Fall 12-16-2016

Global gene expression profiling of healthy human brain and its application in studying neurological disorders

Simarjeet K. Negi
University of Nebraska Medical Center

Global gene expression profiling of healthy human brain and its application in studying neurological disorders

by

Simarjeet K. Negi

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Biomedical Informatics Graduate Program

Under the Supervision of Dr. Chittibabu (Babu) Guda

University of Nebraska Medical Center
Omaha, Nebraska

November, 2016

Supervisory Committee

Howard Fox, M.D., Ph.D.	Fang Yu, Ph.D.
James Eudy, Ph.D.	Sanjukta Bhowmick, Ph.D.

To Biji and Nani ma

Expression map of healthy adult human brain and its application to study neurological disorders

Simarjeet K. Negi, Ph.D.

University of Nebraska Medical Center, 2016

Advisor: Chittibabu Guda, Ph.D.

The human brain is the most complex structure known to mankind and one of the greatest challenges in modern biology is to understand how it is built and organized. The power of the brain arises from its variety of cells and structures, and ultimately where and when different genes are switched on and off throughout the brain tissue. In other words, brain function depends on the precise regulation of gene expression in its sub-anatomical structures. But, our understanding of the complexity and dynamics of the transcriptome of the human brain is still incomplete. To fill in the need, we designed a gene expression model that accurately defines the consistent blueprint of the brain transcriptome; thereby, identifying the core brain specific transcriptional processes conserved across individuals. Functionally characterizing this model would provide profound insights into the transcriptional landscape, biological pathways and the expression distribution of neurotransmitter systems.

Here, in this dissertation we developed an expression model by capturing the similarly expressed gene patterns across congruently annotated brain structures in six individual brains by using data from the Allen Brain Atlas (ABA). We found that 84% of genes are expressed in at least one of the 190 brain structures. By employing hierarchical clustering we were able to show that distinct structures of a bigger brain region can cluster together while still retaining their expression identity. Further, weighted correlation network analysis identified 19 robust modules of coexpressing genes in the brain that demonstrated a wide range of functional associations. Since signatures of local phenomena can be masked by larger signatures, we performed local analysis on each distinct brain structure. Pathway and gene ontology enrichment analysis on these structures showed, striking enrichment for brain region specific processes. Besides, we also mapped the structural distribution of the gene expression profiles of genes associated with major neurotransmission systems in the human. We also postulated the utility of healthy brain tissue gene expression to predict potential genes involved in a neurological disorder, in the absence of data from diseased tissues. To this end, we developed a supervised classification model, which achieved an accuracy of 84% and an AUC (Area Under the Curve) of 0.81 from ROC plots, for predicting autism-implicated genes using the healthy expression model as the baseline. This study represents the first use of healthy brain gene expression to predict the scope of genes in autism implication and this generic methodology can be applied to predict genes involved in other neurological disorders.

ACKNOWLEDGEMENTS

This work would not have been possible without the tremendous amount of support I received from many individuals.

First, I would like to thank my graduate program and my thesis advisor Dr. Chittibabu (Babu) Guda. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. Besides my advisor, I would like to thank the rest of my thesis committee, Dr. Howard Fox, Dr. James Eudy, Dr. Fang Yu and Dr. Sanjukta Bhowmick, not only for their insightful comments and encouragement, but also for the hard questions that challenged me to widen my research perceptions.

The members of the Guda group have contributed immensely to my personal and professional time at UNMC. The group has been a source of friendships as well as good advice and collaboration. A heartfelt thanks to everyone in the lab.

Most importantly, I would like to thank my parents, Harvinder and Bhupinder Negi, and my sister Dapinder, for their unwavering support, guidance, and inspiration. It is with their belief and blessings that I have been able to come so far in my life. Also, a special thanks to my brother-in-law Jagdeep and my adorable niece Kulnoor and nephew Angad who have never failed to cheer me up with their cute little smiles. You all are the greatest source of joy in my life.

TABLE OF CONTENTS

List of Figures and Tables

List of Abbreviations

ABA Allen Brain Atlas

CNS central nervous system

BS brain structure

MD model

WGCNA weighted gene coexpression network analysis

kME module membership

ME module eigengene

DG dentate gyrus

CA4 cornus ammonis

LTP long term potentiation

AUC area under the curve

ML machine learning

RF random forest

ROC receiver operating characteristic

TN true negative

TNR true negative rate

TP true positives

TPR true positive rate

WEKA waikato environment for knowledge analysis

ASD autism spectrum disorders

CDC Center for Disease Control and Prevention

BA Broadman's area

CGH comparative genomic hybridization

CNV copy number variation

DSM Diagnostic and Statistical Manual of Mental Disorders

mRNA messenger RNA

FDR false discovery rate

GABA gamma aminobutyric acid

GO gene ontology

Hipp, Hip hippocampus

Amyg, Amy amygdala

Cere cerebellum

CPLV choroid plexus

IPA ingeunuity pathway analysis

CHAPTER 1

INTRODUCTION

Introduction of Research Area

Human brain gene expression

Human brain is one of the most perplexing organs known to mankind with over 86 billion neurons distinctly classified over 100 different types (1). At any given time, more than 80% of the genes are expressed in the brain, which is far more than any other body tissue (2). Also, compared to other species, human brains express mRNA transcripts at much higher levels and with much greater complexity (3, 4 and 5). It has been hypothesized that this increased level of gene expression in the human brain is in part responsible for the higher level of neuronal activity and higher order reasoning and function in humans (6). Brain's complexity arises due to its multilayered complex anatomy and computational capabilities, owing to different cell types and the connections between them. The incredible gene expression complexity of a given brain region along with the importance of relationships amongst these distinct brain regions cannot be undermined. Various neurological disorders; especially neuropsychiatric and neurodevelopmental disorders, occur due to the disengagement in communication between two or more brain regions (7, 8). Underscoring the importance of region-specific expression have revealed that gene expression differences between any two brain areas within one individual are more pronounced than the gene expression differences between two different individuals within the same brain region (5, 9 and 10). Also, the number of highly expressed genes in brain surpasses any other tissue type by far (11) and it is one such organ which consumes more than 20% of the inhaled oxygen (12) due to its exceedingly active state because of gene transcription, translation and metabolic process accompanied by specific brain functions.

Post genomic era

We live in a new era of biological research – the post-genomic era. The genome, defined as the complete genetic material (DNA and RNA) of a species, has been sequenced for numerous species. Genomics has brought forth a variety of methods for generating and analyzing these data that have revolutionized biomedical research. However, genomic methods are not without limitation, sometimes resulting in noisy data, inconclusive results, and often are expense on the order of thousands of dollars (13). To study the biological system quantitatively, several techniques have been developed to measure the expression levels of mRNAs and proteins. These techniques include Western Blot, ELISA, Mass Spectrometry (MS), quantitative RT-PCR, and DNA Microarrays for mRNA expression measurement.

The microarray approach has made it feasible to carry out genome-wide expression studies (14). Transcriptomic studies utilizing microarray techniques have established themselves to be invariably beneficial tools in the understanding of gene expression in healthy states as well as uncovering the basis of CNS malfunction (15). We can gain insights into the functionality of human brain by combining high throughput expression data sets with computational analyses. Microarray analyses of gene-expression in humans have allowed researchers to uncover the characteristics that define human brain at the molecular level (16). Connecting such data to phenotypes of interest, particularly human brains susceptibility to neurological disorders can now be carefully conducted studied via genomics approaches (16). In addition, the advent of genome wide studies in the human brain at finest possible anatomical resolutions bearing an ‘all- gene, all- structure’ approach (17) has led to the rise of a new wave of transcriptional data accompanied by in-depth pathway and functional analysis.

Microarrays as a data resource have helped progress important discoveries in many biological studies, especially providing insights about evolution, development and functioning of human brain (18). Many studies have surveyed the global expression patterns for anatomically distinct sites of the human nervous system and showed that the expression patterns of CNS were significantly different from non-CNS tissues (8, 19). Other interesting studies have analyzed gene expression data from various tissues of both humans and chimpanzees, and humans and mice to get a deeper understanding of differences and similarities in the transcriptional regulation between species (19, 20). Such publicly available datasets have been used to elaborate on already existing knowledge of brain disorders (21, 22, and 23). Thus, these publicly accessible repositories provide highly valuable data and have vast potential to make note-worthy contributions to brain science (23, 24).

Genetics in neurological disorders

Diseases of the central nervous system remain among the most gripping illnesses known to humankind. This is because neurological disorders are typically overwhelming to the patients and their families, often depriving individuals of the quality of life and because the vast majority of neurological disorders lack effective therapies (25). Recent estimates suggest that approximately 25% of adults in the U.S. are diagnosable in a given year for one or more mental disorders (26). Knowledge of genetics in neurological disorders has emerged from the advances made in molecular biology, genetics and a constant quest to understand the relationships among genes, brain, behavior and neurological disorders (27). The advent of advanced technologies in the 2000s have made genetic analyses readily available. Thus, the last two decades have seen a noticeable increase in recognizing the precise role genes played in relation to neurological disorders. Advancements were made in but not limited to: Fragile X syndrome, Alzheimer's, Parkinson's, epilepsy and ALS. While the genetic basis of simple diseases and disorders has been accurately pinpointed, the genetics behind more complex neurological disorders; like autism is still a source of ongoing research (28). With the expansion of neurogenetics a better understanding of specific neurological disorders and their respective phenotypes has soared. For severe disorders such as epilepsy, brain malformations, or mental retardation a single gene has been identified 60% of the time; however, the milder the intellectual handicap the lower chance a specific genetic cause has been pinpointed (29). As an example, autism is linked to a specific mutated gene only in 15-20% of the cases, and the mildest forms of mental handicaps are linked to genetics only in less than 5% of the cases (30). Moreover, such implications are usually not restricted to a single brain region. For example, multiple studies have shown that the cerebellum contains the most unique gene expression pattern compared to other brain structures (3, 8, 9), which is of consequence to autism in particular, as this region has been consistently implicated in the pathogenesis of this disorder (31).

In 1944, Dr. Leo Kanner, defined the childhood neuro-psychiatric disorder called autism, a phenotype which is now known to affect ~1/100 individuals (32). However, it was not until the 1970s that there was even one scientific publication mentioning both the words gene and autism. By 1989, a twin study provided the first evidence for a genetic predisposition to autism (33). With the completion of The Human Genome Project in 2003, the discovery of autism genes has increased exponentially and has helped to elucidate a number of key genes and pathways involved in this complex disorder (34) (Figure 1). My dissertation work started in 2012 and is now culminating in 2016, a time in which the discovery of autism genes has been occurring at a very rapid rate.

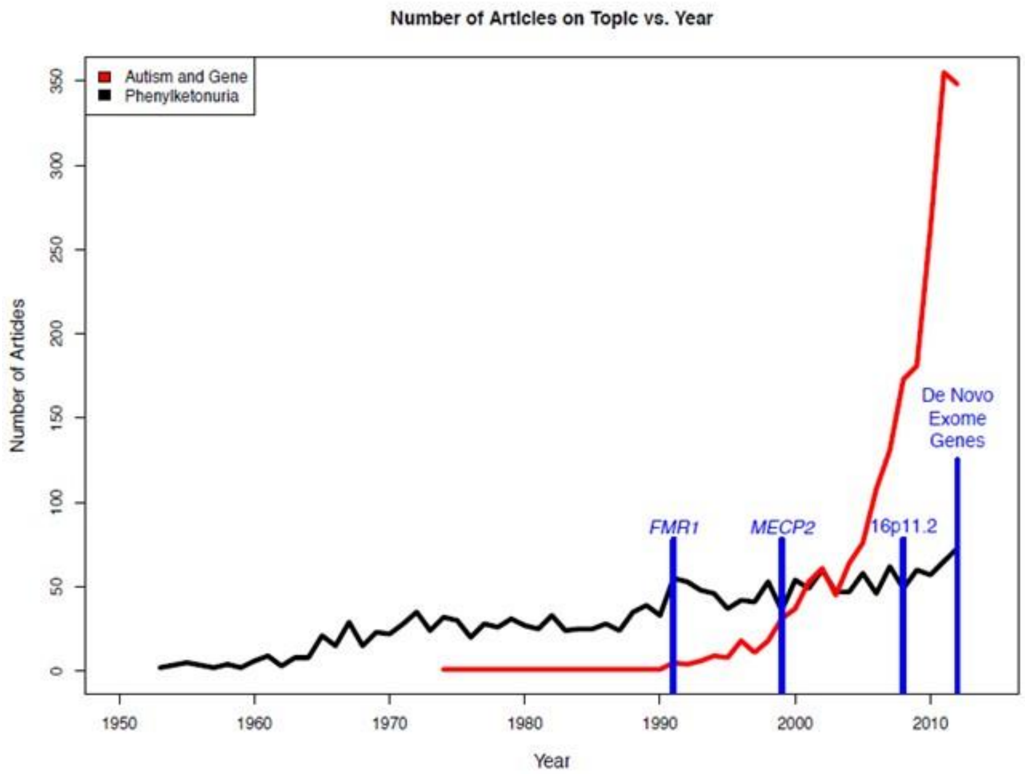


Figure 1- Investigation of the autism genetic literature. The growth curve depicts an exponential increase in the number of articles containing the words autism and gene starting at the time point of the human genome project's completion. A few key points in autism genetic history are highlighted: (1) FMR1 in Fragile X Syndrome (2) MECP2 in Rett Syndrome (3) 16p11.2 deletion and (4) de novo single nucleotide variants discovered in exome sequencing studies

With all the complexities in the nature of neurological diseases, there is an immense scope for incorporating more technology solutions in the transcriptomics. One such important resolution can be leveraging the big data and machine learning technologies to predict disease implicated genes and aid in the complex and time consuming processes of finding genes using the traditional molecular biology techniques. Development of methodologies which utilizes immense volume of gene expression data available towards predictive modeling will be greatly valuable to the scientific community.

Major unanswered questions

Human brain is an extremely dexterous organ and to comprehend its multi-layered functioning is one of the greatest challenges (34). Owing to the complex nature of the human brain, one of the major but urgent needs is to understand the gene expression patterns in the brain. The recent availability of comprehensive and detailed gene expression data from healthy brain tissues has now made it possible to discover global patterns (17). The possible generation of a map of the transcriptional landscape of the human brain will possibly help achieve closer look at the expression signature of unique brain structures and lend an insight into the molecular functions and the cross-talk between the distinct brain regions. Our method provides a novel way to systematically integrate high throughput transcriptome profiling data from different human brains and functional analysis results. Also, our method is of high utility to make advances in the current understanding about neurological disorders.

Studying the transcriptional communication between the local structures of a higher order brain region as well as global transcriptional organization is vital to understanding the underlying mechanisms of the human brain. It is known that the expression profiles of genes is reasonably stereotyped between individuals and the humungous amount of publicly available data can be put to good use by generating a duplicable framework of the healthy human brain transcriptome consistent across individuals and then using it as the baseline to predict the possible association of genes with neurological disorders.

Although there are several published studies in which transcriptional profiling has been used to examine gene expression in neurological diseases (23, 35, 36), there has not been considerable reports (37) that focuses exclusively on utilizing healthy tissue expression data from sources like the Allen Brain Atlas (ABA). The ABA provides comprehensive gene expression data at high neuroanatomical resolution of brain structures using microarray technology. In this study, we have used data from the ABA (23) to develop a framework for the transcriptional machinery of the healthy adult human brain followed by its application to studying disorders like autism. The basic premise of the study is that gene expression data from multiple healthy individuals can be used to design an expression based model that precisely describes the expression relationships of various brain regions. This model can serve as an expression map template for understanding the genetic underpinnings of highly conserved features of brain organization. Finally, we can exercise the utility of this model as a baseline to develop a prediction model for identifying potentially new candidate genes implicated in neurological diseases using machine learning. Our methodology has a generic approach that can be applied to any other brain disorders like Schizophrenia, Alzheimer's etc. Also, the inherent application of the model to diseased brain states would help improve the quality of life of people suffering from neuropsychiatric disorders as well as help control the escalating economic cost associated with these disorders.

Challenges

A human brain is composed of roughly 86 billion neurons, which form a very intricate but also incredibly dynamic network. Most brain studies are done in model organisms like mice where over 90% similarity (39) is reported in the genes (90% of the mouse genome could be lined up with a region on the human genome, 99% of mouse genes turn out to have analogues in humans). Extrapolation from animal models to human patients is always uncertain, but this is especially true for brain disorders given the profound anatomical differences between the brains of humans and rodents (40). The use of non-human models i.e., mouse or primates for finding, corroborating and replicating human brain transcriptome has inherent shortcomings [41, 42]. In contrast to human brain, distinct brain structures of experimental animals especially mice, can be dissected relatively easily to perform a neuroanatomical profiling, without a considerable postmortem delay. But, a 1000-fold increase in brain size when compared to mouse brings along with it a number of anatomical and structural variations which in turn account for the differences in neural circuits, gene expression intensities and regional dissimilarities in case of mouse modeling. It is nearly impossible to model the affected brain circuits and regions in the model organisms including primate species. This is because, the rate of gene expression changes in the brain is accelerated during human evolution [43, 44], and gene-expression changes in the evolution of the human brain primarily involved increased expression (upregulation).

On the other hand, the human brain samples can only be collected from deceased people as no regular biopsies can be performed in the brain tissue. While cerebrospinal fluid (CSF) and small central nervous system (CNS) tissue samples can be obtained from a living patient, the procurement of whole brains is only possible from deceased individuals. Even if the tissue is available, the degradation of mRNA imposes a huge problem because the use of low quality RNA samples for gene expression profiling is not reliable (45). Degradation of RNA transcripts by the cellular machinery is a complex and highly regulated process. In live cells and tissues, the abundance of mRNA is tightly regulated, and transcripts are degraded at different rates by various mechanisms, partially in relation to their biological functions. In contrast, the fates of RNA transcripts in dying tissue, and the decay of isolated RNA are not part of normal cellular physiology and, therefore, are less likely to be tightly regulated. However, recent advances in the post-mortem tissue acquisition and development of new methods that could be applied in studies of postmortem brain tissue, created opportunities for novel and powerful investigations into the human brain (46). At present, the study of cellular and molecular processes can only be detected through the direct use of postmortem brain tissue. Thus, studies of the postmortem human brain represent a critical and complementary approach to *in vivo* studies, as well as an essential interface between clinical investigations and studies in animal models (46).

In addition to understanding the healthy brain transcriptome, the investigation of neurological disorders presents more unique challenges (47, 48). Availability of diseased human brain tissue with neuro-anatomically high resolution samples continues to be a major issue. Due to this limited availability of brain tissue often studies focus on using blood samples from the patients (35, 49, 50, 51, 52 and 53). Even though, the blood samples are easily accessible and can support large population- based collections, they do not accurately represent the expression profile of a patient's brain (53). To overcome this issue, in the recent years' researchers have attempted to induce pluripotent stem (iPS) cells from individuals with particular disorders and prompt the generation of specific neuronal cell types in order to study these in-vitro (54). However, the iPS technology is still in its infancy with challenges associated to low efficiency and high technical expertise. Taken together, many studies have explored gene expression profiles in neurological disorders (52, 53 and 55), but none of them focuses exclusively on utilizing healthy tissue expression data from sources like ABA and exploring it in a framework of known disease implicated genes.

Most of the large- and wide-scale studies are the amalgamation of several individual smaller studies performed on different microarray technological platforms. The challenge in such studies is maintaining the quality control standards for RNA integrity, array processing, adjusting batch effects, normalization, etc (56, 57). When using data from multiple studies, integration of heterogeneous types of data generated from diverse technology platforms poses the first challenge (58, 59). Although meta-analysis studies have shown lists of differently expressed genes (DEGs), there tends to be inconsistencies among studies due to varying results obtained by different groups, accomplished by different laboratory protocols, microarray platforms and analysis techniques (60, 61). Computational expense of running analysis on these large datasets can also create an impediment, but with appropriate utilization of computational resources the data can be stored and processed using much less storage and running time.

Recruitment challenges for brain tissue donation accompanied by the lack of appropriate bio-banking facilities of the donated tissue also represent a major hurdle to brain research. Besides, human neuroscience has also been engulfed by ethical issues (62) associated with direct manipulation and neuronal recording of the human brain and has in part contributed to the lag in brain research.

Motivation

Knowledge of the transcriptome organization of the healthy brain offers contextual information on the key cellular processes and biological pathways involved in brain's functioning that maintains the equilibrium state of the brain (63). Moreover, characterizing the gene expression in the brain is of utmost importance as brain shows more intricate pattern of gene expression than other body organ (64, 65) and the gene expression patterns of the diseased human brain are expected to contrast those of the healthy brain. Most complex brain disorders involve interaction of multiple genes and the brain cells respond to the diseased state by altering their transcriptional program (65). It has been over a decade since the publication of the first high-throughput gene expression profiling study of the brain (66). Since then, several important studies have come along; however, none provides an integrated view of the brains transcriptome. Most studies have focused on a few brain regions rather than whole brain expression profiling (67, 68 and 69) or used model organisms (70, 71, 72 and 73) to make inferences about the global gene expression profiles in the human central nervous system. While such efforts have generated valuable information, precise catalogue of healthy gene expression is far from complete owing to the difficulty in getting good quality neurotypical human brains as well as the resources to dissect the brain tissue to highest level of precision so as to capture all the know brain structures (representing different functionality). Also, no study is yet known that provides a consistent expression profile of the genes with respect to their differential activity across multiple individuals at a high anatomical brain structure resolution. Therefore, the development of neuroanatomically comprehensive, genome-wide models of gene distributions using the postmortem human tissue are critical to understand the brains functionality.

Taking into account all the shortcomings from the already published studies, the development of a statistically significant expression model followed by its analysis will reveal important features of the human brain transcriptional events in a healthy brain. Using the rich microarray profiling dataset in the Allen human brain atlas, we can capture the highly consistent patterns of transcriptional regulation across brain structures. Also, the genes with consistent anatomical patterning across individuals will very likely be significant for brain function and disease. Moreover, the consistent gene expression patterns (17, 74) found across multiple individuals will provide a means to compare and contrast the similarities and differences between healthy and abnormal neurological states. Besides, our study holds promise to demonstrate that the healthy adult human brain gene expression profile can be effectively used along with machine-learning methods to identify potentially new genes implicated in a neurological disorder like autism. Also, generation of a model that captures the transcriptome organization and represents the consistent expression profile of genes across multiple samples processed on the same platform with same QC measures is of high significance to further the brain transcriptome studies.

Structure of Thesis

This thesis will focus on the development of a unified gene expression model from clinically unremarkable human brains accompanied by its functional characterization and followed by the development of a supervised classification model for predicting disease implicated genes in neurological disorders. First, the challenges and motivation for the research topic would be presented with a brief overview of the research gap. This will be followed by the essential background required for the field of work. A thorough exposition of the methodology will be reported covering the detailed strategy used to create the model, the functional characterization both on the global and local structural landscape of the brains transcriptome and, the supervised machine learning prediction model will be reported. Results will be presented and discussed, ending with a conclusion, as well as future opportunities where this work can be extended and applied.

CHAPTER 2

ESSENTIAL BACKGROUND

Microarrays

In the world of bio-molecules, proteins play the key roles as structural components, enzymes, antibodies, and so on. Genes in DNA molecules carry the encoding information for proteins. The flow of this encoding information from genes to proteins involves two stages: transcription and translation. In transcription, a gene is transcribed into a single stranded sequence of RNA, called messenger RNA (mRNA). Then in translation, the mRNA is translated into a sequence of amino acids, which folds into a functional protein after some modifications. To study the biological system quantitatively, several techniques have been developed to measure the expression levels of mRNAs and proteins. In this dissertation, we would be using microarrays for quantifying (recording) gene expression because the datasets on the whole healthy brain samples are available only from microarray experiments. These data would help understand the molecular mechanisms, gene networks, and signaling pathways that are recurrently observed in healthy human brains.

To measure the expression levels of genes using the DNA Microarray techniques, hundreds of thousands of DNA probes are immobilized on a small glass, plastic, or nylon membrane, which is called an array or a microarray chip. Each probe contains a known quantity of DNA and represents a specific gene or an alternative form or a gene. mRNAs from the sample cells are hybridized with the probes on the array. So by measuring the intensity of the mRNAs hybridized with the probes, we can have the expression levels of the genes that we're interested in. This technique enables us to measure the expression values for hundreds of thousands of genes simultaneously so that we can observe the changes in genes' expression systematically. Also, with the aid of custom microarray chips, more intricate experiments can be designed to predict gene function, infer gene regulatory networks, understand disease mechanisms etc.

Application of Microarray Technology in the Study of Neurobiology

Brain represents the most complex organ in the human body and not surprisingly, the most number of genes are expressed in the brain tissue than any other human tissue, indicating the intense and dynamic role of transcriptome in the brain function. New NGS technologies can add more knowledge to our current understanding of the brain transcriptomics. However, due to the unavailability of RNA-seq data at a high structural resolution of brain anatomy, we used the microarray datasets. Microarray technology has been used widely in many studies of the CNS including brain development (83), behavior (84) and neurological diseases (85). Those diseases include Alzheimer's disease (86), Schizophrenia (87), Parkinson's disease (88), Huntington's disease (89), bipolar disorder (90), multiple sclerosis (91) and autism (92).

Array types

There are a number of microarray technologies for large-scale gene expression measurements. Among them, cDNA arrays and oligonucleotide arrays are the most popular approaches. Although they use the same principle, they differ in many aspects.

In a typical cDNA array experiment, mRNAs from two different samples are extracted and reverse-transcribed into cDNAs, which are labeled with dyes of different colors if they're in different samples. Then equal amount of labeled cDNA samples are mixed together and hybridized with the probes on the array. The probes are spotted cDNA of hundreds of nucleotides in length. After the hybridization, a laser scanner measures dye fluorescence of each color at a fine grid of pixels. Higher fluorescence indicates higher amount of hybridized cDNA and hence higher gene expression in the corresponding sample. After the scanning, typically two intensities for spotted cDNA of two colors and two intensities for the background of two colors are obtained. So there're at least four quantities for each probe on the cDNA array. Sometimes, these are accompanied with quantities that measure the quality of the spot, e.g. the variability of the pixel intensity. Since samples are labeled with different colors and hybridized competitively to the same set of probes, the cDNA array is also called two-channel array. The two channel array allows measurement of the relative gene expression in the two samples, i.e. the ratios of the two colors for each spot. The cDNA arrays are available from Agilent etc. The array used to generate gene expression data in this dissertation is cDNA array from the Agilent technologies.

The oligonucleotide arrays are available commercially from several companies, such as Affymetrix, Illumina, NimbleGen, Agilent, etc. Although each vendor uses different techniques, they have one thing in common: the short oligonucleotide sequences are used as probes. For example, in Affymetrix array, each gene is represented by one or more probe sets, each composed of 11-20 pairs of 25bps long oligonucleotide. Each pair consists of a perfect match and a mismatch. The mismatch is created by changing the middle (13th) base of the perfect match sequence to reduce the specificity of binding of mRNA for that gene. The goal of the mismatch is to control experimental variation and nonspecific binding of other mRNAs with the probe [Aff01]. Unlike the two-channel cDNA array, oligonucleotide array is often one-channel: mRNA from only one sample is prepared, labeled with a fluorescent dye, and hybridized to the probes on an array. After the hybridization, arrays are scanned, and images are produced and analyzed to obtain a fluorescence intensity value for each probe. In the probe set level, the typical output for a probe set includes two vectors of intensity readings, one for perfect matches and the other for mismatches.

Normalization techniques

Experimental variations, such as RNA quality, probe labeling, hybridization condition, washing, signal and background detection in the scanning process, slide and block effects, pose significant challenges in the analysis of microarray data. The first step in microarray analysis is to remove the systematic biases due to the variations in experimental conditions so as to make multiple array analyses comparable and meaningful. These efforts are collectively referred to as the normalization of microarray data in the literature. Normalization means to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes.

A number of useful normalization protocols for cDNA arrays have been proposed based on different assumptions. These include the global normalization (93), rank invariant normalization (94), LOWESS normalization (95), Semi-Linear In-slide Model (96), Two-way Semi-Linear Models, robust TW-SLM (96), non-parametric regression, RMA normalization, seminal normalization methods and normalization of small diagnostic microarrays (97). All of the aforementioned methods are based on some statistical and biological assumptions.

Several normalization approaches that are most frequently used for oligonucleotide arrays include Loess normalization (95). It is based on the M vs A methodology where M is the difference in log expression values and A is the average of the log expression values. The underlying rationale is that very few genes will have different expressions in two arrays. So an M vs A plot for the normalized data should have a point cloud centered on the M = 0 axis. Next is the contrast normalization. It is also based on the M vs A methodology, but this method transforms the data into a set of contrasts before the normalization. In the Quantiles normalization, the goal is to achieve the same distribution of probe intensities for each array in the dataset. If two data vectors have the same distribution, the Q-Q plots of them are a straight diagonal line. The other two less popular approaches are the Qspline normalization (98) and Invariant set normalization (99)

Microarray data analysis

Despite the high throughput and high efficiency of microarray technologies, high level of noises and complex experimental artifacts are associated with microarray data, which emphasizes the requirement for statistical and data analytic techniques for all stages of experimentation. Microarray data analysis can roughly be classified into three levels: low, middle, and high level, according to the stage of experimentation and involvement of other data sources.

Low level of data analysis, also termed as signal extraction, includes image analysis, gene filtering, background correction, probe level analysis and gene summarization for oligonucleotide arrays, as well as between-array normalization and removal of artifacts for comparisons across arrays. Middle level of data analysis includes selection of differentially expressed genes between experimental conditions, clustering/classification of biological samples or genes, construction of gene co-expression network, etc. High level analysis includes those approaches that integrate microarray data sets from different platforms or combine microarray data with other data sources, such as Gene Ontology information, pathway information, and so on. Great success has been achieved in the past few years by performing high level microarray data analysis and, this dissertation primarily focuses on middle & high level analysis and beyond as the scope goes much further than merely tagging the analysis with these two levels.

Functional genomics of human brain

The field of neuroscience has been slow to adopt functional genomic and genetic methods and the large-scale databases and resources that ideally result from their use. For example, neuroscience has consistently lagged behind cancer biology in the adoption of new molecular and genetic methods, starting with molecular cloning and continuing to functional genomics and genetics today. There are legitimate reasons for this, including the extreme cellular heterogeneity and complexity of neural circuits relative to most non-neural tissues, and the reliance on post-mortem materials for most human studies. Another obstacle is the generation of enormous amounts of data. The integration of computational biology or bioinformatics in modern neuroscience laboratories or groupings have become even more critical as more powerful technologies now generate many orders of magnitude more data, than the traditional molecular biology approaches.

Transcription of the inherited DNA sequence into copies of messenger RNA (mRNA) is the most fundamental process by which the genome functions to guide development and maintain functions. Furthermore, encoded sequence information, inherited epigenetic marks, and environmental influences all converge at the level of mRNA gene expression to allow for cell type-specific, tissue-specific, spatial, and temporal patterns of expression. Thus, the transcriptome represents a complex interplay between inherited genomic structure, dynamic experiential demands, and external signals. This property makes transcriptome studies uniquely positioned to provide insights into complex genetic-epigenetic-environmental processes such as non-mendelian genetic etiologies such as autism spectrum disorders.

Human brain gene expression has been demonstrated to be particularly unique compared to other human tissues, and in its complex regulatory processes, underscoring the need to understanding its functional genomics is crucial. In the brain, an individual gene can be expressed in multiple ways depending on the particular tissue (cell-type) of habitation, state of brain, and local or long distance signaling mechanisms being received. Therefore, in order to understand how a gene or large groups of genes may contribute towards the overall functioning of human brain, it is critical to assess the expression and function in the appropriate brain structures and find the uniformity in expression patterns across individuals.

Gene expression in adult human brain

Compared to other species, human brains express mRNA transcripts at much higher levels and with much greater complexity. For instance, comparisons of human brain gene expression with both mouse (100, 3) and primates (4, 5) have demonstrated that most of the differentially expressed genes between the species are up-regulated in humans, but this phenomenon is not apparent in other tissues. Additionally, the human brain expresses ~80% of all genes encoded in the human genome at some point during it's development (68), which is greater than any other individual tissue type. It is hypothesized that this increased level of gene expression in the human brain is at least partially responsible for the higher level of neuronal activity and overall cognitive function in humans.

Within humans specifically, the brain also displays a distinct gene expression profile from other tissues. Using both the array and sequencing-based techniques (101), the brain has been shown to have higher expression levels and greater transcriptome complexity than other human tissue and cell types. In particular, human brain gene expression displays a high level of alternative splicing and a unique diversity of noncoding RNA types expressed. For example, studies have demonstrated that the human brain transcriptome has an unusually high level of alternatively spliced transcripts compared to other tissues (102, 103, 104), and the set of isoforms produced in brain differs considerably from other tissue types (101, 102). In addition to increased numbers and types of spliced mRNAs, the human brain transcriptome also displays a uniquely high abundance of transcribed noncoding RNAs (ncRNAs). In fact, the brain displays the greatest abundance of transcribed ncRNAs among all tissues studied thus far (105). Both short ncRNAs, such as microRNAs (miRNAs), and long non-coding RNAs (lncRNAs) are highly enriched in the brain (106, 107, 108). As ncRNAs are becoming increasingly recognized as important regulatory elements in genome processing during neurodevelopment and in the pathogenesis of neurodevelopmental disorders (109), their abundance in the brain further highlights the uniqueness of neurodevelopmental functional genomics (discussed further below).

While within a given brain region the human transcriptome has been shown to be incredibly complex, perhaps unsurprisingly, there is strong evidence that distinct regions of the human brain have distinct gene expression profiles. Animal studies have suggested that this variation is related to both structural and functional differences (110). For instance, a microarray study of twenty distinct brain and spinal cord sites showed that expression profiles can cluster samples from different donors by anatomical origin, and that some anatomical regions have up to 2,000 region-specific genes (8). Multiple studies have shown that the cerebellum contains the most unique gene expression pattern compared to other brain structures (3, 8, 9), which is of consequence to autism in particular, as this region has been consistently implicated in the pathogenesis of the disorder (111). Even just within the neocortex, different cortical layers each express a detectably distinct profile of mRNA transcripts. Underscoring the importance of region-specific expression are results that have shown gene expression differences between any two brain areas within one individual are more pronounced than are gene expression differences between two different individuals within the same brain region (5, 9, 10).

In summary, the human brain has been demonstrated to have a unique pattern and complexity of gene expression compared to other species as well as compared to other human tissues, including region-specific gene expression patterns. This highlights the importance of understanding human neuropsychiatric disorders, such as ASD, in the context of human brain gene expression specifically, as it is likely that animal, cellular, and other models do not recapitulate the uniqueness of human brain functional genomics with the appropriate level of fidelity.

Gene expression during human neurodevelopment

The developing human brain grows remarkably fast—the weight of a newborn's brain is approximately 25% of its adult weight, but within two years, it nearly reaches its adult size (112). During this time, the brain grows mainly through glial multiplication, myelination, formation of new synaptic connections, and pruning of unused synaptic connections. While the human brain continues to mature up to the age of 25 years (113), the greatest changes occur only during the periods of infancy and early childhood. Coincidentally, most neurodevelopmental disorders, including autism spectrum disorders, become clinically recognizable around this age.

Underlying these dramatic early changes in brain development are complex and dynamic broad patterns of gene expression, which have only recently begun to be understood. The most comprehensive study to date of the developing human brain transcriptome (68) documented that genome-wide patterns of gene expression correspond closely to the major stages of clinical development (namely prenatal, early infancy, childhood, adolescence, and adulthood), and that the molecular profile of these stages are distinct from each other. The most striking observation was that the greatest shifts in gene expression occur around the period of birth, where the authors found almost 60% of genes change their expression patterns in the neocortex (68). Other studies have demonstrated similar changes, and have showed that many of the genes identified during this shift are known to be involved in cortical development and higher order cognitive functioning (69).

Changes in gene expression in early post-natal life have also been shown to have greater amplitude of change (114, 115, and 116). In fact, it was shown that many genes actually reverse their expression trajectory in early life (114), mostly shifting from a pattern of increasing expression in fetal life and infancy to a decrease in expression beginning in childhood. Moreover, as the brain begins to mature, the gene expression profile within each anatomical region becomes more similar to other regions, with the exception of the cerebellum, suggesting that most of the region specific development is completed early in life. Interestingly, these broad gene expression patterns appear to reverse themselves in older age, at least in the prefrontal cortex (116).

Gene expression dynamics in human brain development are clearly both spatially and temporally specific. This suggests not only that they are highly regulated, but that different genes and gene networks will have dynamic expression throughout space and time.

Gene expression based networks in the human brain

While assessing the genome is an important approach to comprehensively understand the complex functional genomics of human brain, it is equally important to consider how disparate genomic elements may work in concert with each other to produce biological effects that are emergent only after their interaction. The study of genetic interactions can be done by modeling large gene sets as networks of interacting nodes and edges, allowing for a statistical assessment of relationships among and between genes, as opposed to the study of individual genes themselves. Such approaches are particularly important in complex genetic syndromes like autism spectrum disorder, as genome wide association studies have consistently demonstrated that most individual variants in ASD have only very small effects by themselves.

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 (117, 118). Several studies have demonstrated that genes with similar brain co-expression patterns are likely to function together in common cellular pathways (119, 120). These transcriptional co-expression relationships are particularly relevant to the functional genomics of the human brain, as the precise regulation of gene expression across brain regions instructs the exquisite specialization and connectivity within the brain. For instance, if two genes are expressed with similar patterns (i.e. they have a similar magnitude and direction of expression change across developmental time), they would have a higher correlation than two genes whose expression appears to be randomly related to one another. In this network, edges would link genes with similar expression profiles, whereas unrelated genes would not share an edge (120) (Figure 2). Defining edges between genes in this way allows the conclusion that the two nodes share related biological function, and can be used to derive and study large-scale genetic interaction networks.

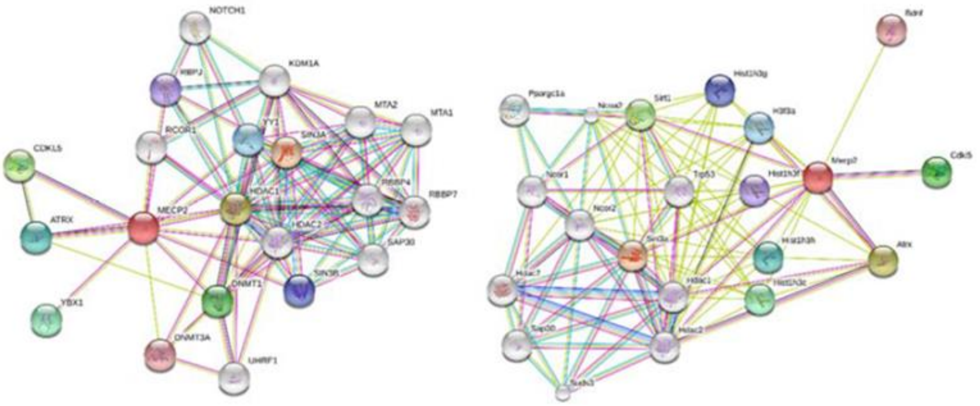


Figure 2- Examples of gene interaction networks. Properties of networks as a whole can become apparent that would not be appreciated by studying individual genes. For example, the network on the left represents known 2nd degree protein-protein interactions with the gene Mecp2 in humans, whereas the network on the right represents the known Mecp2 interactions in mice. As can be seen, the human network is much more densely connected, suggesting Mecp2 has more known interactions in humans.

Another widely used approach to infer interaction networks is to draw upon experimentally determined protein-protein interactions. Studies have demonstrated that protein interaction networks are conserved evolutionarily (121), and that proteins in the network with high degrees of connectedness are more important for organismal survival and fitness than those with lesser connectivity (122). This suggests that information on the importance of individual genes/proteins in a network can be inferred by studying the overall structure of the network as a whole.

Recent largescale proteomics efforts have shown that protein co-expression patterns are slightly better predictors of protein interactions than are mRNA co-expression patterns. However, obtaining comprehensive and unbiased datasets of protein co-expression is much more technically challenging than obtaining genome-wide RNA expression levels. Consequently, understanding gene co-expression patterns in brain is an important first step that could provide insights into the complex functional genomics of the human brain.

Autism and classification/prediction algorithms

The autism spectrum disorders are a heterogeneous set of syndromes defined by impairments in verbal and non-verbal communication, restricted social interaction, and the presence of stereotyped patterns of behavior. Autism spectrum disorders are one of the most common problems affecting children in the Western world. The most recent estimates have shown that ASD affects between 1 in 88 children (Centers for Disease Control and Prevention (CDC) 2012). Boys are at least four times more likely to receive a diagnosis of ASD as compared to girls (CDC 2012), and this ratio increases significantly when only mildly affected children are considered (123). The costs associated with autism are similarly great. Perhaps more importantly, the emotional toll placed on parents and caregivers of children with autism is immense, unrelenting (124).

Evidence for a strong heritable risk of ASD was initially described in twin and sibling epidemiological studies of autism (125), and has since been firmly established through multiple genetic approaches (126). The first twin studies in ASD demonstrated a concordance rate approaching 90% in monozygotic twins and 10% in dizygotic twins (127, 33, 128). Subsequently, larger studies have shown the dizygotic concordance rate to be greater than 20% (129).

The development of microarray technology such as comparative genomic hybridization (CGH) allowed the unbiased assessment of the genomic architecture of ASD. The first of these analysis indicated that individuals with ASD had 10-20 times the number of CNVs as controls (130, 131). Different brain regions have been implicated in both post-mortem and neuroimaging studies, notably the prefrontal and temporal cortices, and the cerebellum (132). Intriguingly, many of the genes known to be integral to these processes have been independently linked to autism in genetics studies. For instance, the Shank family of proteins, notably Shank 1 and Shank 3, has been repeatedly implicated in ASD (133).

In most studies, classification or prediction-based analysis in autism has been performed to distinguish the autistic patients from the normal controls. In some more recent studies, this is usually achieved by training a classifier on the gene expression signatures (usually from the peripheral blood) of the cases compared to controls. Integrating multi-parametric functional and structural measure from MRI/fMRI based on the patients versus the controls is another widely implemented methodology using the principles of machine learning and graph theory (134, 135). Other methods include text classification where the data sets are related to patient's details in the form of text containing the detailed explanation of the symptoms from which they suffer (136, 137). Further, complementary machine-learning approach based on human brain-specific gene network have also been developed which present a genome-wide prediction of autism risk genes (138). However, none of the prediction methods described above make use of only healthy brain gene expression to predict genes involved in autism implication using supervised classification machine learning models. Evaluating the utility of healthy brain gene expression as a tool to predict the association of new genes for a neurological disorder offers tremendous opportunities to enhance our current knowledge. In the present study, we used the discriminatory power of the expression patterns of known autism genes in healthy individuals to develop a model that could be used to identify new genes with potential association to autism.

Machine learning and data mining

The field of classification is domain-independent, and has many names (e.g., statistical inference, pattern recognition, subtyping, soting, etc.) Regardless of the name, the goal is to develop a model that can be used to assign objects to a category, with the highest accuracy possible. Computational methods for classification draw from research in machine learning and data mining, which themselves are largely based on theory from probability and mathematical statistics. Given a dataset of labeled genes, the numeric expression values representing distinct brain structures act as the features and the labels are the different classes (autism genes versus non-autism genes), then the class-specific features can be extracted to build models that can predict potentially new genes for the neurological disorder.

Classification in machine learning is often broken up into three distinct areas: (i) *supervised learning*, (ii) *unsupervised learning*, and (iii) *semi-supervised learning*. In supervised learning, a set of data that is labeled with the concept being learned is used, with the aim of inducing a model that can label new data with a high level of accuracy. In most cases, the algorithm outputs a discriminative function that maps data to a concept. Decision trees, artificial neural networks, and support vector machines are all examples of supervised learning algorithms. In unsupervised learning, a model of the input is induced, but no mapping to any output is created, as the input data is not labeled. This is appropriate in cases where a model is desired that learns natural groups or arrangements of the data. Clustering is a common form of unsupervised learning. In semi-supervised learning, a model is induced from both labeled and unlabeled data, again with the intent of labeling new data. These methods are useful for datasets with a limited labeled data and a large size unlabeled data of the same kind. Expectation maximization methods used in conjunction with supervised learning methods are often used in semi-supervised learning.

In the context of learning, statistical inference deals with problems where a set of data is given, and the method learns about the underlying probability distribution that generated the data. There may be one, several, or an infinite number of possible distributions that may have generated the data. Some methods may simply choose the most likely distribution to model the data, while other methods may consider all possible distributions in the hope of obtaining a better overall model. Irrespective of the method, the resulting model can be used to make inferences about tested and untested data.

The rest of this chapter is focused on establishing the essential framework for classification.

Steps for classification

In general, there are five important steps that must be considered for successful classification:

Data Collection – Let D denote the domain of interest. (In this research, D represents the set of all possible genes, and will be referred to as the gene space). A sufficient number of instances from D must be collected for learning, or training a model of the data. The dataset of examples which will be used for training is called *training data*, and will be denoted as D. Thus, $D \subseteq \mathrm{D}$. Let d_i represent the i^{th} instance in D. Each instance will belong to a class (or category), denoted as y_j, from a set of possible classes, denoted as C.

Input Representation – A set of features must be selected to represent instances that are to be classified. The feature set must be easy to extract from the data, and highly discriminatory among the different classes, yet general enough to allow representation of data in the domain that is yet to be seen. Let X represents the *feature space* – the space of all possible features that could be used to represent any instance in D. The function that maps instances in D to features over X is assumed to be neither injective nor surjective; that is, there may be distinct instances in D that map to the same feature vector, *and* there may be feature vectors over X for which no instance in D could possibly exist in nature.

The only assumption is that each unique instance in D maps to at least one feature vector over X. Let $\mathbf{x}_i$ represent the vector of values of features representing instance d_i in D. Then, D is a set of ordered pairs ($\mathbf{x}_i$, y_j). However, the notation $d_i \in D$ and $\mathbf{x}_i \in D$ are both used in this research to refer to the same i^{th} instance in the training data D; the representation is implied in the notation.

Classification Algorithm/Model – Machine learning and data mining offer many choices of classification algorithm. It is not always clear which algorithm is the best for which problem. The efficiency of an algorithm depends on the nature and distribution of the training dataset. Regardless of the choice, the classification algorithm will provide the framework for learning.

Learning – Once the algorithm is established and the form of the model is known, the learning takes place. In this step, the training data is analyzed to establish patterns between the data and the classes the data is assigned to. Successful learning results in the induction of a hypothesis that best models (or explains) the data being analyzed. The hypothesis is output as a target function, denoted as h, which maps instances represented by their features to a class, denoted as:

$$h\colon \mathrm{X} \rightarrow \mathrm{C}$$

One can think of the feature vector as a uniform description of the object to be classified. The classifier evaluates the description (features) of the new object, and assigns the category that belongs to the best matching descriptions of known objects. The classifier is a hypothesis that best explains the differences in the descriptions between objects belonging to each class. Learning involves determining what descriptors make the objects unique to the class, and adjusting the model to best exploit those descriptors.

Evaluation – In reality, the hypothesis will not perfectly explain everything. The classifier will likely make mistakes. We want to select the model that makes the fewest mistakes which is generally measured by the number of correctly predicted known positives (sensitivity) and the number of correctly predicted known negatives (specificity). One interpretation of machine learning is focused on the development of computational methods that optimize a *performance measure* using example data or past experience (139). The performance measure for a hypothesis is usually a quantitative measure of the error between the hypothesis and the true concept being learned on some data being tested (Machine learning algorithms "learn" in a sense that the performance of the method is designed to improve as more data become available).

CHAPTER 3

MATERIALS AND METHODS

Data sets and data representation

The data used to generate healthy human brain gene expression model in this dissertation originates from ‘Allen Brain Atlas (ABA)’ (17), which comprises of a global ‘all-genes, all-structures’ survey of gene expression using microarrays. ABA contains data from six neurotypical adult individual brains, where each brain was dissected into hundreds of precise anatomic structures. Throughout the process, anatomic data (MRI, blockface images, histology) were annotated to enable the collection of anatomically defined samples for microarray. After successful screening and QC of the tissue, a multipart dissection protocol was utilized to process the fresh frozen human brain tissue. Large format histology data was collected from each tissue slab with 4.65 µm/pixel digital image resolution. Each tissue slab was then subdivided into smaller tissue blocks categorized according to whether they contained primarily cortical or subcortical brain structures. These tissue blocks were sectioned for histology data with a final digital image resolution of 1 µm/pixel. If the blocks contained subcortical structures, additional sections were collected onto membrane slides that would allow laser microdissection (LMD) of these structures. Anatomically defined samples were collected for microarray analysis by either manual macrodissection of the remaining tissue from each block (cortical and some subcortical structures) or by laser-based microdissection (subcortical and brainstem areas). A schematic of the above mentioned process is shown in Figure 3 (17), summarizing the experimental strategy to subdivide intact brains and isolate precise anatomical samples. The microarray datasets contain approximately 400–500 tissue samples per hemisphere, often with multiple samples per structure for each individual brain. Also, due to normal variation in brain size and morphology, the number of samples per structure varies across the individual brains.

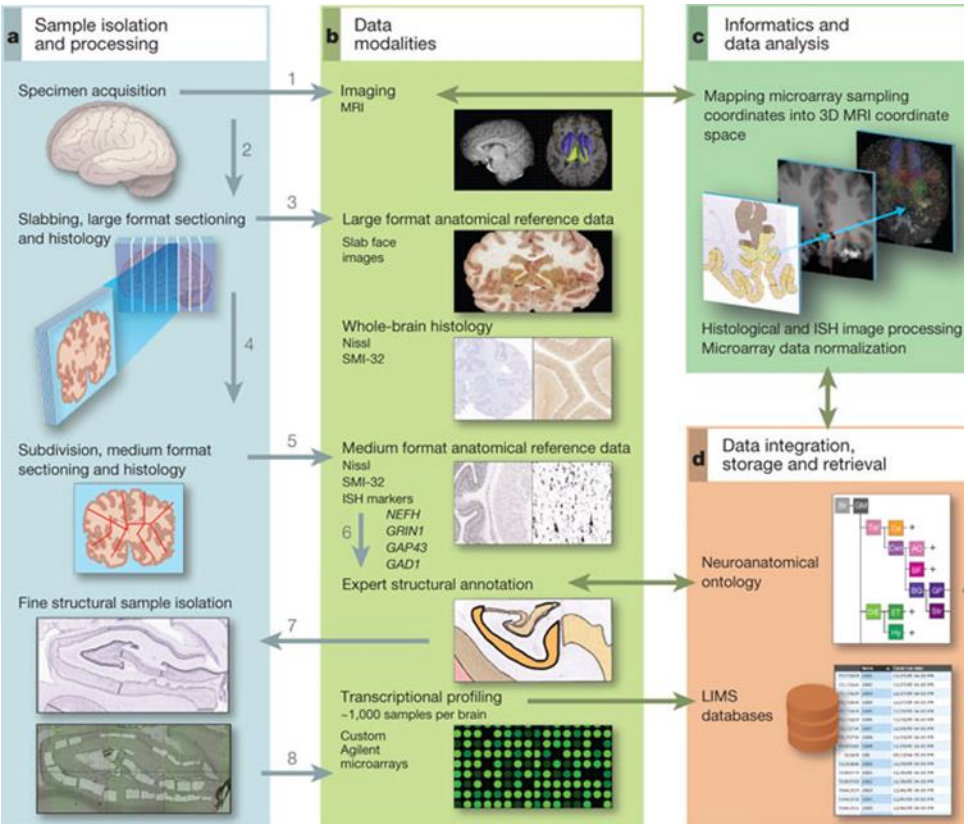
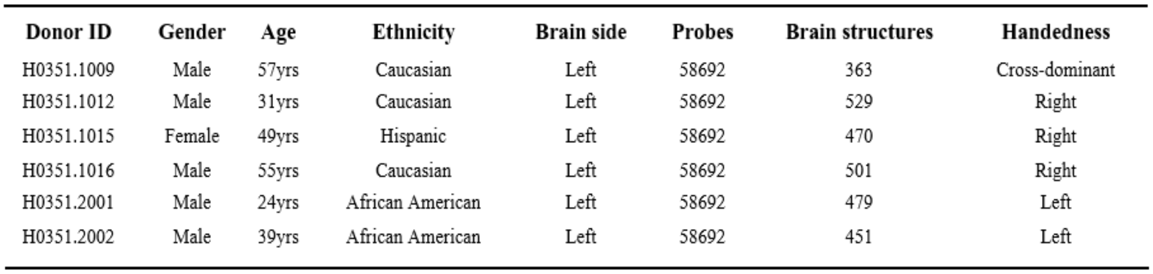

a
Sample isolation and processing
Specimen acquisition
1
2
Slabbing, large format sectioning and histology
3
4
Subdivision, medium format sectioning and histology
5
Fine structural sample isolation
7
8
b
Data modalities
Imaging
MRI
Large format anatomical reference data
Slab face images
Whole-brain histology
Nissl
SMI-32
Medium format anatomical reference data
Nissl
SMI-32
ISH markers
NEFH
GRIN1
GAP43
GAD1
6
Expert structural annotation
Transcriptional profiling
~1,000 samples per brain
Custom Agilent microarrays
c
Informatics and data analysis
Mapping microarray sampling coordinates into 3D MRI coordinate space
Histological and ISH image processing
Microarray data normalization
d
Data integration, storage and retrieval
Neuroanatomical ontology
LIMS databases

Figure 3 - Process schematic of primary steps in the creation of the whole brain microarray survey of the Allen Human Brain Atlas

The profiles of individual donors and the statistics on the brain structures are shown in Table 1. Note, only for two brains (H0351.2001 and H0351.2002) the samples were collected from both hemispheres. Otherwise, samples for microarray were collected from the left brain hemisphere alone. Therefore, to eliminate bias arising due to hemispheric specificity of gene expression, only left hemisphere of the donor brains was used for our analysis. Agilent cDNA array technology was used to generate the microarray data.

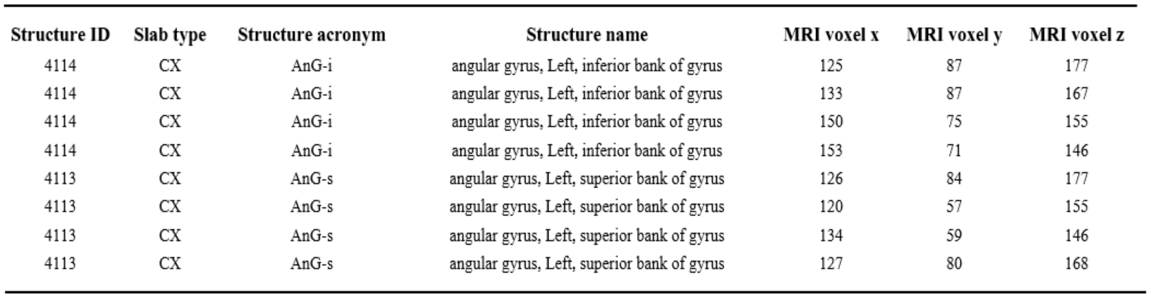


Donor ID	Gender	Age	Ethnicity	Brain side	Probes	Brain structures	Handedness
H0351.1009	Male	57yrs	Caucasian	Left	58692	363	Cross-dominant
H0351.1012	Male	31yrs	Caucasian	Left	58692	529	Right
H0351.1015	Female	49yrs	Hispanic	Left	58692	470	Right
H0351.1016	Male	55yrs	Caucasian	Left	58692	501	Right
H0351.2001	Male	24yrs	African American	Left	58692	479	Left
H0351.2002	Male	39yrs	African American	Left	58692	451	Left

Table 1 - Donor profile representing their age, gender, ethnicity and handedness

Generation of gene expression based model

Since multiple samples per brain structure can be present, we merged the samples (brain structures) that are identically annotated according to the anatomical information and averaged the expression values of all genes (Table 2). This reduced the number of samples and made the structure annotations unique (Table 1). To estimate the expression profile of each gene in every distinct brain structure (sample) w.r.t other structures, we calculated the z-scores. Statistically, the standard score or z-score is the signed number of standard deviations by which an observation deviates from the mean. A z-score is used for making norm-referenced interpretations, for which the mean and standard deviation are selected to simplify interpretations. On similar lines, p-values are probabilities and both these statistics are associated with the standard normal distribution. This distribution relates standard deviations with probabilities and allows significance and confidence to be attached to z-scores and p-values. Very high or a very low (negative) z-scores, associated with very small p-values, are found in the tails of the normal distribution (Figure 4). We compute the z-score for each probe independently over all samples and independently for all donors and to identify the brain structures whose expression levels fall into the tails of the normal distribution. Thereby, each gene has a unique z-score for each individual brain structure in each donor. We used a cut-off of '-2.0 ≥ z-score ≥ 2.0' to retain only the significantly differentially expressed genes in the matrix for each of the six brains independently. Cells in the matrix with missing values (that do not meet the z-score cutoff) were populated with 'zeroes'. Subsequently, we designed four different pattern selection criteria (summarized in Figure 5) to select reproducible gene expression patterns across six whole brain gene expression matrices. Once, we had identified the reproducible genes across six individuals based on their expression profile in discrete anatomical structures we averaged the z-scores. Now, we had a blueprint expression for each gene for every brain structure. However, in the microarray data, each gene is assayed with multiple probes (58,692 probes covering 29,165 genes). To select the best representative probe for each gene, the probe that showed maximum variance across brain regions was selected (140). The general schematic of the expression model is summarized in the flowchart in Figure 6.



Structure ID	Slab type	Structure acronym	Structure name	MRI voxel x	MRI voxel y	MRI voxel z
4114	CX	AnG-i	angular gyrus, Left, inferior bank of gyrus	125	87	177
4114	CX	AnG-i	angular gyrus, Left, inferior bank of gyrus	133	87	167
4114	CX	AnG-i	angular gyrus, Left, inferior bank of gyrus	150	75	155
4114	CX	AnG-i	angular gyrus, Left, inferior bank of gyrus	153	71	146
4113	CX	AnG-s	angular gyrus, Left, superior bank of gyrus	126	84	177
4113	CX	AnG-s	angular gyrus, Left, superior bank of gyrus	120	57	155
4113	CX	AnG-s	angular gyrus, Left, superior bank of gyrus	134	59	146
4113	CX	AnG-s	angular gyrus, Left, superior bank of gyrus	127	80	168

Table 2 - Structural annotation based amalgamation of brain regions

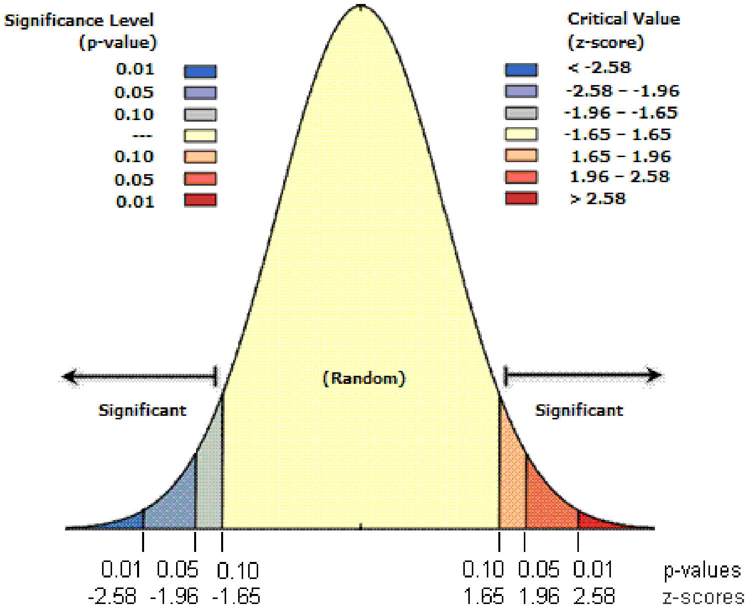
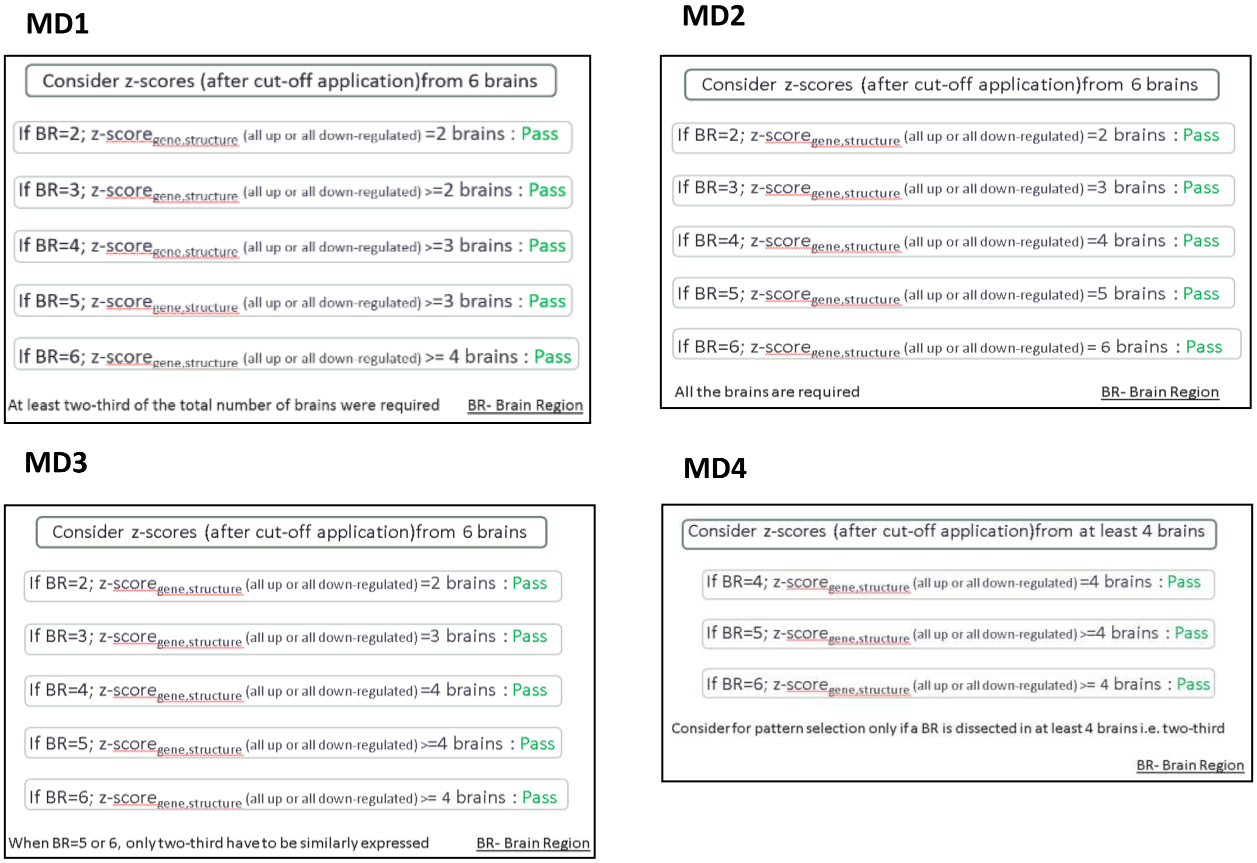


Figure 4 - z-score statistical representation and its relation to p-value



MD1

Consider z-scores (after cut-off application)from 6 brains

If BR=2; z-score$_{gene,structure}$ (all up or all down-regulated) =2 brains : Pass

If BR=3; z-score$_{gene,structure}$ (all up or all down-regulated) >=2 brains : Pass

If BR=4; z-score$_{gene,structure}$ (all up or all down-regulated) >=3 brains : Pass

If BR=5; z-score$_{gene,structure}$ (all up or all down-regulated) >=3 brains : Pass

If BR=6; z-score$_{gene,structure}$ (all up or all down-regulated) >= 4 brains : Pass

At least two-third of the total number of brains were required

BR- Brain Region

MD2

Consider z-scores (after cut-off application)from 6 brains

If BR=2; z-score$_{gene,structure}$ (all up or all down-regulated) =2 brains : Pass

If BR=3; z-score$_{gene,structure}$ (all up or all down-regulated) =3 brains : Pass

If BR=4; z-score$_{gene,structure}$ (all up or all down-regulated) =4 brains : Pass

If BR=5; z-score$_{gene,structure}$ (all up or all down-regulated) =5 brains : Pass

If BR=6; z-score$_{gene,structure}$ (all up or all down-regulated) = 6 brains : Pass

All the brains are required

BR- Brain Region

MD3

Consider z-scores (after cut-off application)from 6 brains

If BR=2; z-score$_{gene,structure}$ (all up or all down-regulated) =2 brains : Pass

If BR=3; z-score$_{gene,structure}$ (all up or all down-regulated) =3 brains : Pass

If BR=4; z-score$_{gene,structure}$ (all up or all down-regulated) =4 brains : Pass

If BR=5; z-score$_{gene,structure}$ (all up or all down-regulated) >=4 brains : Pass

If BR=6; z-score$_{gene,structure}$ (all up or all down-regulated) >= 4 brains : Pass

When BR=5 or 6, only two-third have to be similarly expressed

BR- Brain Region

MD4

Consider z-scores (after cut-off application)from at least 4 brains

If BR=4; z-score$_{gene,structure}$ (all up or all down-regulated) =4 brains : Pass

If BR=5; z-score$_{gene,structure}$ (all up or all down-regulated) >=4 brains : Pass

If BR=6; z-score$_{gene,structure}$ (all up or all down-regulated) >= 4 brains : Pass

Consider for pattern selection only if a BR is dissected in at least 4 brains i.e. two-third

BR- Brain Region

Figure 5 - Pattern selection criteria to discover reproducible gene patterns in identically annotated brain structures across individuals to generate four distinctive models

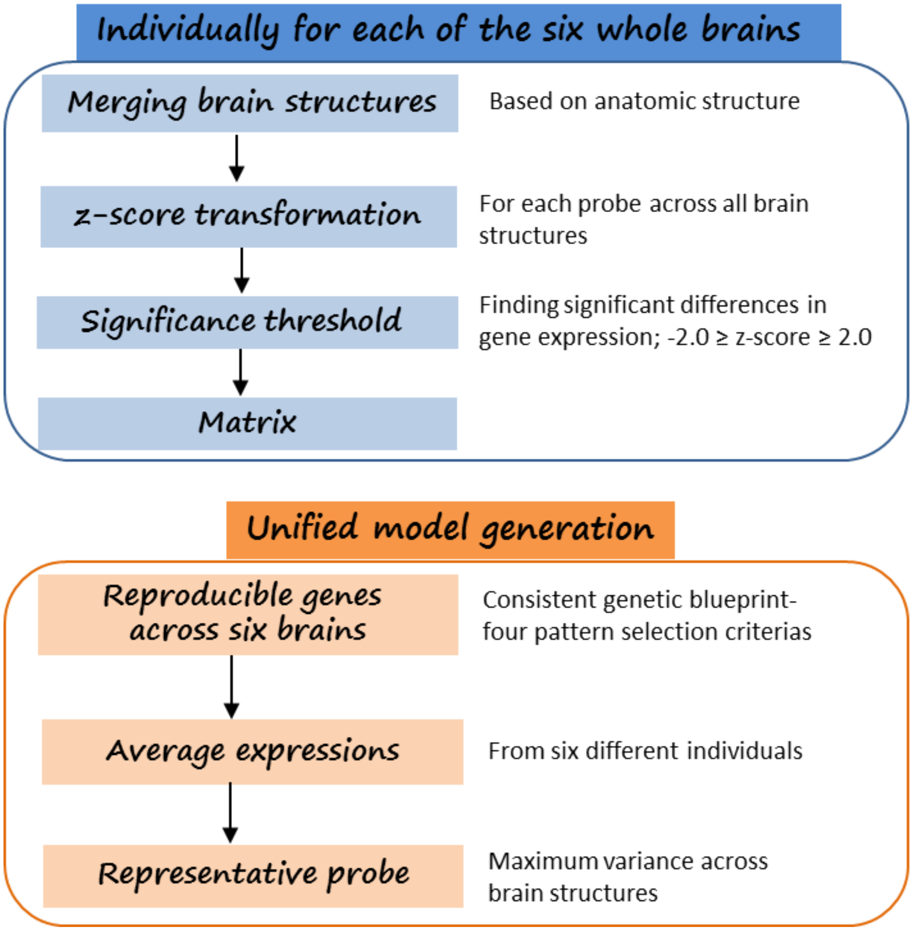

Individually for each of the six whole brains
Merging brain structures
Based on anatomic structure
z-score transformation
For each probe across all brain structures
Significance threshold
Finding significant differences in gene expression; -2.0 ≥ z-score ≥ 2.0
Matrix
Unified model generation
Reproducible genes across six brains
Consistent genetic blueprint- four pattern selection criterias
Average expressions
From six different individuals
Representative probe
Maximum variance across brain structures

Figure 6 - Schematic of the gene expression model generated for the healthy adult human brain

Hierarchical clustering

The hierarchical clustering algorithm used is based closely on the average-linkage method of Sokal and Michener (141), which was developed for clustering matrices such as those used here. The object of this algorithm is to compute a dendrogram that assembles all elements into a single tree. For any set of n genes, an upper-diagonal similarity matrix is computed by using the metric described above, which contains similarity scores for all pairs of genes. The matrix is scanned to identify the highest value (representing the most similar pair of genes). A node is created joining these two genes, and a gene expression profile is computed for the node by averaging observation for the joined elements (missing values are omitted and the two joined elements are weighted by the number of genes they contain). The similarity matrix is updated with this new node replacing the two joined elements, and the process is repeated n-1 times until only a single element remains. Software implementation of this algorithm can be obtained from the authors at 'http://bonsai.hgc.jp/~mdehoon/software/cluster/' and the clustered files were visualized using 'TreeView' software (142).

Hierarchical clustering of all the 191 samples (brain structures) was performed using uncentered correlation as the similarity measure and average linkage as the clustering method. Two-way clustering was used, where both the genes and the samples were clustered. Since, clustering the full complement of genes diminishes the significance of brain specific gene expression; we applied a cut-off to our gene set to focus on genes that are highly biologically relevant with enrichment for brain-related annotations and disease associations. A gene had to be significantly differentially expressed in at least five brain structures to be included in the gene set. This set of genes was also typically associated with high expression variance. This reduced the number of genes to 6,984 and also the sample size of distinct brain structures was reduced to 166. Same clustering algorithm with identical parameters was applied to this reduced but theoretically brain specific genes.

Clusters of genes coordinately expressed were achieved and we used ‘DAVID’ (143) software for the characterization of the clusters. The functional characterization was also validated using a second tool named WebGestalt (144).

Weighted gene coexpression network analysis (WGCNA)

WGCNA (145) was performed in the R and a coexpression network was constructed on the basis of 6,984 genes. For all possible pairs of the variable genes, Pearson correlation coefficients were calculated across all samples. The correlations matrix was raised to the power 9, thus producing a weighted network. The weighted network was transformed into a network of topological overlap (TO)—an advanced coexpression measure that considers not only the correlation of 2 genes with each other, but also the extent of their shared correlations across the weighted network. Genes were hierarchically clustered on the basis of their TO. Modules were identified on the dendrogram using the Dynamic Tree Cut algorithm. For each gene, we determined its connectivity within its module of residence by summing up the TOs of the gene with all the other genes in the module. By definition, highly connected (hub) genes display expression profiles highly characteristic for their module of residence (146). To obtain a condensed representative expression profile of each module (module eigengene, ME), we summarized expression levels of the top hub genes in the module.

Functional annotation of the modules was performed on the basis of analysis of their gene composition. We used DAVID (http://david.abcc.ncifcrf.gov/) to test each module for enrichment in genes with particular GO terms and biological pathways compared with the background list of all genes on the array. The functional characterization of the modules was also authenticated using another tool named 'WebGestalt'.

Neurotransmitter maps

We used Kyoto Encyclopedia of Genes and Genomes (KEGG) (147), the most comprehensive database source that integrates genomic, proteomic and systemic functional information for investigating the distribution of neurotransmitter receptors and generate their pathway maps. The KEGG Pathway suite is a collection of manually drawn maps demonstrating the existing knowledge on the molecular interaction and reaction networks. KEGG pathways were used as reference pathways to map human neurotransmitter systems

AutDB

AutDB (38) is a publicly available web-portal for on-going collection, manual annotation and visualization of genes linked to autism. We downloaded all the 845 genes associated with ASD, including both rare mutations and common variants from the AutDB database. Genes having a predisposition to autism in the context of a syndromic disorder and genes demonstrating strong evidence for replication in an independent experiment after a rigorous statistical comparison between cases and controls were retained. All the other genes with minimal evidence and hypothesized but untested evaluations were filtered out. Finally, we had a list of 219 autism implicated genes with high confidence as summarized in Table 3

C15orf43	ABCA10	ERBB2IP	MBD3	RBFOX1	TBC1D5	HTR1B
CA6	ADORA2A	FBN1	MDGA2	REEP3	TBL1X	IL1R2
CTNND2	AGBL4	FBXO40	MED13L	RELN	TCF7L2	INTS6
CYP11B1	AHI1	FHIT	MFRP	RNF135	TERF2	KATNAL2
DDX53	ANK2	GLIS1	MIB1	ROBO2	TRIO	LRRC1
ELAVL3	APH1A	GLO1	MNT	RPS6KA2	TRIP12	MBD5
FABP5	ASMT	GNB1L	MSR1	RPS6KA3	TRPC6	MET
FAM92B	ASTN2	GPR37	MYO16	SBF1	TSC2	MYO9B
GAS2	ASXL3	GPX1	NAALADL2	SCFD2	TSPAN7	NCKAP1
GRID2	BRCA2	GRIP1	NBEA	SCN4A	UBE2H	NF1
HSD11B1	CACNA1F	GUCY1A2	NDUFA5	SDC2	UBR5	NLGN1
IL1RAPL1	CACNB2	HDAC4	NINL	SEMA5A	UPF3B	NUAK1
KCNQ3	CCDC91	HEPACAM	NLGN4Y	SETD5	USP45	PACS1
KIRREL3	CDH9	HLA-B	NSD1	SETDB1	VASH1	PRODH
MYT1L	CHKB	HRAS	NTNG1	SGSH	VPS13B	PTK7
NLGN3	CLTCL1	HS3ST5	NTRK3	SGSM3	VSIG4	RAPGEF4
NRXN1	CNR1	HTR3A	ODF3L2	SLC1A1	WAC	RPL10
NRXN2	CTCF	HYDIN	OR2M4	SLC22A15	ZBTB20	SCN1A
RIMS3	CTNNA3	ICA1	P2RX5	SLC27A4	ZMYND11	SETBP1
SLC12A5	CTTNBP2	ILF2	P4HA2	SLC38A10	ZNF559	SHANK3
SLC1A2	CUL3	ITGB3	PARK2	SLC7A3	ADK	SNX14
SLC4A10	CUL7	JMJD1C	PAX5	SLC9A6	ATP10A	ST7
STXBP1	DAPP1	KAT2B	PAX6	SMARCA2	CACNA1D	SUV420H1
SYN1	DDX3X	KAT6A	PCDH15	SMARCC2	CACNA2D3	TAF1
TTN	DIP2A	KCNQ2	PER1	SND1	CHD2	TBX1
ARX	DLX2	LAMA1	PIK3R2	SNTG2	DEAF1	TMLHE

CNTNAP2	DMD	LAMB1	PLCB1	SOX5	DNMT3A	TSC1
LEP	DMPK	LRBA	PPP1R1B	SPARCL1	EIF4E	UBE3A
DMXL2	LRP2	PRICKLE2	SPAST	USP7	DRD3	SYT17
DNER	LZTS2	PRKCB	SSPO	WDFY3	DVL3	MACROD2
DPP10	EP400	PRKD1	STXBP5	FBXO33	DYRK1A	PTEN
DPP4	MBD1	PTCHD1	STYK1	FOXP1	EHMT1	PTPN11
GIGYF2	EN2	GRIK2	HECW2	RAB39B		

Table 3 - List of 229 autism implicated genes curated from AutDB database

The Human Protein Atlas

The Human protein atlas (148) provides expression for all protein-coding genes in all major tissues and organs in the human body including the brain. A total of 1223 genes were shown to demonstrate elevated expression in brain compared to other tissue types in the Human Protein Atlas. Of these 1223 genes we filtered out any overlapping genes in the autism dataset derived from AutDB. Also, the ‘group enriched’ genes (genes which shown at least five-fold higher mRNA levels in a group of 2-7 tissues) as defined by the protein Atlas were filtered out, leaving us with a set of 830 genes with elevated expression in the human brain.

Development of Supervised Classification Model

To carryout feature selection, classification model generation using ML algorithms and performance measurements, we used the Waikato Environment for Knowledge Analysis (WEKA) (149) framework, which is an open-source, Java-based framework. We used three diverse and most popular ML algorithms; namely RF (150), BayesNet (151), J48 (152) and SVM, to build classification models.

To evaluate the performance of the method, we apply a standard validation technique called ten-fold cross validation, where sequences from each class are divided into ten parts – the model is built using nine parts, and predictions are generated and evaluated on the data contained in the remaining part. This process is repeated for all ten possible combinations. After cross-validation, we assessed the performance of the fully trained classifier models using the test set (20% of original data) that were hidden from the classifiers. We report standard performance measures over each enzyme class including the following:

- True positives (TP) - the number of sequences that are correctly identified in a class that belongs to them;
- False negatives (FN) - the number of sequences that are not identified in a class that belongs to them;
- True negatives (TN) - the number of sequences that are not found in a class that does not belong to them;
- False positives (FP) - the number of sequences that are identified in a class that does not belong to them;

Using these four quantitative measures, we report a number of standard measurements in judging classifier performance:

- *Overall Accuracy* – a measure of the overall classifier performance. It is defined as the fraction of the data tested that is classified correctly. Though it is a poor measure to consider on highly unbalanced datasets, it is still reported it as a general overall comparative measure
- *Sensitivity* (a.k.a Recall, TP-rate) - the proportion of true positives that are predicted as positives. This gives a measure of individual class accuracy. Poor sensitivity indicates that the classifier is under-predicting class
- *Specificity* - the proportion of true negatives that are predicted as negatives.
- *False Positive Rate* – the fraction of data not in a class that was incorrectly predicted to be in that class. The sensitivity and specificity are given by,

$$\text{Sensitivity} = \text{TP}/\ (\text{TP} + \text{FN});$$

$$\text{Specificity} = \text{TN}/\ (\text{TN} + \text{FP}).$$

False Positive Rate = 1-Specificity

We optimize and validate the accuracy of our prediction model by selecting the optimal model that has maximum true positive rate (sensitivity) and minimum false positive rate (1- specificity). A receiver operating characteristic (ROC) curve depicts the relationship between specificity and sensitivity for a single class. The ROC curve for the perfect classifier would result in a straight line up to the top left corner, and then straight to the top right corner, indicating that a single score threshold can be chosen to separate all of the positive examples of a class from all of the negative examples. Each point in the curve is plotted based on different confidence score thresholds. The area under the curve (AUC) is a numeric measure of performance depicted by ROC curves.

CHAPTER 4

RESULTS

Model generation

Generation of gene expression model

To identify genes with expression pattern sustenance across healthy adult human brains, we developed four versions of the gene expression models using microarray data from the ABA (26). The profiles of six donors from ABA representing their age, gender, ethnicity, handedness and the number of samples obtained per left hemisphere has been summarized in Table 2. Briefly, five males and one female brain of age range 24 years to 57 years who died a natural death constitute the donors in the ABA. About 400- 500 tissue samples per hemisphere were dissected for microarray data generation from each of the donors. Due to the normal variation in the brain size and morphology, some of the brain structures were sampled multiple times. Since these brain samples map back to the same neuroanatomical structural annotation, we averaged their expression values to achieve a unique anatomical annotation.

For the purposes of data consistency, left hemisphere of the brain were used for current study as the right hemispheric microarray data were available for only two donors. Also, the brain structures that were sampled in any one individual brain only were not considered to be part of the model. Finally, comparison was made across a total of 212 brain structures expanding over 6 individuals. The fundamental principles to quantify each of the 58,692 probes with their differential expression in every brain structure with reference to all the remaining brain structures are steadfast for each donor. However, four distinctive criteria implemented to find globally conserved transcriptional profiles for precise anatomical locations across the six donors as summarized in the flowchart Figure 6. These criteria (Figure 5) differ in capturing the gene patterns in identically annotated brain substructures across individuals (increasing pattern selection stringency from MD1 to MD4). Recapitulation of expression in identically labeled brain structures for every gene across individuals was crucial in establishing reliable expression profiles.

For our first model (MD1), we observed that about 85% (24,863) of the genes demonstrate synonymous expression profile in at least one brain structure for our first model (MD1). These results corroborate with the results from previous studies which have shown 84% of genes to be expressed in the adult human brain (17) and 80% in the mouse brain (70). Comparison of expression patterns resulted in 190 brain structures carrying similarly pattered gene expression for at least one gene. The frequency distribution of genes with reproducible expression patterns across distinct brain structures for MD1 is shown in Figure 7. Our results demonstrate the consistency of expression patterns across six individuals in all four models of healthy individuals. However, due to the stringency of filtering criteria, the frequency distribution of similarly expressed genes across distinct brain structures were reduced from MD1 in all the other models. About 54% of the genes are expressed in at least one brain structure in MD2, 69% in MD3 and for our last model MD4, only 59% of the genes were expressed in at least one brain structure (Figure 7). All the following results in the current study are based on MD1. It is the optimal selection since MD1 ensures minimum loss of gene expression information and retains the most number of brain structures with differentially expressed genes when compared to the other three models. Number of unique brain structures passing the gene expression pattern selection criteria implemented by each of the four different models has been shown in Figure 8.

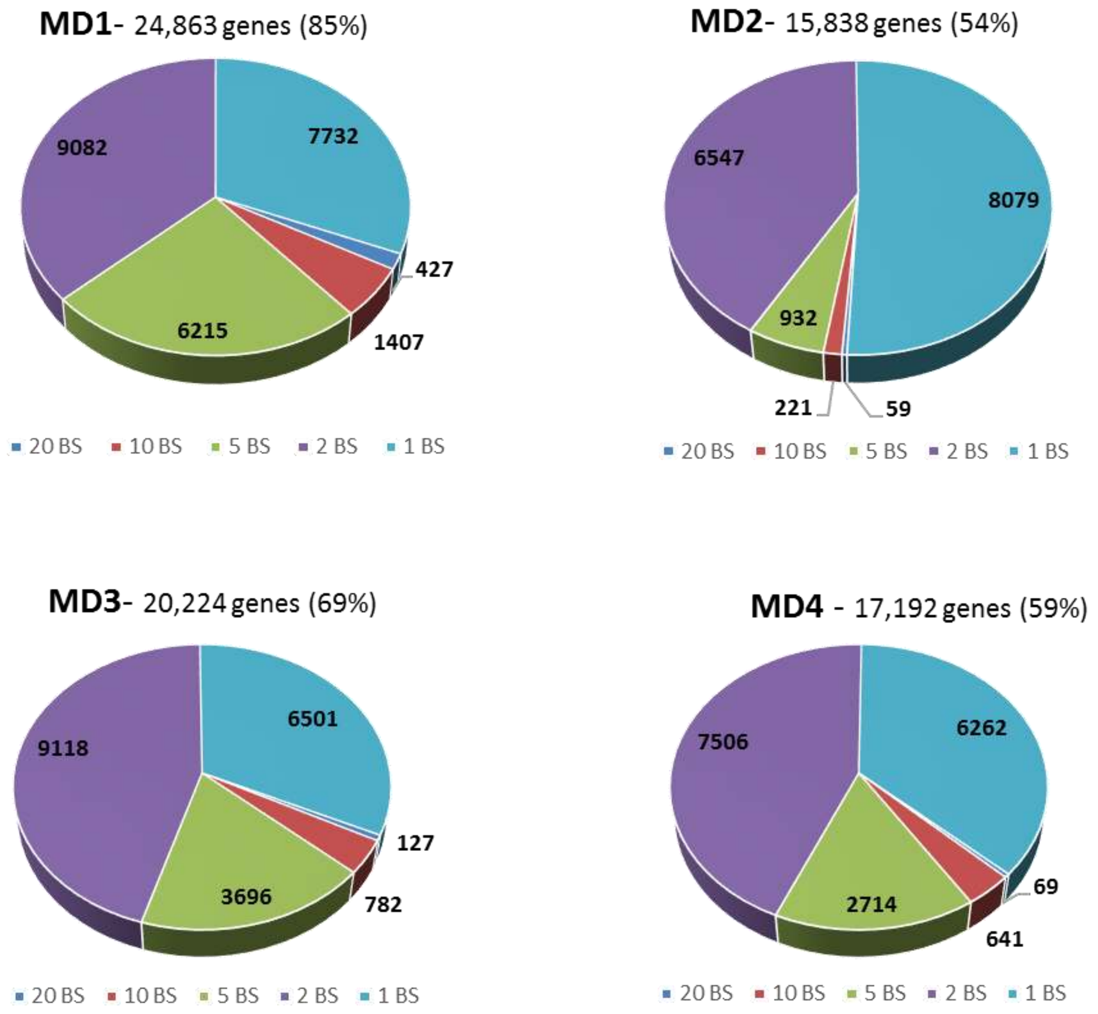


Figure 7 - Frequency distribution of reproducible genes across distinct brain structures (BS) for MD1, MD2, MD3 and MD4

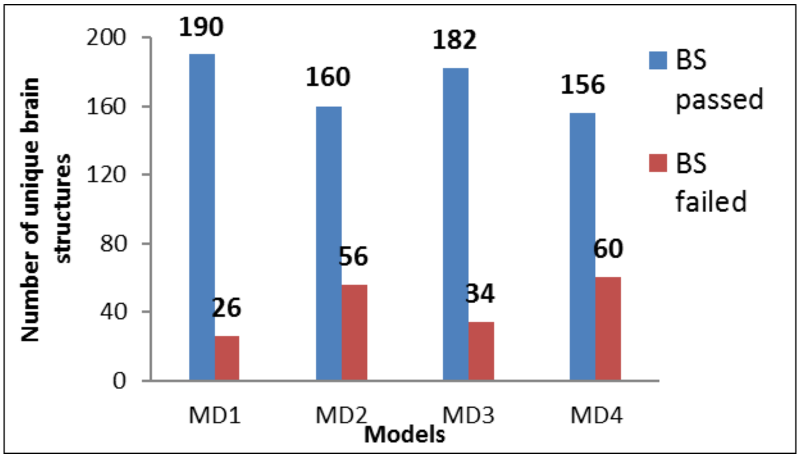


Figure 8 - Number of unique brain structures passing the gene expression pattern selection criteria implemented by each of the four different models

Brain structure based exhibit of gene expression established the blood brain barrier (CPLV) (153, 154) as the most transcriptionally active brain structure followed closely by paraventricular nuclei and corpus callosum. The bar graph in Figure 9 reports the top 10 brain structures that have the most number of differentially expressed genes among all the models. Across the four models, very often, the same brain structures retain the most number of differentially expressed genes, however, with observable differences in the gene numbers. Upon inspecting, most of the genes conserved across different models for precise brain structures were mostly related to the fundamental functioning of the human brain; with functions highly specific to the brain and generic cell process.

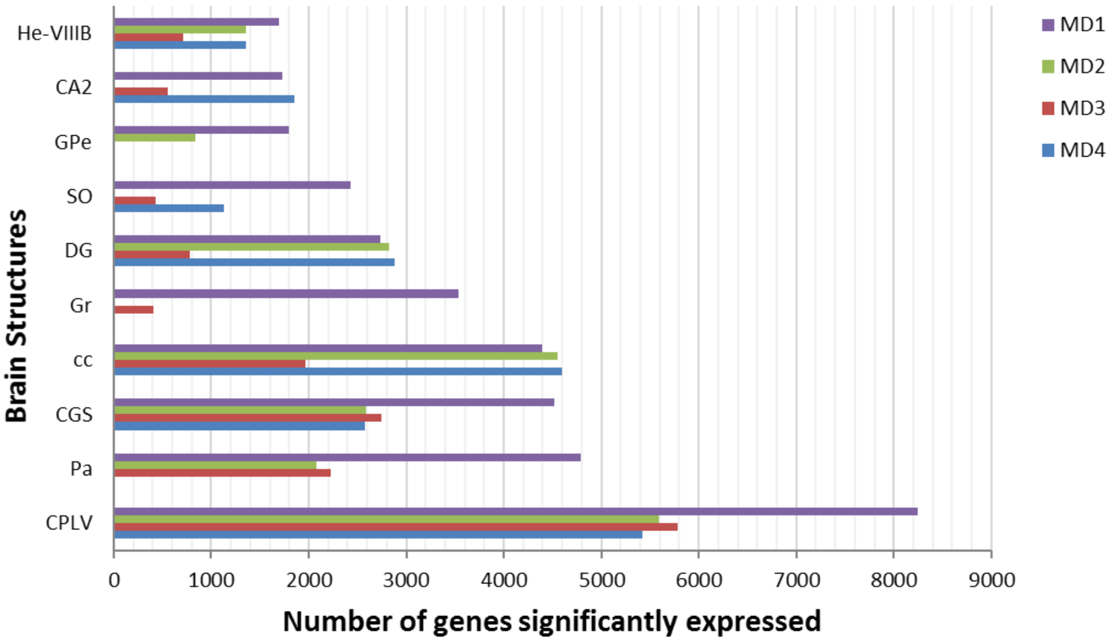


Figure 9 - Brain structure based gene enrichment. Structures with significant differential expression for maximum number of genes across four models

Functional genomics of the healthy adult human brain

Spatial organization of transcriptome

Hierarchical clustering (79) of gene expression data provides a holistic view of the transcriptome organization in a full set of brain samples. To gain better insights into the structural and functional similarities between distinct anatomical locations, we checked for spatial grouping of 190 brain structures based on the expression of 6,984 genes showing significant differential expression in at least five brain structures Figure 10a. Clustering using the full complement of expressed genes has been reported in the Figure 10b.

As shown in Figure 10a, using hierarchical clustering we were able to show that distinct sub-structures of a bigger brain region can cluster together while still retaining their expression identity. Sets of genes expressed largely across the brain were identified, suggesting housekeeping functions. Cliques of genes in functionally related brain structures were also identified. Besides, we also observed sets of genes that express only in a singular brain location, signifying their functional significance in discreet brain locations. By performing enrichment analyses on these gene groups, we identified biological processes and molecular functions which are either specific to the brains functional and anatomical groups or represent generic cellular processes.

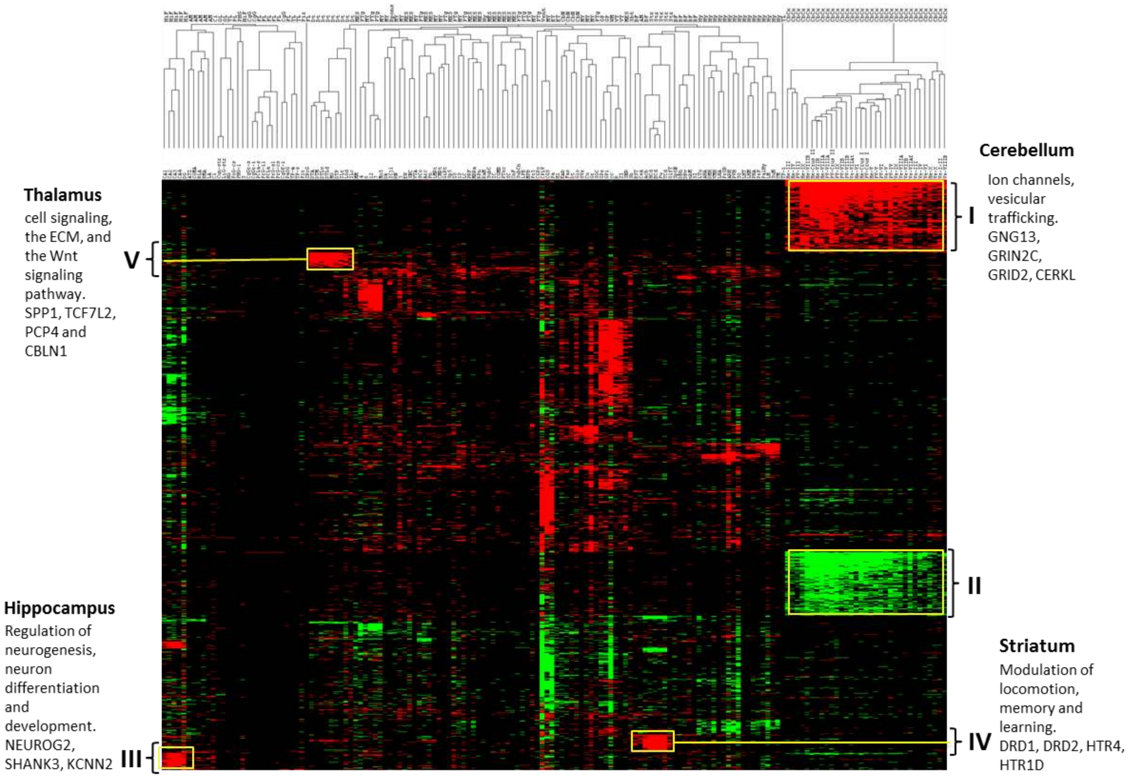


Figure 10a - Dendrogram and heat-map overview of the two-way unsupervised hierarchical cluster analysis of gene expression data from 160 samples using 6,984 genes. Columns represent individual brain structures and rows represent each gene and the z-scores were calculated across rows. The expression level of each feature (gene) in every brain structures is represented as a cell in a two dimensional matrix. Red and green reflects high and low expression levels, respectively, in a given brain structure with respect to all the other brain structures.

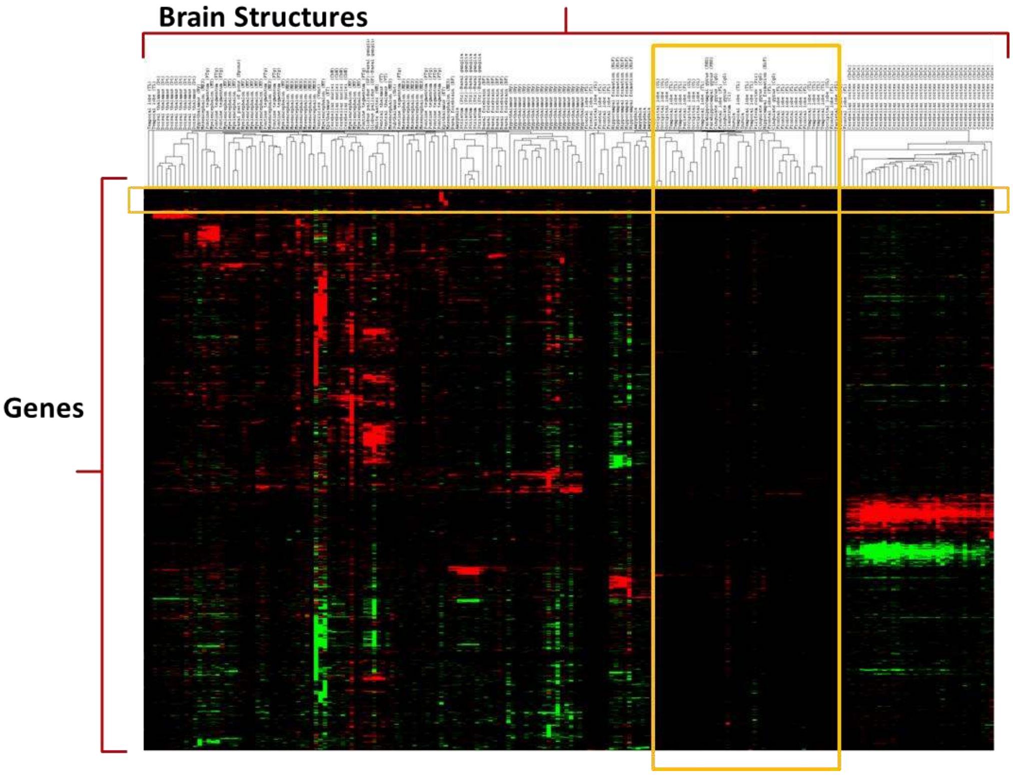


Figure 10b- Dendrogram and heat-map overview of the two-way unsupervised hierarchical cluster analysis of gene expression data from 190 samples using full complement of genes

Most genes across brain regions are not selective for a single major brain region; rather, they are expressed in multiple regions in a non-uniform fashion. This suggests that a number of genes are pleiotropic to brain functionality (17). Our data suggests that the cerebellum has the most distinguishable gene expression patterns of all the brain regions and displays the least internal heterogeneity (8, 3, and 155). The total number of genes clustered in cerebellum is about 1,600. Most of these genes when studied with enrichment analysis showed the enrichment for ion channels as the dominant GO classes for the upregulated gene cluster "I" (Figure 10a), which consists of 865 genes. It has been previously reported that the ion channels are enriched in cerebellum (156, 157). Also, enrichment of genes associated with vesicular trafficking as well as E3-ubiquitin-protein ligases that play a role in DNA damage signaling was found in our results for cluster "I" and is corroborated with existing knowledge (158). Critical role of transmembrane ion flux via transporters and channels in various functions of cerebellum is well established, including neuronal signal transmission and electrolyte homeostasis (159), and we found our results are in support of this concept. Some of the major gene players in the upregulated cerebellum cluster were GNG13, CERKL, GRIN2C, and GRID2. For example, the upregulation of GRIN2C in the adult cerebellum has been previously reported during the innervation of mosey fibers into granule cells (160, 161).

About 740 genes were downregulated in cerebellum in the cluster labelled "II". We determined that the transcriptome of cerebellum possess a rich homogenous gene expression structure which might reflect the underlying cellular composition of the brain tissue. Other brain structures that show conserved patterns include hippocampus, amygdala, hypothalamus, dorsal thalamus, striatum and cerebellar nuclei. With the exception of the above mentioned brain structures, not many genes showed differential expression amongst the cortical regions such as occipital, parietal, frontal and temporal lobes and these results are agreeable with existing literature (8). The hippocampal cluster labeled “III” (Figure 10a) showed significant over-representation of terms like neuron differentiation, neuron development, cell morphogenesis involved in neuron differentiation, regulation of neurogenesis etc. This is expected because hippocampus is the site for adult neurogenesis (162). Genes such as NEUROG2, a transcriptional regulator involved in neuronal differentiation (162, 163); SHANK3, a gene known to be implicated in schizophrenia and autism (164); and GRIA1 and KCNN2 were all part of the cluster “III”. Dorsal thalamus showed association with nicotine related GO categories such as nicotine acetylcholine gated receptor-channel complex and behavioral response to nicotine. Local clusters for striatum and its sub-divisions show high expression for dopamine receptors DRD1, DRD2, DRD3. Striatum is a brain structure where dopamine exerts its maximum effect as the dopamine producing neurons have their cell bodies in substantia nigra, which projects into the striatum (165). High expression of HTR isoforms were also found in striatum. HTR and its isoforms are known to regulate the release of dopamine and regulation of extracellular dopamine, thereby affecting the neural activity (166). Significant enrichment of PDE10A was found in the striatum sub-structures. An association has been established between the striatal expression of PDE10A gene and bipolar disorder patients (167). The highly upregulated cluster in hypothalamus was enriched in molecular functions such as hormone activity and response to endogenous stimulus. Most prominent genes in this category were TRH, CRHBP, and GHRH. Other GO categories showing over-representation in hypothalamus included steroid hormone stimulus, neuropeptide hormone activity, response to corticosteroid stimulus, estrogen stimulus and response to glucocorticoid stimulus.

Taken together, the transcriptional profiles of the sub-divisions of bigger structural units are well conserved across six individuals; thus demonstrating transcriptional similarity amongst functionally related sub-structures. Also, for both the full complement of expressed genes as well as the smaller set of 6,984 genes, the supervised hierarchical clustering is satisfactorily robust owing to the distinctive expression patterns.

Co-expression network construction

There are numerous ways to analyze multi-dimensional gene expression data; however, correlation networks provide a comprehensive outlook on the intrinsic organization of a transcriptome. Gene co-expression networks investigate gene-to-gene relationships in an unsupervised way and cluster coordinately expressed genes into modules. This provides a framework to better understand gene expression patterns in distinct brain structures, which may be driven by distinct cellular and biological processes (Figure 11). WGCNA is a package in R that helps construct such networks.

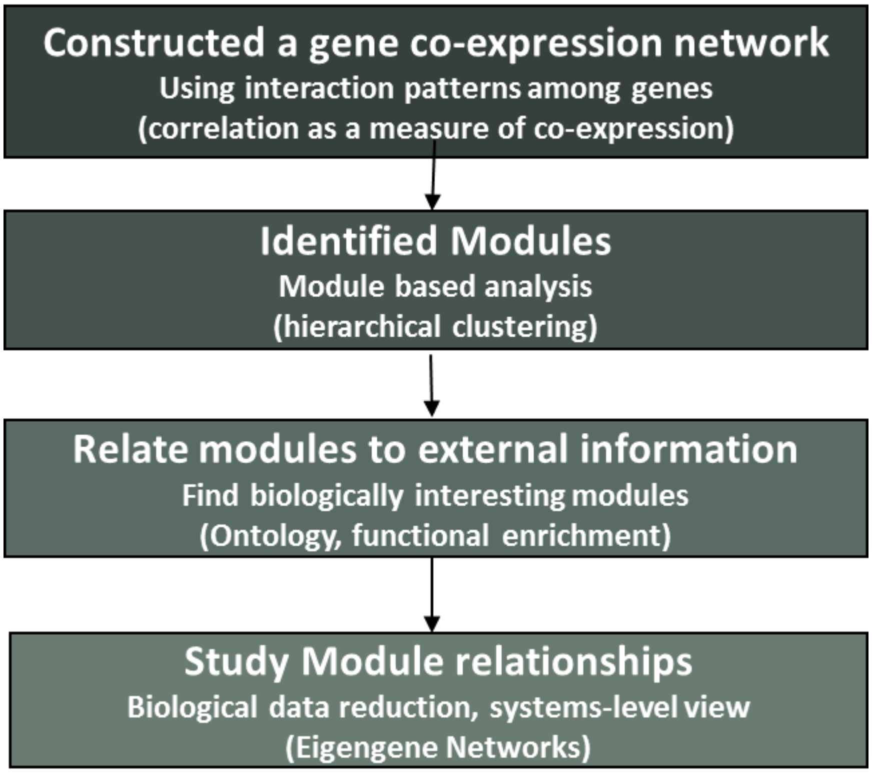

Constructed a gene co-expression network
Using interaction patterns among genes
(correlation as a measure of co-expression)
Identified Modules
Module based analysis
(hierarchical clustering)
Relate modules to external information
Find biologically interesting modules
(Ontology, functional enrichment)
Study Module relationships
Biological data reduction, systems-level view
(Eigengene Networks)

Figure 11 - Overview of weighted gene co-expression network analysis methodology

Same set of 6,984 genes as previously discussed was used to construct the weighted gene coexpression network. Based on the TO (topological overlap), WGCNA evaluated the coexpression for every pair of genes, while simultaneously considering the degree of shared neighbors for every gene pair across the whole network. This results in the discovery of consistent gene coexpression patterns in the transcriptome. Our analysis resulted in 19 groups (modules) of highly co-expressed genes (Figure 12). Modules are groups of genes exhibiting high intra-module topological overlap. Each module was assigned a unique color and the number of genes assigned to a module varies, ranging from 67 to 1469. Each module is prefixed with a module eigengene (ME). To further explore the co-expression relationships amongst the distinct modules, the first principal component of all the modules was summarized. To identify genes with the highest connectivity inside a module, the eigengene based connectivity measure (kME) was used. kME measures the strength (0 to 1) of connection based on gene co-expression in a given module. Genes participating towards high module membership or highly connected genes are referred to as hub genes. Hub genes are representative of their resident module and point towards key, biological processes for the module.

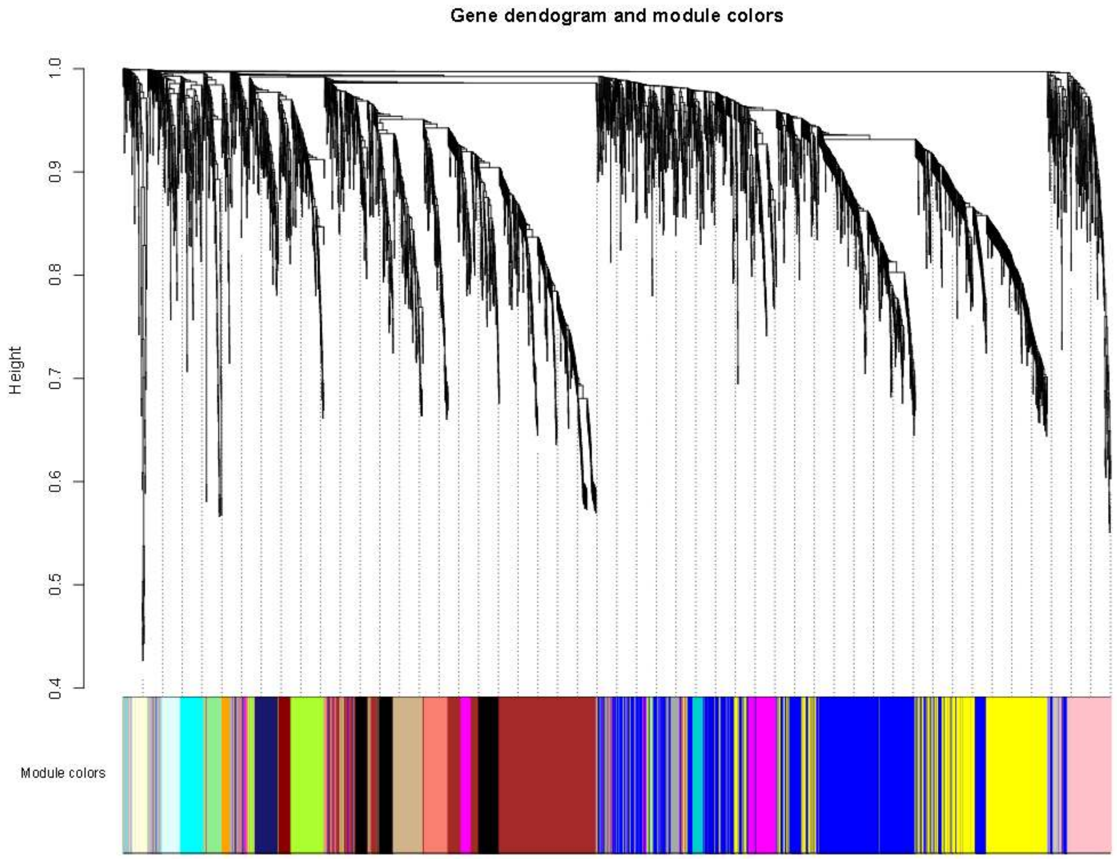
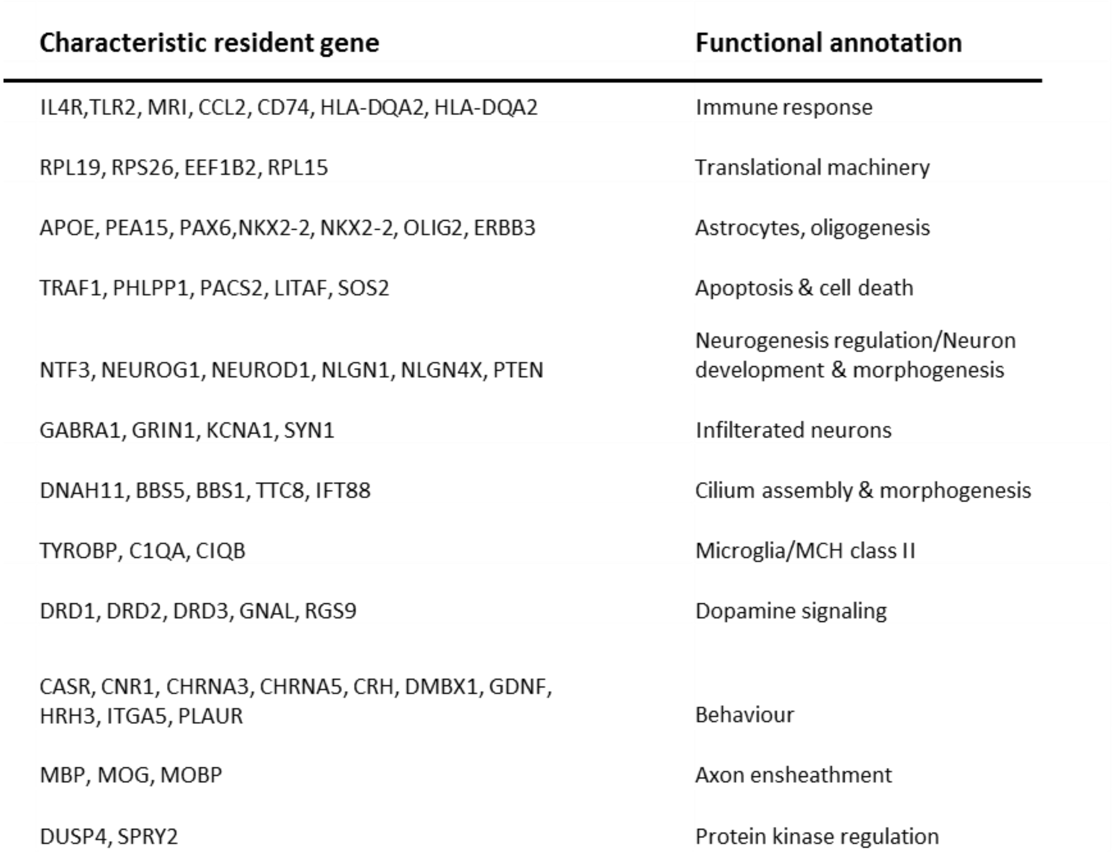


Figure 12 - Brain transcription coexpression network and gene modules. The genes were clustered by expression patterns as represented by the dendrogram and correlation heat map

Biological relevance of network modules

The functional relevance of the gene module was assessed to make sure that the modules designed by co-expressed genes convey biologically relevant information. Gene ontology enrichment analysis was performed to examine the ontology terms over-represented in the modules. Also, KEGG orthology based enrichment analysis was performed using hypergeometric tests for each module. A wide range of functional association was configured with the statistical analyses of gene composition, which can be grouped into several categories as summarized in Table 4.



Characteristic resident gene	Functional annotation
IL4R,TLR2, MRI, CCL2, CD74, HLA-DQA2, HLA-DQA2	Immune response
RPL19, RPS26, EEF1B2, RPL15	Translational machinery
APOE, PEA15, PAX6,NKX2-2, NKX2-2, OLIG2, ERBB3	Astrocytes, oligogenesis
TRAF1, PHLPP1, PACS2, LITAF, SOS2	Apoptosis & cell death
NTF3, NEUROG1, NEUROD1, NLGN1, NLGN4X, PTEN	Neurogenesis regulation/Neuron development & morphogenesis
GABRA1, GRIN1, KCNA1, SYN1	Infilterated neurons
DNAH11, BBS5, BBS1, TTC8, IFT88	Cilium assembly & morphogenesis
TYROBP, C1QA, CIQB	Microglia/MCH class II
DRD1, DRD2, DRD3, GNAL, RGS9	Dopamine signaling
CASR, CNR1, CHRNA3, CHRNA5, CRH, DMBX1, GDNF, HRH3, ITGA5, PLAUR	Behaviour
MBP, MOG, MOBP	Axon ensheathment
DUSP4, SPRY2	Protein kinase regulation

Table 4 - Functional annotation of the modules where the genes that are typically representative of their module of residence have been carefully noted in the table along with the top overrepresented functional terms

Notably, a significant over-representation of the dopamine receptor signaling pathway, G-protein coupled receptor signaling pathway and catecholamine binding was observed for the pink module consisting of 225 genes. Structurally, Amygdala and striatum sub-structures are the predominant regions accommodating the pink module genes. Role of dopamine in amygdala has been well studied specially with regards to D1 and D2 receptors (165, 168). Yellow module was enriched for biological processes like cell death, regulation of apoptosis and regulation of programmed cell death. Top genes for the yellow module included TRAF1, PHLPP1, PACS2, LITAF, PREX1 and SOS2. Also, the yellow module was enriched for chemokine signaling pathway and neurotrophin signaling pathway, suggesting reduced neuronal support to maintain homeostasis in the adult human body (169). The role of primary cilia in neuronal functions has been well recognized (170, 171). The top 200 genes from the brown module with the highest K-within (intramodular connectivity) showed significant enrichment of cilium organization and morphogenesis, and cilium assembly. The major gene players in the brown module included DNAH11, BBS5, BBS1, TTC8 and IFT88. Biological processes like neuron development & morphogenesis and regulation of neurogenesis were significantly enriched in the turquoise module, which also happens to be the largest with 1,469 genes. Genes like NEUROG1, NTF3, NEUROD1, and NLGN1 were the highly connected nodes in this module. Turquoise module genes could be mapped back predominantly to cerebellum and hippocampus. Blue module was most strongly enriched for GO terms such as immune response, microglia and MCH class II.

Neuroanatomically indigenous functional annotation

Since signatures of local phenomena can be masked by larger global signatures, we performed local analysis on smaller brain sub-structures of a bigger brain region. Here, we show in-depth biologically enriched similarities/differences between the sub-structures of hippocampus. For similar analyses on other brain structures please refer to supplementary data.

Gene ontology and pathway (Figure 13) enrichment was performed on each hippocampus sub-structure using the significantly differentially expressed genes, respective of each structure. One of the interesting findings was Neurogenesis as the top biological process for dentate gyrus (DG) and cornus ammonis (CA4) subfields. It has been previously established that adult hippocampus is the sight of neurogenesis (172); essentially DG and CA4. Furthermore, long-term potentiation (LTP) enhances the neurogenesis process in DG (173), and interestingly LTP was the leading pathway in our results. Upon inspecting the list of genes enriched in DG and correspondingly in LTP, we found AMPAR, Plcb1, Mapk1/3, Ascl1, Adcy1 and IP3 to be well represented in our data. We mapped these genes on the long-term potentiation pathway as shown in Figure 14. AMPAR is known for its role in postnatal hippocampal neurogenesis (174). During LTP, dendritic NMDA and AMPAR receptors are involved in the development of new synapses. Similarly, MAPK1/3 and Plcb1 have been documented (175) as genes regulating adult hippocampal neurogenesis and neuronal differentiation. Also, Ascl1 has been widely recognized to regulate gene expression during neurogenesis and neuronal differentiation (176) and Adcy sub-types have been connected with memory processes. Similarly, robust regional patterns of biological importance were observed in basal ganglia, striatum, amygdala and dorsal thalamus.

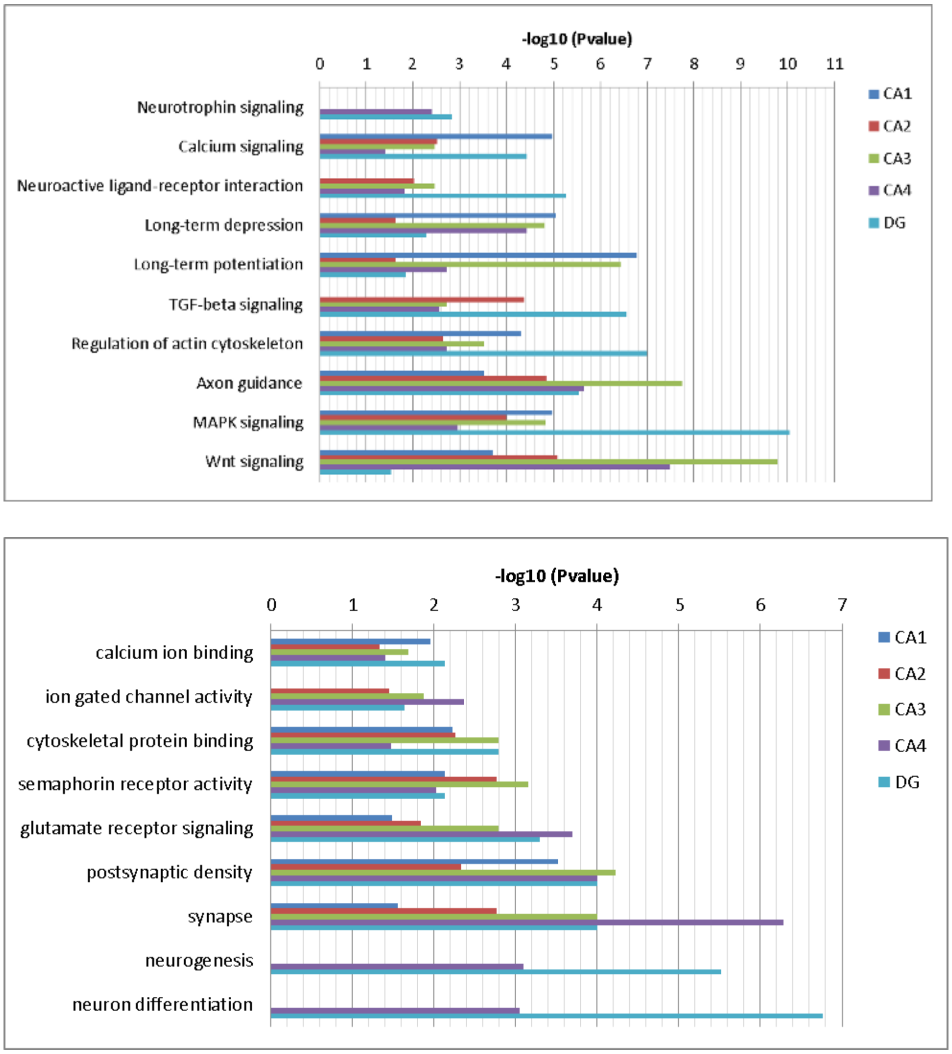


Figure 13 - Bar graphs summarizing the pathway and GO enrichment analysis for the different substructures of hippocampus

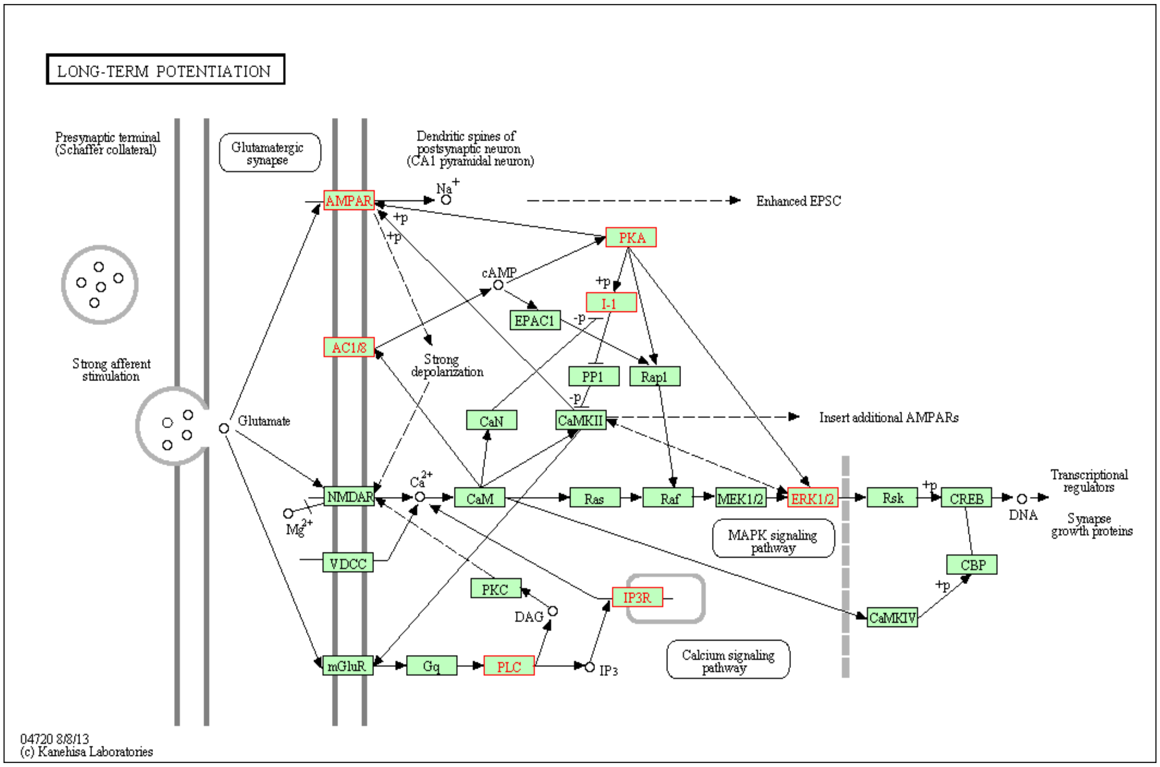


Figure 14 - Long term potentiation pathway mapped with the genes enriched in DG and known for their role in neurogenesis and other hippocampus specific functions. The genes of interest have been highlighted in red color

Neurotransmitter system maps

Neurotransmitters play a very important role in the overall machinery of brain function. Mapping the structural distribution of the major neurotransmitter receptors can provide novel and functionally more relevant insights into the spatial organization of the human brain. We mapped the pathway distribution of five major neurotransmitter systems- serotonin, dopamine, choline, GABA and glutamate.

In serotonergic synapse the major molecule serotonin (5-Hydroxytryptamine, 5-HT) is a monoamine neurotransmitter playing an important role in physiological functions such as learning and memory, pain, endocrine secretion, as well as states of abnormal mood and bad cognition (177). The serotonin 5-HT6 receptors are located primarily in the striatum (178), and receptor mapping in our study shows enrichment patterns consistent with the literature. We also found other HTR receptors like HTR7, HTR4, HTR1D and HTR1A to be significantly present in striatum. Another interesting finding was the enrichment of HTR1A and HTR1B in substantia nigra, hippocampus and hypothalamus. The detailed map of serotonin synapse participating molecules can be seen in Figure 15a.

For the dopaminergic synapse, a noticeable enrichment of tyrosine hydroxylase in substantia nigra pars compacta was observed as shown in Figure 15b. Dopamine serves as a precursor for noradrenaline for the neurons in these locations (17). Also, high expression of DRD1, DRD2 and DRD3 was seen in striatum. DRD2 was also enriched in substia nigra (SNC), ventral tegmental area (VTA) and hypothalamus. Since, VTA is the origin of the dopaminergic cell bodies of the dopamine system significant expression of DRD receptors is expected to be high. (179).

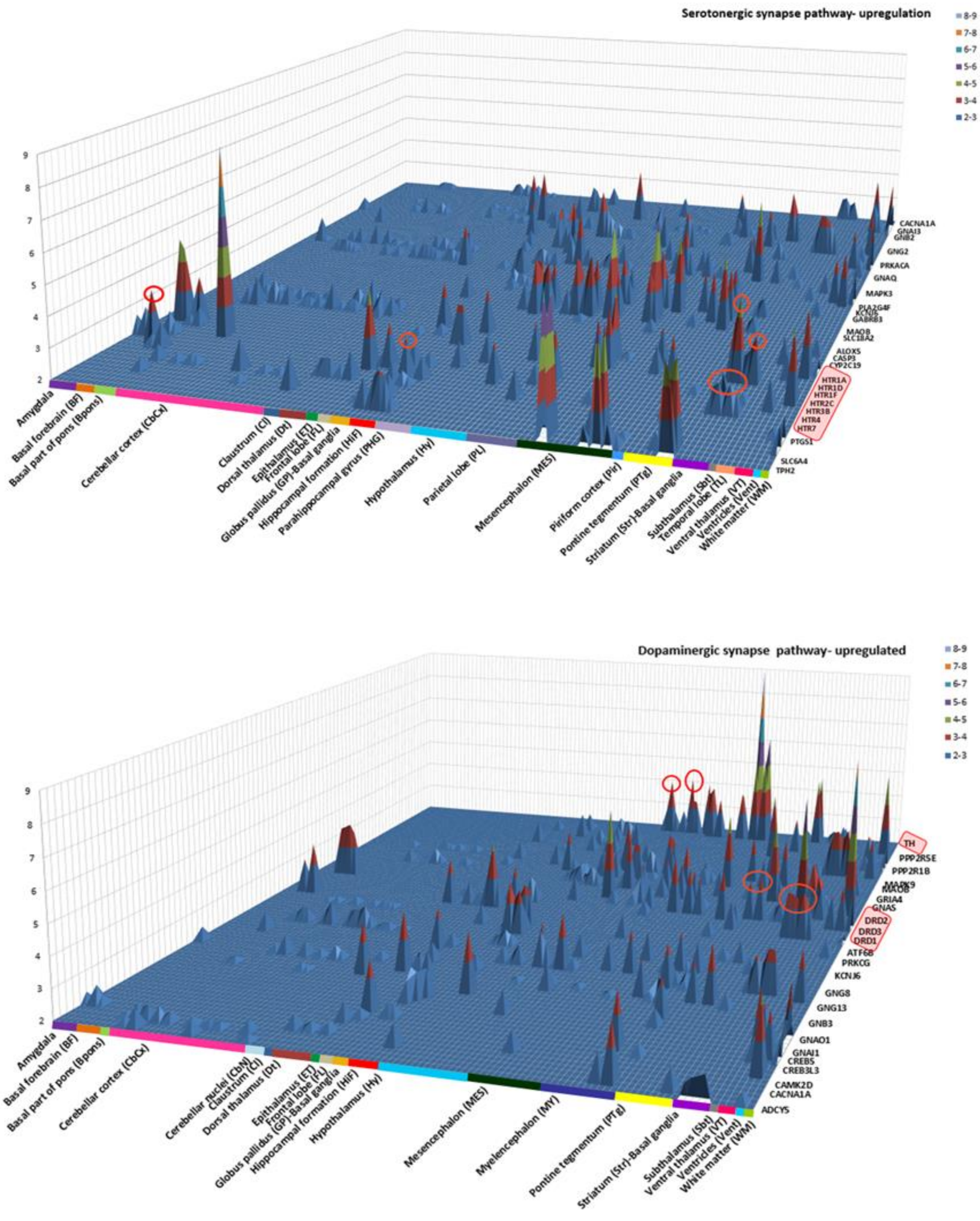


Figure 15 - Structural distribution of gene expression in neurotransmitter systems. The distribution patterns of receptors shed light on the relation between anatomical units and their functions

Spatial distribution of neurotransmitter systems reflects role of multiple receptors in their respective regions of high expression. Additionally, collective mapping of multiple receptors provides a multifaceted view of the anatomical, functional and biology driven organizational principles of the human brain. These maps also serve as basis for pharmacological studies to better understand brain diseases.

Prediction model based on healthy brain gene expression

Evaluating the utility of healthy brain gene expression as a tool to predict potentially new genes for a neurological disorder offers a massive scope in incrementing our current knowledge. In the present study, we used the discriminatory power of the expression patterns of known autism genes in a healthy individual and developed a prediction model to identify new genes that may be potentially associated to autism. Based on random forest and three other classification algorithms an overall class prediction accuracy of 84% was achieved. The sensitivity and specificity was 0.84 and 0.60 respectively.

For the prediction model building, 219 autism implicated genes from the AutDB (38) database constituted the positive dataset. Their expression profiles across the 190 brain structures were extracted from our healthy brain expression model and served as the feature vectors. For the negative dataset, 830 brain enriched genes were selected from the Protein Atlas (148). Similarly, their expression profiles were extracted from the healthy brain expression model. Note that both the positive and the negative training datasets come from the gene expression profiles of the brain. Using these datasets, we labeled each gene with its assigned class (autism-associated and non-autism associated) and developed a classification model to predict the classes of unseen or novel autism associated genes. Three popular machine learning

(ML) algorithms were tested; Random Forest (RF) (150), BayesNet (151) and J48 (152), to find the most appropriate algorithm for our dataset. Random division of all the feature vectors was conducted to generate 80 and 20 percent subsets for training and testing, respectively. Since the datasets are unbalanced across classes, using stratified partitioning we preserved the approximate class distributions for training and testing sets. A two-step validation technique was used. In step one; we determined 10-fold cross-validation accuracy on the training set. For step two; using the testing dataset that is not a part of the training data we determined the testing accuracy of the model. We also report standard performance measures of each class, including true positive rate (TPR), false positive rate (FPR), and receiver operating characteristic (ROC) curves and the area under the curve (AUC).

Table 5 shows the performance measures of each ML algorithm. Out of the aforementioned ML algorithms, we selected RF method for further use in this study owing to its superior performance. Also, J48 algorithm achieved a close accuracy with a slight loss in TPR and FPR measure. Figure 16 illustrates the ROC curves showing the relationship between TPR (sensitivity) and FPR (1-specificity). In an ideal scenario, ROC curve goes straight up on the Y-axis and then to the right parallel to the X-axis; thereby maximizing the area under the curve (AUC). An AUC close to 1 indicates that the classifier is predicting with maximum TP and minimum FP. We calculated an AUC of 0.81, indicating that the classification model can markedly differentiate between the autism versus non-autism associated genes.

Classifier	TPR	FPR	TNR	FNR	ROC area
Random Forest	0.84	0.39	0.83	0.84	0.81
J48 (C4.5)	0.82	0.47	0.80	0.81	0.77
BayesNet	0.74	0.28	0.80	0.76	0.78

Table 5 - Performance measures of each ML algorithm

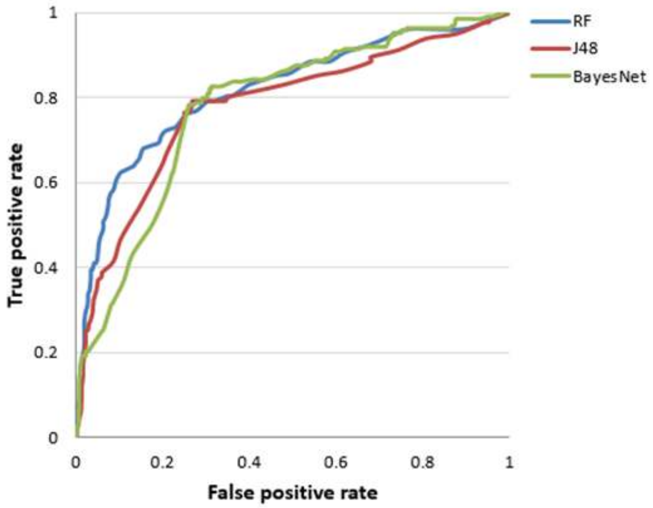


Figure 16 - ROC curves showing the relationship between TPR (sensitivity) and FPR (1-specificity) for the three ML algorithms

Discussion

Structured organization of hundreds of neuroanatomic regions, each with its specific molecular underpinnings gives rise to the complexity of whole brain function. We have developed a new framework to systematically integrate high throughput transcriptome profiling data from healthy human brains, complemented with various functional analysis techniques to develop a gene functional map on the spatial dimension of human brain. The brain structure specific gene expression profile generated from this framework served as a baseline reference for the development of a generic prediction model to find new disease implicated genes for any given neurological disorder.

This study was motivated by three objectives. First, we wanted to reduce the complex brain gene expression into recurring patterns where the spatial information is conserved. Since reproducible expression patterns across neuroanatomical structures in different individuals tend to have properties fundamental to brains functioning (17, 180), defining these brain specific and general transcriptional patterns is essential. In our second objective, we wanted to functionally characterize the transcriptional landscape of the human brain as captured in our gene expression model. Execution of clustering, network analysis and receptor mapping suggested striking features of the global transcription patterns as well as local patterns. The third impetus for this study comes from the paucity of brain expression data to study neurological conditions. So, in this study we were also able to move towards developing a prediction model, which would be effective to identify new genes for their potential association or non-association to any given neurological disorder using the reproducible gene set of healthy brain gene expression. As an example, we have demonstrated the utility of this method to predict potentially new genes involved in autism and evaluated the performance measures of the classifier. This to the best of our knowledge has never been presented before.

While building the expression model we noticed recurrence of similar expression profiles across the similarly annotated sample from six individuals suggesting high recapitulation of the brain's transcriptome (16, 69, 70). Genes in the adult human brain show significant spatial heterogeneity, however with distinct patterns in higher level anatomic structures (Figure 10). By performing hierarchical clustering, the bunching of relatively unknown/ill characterized genes with genes of known function helps provide a meaningful context to the functionality of these genes (79). Also, groups of genes based on similarity in gene expression profile were described that would 'occur naturally' and can point towards putative co-regulated genes. We found that there are some such coherent gene groups (cerebellum, hippocampus, striatum, dorsal thalamus, amygdala), but many, of the genes exhibit high variation over different brain structures and do not lie in any of these groups. We noticed that the sensory and motor regions have the most distinct whole-transcriptome signatures, probably related to their specialized cellular and functional architecture. Additionally, mapping of the structural distribution of neurotransmitter systems revealed interesting patterns of enrichment in localized regions for various receptors. We found a high expression of HTR2B in in cerebellum in the serotonergic neurotransmitter system which is in contrast to its detection at lower expression levels in cerebellum along with occipital and frontal cortex as previously reported (178). Neurotransmitters not only are associated with the communication between differentiated neurons but rather offer a crosstalk between various other functions like neurogenesis, regeneration and neuroplasticity. The structural distribution maps provide an overview of anatomical, functional and molecular organization of the receptors.

Here, we also performed an exhaustive assessment of co-expression patterns in the human brain expression map. We found that coexpression of genes is usually associated with a discrete spatial location, probably specific to cell type functions. For example, the pink module, which showed properties related to dopaminergic synaptic function (Table 4) was identified in striatum and amygdala. This finding is consistent with the predominant neuronal phenotype made of medium spiny neurons that have been associated with high expression of DRD receptors (179). To summarize, each module appeared to display a particular biological function: the magenta module reflects the translational machinery with resident genes such as RPL19, RPS26, EEF1B2f, and the cyan module reflects glial cell differentiation/oligodendrocyte differentiation and astrocytes enriched with genes like NKX2-2, NKX6-2, OLIG2, APOE, PEA15 and PAX6. Functional enrichment of genes in each of the coexpression modules thus reflects the molecular events in the human brain.

There is sufficient evidence which points to a strong genetic basis of autism and the number of associated genes has been estimated to be close to a thousand. However, only a few hundred genes are currently known to be certainly implicated in autism, mostly being targets of rare mutations. Most studies focus on characterizing the molecular interactions of known autism spectrum disorder genes (181, 182, and 183) and no study, as of now, has put in effort into predicting entirely novel autism genes. In the current study, we used machine learning classifier that learns the expression patterns of known ASD genes in the distinct brain structures in a healthy state and then uses the expression driven indicators specific to ASD genes to predict novel genes from the full complement of the genome. By this approach, we can predict new genes which have not been implicated in autism before. We explored the predictive capabilities of 219 autism genes in the framework of healthy brain expression data. The classification performance based on our model is encouraging, especially given the use of healthy brain data to make predictions. To the best of our knowledge, prediction of potentially new disease implicated genes using healthy brain tissue gene expression has never been shown before.

CHAPTER 5

CONCLUSIONS

Brain function is governed by precise regulation of gene expression across its anatomically distinct sub structures. However, the expression patterns of genes across hundreds of brain sub structures is not clearly understood. Here, we describe an expression model, which is representative of the healthy adult human brain transcriptome by using data from the Allen Brain Atlas. This gene expression model captures the canonical signatures that are reproducible across individuals. Our in-depth analysis revealed that 84% of genes are expressed in at least one of the 190 brain structures studied. Hierarchical clustering based on gene expression delineated the brain regions into structurally tiered spatial groups. Gene enrichment and pathway analysis of differentially expressed genes in the higher-order brain regions showed striking enrichment for region-specific processes. Further, weighted correlation network analysis of this model identified robust modules of coexpressing genes in the brain that demonstrated wide range of functional associations. Also, the structural distribution of major neurotransmission systems was plotted. Finally, we developed a supervised classification model, which achieved an accuracy of 84% with an ROC of 0.81 for predicting autism-implicated genes using our expression model as a baseline. This study represents the first use of only healthy brain gene expression data to predict potentially new disease implicated genes and this generic methodology can be applied to other neurological diseases.

A major strength of our study is the utilization of the Allen Human Brain Atlas to build the blueprint of a gene expression model representing a healthy state of human brain. Conceptualized on the ‘all gene- all structure’ approach, ABA allows to explore the transcriptome properties at a high anatomical resolution of the human brain. Also, this study is focused on only those genes that exhibit compelling experimental evidence to autism susceptibility rather than the complete gene set catalogued in AutDB. There are also some pitfalls in our study. First, the sample size is relatively moderate constituting only five males and one female donor. Also, sex biased transcriptional variation may be detrimental to our method. However, since we are trying to capture the common core components of gene expression in human brains, sex as a confounding factor can be disregarded. Second, we implemented stringent expression pattern selection parameters with an aim to capture brain specific transcription patterns, rather than capture all the changes that may occur. In course, we anticipate losing some expression information. Finally, the brain structure specific profiles have been established at RNA level and these may not translate at protein levels. Also, the influence of positive training instances cannot be discounted. For a ML algorithm to perform efficiently, the ratio of positive to negative training instances is crucial. Since, the number of genes that have been experimentally validated in disease etiology are limited, a thorough review of literature and databases is essential to make a comprehensive list of disease implicated genes to ensure the balance between the positive and negative datasets in the training classes. For this reason, we chose BayesNet (a modified Naïve Bayes) and Random Forest MLs since they are the least sensitive to changes in the number of instances in the training set. Nevertheless, the accuracy we have obtained in the current study is an important step towards predicting potentially disease implicated genes using expression data from healthy brain tissues. In conclusion, this study describes gene expression patterns conserved across distinct brain structures among healthy population. The strategy employed in this study constitutes a generic approach involving a pipeline of functional analysis tools and algorithms. Finally, to the best of our knowledge, this study demonstrates the first comprehensive utilization of healthy brain tissue gene expression models that could predict disease implicated genes for a given neurological disorder.

APPENDIX

A.1 Mathematical Notation

A list of all mathematical and symbolic notation used is indicated below

Table 17 - Mathematical Notation

di	The ith protein sequence in the dataset
yj	The jth category that a protein may be assigned to
X	Feature space used for representing objects being classified
C	Set of categories that each data instance may be assigned to objects
xi	The ith instance, represented by features
xj	The jth feature
yj	The jth category
D	The space of all possible proteins
D	Set of training data used to train the classifier
$h:X \rightarrow Y$	The classification model, mapping objects to their classification

REFRENCES

1. Ascoli GA, Donohue DE, Halavi M. NeuroMorpho.Org: a central resource for neuronal morphologies.J Neurosci., 27(35):9247-51 (2007)

2. Shen, E. H., Overly, C. C. & Jones, A. R. The Allen Human Brain Atlas. Comprehensive gene expression mapping of the human brain. Trends in Neurosciences 35, 711–714 (2012).

3. Enard, W. et al. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418, 869–72 (2002).

4. Caceres, M. et al. Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl. Acad. Sci. U. S. A. 100, 13030–13035 (2003).

5. Lockhart, D. J. & Barlow, C. Expressing what's on your mind: DNA arrays and the brain. Nat. Rev. Neurosci. 2, 63–68 (2001).

6. Sherwood, C. C., Subiaul, F. & Zawidzki, T. W. A natural history of the human mind: Tracing evolutionary changes in brain and cognition. in Journal of Anatomy 212, 426–454 (2008).

7. Geschwind, D. H. & Levitt, P. Autism spectrum disorders: developmental disconnection syndromes. Current Opinion in Neurobiology 17, 103–111 (2007).

8. Roth, R. B. et al. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics 7, 67–80 (2006).

9. Strand, A. D. et al. Expression profiling of Huntington's disease models suggests that brain-derived neurotrophic factor depletion plays a major role in striatal degeneration. J. Neurosci. 27, 11758–68 (2007).

10. Naumova, O. Y. et al. Age-related changes of gene expression in the neocortex: preliminary data on RNA-Seq of the transcriptome in three functionally distinct cortical areas. Dev. Psychopathol. 24, 1427–42 (2012).

11. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40, 1413–5 (2008).

12. Peterson, E. C., Wang, Z. & Britz, G. Regulation of cerebral blood flow. International Journal of Vascular Medicine 2011, (2011).

13. Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. Genome Res. 15, 945–953 (2005).

14. Macgregor, P. F. & Squire, J. A. Application of microarrays to the analysis of gene expression in cancer. Clinical Chemistry 48, 1170–1177 (2002).

15. Winkler, J. M., Chaudhuri, A. D. & Fox, H. S. Translating the brain transcriptome in neuroAIDS: From non-human primates to humans. Journal of Neuroimmune Pharmacology 7, 372–379 (2012).

16. Preuss, T. M., Cáceres, M., Oldham, M. C. & Geschwind, D. H. Human brain evolution: insights from microarrays. Nat. Rev. Genet. 5, 850–860 (2004).

17. Hawrylycz, M. J. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489, 391–9 (2012).

18. Geschwind, D. H. & Konopka, G. Neuroscience in the era of functional genomics and systems biology. Nature 461, 908–915 (2009).

19. Lin, S. et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. 111, 201413624 (2014).

20. Zheng-Bradley, X., Rung, J., Parkinson, H. & Brazma, A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. 11, R124 (2010).

21. Zeng, J. et al. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am. J. Hum. Genet. 91, 455–465 (2012).

22. Ch'ng, C., Kwok, W., Rogic, S. & Pavlidis, P. Meta-analysis of gene expression in Autism spectrum disorder. Autism Res. 8, 593–608 (2015).

23. Winkler, J. M. & Fox, H. S. Transcriptome meta-analysis reveals a central role for sex steroids in the degeneration of hippocampal neurons in Alzheimer's disease. BMC Syst. Biol. 7, 51 (2013).

24. Neueder, A. & Bates, G. P. A common gene expression signature in Huntington's disease patient brain regions. BMC Med. Genomics 7, 60 (2014).

25. McCabe, M. P. & O'Connor, E. J. A longitudinal study of economic pressure among people living with a progressive neurological illness. Chronic Illn. 5, 177–83 (2009).

26. Perou, R. et al. Mental health surveillance among children--United States, 2005-2011. Morb. Mortal. Wkly. report. Surveill. Summ. 62, 1–35 (2013).

27. Guo, S. Linking genes to brain, behavior and neurological diseases: What can we learn from zebrafish? Genes, Brain and Behavior 3, 63–74 (2004).

28. El-Fishawy, P. & State, M. W. The genetics of autism: key issues, recent findings, and clinical implications. Psychiatr. Clin. North Am. 33, 83–105 (2010).

29. Walsh, C. A. & Engle, E. C. Allelic diversity in human developmental neurogenetics: Insights into biology and disease. Neuron 68, 245–253 (2010).

30. Miles, J. H. Autism spectrum disorders--a genetics review. Genet. Med. 13, 278–294 (2011).

31. Fatemi, S. H. et al. Prenatal viral infection of mice at E16 causes changes in gene expression in hippocampi of the offspring. Eur. Neuropsychopharmacol. 19, 648–653 (2009).

32. Kogan, M. D. et al. Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. Pediatrics 124, 1395–1403 (2009).

33. Steffenburg, S. et al. A Twin Study of Autism in Denmark, Finland, Iceland, Norway and Sweden, A Twin Study of Autism in Denmark, Finland, Iceland, Norway and Sweden. J. Child Psychol. Psychiatry, J. Child Psychol. Psychiatry 30, 30, 405, 405–416, 416 (1989).

34. Turner, N. T. Elucidating the Etiology of Autism Using Genomic Methods (Doctoral dissertation). Retrieved from Johns Hopkins DSpace repository (2013)

35. Campbell, M. G., Kohane, I. S. & Kong, S. W. Pathway-based outlier method reveals heterogeneous genomic structure of autism in blood transcriptome. BMC Med. Genomics 6, 34 (2013).

36. de Jong, S. et al. A gene co-expression network in whole blood of Schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. PLoS One 7, (2012).

37. Kumar, A. et al. A brain region-specific predictive gene map for autism derived by profiling a reference gene set. PLoS One 6, (2011).

38. Basu, S. N., Kollu, R. & Banerjee-Basu, S. AutDB: A gene reference resource for autism research. Nucleic Acids Res. 37, (2009).

39. Buckley, F. Modelling Down syndrome. Downs. Syndr. Res. Pract. 12, 98–102 (2008).

40. Jennings, C. G. et al. Opportunities and challenges in modeling human brain disorders in transgenic primates. Nat. Neurosci. 19, 1123–30 (2016).

41. Emery, B. & Barres, B. A. Unlocking CNS Cell Type Heterogeneity. Cell 135, 596–598 (2008).

42. Montaño, C. M. et al. Measuring cell-type specific differential methylation in human brain tissue. Genome Biol. 14, R94 (2013).

43. Wang, H. Y. et al. Rate of evolution in brain-expressed genes in humans and other primates. PLoS Biol. 5, 0335–0342 (2007).

44. Bart van der Worp, H. et al. Can animal models of disease reliably inform human studies? PLoS Med. 7, 1–8 (2010).

45. Gallego Romero, I., Pai, A. a, Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 12, 42 (2014).

46. Lewis, D. A. The human brain revisited: Opportunities and challenges in postmortem studies of psychiatric disorders. Neuropsychopharmacology 26, 143–154 (2002).

47. Hayashi-Takagi, A., Vawter, M. P. & Iwamoto, K. Peripheral biomarkers revisited: Integrative profiling of peripheral samples for psychiatric research. Biological Psychiatry 75, 920–928 (2014).

48. Liu, L., Lei, J., Sanders, S. & Willsey, A. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism 5, 1–18 (2014).

49. Glatt, S. J. et al. Blood-based gene expression signatures of infants and toddlers with autism. J. Am. Acad. Child Adolesc. Psychiatry 51, (2012).

50. Zaman, S. et al. A Search for Blood Biomarkers for Autism: Peptoids. Sci. Rep. 6, 19164 (2016).

51. Huang, F. et al. miRNA profiling in autism spectrum disorder in China. Genomics Data 6, 108–109 (2015).

52. Kong, S. W. et al. Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders. PLoS One 7, (2012).

53. Wang, Y. et al. Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. Transl. Psychiatry 5, e660 (2015).

54. Yuhas, J. et al. Brief report: Sensorimotor gating in idiopathic autism and autism associated with fragile X syndrome. J. Autism Dev. Disord. 41, 248–253 (2011).

55. Li, X., Zou, H. & Brown, W. T. Genes associated with autism spectrum disorder. Brain Research Bulletin 88, 543–552 (2012).

56. Larsen, M. J., Thomassen, M., Tan, Q., Sørensen, K. P. & Kruse, T. A. Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency. Biomed Res. Int. 2014, (2014).

57. Viljoen, K. S. & Blackburn, J. M. Quality assessment and data handling methods for Affymetrix Gene 1.0 ST arrays with variable RNA integrity. BMC Genomics 14, 1–13 (2013).

58. Waters, K. M., Pounds, J. G. & Thrall, B. D. Data merging for integrated microarray and proteomic analysis. Briefings in Functional Genomics and Proteomics 5, 261–272 (2006).

59. Gomez-Cabrero, D. et al. Data integration in the era of omics: current and future challenges. BMC Syst. Biol. 8 Suppl 2, I1 (2014).

60. Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. Trends in Genetics 22, 101–109 (2006).

61. Yang, Z. et al. Meta-analysis of differentially expressed genes in osteosarcoma based on gene expression data. BMC Med. Genet. 15, 80 (2014).

62. Parsons, T. D. Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences. Front. Hum. Neurosci. 9, 660 (2015).

63. Fougerousse, F. et al. Human-mouse differences in the embryonic expression patterns of developmental control genes and disease genes. Hum. Mol. Genet. 9, 165–173 (2000).

64. Konopka, G. et al. Human-Specific Transcriptional Networks in the Brain. Neuron 75, 601–617 (2012).

65. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. Nat. Rev. Genet. 7, 693–702 (2006).

66. A.C. Mitchell, K. Mirnics. Gene expression profiling of the brain: Pondering facts and fiction. Neurobiol Dis, 45, pp. 3–7 (2011).

67. Ponomarev, I., Wang, S., Zhang, L., Harris, R. A. & Mayfield, R. D. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. J Neurosci 32, 1884–1897 (2012).

68. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483–9 (2011).

69. Johnson, M. B. et al. Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron 62, 494–509 (2009).

70. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. Nature 508, 199–206 (2014).

71. Mirnics, K., Levitt, P. & Lewis, D. A. DNA Microarray Analysis of Postmortem Brain Tissue. Int. Rev. Neurobiol. 60, 153–181 (2004).

72. Ball, S., Gilbert, T. L. & Overly, C. C. The Human Brain Online: An Open Resource for Advancing Brain Research. PLoS Biol. 10, (2012).

73. McCullumsmith, R. E. & Meador-Woodruff, J. H. Novel approaches to the study of postmortem brain in psychiatric illness: Old limitations and new challenges. Biological Psychiatry 69, 127–133 (2011).

74. van der Staay, F. J., Arndt, S. S. & Nordquist, R. E. Evaluation of animal models of neurobehavioral disorders. Behav. Brain Funct. 5, 11 (2009).

75. Klingseisen, A. & Jackson, A. P. Mechanisms and pathways of growth failure in primordial dwarfism. Genes and Development 25, 2011–2024 (2011).

76. Oldham, M. C., Horvath, S. & Geschwind, D. H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc. Natl. Acad. Sci. 103, 17973–17978 (2006).

77. Obayashi, T. et al. COXPRESdb: A database of comparative gene coexpression networks of eleven species for mammals. Nucleic Acids Res. 41, (2013).

78. Oldham, M. C., Horvath, S. & Geschwind, D. H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A 103, 17973–17978 (2006).

79. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. Bioinformatics 20, 1453–1454 (2004).

80. Gillis, J. & Pavlidis, P. &quot;Guilt by association&quot; is the exception rather than the rule in gene networks. PLoS Comput Biol 8, e1002444 (2012).

81. Matuszek, G. & Talebizadeh, Z. Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. BMC Med. Genet. 10, 102 (2009).

82. Xu, L. M. et al. AutismKB: An evidence-based knowledgebase of autism genetics. Nucleic Acids Res. 40, (2012).

83. Shen-Orr, S. S. et al. Cell type-specific gene expression differences in complex tissues. Nat Methods 7, 287–289 (2010).

84. Eggers, K. et al. Polysialic acid controls NCAM signals at cell-cell contacts to regulate focal adhesion independent from FGF receptor activity. Journal of Cell Science 124, 3279–3291 (2011).

85. Nisenbaum, L. K. The ultimate chip shot: Can microarray technology deliver for neuroscience? Genes, Brain and Behavior 1, 27–34 (2002).

86. Ginsberg, S. D. et al. Microarray analysis of hippocampal CA1 neurons implicates early endosomal dysfunction during Alzheimer’s disease progression. Biol. Psychiatry 68, 885–893 (2010).

87. Mirnics, K., Middleton, F., Marquez, A., Lewis, D. & Levitt, P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. Neuron 28, 53–67 (2000).

88. Grünblatt, E., Mandel, S. & Youdim, M. B. Neuroprotective strategies in Parkinson's disease using the models of 6-hydroxydopamine and MPTP. Ann. N. Y. Acad. Sci. 899, 262–273 (2000).

89. Desplats, P. A., Lambert, J. R. & Thomas, E. A. Functional roles for the striatal-enriched transcription factor, Bcl11b, in the control of striatal gene expression and transcriptional dysregulation in Huntington's disease. Neurobiol. Dis. 31, 298–308 (2008).

90. Fagiolini, A. et al. Functional impairment in the remission phase of bipolar disorder. Bipolar Disord. 7, 281–285 (2005).

91. Mycko, M. P., Papoian, R., Boschert, U., Raine, C. S. & Selmaj, K. W. cDNA microarray analysis in multiple sclerosis lesions: detection of genes associated with disease activity. Brain 126, 1048–1057 (2003).

92. Sappok, T. et al. The Diagnostic Behavioral Assessment for autism spectrum disorder—Revised: A screening instrument for adults with intellectual disability suspected of autism spectrum disorders. Res. Autism Spectr. Disord. 8, 362–375 (2014).

93. Lockhart, D. J., Lockhart, D. J., Winzeler, E. a & Winzeler, E. a. Genomics, gene expression and DNA arrays. Nature 405, 827–836 (2000).

94. Hu, J. & He, X. Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. Biometrics 63, (2007).

95. Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sin. 12, 111–139 (2002).

96. Fan, J. & Niu, Y. Selection and validation of normalization methods for c-DNA microarrays using within-array replications. Bioinformatics 23, 2391–2398 (2007).

97. Johnstone, D. et al. Evaluation of Different Normalization and Analysis Procedures for Illumina Gene Expression Microarray Data Involving Small Changes. Microarrays 2, 131–152 (2013).

98. Piccolo, S. R. et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics 100, 337–344 (2012).

99. Pelz, C. R., Kulesz-Martin, M., Bagby, G. & Sears, R. C. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics 9, 520 (2008).

100. Enard, W. et al. Intra- and interspecific variation in primate gene expression patterns. Science 296, 340–343 (2002).

101. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5, (2009).

102. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. Genome Biol 5, R74 (2004).

103. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–6 (2008).

104. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628 (2008).

105. Qureshi, I. A. & Mehler, M. F. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat. Rev. Neurosci. 13, 528–41 (2012).

106. Chodroff, R. A. et al. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol. 11, R72 (2010).

107. Kuss, A. W. & Chen, W. MicroRNAs in brain function and disease. Curr Neurol Neurosci Rep 8, 190–197 (2008).

108. Aprea, J. & Calegari, F. Long non-coding RNAs in corticogenesis: deciphering the non-coding code of the brain. EMBO J. 34, 2865–2884 (2015).

109. Qureshi, I. A. & Mehler, M. F. Non-coding RNA networks underlying cognitive disorders across the lifespan. Trends in Molecular Medicine 17, 337–346 (2011).

110. Nadler, J. J. et al. Large-scale gene expression differences across brain regions and inbred strains correlate with a behavioral phenotype. Genetics 174, 1229–1236 (2006).

111. Fatemi, S. H. et al. Consensus paper: Pathological role of the cerebellum in Autism. Cerebellum 11, 777–807 (2012).

112. Dekaban, A. S. & Sadowsky, D. Changes in brain weights during the span of human life: Relation of brain weights to body heights and body weights. Ann. Neurol. 4, 345–356 (1978).

113. Sowell, E. R., Thompson, P. M. & Toga, A. W. Mapping changes in the human cortex throughout the span of life. Neuroscientist 10, 372–392 (2004).

114. Colantuoni, C. et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature 478, 519–523 (2011).

115. Somel, M. et al. Transcriptional neoteny in the human brain. Proc. Natl. Acad. Sci. U. S. A. 106, 5743–8 (2009).

116. Somel, M. et al. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. Genome Res. 20, 1207–1218 (2010).

117. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (80-.). 302, 249–255 (2003).

118. Wolfe, C. J., Kohane, I. S. & Butte, A. J. Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. BMC Bioinformatics 6, 227 (2005).

119. Oldham, M. C. et al. Functional organization of the transcriptome in human brain. Nat. Neurosci. 11, 1271–1282 (2009).

120. Ziats, N. M. et al. Functional genomics studies of human brain development and implications for autism spectrum disorder. Translational Psychiatry 5, (2015).

121. Pérez-Bercoff, Å., Hudson, C. M. & Conant, G. C. A Conserved Mammalian Protein Interaction Network. PLoS One 8, (2013).

122. Baryshnikova, A. et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat. Methods 7, 1017–1024 (2010).

123. Gillberg, C., Cederlund, M., Lamberg, K. & Zeijlon, L. Brief report: 'The autism epidemic'. The registered prevalence of autism in a Swedish urban area. J. Autism Dev. Disord. 36, 429–435 (2006).

124. Rao, P. A. & Beidel, D. C. The Impact of Children with High-Functioning Autism on Parental Stress, Sibling Adjustment, and Family Functioning. Behav. Modif. 33, 437–451 (2009).

125. Folstein, S. & Rutter, M. Infantile autism: a genetic study of 21 twin pairs. J. Child Psychol. Psychiatry. 18, 297–321 (1977).

126. Krumm, N., O'Roak, B. J., Shendure, J. & Eichler, E. E. A de novo convergence of autism genetics and molecular neuroscience. Trends in Neurosciences 37, 95–105 (2014).

127. Bailey, a. et al. Autism as a strongly genetic disorder: evidence from a British twin study. Psychol. Med. 25, 63 (2009).

128. Simonoff, E. et al. Psychiatric Disorders in Children With Autism Spectrum Disorders: Prevalence, Comorbidity, and Associated Factors in a Population-Derived Sample. J. Am. Acad. Child Adolesc. Psychiatry 47, 921–929 (2008).

129. Hallmayer, J. et al. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. Arch. Gen. Psychiatry 68, 1095–1102 (2011).

130. Jacquemont, M.-L. et al. Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. J. Med. Genet. 43, 843–9 (2006).

131. Sebat, J. et al. Strong association of de novo copy number mutations with autism. Science 316, 445–9 (2007).

132. Abrahams, B. S. & Geschwind, D. H. Connecting genes to brain in the autism spectrum disorders. Arch. Neurol. 67, 395–9 (2010).

133. Bourgeron, T. A synaptic trek to autism. Current Opinion in Neurobiology 19, 231–234 (2009).

134. Filippone, M. et al. Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. Ann. Appl. Stat. 6, 1883–1905 (2012).

135. Zhou, Y., Yu, F. & Duong, T. Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning. PLoS One 9, (2014).

136. Duda, M., Kosmicki, J. A. & Wall, D. P. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Transl. Psychiatry 4, e424 (2014).

137. Wall, D. P., Dally, R., Luyster, R., Jung, J. Y. & DeLuca, T. F. Use of artificial intelligence to shorten the behavioral diagnosis of autism. PLoS One 7, (2012).

138. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. (2016). doi:10.1038/nn.4353

139. Mitchell, T. M. Does Machine Learning Really Work? AI Mag. 18, 11 (1997).

140. Miller, J. a et al. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics 12, 322 (2011).

141. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, 1409–1437 (1958).

142. Saldanha, A. J. Java Treeview--extensible visualization of microarray data. Bioinformatics 20, 3246–3248 (2004).

143. Huang, D. W., Lempicki, R. a & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).

144. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 41, (2013).

145. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).

146. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. 4, (2008).

147. Kanehisa, M. & Goto, S. Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30 (2000).

148. Uhlen, M. et al. Tissue-based map of the human proteome. Science (80-.). 347, 1260419–1260419 (2015).

149. Hall, M. et al. The WEKA data mining software: An update. SIGKDD Explor. 11, 10–18 (2009).

150. Breiman, L. Random forests. Mach. Learn. 45, 5–32 (2001).

151. Rish, I. An empirical study of the naive Bayes classifier. IJCAI 2001 Work. Empir. methods Artif. Intell. 22230, 41–46 (2001).

152. Quinlan, J. R. J. Ross Quinlan_C4.5_ Programs for Machine Learning.pdf. Morgan Kaufmann 5, 302 (1993).

153. Weiss, N., Miller, F., Cazaubon, S. & Couraud, P. O. The blood-brain barrier in brain homeostasis and neurological diseases. Biochimica et Biophysica Acta - Biomembranes 1788, 842–857 (2009).

154. Huntley, M. A., Bien-Ly, N., Daneman, R. & Watts, R. J. Dissecting gene expression at the blood-brain barrier. Front. Neurosci. 8, (2014).

155. Strand, A. D. et al. Conservation of regional gene expression in mouse and human brain. PLoS Genet 3, e59 (2007).

156. Shao, Y., Yamamoto, M., Figeys, D., Ning, Z. & Chan, H. M. Proteomic analysis of cerebellum in common marmoset exposed to methylmercury. Toxicol. Sci. 146, 43–51 (2015).

157. Vriend, J., Ghavami, S. & Marzban, H. The role of the ubiquitin proteasome system in cerebellar development and medulloblastoma. Mol. Brain 8, 64 (2015).

158. Scholl, U. I. et al. Seizures, sensorineural deafness, ataxia, mental retardation, and electrolyte imbalance (SeSAME syndrome) caused by mutations in KCNJ10. Proc. Natl. Acad. Sci. U. S. A. 106, 5842–5847 (2009).

159. Ozaki, M., Sasner, M., Yano, R., Lu, H. S. & Buonanno, A. Neuregulin- ␤ induces expression of an NMDA-receptor subunit. Nature 390, 691–694 (1997).

160. Rieff, H. I. et al. Neuregulin induces GABA(A) receptor subunit expression and neurite outgrowth in cerebellar granule cells. J Neurosci 19, 10757–10766 (1999).

161. Busskamp, V. et al. Rapid neurogenesis through transcriptional activation in human stem cells. Mol Syst Biol 10, 760 (2014).

162. Huang, H. S. et al. Transcriptional Regulatory Events Initiated by Ascl1 and Neurog2 During Neuronal Differentiation of P19 Embryonic Carcinoma Cells. J. Mol. Neurosci. 55, 684–705 (2014).

163. Kouser, M. et al. Loss of predominant Shank3 isoforms results in hippocampus-dependent impairments in behavior and synaptic transmission. J. Neurosci. 33, 18448–68 (2013).

164. Surmeier, D. J., Ding, J., Day, M., Wang, Z. & Shen, W. D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends in Neurosciences 30, 228–235 (2007).

165. Groux, H., Fournier, N. & Cottrez, F. Role of dendritic cells in the generation of regulatory T cells. Seminars in Immunology 16, 99–106 (2004).

166. MacMullen, C. M., Vick, K., Pacifico, R., Fallahi-Sichani, M. & Davis, R. L. Novel, primate-specific PDE10A isoform highlights gene expression complexity in human striatum with implications on the molecular pathology of bipolar disorder. Transl. Psychiatry 6, e742 (2016).

167. Missale, C., Nash, S. R., Robinson, S. W., Jaber, M. & Caron, M. G. Dopamine receptors: from structure to function. Physiol. Rev. 78, 189–225 (1998).

168. Ryu, J. R. et al. Control of adult neurogenesis by programmed cell death in the mammalian brain. Mol. Brain 9, 43 (2016).

169. Louvi, A. & Grove, E. A. Cilia in the CNS: The quiet organelle claims center stage. Neuron 69, 1046–1060 (2011).

170. Lee, J. H. & Gleeson, J. G. The role of primary cilia in neuronal function. Neurobiology of Disease 38, 167–172 (2010).

171. Eriksson, P. S. et al. Neurogenesis in the adult human hippocampus. Nat. Med. 4, 1313–1317 (1998).

172. Bruel-Jungerman, E., Davis, S., Rampon, C. & Laroche, S. Long-Term Potentiation Enhances Neurogenesis in the Adult Dentate Gyrus. J. Neurosci. 26, 5888–5893 (2006).

173. Schmidt-Salzmann, C., Li, L. & Bischofberger, J. Functional properties of extrasynaptic AMPA and NMDA receptors during postnatal hippocampal neurogenesis. J. Physiol. 592, 125–40 (2014).

174. Overall, R. W., Paszkowski-Rogacz, M. & Kempermann, G. The Mammalian Adult Neurogenesis Gene Ontology (MANGO) Provides a Structural Framework for Published Information on Genes Regulating Adult Hippocampal Neurogenesis. PLoS One 7, (2012).

175. Vasconcelos, F. F. et al. Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. Cell Rep. 10, 1544–1556 (2015).

176. Daubert, E. A. & Condron, B. G. Serotonin: A regulator of neuronal morphology and circuitry. Trends in Neurosciences 33, 424–434 (2010).

177. Navailles, S. & De Deurwaerdère, P. Presynaptic control of serotonin on striatal dopamine function. Psychopharmacology 213, 213–242 (2011).

178. Beaulieu, J.-M. & Gainetdinov, R. R. The physiology, signaling, and pharmacology of dopamine receptors. Pharmacol. Rev. 63, 182–217 (2011).

179. Hawrylycz, M. et al. Canonical genetic signatures of the adult human brain. Nat. Neurosci. 18, 1832–1844 (2015).

180. Parikshak, N. N. et al. XIntegrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell 155, (2013).

181. Willsey, A. J. et al. XCoexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell 155, (2013).

182. Uddin, M. et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. Nat. Genet. 46, 742–747 (2014).