the murine central nervous system: An immunocytochemical study. *Neuroscience* 54(1):167–187 (1993).

[249] W. M. Snow, M. Fry, J. E. Anderson. Increased density of dystrophin protein in the lateral versus the vermal mouse cerebellum. *Cellular and Molecular Neurobiology* 33(4):513–520 (2013).

[250] B. Lenhard, A. Sandelin, P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 13(4):233–245 (2012).

[251] M. Lizio, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* 16(1):22 (2015).

[252] A. R. R. Forrest, et al. A promoter-level mammalian expression atlas. *Nature* 507(7493):462–70 (2014).

[253] X. Zhou, et al. The Human Epigenome Browser at Washington University. (2011).

[254] H. G. Lidov. Dystrophin in the nervous system. *Brain pathology* 6(1):63–77 (1996).

[255] D. C. Górecki, et al. Expression of four alternative dystrophin transcripts in brain regions regulated by different promoters. *Human molecular genetics* 1(7):505–10 (1992).

[256] V. Ricotti, et al. Ocular and neurodevelopmental features of Duchenne muscular dystrophy: a signature of dystrophin function in the central nervous system. *European journal of human genetics : EJHG* (2015).

[257] S. E. Cyrulnik, V. J. Hinton. Duchenne muscular dystrophy: A cerebellar disorder? *Neuroscience and Biobehavioral Reviews* 32(3):486–496 (2008).

[258] R. Chaussenot, et al. Cognitive dysfunction in the dystrophin-deficient mouse model of Duchenne muscular dystrophy: A reappraisal from sensory to executive processes. *Neurobiology of learning and memory* 124:111–122 (2015).

[259] C. Vaillend, J. M. Billard, S. Laroche. Impaired long-term spatial and recognition memory and enhanced CA1 hippocampal LTP in the dystrophin-deficient Dmdmdx mouse. *Neurobiology of Disease* 17(1):10–20 (2004).

[260] M. D. Al-Sayed, et al. Mutations in NALCN cause an autosomal-recessive syndrome with severe hypotonia, speech impairment, and cognitive delay. *American journal of human genetics* 93(4):721–6 (2013).

[261] Ç. Köroğlu, M. Seven, A. Tolun. Recessive truncating NALCN mutation in infantile neuroaxonal dystrophy with facial dysmorphism. *Journal of medical genetics* 50(8):515–20 (2013).

[262] K. Bushby, et al. Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management. *The Lancet Neurology* 9(1):77–93 (2010).

[263] S. Mansour, et al. Van Maldergem syndrome: further characterisation and evidence for neuronal migration abnormalities and autosomal recessive inheritance. *European journal of human genetics : EJHG* 20(10):1024–31 (2012).

[264] D. A. Parry, et al. Loss of the metalloprotease ADAM9 leads to cone-rod dystrophy in humans and retinal degeneration in mice. *American journal of human genetics* 84(5):683–691 (2009).

[265] M. Gliem, et al. Sorsby Fundus Dystrophy: Novel Mutations, Novel Phenotypic Characteristics, and Treatment Outcomes. *Investigative Opthalmology & Visual Science* 56(4):2664 (2015).

[266] J. M. Wheway, R. G. Roberts. The dystrophin lymphocyte promoter revisited: 4.5-Megabase intron, or artefact? *Neuromuscular Disorders* 13(1):17–20 (2003).

[267] J. Chen, E. E. Bardes, B. J. Aronow, A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* 37(SUPPL. 2):305–311 (2009).

[268] Y. Benjamini, Y. Hochberg. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57(1):289–300 (1995).

[269] A. Rauch, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: An exome sequencing study. *The Lancet* 380(9854):1674–1682 (2012).

[270] J. de Ligt, et al. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* 367(20):1921–1929 (2012).

[271] S. N. Basu, R. Kollu, S. Banerjee-Basu. AutDB: a gene reference resource for autism research. *Nucleic acids research* 37(Database issue):6 (2009).

[272] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, D.C. (2000).

[273] Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders–Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. Tech. Rep. 3 (2012).

[274] S. Smalley, R. Asarnow, M. Spence. Autism and genetics. A decade of research. *Archives of general psychiatry* 45(10):953–961 (1988).

[275] E. Ritvo, et al. The UCLA-University of Utah epidemiologic survey of autism: recurrence risk estimates and genetic counseling. *The American journal of psychiatry* 146(8):1032–1036 (1989).

[276] S. Steffenburg, et al. A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *Journal of child psychology and psychiatry, and allied disciplines* 30(3):405–416 (1989).

[277] A. Bailey, et al. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological medicine* 25(1):63–77 (1995).

[278] J. Hallmayer, et al. Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of general psychiatry* 68(11):1095–1102 (2011).

[279] D. H. Geschwind, P. Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology* 17(1):103–111 (2007).

[280] J. M. Berg, D. H. Geschwind. Autism genetics: searching for specificity and convergence. *Genome biology* 13(7):247 (2012).

[281] K. Wang, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459(7246):528–533 (2009).

[282] L. Weiss, et al. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 461(7265):802–808 (2009).

[283] R. Anney, et al. A genome-wide scan for common alleles affecting risk for autism. *Human molecular genetics* 19(20):4072–4082 (2010).

[284] J. Sebat, et al. Strong association of de novo copy number mutations with autism. *Science (New York, NY)* 316(5823):445–449 (2007).

[285] P. Szatmari, et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature genetics* 39(3):319–28 (2007).

[286] C. Marshall, et al. Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics* 82(2):477–488 (2008).

[287] D. Pinto, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466(7304):368–372 (2010).

[288] D. Levy, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 70(5):886–897 (2011).

[289] S. Sanders, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70(5):863–885 (2011).

[290] I. Iossifov, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* (2014).

[291] C. Wolfe, I. Kohane, A. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics* 6:227 (2005).

[292] K. Winden, et al. The organization of the transcriptional network in specific neuronal classes. *Molecular systems biology* 5:291 (2009).

[293] E. Ben-David, S. Shifman. Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. *PLoS genetics* 8(3):e1002556 (2012).

[294] T. Sun, et al. Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. *Science (New York, NY)* 308(5729):1794–1798 (2005).

[295] B. Abrahams, et al. Genome-wide analyses of human perisylvian cerebral cortical patterning. *Proceedings of the National Academy of Sciences* 104(45):17849–17854 (2007).

[296] M. Johnson, et al. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 62(4):494–509 (2009).

[297] B. Ip, et al. Investigating gradients of gene expression involved in early human cortical development. *Journal of anatomy* 217(4):300–311 (2010).

[298] M. Somel, et al. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome research* 20(9):1207–1218 (2010).

[299] D. W. Huang, B. Sherman, R. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4(1):44–57 (2009).

[300] D. Lioy, et al. A role for glia in the progression of Rett's syndrome. *Nature* 475(7357):497–500 (2011).

[301] F. Cao, et al. Alteration of astrocytes and Wnt/β-catenin signaling in the frontal cortex of autistic subjects. *Journal of neuroinflammation* 9(1):223 (2012).

[302] A.-L. Barabási, N. Gulbahce, J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews Genetics* 12(1):56–68 (2011).

[303] M. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* 27(3):431–432 (2011).

[304] G. Huguet, E. Ey, T. Bourgeron. The Genetic Landscapes of Autism Spectrum Disorders. *Annu Rev Genomics Hum Genet* (2013).

[305] H. Zoghbi. Postnatal neurodevelopmental disorders: meeting at the synapse? *Science (New York, NY)* 302(5646):826–830 (2003).

[306] R. A. Carper, E. Courchesne. Localized enlargement of the frontal cortex in early autism. *Biol Psychiatry* 57(2):126–133 (2005).

[307] E. Courchesne, K. Pierce. Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Current opinion in neurobiology* 15(2):225–230 (2005).

[308] T. Bourgeron. A synaptic trek to autism. *Current opinion in neurobiology* 19(2):231–234 (2009).

[309] J. Rubenstein, M. Merzenich. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes, brain, and behavior* 2(5):255–267 (2003).

[310] W. Spooren, L. Lindemann, A. Ghosh, L. Santarelli. Synapse dysfunction in autism: a molecular medicine approach to drug discovery in neurodevelopmental disorders. *Trends in pharmacological sciences* 33(12):669–684 (2012).

[311] R. C. Paolicelli, et al. Synaptic pruning by microglia is necessary for normal brain development. *Science (New York, NY)* 333(6048):1456–8 (2011).

[312] A. Sheikh, et al. Cathepsin D and apoptosis related proteins are elevated in the brain of autistic subjects. *Neuroscience* 165(2):363–370 (2010).

[313] I. Maezawa, M. Calafiore, H. Wulff, L.-W. Jin. Does microglial dysfunction play a role in autism and Rett syndrome? *Neuron glia biology* 7(1):85–97 (2011).

[314] R. Kelleher, M. Bear. The autistic neuron: troubled translation? *Cell* 135(3):401–406 (2008).

[315] J. T. Glessner, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459(7246):569–573 (2009).

[316] R. Smith, W. Sadee. Synaptic signaling and aberrant RNA splicing in autism spectrum disorders. *Frontiers in synaptic neuroscience* 3:1 (2011).

[317] A. Piton, et al. Analysis of the effects of rare variants on splicing identifies alterations in GABA(A) receptor genes in autism spectrum disorder individuals. *European journal of human genetics : EJHG* (2012).

[318] M. Neves-Pereira, et al. Deregulation of EIF4E: a novel mechanism for autism. *Journal of medical genetics* 46(11):759–765 (2009).

[319] C. Gkogkas, et al. Autism-related deficits via dysregulated eIF4E-dependent translational control. *Nature* 493(7432):371–377 (2013).

[320] E. Santini, et al. Exaggerated translation causes synaptic and behavioural aberrations associated with autism. *Nature* 493(7432):411–415 (2013).

[321] D. Rossignol, R. Frye. Mitochondrial dysfunction in autism spectrum disorders: a systematic review and meta-analysis. *Molecular psychiatry* 17(3):290–314 (2012).

[322] A. Anitha, et al. Downregulation of the Expression of Mitochondrial Electron Transport Complex Genes in Autism Brains. *Brain Pathol* (2012).

[323] A. Anitha, et al. Brain region-specific altered expression and association of mitochondria-related genes in autism. *Molecular autism* 3(1):12 (2012).

[324] Z.-H. Sheng, Q. Cai. Mitochondrial transport in neurons: impact on synaptic homeostasis and neurodegeneration. *Nature reviews Neuroscience* 13(2):77–93 (2012).

[325] R. Bernier, et al. Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. *Cell* 1–14 (2014).

[326] D. Pinto, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American journal of human genetics* 94(5):677–94 (2014).

[327] C. Lintas, R. Sacco, A. M. Persico. Genome-wide expression studies in Autism spectrum disorder, Rett syndrome, and Down syndrome (2012).

[328] I. Voineagu. Gene expression studies in autism: moving from the genome to the transcriptome and beyond. *Neurobiology of disease* 45(1):69–75 (2012).

[329] M. L. Chow, et al. Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. *PLoS genetics* 8(3):e1002592 (2012).

[330] K. Garbett, et al. Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* 30:303–311 (2008).

[331] A. E. Purcell, O. H. Jeon, A. W. Zimmerman, M. E. Blue, J. Pevsner. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* 57:1618–1628 (2001).

[332] M. Smith, et al. Mitochondrial and ion channel gene alterations in autism. In *Biochimica et Biophysica Acta - Bioenergetics*, vol. 1817, 1796–1802 (2012).

[333] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5(7):621–628 (2008).

[334] G. Matuszek, Z. Talebizadeh. Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. *BMC medical genetics* 10:102 (2009).

[335] L.-M. Xu, et al. AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic acids research* 40(Database issue):22 (2012).

[336] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64(3):479–498 (2002).

[337] V. Stanisić, D. M. Lonard, B. W. O'Malley. Modulation of steroid hormone receptor activity. *Progress in brain research* 181(08):153–176 (2010).

[338] E. R. de Kloet, M. Joëls, F. Holsboer. Stress and the brain: from adaptation to disease. *Nature reviews Neuroscience* 6(6):463–475 (2005).

[339] S. Toffoletto, R. Lanzenberger, M. Gingnell, I. Sundström-Poromaa, E. Comasco. Emotional and cognitive functional imaging of estrogen and progesterone effects in the female human brain: a systematic review. *Psychoneuroendocrinology* 50:28–52 (2014).

[340] S. John, et al. Chromatin accessibility predetermines glucocorticoid receptor binding patterns. *Nature genetics* 43(3):264–268 (2011).

[341] S. A. Krum, et al. Unique ERalpha cistromes control cell type-specific gene regulation. *Molecular endocrinology (Baltimore, Md)* 22(11):2393–2406 (2008).

[342] T. Ravasi, et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell* 140(5):744–752 (2010).

[343] N. a. Datson, et al. The transcriptional response to chronic stress and glucocorticoid receptor blockade in the hippocampal dentate gyrus. *Hippocampus* 22(2):359–71 (2012).

[344] J. M. H. M. Reul, E. R. De Kloet. Two receptor systems for corticosterone in rat brain: Microdistribution and differential occupation. *Endocrinology* 117(6):2505–2511 (1985).

[345] S. E. Pérez, E.-Y. Chen, E. J. Mufson. Distribution of estrogen receptor alpha and beta immunoreactive profiles in the postnatal rat brain. *Brain research Developmental brain research* 145(1):117–139 (2003).

[346] K. N. Nilaweera, et al. G protein-coupled receptor 101 mRNA expression in supraoptic and paraventricular nuclei in rat hypothalamus is altered by pregnancy and lactation. *Brain Research* 1193:76–83 (2008).

[347] S. Kanno, S. Hirano, F. Kayama. Effects of the phytoestrogen coumestrol on RANK-ligand-induced differentiation of osteoclasts. *Toxicology* 203(1-3):211–220 (2004).

[348] E. De Marinis, et al. 17β-Oestradiol anti-inflammatory effects in primary astrocytes require oestrogen receptor β-mediated neuroglobin up-regulation. *Journal of Neuroendocrinology* 25(3):260–270 (2013).

[349] K. A. Baltgalvis, S. M. Greising, G. L. Warren, D. A. Lowe. Estrogen regulates estrogen receptors and antioxidant gene expression in mouse skeletal muscle. *PLoS ONE* 5(4) (2010).

[350] M. T. K. Zia, et al. Postnatal glucocorticoid-induced hypomyelination, gliosis, and neurologic deficits are dose-dependent, preparation-specific, and reversible. *Experimental neurology* 263:200–13 (2015).

[351] N. a. Datson, et al. Previous history of chronic stress changes the transcriptional response to glucocorticoid challenge in the dentate gyrus region of the male rat hippocampus. *Endocrinology* 154(9):3261–3272 (2013).

[352] E. Dias-Ferreira, et al. Chronic stress causes frontostriatal reorganization and affects decision-making. *Science (New York, NY)* 325(5940):621–625 (2009).

[353] S. Lachize, et al. Steroid receptor coactivator-1 is necessary for regulation of corticotropin-releasing hormone by chronic stress and glucocorticoids. *Proceedings of the National Academy of Sciences* 106(19):8038–8042 (2009).

[354] J. C. Nwachukwu, et al. Resveratrol modulates the inflammatory response via an estrogen receptor-signal integration network. *eLife* 2014(3) (2014).

[355] N. Kotaja, S. Aittomäki, O. Silvennoinen, J. J. Palvimo, O. A. Jänne. ARIP3 (androgen receptor-interacting protein 3) and other PIAS (protein inhibitor of activated STAT) proteins differ in their ability to modulate steroid receptor-dependent transcriptional activation. *Molecular endocrinology (Baltimore, Md)* 14(12):1986–2000 (2000).

[356] P. Alen, et al. Interaction of the putative androgen receptor-specific coactivator ARA70/ELE1alpha with multiple steroid receptors and identification of an internally deleted ELE1beta isoform. *Molecular endocrinology (Baltimore, Md)* 13(1):117–128 (1999).

[357] I. Zalachoras, et al. Differential targeting of brain stress circuits with a selective glucocorticoid receptor modulator. *Proceedings of the National Academy of Sciences* 110(19):7910–5 (2013).

[358] M. Niwa, et al. Adolescent stress-induced epigenetic control of dopaminergic neurons via glucocorticoids. *Science (New York, NY)* 339(6117):335–9 (2013).

[359] T. D. Purves-Tyson, et al. Testosterone induces molecular changes in dopamine signaling pathway molecules in the adolescent male rat nigrostriatal pathway. *PLoS ONE* 9(3) (2014).

[360] K. Schauwaers, et al. Loss of androgen receptor binding to selective androgen response elements causes a reproductive phenotype in a knockin mouse model. *Proceedings of the National Academy of Sciences* 104(12):4961–4966 (2007).

[361] J. A. E. Polman, E. R. De Kloet, N. a. Datson. Two populations of glucocorticoid receptor-binding sites in the male rat hippocampal genome. *Endocrinology* 154(5):1832–1844 (2013).

[362] L. M. Shin, I. Liberzon. The neurocircuitry of fear, stress, and anxiety disorders. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 35(1):169–191 (2010).

[363] S. R. Lee, et al. AR and ER interaction with a p21-activated kinase (PAK6). *Molecular endocrinology (Baltimore, Md)* 16(1):85–99 (2002).

[364] T. Nekrasova, M. L. Jobes, J. H. Ting, G. C. Wagner, A. Minden. Targeted disruption of the Pak5 and Pak6 genes in mice leads to deficits in learning and locomotion. *Developmental biology* 322(1):95–108 (2008).

[365] E. Atucha, et al. A mixed glucocorticoid/mineralocorticoid selective modulator with dominant antagonism in the male rat brain. *Endocrinology* (September):en.2015–1390 (2015).

[366] R. E. Mercer, R. Wevrick. Loss of Magel2, a candidate gene for features of Prader- Willi syndrome, impairs reproductive function in mice. *PLoS ONE* 4(1) (2009).

[367] U. Eiholzer, et al. Hypothalamic and gonadal components of hypogonadism in boys with Prader-Labhart-Willi syndrome. *Journal of Clinical Endocrinology and Metabolism* 91(3):892–898 (2006).

[368] M. Sadagurski, X. C. Dong, M. G. Myers, M. F. White. Irs2 and Irs4 synergize in non-LepRb neurons to control energy balance and glucose homeostasis. *Molecular Metabolism* 3(1):55–63 (2014).

[369] A. Frank, L. M. Brown, D. J. Clegg. The role of hypothalamic estrogen receptors in metabolic regulation (2014).

[370] N. a. Datson, et al. Specific regulatory motifs predict glucocorticoid responsiveness of hippocampal gene expression. *Endocrinology* 152(10):3749–57 (2011).

[371] J. T. Robinson, et al. Integrative genomics viewer. *Nat Biotech* 29(1):24–26 (2011).

[372] E. Y. Chen, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* 14(1):128 (2013).

[373] N. A. Datson, et al. A molecular blueprint of gene expression in hippocampal subregions CA1, CA3, and DG is conserved in the brain of the common marmoset. *Hippocampus* 19(8):739–752 (2009).

[374] M. R. Boon, et al. Peripheral cannabinoid 1 receptor blockade activates brown adipose tissue and diminishes dyslipidemia and obesity. *The FASEB Journal* 28(12):5361–5375 (2014).

[375] H. Li, R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25(14):1754–60 (2009).

[376] Y. Zhang, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9(9):R137 (2008).

[377] M. B. Gerstein, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91–100 (2012).

[378] X. Dong, C. Li, Y. Chen, G. Ding, Y. Li. Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC genomics* 11(1):704 (2010).

[379] D. Homouz, A. S. Kudlicki. The 3D Organization of the Yeast Genome Correlates with Co-Expression and Reflects Functional Relations between Genes. *PLoS ONE* 8(1) (2013).

[380] X. Lan, et al. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic acids research* 40(16):7690–704 (2012).

[381] F. Jin, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503(7475):290–4 (2013).

[382] M. Botta, S. Haider, I. X. Y. Leung, P. Lio, J. Mozziconacci. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Molecular systems biology* 6:426 (2010).

[383] D. Rieder, Z. Trajanoski, J. G. McNally. Transcription factories. *Frontiers in Genetics* 3(OCT):1–12 (2012).

[384] Y. Shen, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* 488(7409):116–120 (2012).

[385] S. Schoenfelder, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics* 42(1):53–61 (2010).

[386] A. Papantonis, P. R. Cook. Transcription factories: Genome organization and gene regulation. *Chemical Reviews* 113(11):8683–8705 (2013).

[387] J. Dekker, M. a. Marti-Renom, L. a. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews Genetics* 14(6):390–403 (2013).

[388] E. Lieberman-Aiden, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, NY)* 326(5950):289–93 (2009).

[389] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* 30(1):90–8 (2012).

[390] D. Chen, et al. Dissecting the chromatin interactome of microRNA genes. *Nucleic Acids Research* 42(5):3028–3043 (2014).

[391] K. S. Sandhu, et al. Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks. *Cell Reports* 2(5):1207–1219 (2012).

[392] R. E. Boulos, A. Arneodo, P. Jensen, B. Audit. Revealing Long-Range Interconnected Hubs in Human Chromatin Interaction Data Using Graph Theory. *Physical Review Letters* 111(11):118102 (2013).

[393] W. Winterbach, P. Van Mieghem, M. Reinders, H. Wang, D. de Ridder. Topology of molecular interaction networks. *BMC systems biology* 7(1):90 (2013).

[394] E. Estrada. Generalized walks-based centrality measures for complex biological networks. *Journal of theoretical biology* 263(4):556–65 (2010).

[395] J. I. Fuxman Bass, et al. Using networks to measure similarity between genes: association index selection. *Nature methods* 10(12):1169–76 (2013).

[396] H. W. Ma, X. M. Zhao, Y. J. Yuan, A. P. Zeng. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* 20(12):1870–1876 (2004).

[397] E. Yaffe, A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* 43(11):1059–1065 (2011).

[398] E. de Wit, W. de Laat. A decade of 3C technologies: Insights into nuclear organization. *Genes and Development* 26(1):11–24 (2012).

[399] J. R. Dixon, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–380 (2012).

[400] J. Paulsen, et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic acids research* 41(10):5164–74 (2013).

[401] Z. Duan, et al. A three-dimensional model of the yeast genome. *Nature* 465(7296):363–367 (2010).

[402] S. Babaei, W. Akhtar, J. de Jong, M. Reinders, J. de Ridder. 3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. *Nature Communications* 6:6381 (2015).

[403] J. Lafferty, R. I. Kondor. Diffusion Kernels on Graphs and Other Discrete Input Spaces. *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning* 315—-322 (2002).

[404] S. Babaei, M. Hulsman, M. Reinders, J. de Ridder. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC bioinformatics* 14:29 (2013).

[405] N. D. Heintzman, B. Ren. Finding distal regulatory elements in the human genome. *Current opinion in genetics & development* 19(6):541–549 (2009).

[406] M. Imakaev, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. - Supplement. *Nature methods* 9(10):999–1003 (2012).

[407] J. L. Ronan, W. Wu, G. R. Crabtree. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet* 14(5):347–359 (2013).

[408] R. P. Duin. PRTools (2004).

[409] J. S. Carroll, et al. Genome-wide analysis of estrogen receptor binding sites. *Nature genetics* 38(11):1289–1297 (2006).

[410] G. T. Hart, A. K. Ramani, E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome biology* 7:120 (2006).

[411] L. Hakes, J. W. Pinney, D. L. Robertson, S. C. Lovell. Protein-protein interaction networks and biology–what's the connection? *Nature biotechnology* 26(1):69–72 (2008).

[412] G. M. Cooper, J. Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* 12(9):628–640 (2011).

[413] M. Kircher, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 46(3):310–5 (2014).

[414] L. Liu, J. Lei, S. Sanders, A. Willsey. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism* 5(22):1–18 (2014).

[415] X. He, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics* 9(8):e1003671 (2013).

[416] T. Kling, et al. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research* 43(15):e98–e98 (2015).

[417] H. Wang, et al. From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics (Oxford, England)* 28(18):i619–i625 (2012).

[418] N. K. Batmanghelich, et al. Generative Method to Discover Genetically Driven Image Biomarkers. *information processing in medical imaging* 9123:30–42 (2015).

SUMMARY

Atlases of gene expression across the mammalian brain provide invaluable information to unravel the complex role of genetic variation in the structural and functional organization of the brain. Information about where in the brain and at which developmental stage a certain gene is expressed can help identifying the functional role of that gene in the brain. Nevertheless, neurodevelopmental processes as well as neurological disorders are complex processes governed by hundreds of genes. Given the cellular diversity of the brain, the high spatial resolution of these atlases allows the identification of region-specific interactions between genes by means of co-expression analysis.

Brain transcriptome atlases are, however, multivariate datasets with a high number of dimensions across genes, brain regions, and developmental stages. To derive intuition on the relationships between pairs of variables (either genes, brain regions, or developmental stages), it is beneficial to explore the data in two-dimensional (2D) maps. Co-expression networks that encode the similarity between the spatial and temporal expression patterns of genes can be used to identify groups of genes that signal through similar pathways, share common regulatory elements, or are involved in the same biological process. The brain is hierarchically organized both at the molecular level (transcriptomic, proteomic, and epigenomic) as well as the neurobiological level (cells, circuits, and functional networks). This hierarchical organization requires a multi-scale network analysis in order to integrate data across multiple levels, which is necessary to acquire a full understanding of the mechanisms underlying complex biological processes in the brain.

This thesis introduces data-driven computational algorithms to analyze brain transcriptome atlases to get insight into the role of genetics in the structural and functional organization of the brain. Our methods have enhanced our understanding of the shared pathways among hundreds of genes related to autism and the role of dystrophin in the brain that might explain the high incidence of learning and behavioral problems in Duchene muscular dystrophy. In addition, we have contributed new insights into brain modulation by steroid hormones. These insights can help identify therapeutic targets for selective activation of brain circuits in research and clinical settings.

SAMENVATTING

Gen-expressie-atlassen van de hersenen bevatten waardevolle informatie voor het begrijpen van de rol van genetische variatie in de functie en structuur van het brein. Waar en wanneer (in de ontwikkeling) een gen tot expressie komt geeft informatie over de functie van dat gen in het brein. De ontwikkeling van de hersenen en het ontstaan van hersenziektes zijn echter zeer complexe processen, die worden gestuurd door honderden genen tegelijkertijd. Vanwege de cellulaire diversiteit van het brein, is een hoge spatiële resolutie van deze atlassen noodzakelijk om inzicht te geven in regio-specifieke interacties tussen genen.

Deze gen-expressie-atlassen zijn echter complexe datasets met vele variabelen en een groot aantal dimensies in zowel genen, hersenregio's, als ontwikkelingsstadia. Om intuïtief begrip te krijgen van de relaties tussen sets van variabelen (genen, regio's of stadia), kan het helpen om de data te verkennen in tweedimensionale diagrammen. Co-expressienetwerken beschrijven daarnaast de gelijkenis tussen genen met betrekking tot ruimtelijke en temporele patronen van expressie, en kunnen gebruikt worden om groepen genen te identificeren met gedeelde biologische functies of regulerende elementen. De hersenen zijn hiërarchisch georganiseerd, zowel op moleculair niveau (van expressie, eiwitten en epigenetica) als op neurobiologisch niveau (cellen en functionele netwerken). Deze hiërarchische organisatie vereist een aanpak waarbij data op meerdere schalen wordt bekeken. Op deze wijze kunnen we inzicht krijgen in de mechanismen die ten grondslag liggen aan de complexe biologische processen in het brein.

Dit proefschrift introduceert data-gedreven algoritmen voor het analyseren van gen-expressie-atlassen van het brein, om de rol van genetica in de structurele en functionele organisatie van de hersenen te doorgronden. Onze methoden hebben geholpen bij het begrijpen van de gedeelde *pathways* onder de honderden genen die betrokken zijn bij autisme, en bij het begrijpen van de rol van dystrofine in de leer- en gedragsproblemen die gepaard gaan met de ziekte van Duchenne. Ten slotte hebben we nieuwe inzichten geboden in de effecten van steroïdhormonen op de hersenen. De lessen die wij geleerd hebben kunnen van pas komen bij de zoektocht naar nieuwe medicatie voor het activeren van hersennetwerken, zowel in wetenschappelijk onderzoek als de medische praktijk.

ACKNOWLEDGEMENTS

In the past few years I have had the opportunity to fulfill my enthusiasm for knowledge and deepen my understanding of scientific research through pursuing my PhD. But this was not all that my PhD years were about. Along the way I had the privilege of traveling to spectacular places around the world and meeting fascinating people. The diverse cultures and personalities I encountered have changed the way I perceive the world. I wish to personally thank everyone who has been part of this journey. However there are some people that I would like to mention explicitly for their inspiration and contributions to my knowledge and their help in creating this book.

First and foremost I must thank my promoters and daily supervisors Marcel Reinders and Boudewijn Lelieveldt. Thank you both for your inspiration, guidance, support and persistence. I truly appreciate the time you provided for discussions and lengthy reviews of my articles even during late night hours and holidays. I am specially grateful to both of you for the priceless freedom throughout my work, which is reflected in the wide range of topics covered in this book. Marcel, your challenging discussions and encouragement keep pushing me to always go one step further. Boudewijn, your endless enthusiasm (specially about any work related to tSNE) has always helped me keep my motivation.

Emile Hendrix is one of the very important reasons I'm here today. It is your friendliness and enthusiasm that made me decide to take the step of moving to the Netherlands. Thank you for this amazing opportunity. Furthermore, you and Ans are great hosts. I truly enjoyed the 'gezellig' times in Anjum.

Several chapters in this book have been the result of close collaborations I have truly enjoyed. Sepidah, you are more than just a colleague, you are a true friend. I will always remember my time in Seattle when time difference finally worked in my favour and we worked alternately on chapter 7 of this book. The winter school in Okinawa was one of the highlights of my PhD years. There I met Mark Ziats where we started on a week long project that resulted in chapter 5 of this book. Thank you Mark for you hospitality in Bethesda and for the wonderful time in Cold Spring Harbor and New York.

Sjoerd, thank you for the dutch summary! But not just that of course. Throughout this time you were always there, be that by helping with statistics problems or by music concerts invitations. We now also know that if you ever decide to give up on science, you will make a good tax advisor. Nathalie Doorenweerd, little did I know that after working with you as a master student during the first year of my PhD we will end up writing one of my thesis chapters together and becoming close friends. Thanks to you and Sjoerd for being by my side during my defence.

Beyond individual collaborations, I have benefited enormously from spending my time between two institutes which provided great interactive environments.

At TU Delft, I have deep gratitude for all members of the Delft Bioinformatics Lab that I have met through the years for the challenging discussions during lab meetings, discussion groups, and lab retreats. I would like to particularly thank Marc for always

being reachable; Jeroen for all his guidance on teaching and career development; Erdogan for all the cheerfulness, Friday morning coffee and the December project; Thies for all the fun we had hiking the wrong mountain in France; and Jurgen and Bastian for their help with my EMBO application. For everyone else, thank you for making the days in the office unforgettable.

I would also like to extend my gratitude to the amazing colleagues in the PRB group for all the fun moments we had during 'borrels' and social activities. I would like to specially thank Gorkem for being a true friend all along the journey and for the unforgettable trip to Turkey.

I should also mention here the MSP group on the 10th floor of EWI with whom I have shared many unforgettable moments. Thank you all for your help throughout the first year in Delft and all the fun memories we have. I would like to specially express my gratitude to Zeki and Michele, my first office mates for helping me navigate the Dutch ways and settle in the Netherlands. Finally, I cannot move on from TU Delft without extending my gratitude to Saskia Peters for her support from the start even before I arrived in the Netherlands.

The days at LKEB were never boring! I was lucky to start my journey at the Poortgebouw where I have made very good friends. Vikas, I owe you a great deal for all your help to make my decision to move to the Netherlands an easy one. Thanks to you, Shivani and your loving families for making our trip to India unforgettable. Oleh, as humble as you are, you have been a great friend through the years. Nora, thank you for all the fun we had running and bouldering together, it is always very uplifting! To Martin, Peter, Ece, and Artem, all the fun we had during pancake/movie nights and bowling will forever be in my memory. Rahil and Shahana, thank you for giving me a reason to visit India and for having me at your beautiful wedding. For everyone else at LKEB, the memories are countless. Be it the lunches, 'borrels', BBQ's at Rob's place or the football tournaments, I cherish every moment. A special thank you goes to Marius for all the help with elastix and for the fruitful discussions. Also to Michèle and Anne-Carien for always being reachable and supportive.

Apart from LKEB, I was lucky to have worked with many people around the hospital. Particularly, Louise van der Weerd, our discussions were crucial in shaping my studies during the first year of my PhD. Onno Meijer, I owe you a lot for your enthusiasm and dedication to our work together. For the Lisa's in your group, working with you was always fun and productive. Last but not least, I would like to present my gratitude to Marcel Schaaf for his valuable comments on chapter 6.

Two years ago I had the privilege of spending three months in Evan Eichler's lab at the University of Washington in Seattle. Evan, thank you for the fantastic opportunity. In the time I spent there I met some amazing people and reconnected with old friends. Thank you Fereydoun for your time spent having lengthy discussions on the board. Tonia, you helped make my visit to the lab fun and for that I am grateful. Maike and John, thanks to you I decided to take a climbing course. Michael, I wouldn't have seen and enjoyed the 'Midsummer Night's dream if not for you. Finally, Mahmoud, it was great to catch up with you my friend. Thank you for making my stay in Seattle feel a bit more like home. Also from Seattle, I can't forget to mention Terri Gilbert from the Allen Institute for Brain Sciences for her continuous support.

My Friends in the Netherlands without whom this journey would not have been possible, thank you all for making the Netherlands feel like home. I am specially thankful to Jorge and Hester not just for their unparalleled support but also for the good food and the company. I am lucky to have met both of you.Thanks to Ruba, Hasan, Nur, Osman, Catarina, Iman, Derek, Samira, Rami, Rana, Karim, Donia, Rasha, Nenad for all the lovely gatherings we had along these years. There are so many special memories here that I just can't count them all.

To my friends and family back in Egypt, I owe you tremendous gratitude for your support, encouragement, and understanding. I truly wish you were all here during my defence ceremony.

And of course I cannot miss the chance here to mention my colleagues and friends; Mustafa, Abdallah, Othman, Mohamed and Walid. It has been a long journey since our bachelors but you are all still here and for that I am grateful. As for Hisham, the memories are countless. From our bachelors through the masters and until the trips in Europe. Thank you for sticking around.

I would never be where I am today without my parents. Thank you for making me believe I can go this far and furthermore. The dedication on this thesis is the least I can do to show my love and gratitude. I can never do you justice for all that you have given me. My dear brother Omar, you'll always be my closest of friends. I can always count on you being there, and will forever bug you with my design related questions. The thesis cover wasn't the end of it.

Finally, my beloved little family: Nesma and Reema. Though only my name appears on the cover of this dissertation, Nesma deserves a lot of credit for proof reading every single line I have ever written. I cannot express how thankful I am for your unconditional sacrifices, cheerfulness, kindness, care and love. I couldn't have wished for a better gift for my graduation, our daughter Reema.

Ahmed Mahfouz
Delft, June 2016

CURRICULUM VITÆ

Ahmed Mahfouz was born on January 1st 1987 in Giza, Egypt. In 2008, Ahmed obtained his Bachelor's degree in Systems and Biomedical Engineering from Cairo University, Egypt. Afterwards, Ahmed proceeded with his Master's studies at the School of Communication and Information Technology at Nile University, Egypt. During his time as a master student, Ahmed joined the Medical Imaging and Image Processing lab at Nile University as a research assistant, where he worked on retinal image analysis.

After obtaining his Master's degree in 2010, Ahmed started his PhD studies at the Faculty of Electrical Engineering Mathematics and Computer Science at TU Delft in Delft, The Netherlands. His research project on analyzing spatio-temporal gene expression data from the brain was carried out jointly between the Delft Bioinformatics Lab at TU Delft and the Division of Image Processing (LKEB) at the Leiden University Medical center. In 2014, Ahmed spent three months at the Department of Genome Science at the University of Washington in Seattle, USA. During this research visit, Ahmed worked at the Eichler Lab on developing methods to prioritize autism risk genes. This research visit was supported by a fellowship from the European Molecular Biology Organization (EMBO) and the Leiden University Funds.

Since September 2014, Ahmed works as a post-doctoral researcher at the Imaging Genetics section of LKEB. Ahmed currently works on methods to link neuroimaging data and high-throughput "omics" data to help identify associations between neuroanatomical and neurophysiological observations and the underlying molecular mechanisms in healthy and diseased brains.

LIST OF PUBLICATIONS

JOURNAL PAPERS

- Taskesen E, , Huisman SMH, **Mahfouz A**, Krijthe JH, de Ridder J, Stolpe A, van den Akker E, Verheagh W, Reinders MJT. (2016) Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci Rep* 2016 (April):24949.

- Eising E, Shyti R, 't Hoen PAC, Vijfhuizen LS, Huisman SMH, Broos LAM, **Mahfouz A**, Reinders MJT, Ferrari MD, Tolner E, de Vries B, van den Maagdenberg AMJM. Cortical Spreading Depression Causes Unique Dysregulation of Inflammatory Pathways in a Transgenic Mouse Model of Migraine. *Mol Neurobiol.* (2016) doi:10.1007/s12035-015-9681-5.

- Eising E., Huisman SMH, **Mahfouz A**, Vijfhuizen LS, on behalf of the International Headache Genetics Consortium: Anttila V, Winsvold BS, Kurth T, Ikram MA, Freilinger T, Kaprio J, Boomsma DI, van Duijn CM, Järvelin MRR, Zwart JA, Quaye L, Strachan DP, Kubisch C, Dichgans M, Davey-Smith G, Stefansson K, PalotieA; Chasman DI, Ferrari MD, Terwindt GM, de Vries B, Nyholt DR, Lelieveldt BPF, van den Maagdenberg AMJM, Reinders MJT. Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas. *Hum Genet.* (2016) 135(4):425–439.

- **Mahfouz A**, Lelieveldt BPF, Grefhorst A, van Weert LTCM, Mol IM, Sips HCM, van den Heuvel JK, Datson NA, Visser JA, Reinders MJT, Meijer OC. Genome-wide co-expression of steroid receptors in the mouse brain: identifying signaling pathways and functionally coordinated regions. *Proc Natl Acad Sci* (2016) 113: 2738–2743.

- **Mahfouz A**, Ziats MN, Rennert OM, Lelieveldt BPF, Reinders MJT. Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome. *J Mol Neurosci* (2015) 57(4):580–594.

- Babaei S, **Mahfouz A**, Hulsman M, Lelieveldt BPF, de Ridder J, Reinders M. Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex. *PLOS Comput Biol* (2015) 11: e1004221.

- **Mahfouz A**, van de Giessen M, van der Maaten L, Huisman S, Reinders M, Hawrylycz MJ, Lelieveldt BPF. Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods.* (2015) 73: 79–89.

- **Mahfouz AE**, Fahmy AS. Fast localization of the Optic disc using projection of image features. *Image Process IEEE Trans* (2010) 19:3285–3289.

CONFERENCE PAPERS (PEER-REVIEWED)

- **Mahfouz A**, Ziats MN, Rennert OM, Lelieveldt, BPF, reinders, MJT, "Genomic connectivity networks based on the BrainSpan atlas of the developing human brain," *Proc. SPIE 9034, Medical Imaging 2014: Image Processing*, pp.90344G.

- Fahmy AS, Abdelmoula WM, **Mahfouz AE**, Shah SM, "Segmentation of Choroidal Neovascularization lesions in fluorescein angiograms using parametric modeling of the intensity variation," *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on* , pp.665,668.

- El Kady AM, **Mahfouz AE**, Taher MF, "Mechanical design of an anthropomorphic prosthetic hand for shape memory alloy actuation," *Biomedical Engineering Conference (CIBEC)*, 2010 5th Cairo International, pp.86,89.

- **Mahfouz AE**, Fahmy AS, "Ultrafast Localization of the Optic Disc Using Dimensionality Reduction of the Search Space," *Med Image Comput Comput Assist Interv (MICCAI)* 2009, Part II. LNCS, vol. 5762, pp.985–992, Springer, Heidelberg.

- **Mahfouz AE**, Fahmy AS, "Ultrafast optic disc localization using projection of image features," *Image Processing (ICIP)*, 2009 16th IEEE International Conference on, pp.665,668.