2017

Molecular neuroanatomy: mouse-human homologies and the landscape of genes implicated in language disorders

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Dissertation

MOLECULAR NEUROANATOMY: MOUSE-HUMAN HOMOLOGIES AND THE LANDSCAPE OF GENES IMPLICATED IN LANGUAGE DISORDERS

by

EMMA M. MYERS

B.A., Vassar College, 2004

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader

Jason W. Bohland, Ph.D.
Assistant Professor of Health Sciences

Second Reader

Helen Barbas, Ph.D.
Professor of Health Sciences

Third Reader

Michael Hawrylycz, Ph.D.
Investigator
Allen Institute for Brain Science

ACKNOWLEDGMENTS

This work was supported in part by a 2015 New Century Scholars research grant from the American Speech-Language-Hearing Foundation (PI Jason W. Bohland), and by the Center of Excellence for Learning in Education, Science, and Technology, a National Science Foundation Science of Learning Center (NSF SMA-0835976, PI Barbara Shinn-Cunningham).

The love, and learning, and love of learning that speaks through this document is an echo of the love my family, friends, and teachers have shared. You may see two hundred pages of dry science. I see one chapter of a lifelong adventure--joyful, painful, breathtakingly beautiful, and made with the best and truest companions anyone ever had. I am grateful beyond words.

MOLECULAR NEUROANATOMY: MOUSE-HUMAN HOMOLOGIES AND THE LANDSCAPE OF GENES IMPLICATED IN LANGUAGE DISORDERS

EMMA M. MYERS

Boston University School of Medicine, 2017

Major Professor: Jason W. Bohland, Ph.D., Assistant Professor of Health Sciences

ABSTRACT

The distinctiveness of brain structures and circuits depends on interacting gene products, yet the organization of these molecules (the "transcriptome") within and across brain areas remains unclear. High-throughput, neuroanatomically-specific gene expression datasets such as the Allen Human Brain Atlas (AHBA) and Allen Mouse Brain Atlas (AMBA) have recently become available, providing unprecedented opportunities to quantify molecular neuroanatomy. This dissertation seeks to clarify how transcriptomic organization relates to conventional neuroanatomy within and across species, and to introduce the use of gene expression data as a bridge between genotype and phenotype in complex behavioral disorders.

The first part of this work examines large-scale, regional transcriptomic organization separately in the mouse and human brain. The use of dimensionality reduction methods and cross-sample correlations both revealed greater similarity between samples drawn from the same brain region. Sample profiles and differentially expressed genes across regions in the human brain also showed consistent anatomical specificity in a second human dataset with distinct sampling properties.

The frequent use of mouse models in clinical research points to the importance of comparing molecular neuroanatomical organization across species. The second part of this dissertation describes three comparative approaches. First, at genome scale, expression profiles within homologous brain regions tended to show higher similarity than those from non-homologous regions, with substantial variability across regions. Second, gene subsets (defined using co-expression relationships or shared annotations), which provide region-specific, cross-species molecular signatures were identified. Finally, brain-wide expression patterns of orthologous genes were compared. Neuron and oligodendrocyte markers were more correlated than expected by chance, while astrocyte markers were less so.

The localization and co-expression of genes reflect functional relationships that may underlie high-level functions. The final part of this dissertation describes a database of genes that have been implicated in speech and language disorders, and identifies brain regions where they are preferentially expressed or co-expressed. Several brain structures with functions relevant to four speech and language disorders showed co-expression of genes associated with these disorders. In particular, genes associated with persistent developmental stuttering showed stronger preferential co-expression in the basal ganglia, a structure of known importance in this disorder.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD ..... Alzheimer's Disease

AHBA ..... Allen Human Brain Atlas

AIBS ..... Allen Institute for Brain Science

AMBA ..... Allen Mouse Brain Atlas

Amyg ..... Amygdala

CAF ..... CA Fields

CAS ..... Childhood Apraxia of Speech

Cb ..... Cerebellum

CbCtx ..... Cerebellar Cortex

CbN ..... Cerebellar Nuclei

CN ..... Caudate Nucleus

Ctx ..... Cerebral Cortex

DG ..... Dentate Gyrus

DSM-V ..... Diagnostic and Statistical Manual of Mental Disorders

DVD ..... Developmental Verbal Dyspraxia

DYX ..... Dyslexia

EDA ..... Exploratory Data Analysis

eQTL ..... expression Quantitative Trait Locus

FCTX ..... Frontal Cortex

FDR ..... False Discovery Rate

GP ..... Globus Pallidus

GPe.....External Globus Pallidus

GPi .....Internal Globus Pallidus

Hipp..... Hippocampal Formation

Hyp..... Hypothalamus

ISH ..... *in situ* hybridization

Mb .....Midbrain

Med ..... Medulla

MSigDB .....Molecular Signatures Database

NA.....Nucleus Accumbens

PCC ..... Pearson Product-Moment Correlation Coefficient

PCW ..... post-conceptional weeks

PD .....Parkinson's Disease

PDS .....Persistent Developmental Stuttering

Po .....Pons

Put .....Putamen

SL..... Speech / language

SLDB .....Speech / Language Disorders Database

SLI.....Specific Language Impairment

Str.....Striatum

Sub ..... Subiculum

SVD.....Singular Value Decomposition

TCTX ..... Temporal Cortex

Th ..... Thalamus

t-SNE..... t-Distributed Stochastic Neighbor Embedding

WGCNA ..... Weighted Gene Co-expression Network Analysis

CHAPTER 1: INTRODUCTION

1.1 Gene expression and neuroanatomy

The cytoarchitecture and myeloarchitecture that traditionally define regional brain organization arise from molecular events: the actions and interactions of genes and their products, resulting in proteins and other molecules which are present at varying levels across different cell populations throughout the brain. These events guide the development and differentiation of the brain and are integral to its function throughout life, but are difficult to observe. The technologies to measure mRNA abundance (an approximation of protein level) simultaneously for thousands of sequences were developed in the 1990s, allowing investigators to take a genome-wide "snapshot" of gene expression from a biological sample (see Lockhart and Winzeler, 2000 for an overview of DNA microarrays; and Lennon, 2000 for a brief historical overview). The resulting datasets, known as "high-throughput" datasets for their genome-wide coverage, include gene expression profiles of samples from diverse brain areas (composed of varying cell populations). High-throughput gene expression datasets open the possibility of characterizing the brain's *transcriptome*: the RNA sequences present, which varies by brain tissue as well as time.

The molecular scale that underlies conventional neuroanatomy is reflected by spatial organization (i.e., patterns of common gene expression) within high-throughput gene expression datasets. This dissertation examines global and local transcriptomic organization within three high-throughput, brain-wide, neuroanatomically-specific gene expression datasets, and uses observed transcriptomic correspondence with conventional

neuroanatomy to make necessary initial steps towards connecting the functions of groups of genes with those of brain structures.

1.2 Gene expression and development

The expression datasets studied here are from adult brains (24-57 years old in the human; 56 days old in the mouse). The adult brain, however, must be understood as the outcome of developmental processes at regional, cellular, and molecular scales.

Regional organization in the mature mammalian brain can be traced back to the embryonic *neural tube* (the form assumed by the first neural tissue; see Sanes et al., 2012, Ch. 2, for a review that includes a discussion of molecular mechanisms). Vesicles form along the length of the neural tube, each of which generates cells destined for a different part of the brain. Telencephalic structures share a common developmental origin in the anterior-most vesicle; diencephalic structures all rise from the adjacent vesicle, and so on through three vesicles which generate the midbrain, pons, and medulla, respectively. At the boundary between the midbrain and hindbrain, a transient structure called the *rhombic lip* generates cerebellar neurons (Wingate, 2001; Fink, 2006), as well as some other parts of the hindbrain (Wang et al., 2005). Organization of brain areas along the rostrocaudal axis thus emerges from the order of vesicles along the neural tube. Some connections between these areas begin to form even while neurogenesis and cell migration to the area is still taking place, and this early connectivity further refines areal differentiation. In particular, the developing dorsal thalamus sends projections to the

neocortex (and other parts of the telencephalon), and is vital to cortical arealization (see O'Leary et al., 2007 for a review).

Like the (relatively) stable organization of the mature brain, these developing structures and circuits have cellular and molecular underpinnings. Early brain development involves a fantastically complex set of interacting cellular-scale events. Cells proliferate and migrate to specific destinations while taking on a wide variety of forms, both neurons and glia; cell processes and synapses appear; cells and synapses die in vast numbers; the intricate circuitry of the brain begins to form. These events are effected and regulated, with temporal and spatial precision, by gene products. The first appearance of neural tissue in the embryo requires interactions between gene products working in, at a minimum, three signaling pathways (Stern, 2005; Sanes et al., 2012). The differentiation of that tissue into distinct brain areas is orchestrated by a growing list of transcription factors; for example, morphogens such as *TGF-8* regulate rostrocaudal gradients of expression that influence neuron fate towards forming motor cortex (*Pax2*, *Sp8*) or visual cortex (*Emx2*; see Sansom and Livesey, 2009 for a review). Rapid changes in the developing human brain are reflected by genome-scale transcriptomic change, which is accordingly greatest through infancy (Kang et al., 2011b; also see Fig. 2 in Cahoy et al., 2008). Region-dependent changes continue throughout the lifespan, however, and multiple studies have identified changes related to aging and neurodegenerative disease (Lee et al., 2000; Fraser et al., 2005; Berchtold et al., 2008; Colantuoni et al., 2011).

1.3 Overview of data collection techniques

This section briefly describes the two experimental techniques for measuring mRNA abundance that were used to produce the datasets analyzed in this dissertation (i.e., DNA microarrays and *in situ* hybridization, or ISH). Chapter 2 formally describes these datasets and the normalization procedures applied to them.

Gene expression from neuroanatomically-specific samples in human donor brains can be measured using microarray technology, as is the case in two of the datasets analyzed here (Gibbs et al., 2010; Hawrylycz et al., 2012). A microarray chip consists of a surface (usually made of glass) that is covered in microscopic spots, each containing many synthesized copies of a given DNA sequence, or "probes". "Targets" are created by extracting mRNA strands from the sample tissue and converting them into complementary DNAs (cDNAs). The cDNAs are labelled with fluorescent dye and hybridized to the spots containing the DNA probes. There, they bind to the probe sequences to which they are complementary. After removing cDNAs that failed to bind to any probes, the remaining cDNA at a given spot is revealed by the fluorescent label. Quantification of that signal then yields an estimate of the abundance of the original transcript in the sample tissue. A variety of commercial and custom microarray platforms are now available for performing genome-scale profiling of individual samples.

Microarrays and ISH are essentially complementary ways of expression profiling. Where microarrays involve affixing the probe to a surface and then applying the labelled target, ISH requires affixing the sample tissue and applying the labelled probe. In the ISH method, the probe is complementary RNA, and the target is mRNA present within

the tissue, which has been treated to facilitate the probe's access to the target. After hybridization, the tissue is washed, and the label indicates how much of the probe bound to its target, and thus the abundance of the target transcripts in the tissue. This is a more direct measure of mRNA abundance than microarrays can provide, with a higher signal-to-noise ratio. However, because it requires individual sections of brain tissue, each tested for the presence of a different transcript, ISH is highly labor-intensive and less suitable for high-throughput genomic screening. The small size of the mouse brain, as well as the ability to use many mice, make ISH a more practicable approach for the mouse than the human brain when genome-wide coverage and / or good spatial resolution is desired.

1.4 Large-scale molecular neuroanatomy

The advent of large gene expression datasets from the brains of multiple species offers unprecedented opportunities to investigate molecular neuroanatomy. These datasets show varying gene expression levels across different brain tissues for thousands of genes, allowing us to study not only the pattern of expression of a gene of interest, but the relationships between many genes' spatial expression patterns, or, conversely, the expression profiles of different regions of the brain. Large expression datasets have inspired efforts to map the brain's transcriptome at a large scale, revealing systematic, region-dependent variation of gene expression across the brains of rodents and primates (Lein et al., 2007; Hawrylycz et al., 2012; Ng et al., 2009; Bohland et al., 2010; Bernard

et al., 2012; Ji, 2011), as well as regional differences in the co-expression relationships between different genes (Oldham et al., 2008; Ko et al., 2013; Grange et al., 2014).

The characterization of data pertaining to thousands of genes poses a daunting problem. Conventional neuroanatomy, while sometimes plagued by difficulties in clearly and consistently defining brain regions, is generally based on simple markers and, usually, visually observable features. By contrast, the gene expression profile of a sample is a multidimensional observation of its molecular composition, and the extraction of useful information from these large datasets has demanded the application of a variety of tools from statistics, machine learning, and graph theory. One way to deal with this complexity is to investigate samples one gene at a time to identify genes whose expression patterns distinguish certain samples (see Pavlidis and Noble, 2001, for example, for an examination of these methods). The proportion of genes which are differentially expressed can then be used to quantify dissimilarity between samples from different brain regions (e.g. Hawrylycz et al., 2012; Khaitovich, 2004). Alternatively, the correlation between two samples' expression profiles can be used as a measure of similarity, often followed by clustering of samples based on those similarities (e.g. Hawrylycz et al., 2012; Roth et al., 2006) in order to determine molecularly homogeneous groups of samples and/or samples with discriminable transcriptomes. Similarly, data reduction and visualization methods such as multi-dimensional scaling (MDS), principal components analysis (PCA), or t-Distributed Stochastic Neighbor Embedding (t-SNE) may be used to assign each sample a set of coordinates in a low-dimensional space such that proximity between samples reflects the similarity of the

expression profiles, providing a simple visual representation of the “landscape” of gene expression across samples (Khaitovich, 2004; Mahfouz et al., 2015).

Chapter 3 of this dissertation begins with an overview of the literature on large-scale transcriptomic organization as it relates to neuroanatomy, followed by a study of this organization in the human and mouse, showing relationships between brain samples based on gene expression that correspond to the samples' regions of origin. The chapter also quantitatively compares two expression datasets from the adult human brain, one with high spatial resolution and the other with large sample size, to determine the consistency of anatomically specific expression signatures across datasets and microarray platforms.

1.5 Comparative molecular neuroanatomy

Given the common usage of mouse models to study human neuropathologies, there is a striking paucity of such disease models that have proven to be clinically useful (Le et al., 2014; Duff, 2004; Hardy, 2006). While much is known about the similarities and differences between the human and mouse brain from conventional neuroanatomical methods (based on features such as cytoarchitecture and inter-regional connectivity), mouse models fundamentally rely on cross-species correspondences that are realized at the much smaller, intracellular scale that involves interactions between gene products. Currently, relatively little is known about how these molecular mechanisms, which vary substantially across cell populations and brain regions in each animal, compare across the two species. This lack of an established basis for comparison at the relevant scale may

cause some of the difficulty in developing successful mouse models (Burns et al., 2015).

Chapter 4 briefly reviews cross-species studies of molecular neuroanatomy before presenting three quantitative comparisons of the adult human and mouse transcriptome. A whole-brain, genome-wide comparison based on sample profile correlations is followed by a novel approach to identifying region-specific gene expression signatures that are consistent across the two species. Finally, the brain-wide profiles of orthologous genes are compared between the two species, with a particular focus on genes that mark certain cell types. The ultimate goal of this approach is to inform brain research that relies on mouse models by illuminating gene interactions that are similar, and that impact similar brain areas, in the human and mouse brain.

1.6 Expression of genes associated with speech and language disorders

Transcriptomic data from the brain have the potential to help illuminate how functionally relevant genes impact neuronal and cognitive processes at a larger scale. In the many heritable disorders characterized by behavioral phenotypes, there is currently a large genotype-phenotype gap; knowledge is becoming available about either the genes potentially involved in these disorders or about the brain regions and circuits that are atypically functioning, but little is known about the intermediate pathological mechanisms. Chapter 5 describes an attempt to begin bridging this gap, specifically in the area of language-related functions. Disorders of speech and language are highly heritable, and variants in over three dozen genes to date have been associated with language abilities and disabilities, with varying degrees of evidence (see e.g. Fisher et al.,

2003; Graham et al., 2015). Many of these genes have known roles in brain processes such as cortical migration and pre-synaptic signaling (e.g. Wang et al., 2006; Yang et al., 2011), but little is known of the causal links between variants in a given gene and behavioral traits. Chapter 5 begins with a brief overview of the candidate genes and associated phenotypes, followed by a description of a manually curated database intended to facilitate the integration of genetic and transcriptomic information with neuroimaging results from research into speech and language function. The rest of the chapter examines the expression patterns and co-expression relationships of these genes in regions throughout the adult human brain in an effort to identify the parts of the brain through which the genes may affect language-related functions, and to consider how this might relate to known or suspected functions of the genes at a smaller scale.

Finally, Chapter 6 discusses the data-driven nature of the approach taken in this research, and summarizes the contributions of the dissertation as well as common themes in the future research proposed in the previous chapters.

CHAPTER 2: OVERVIEW OF MATERIALS AND METHODS

2.1 Introduction

This dissertation focuses primarily on two publicly available, high-throughput gene expression datasets. The Allen Human Brain Atlas (AHBA), a microarray dataset from six adult human donors, comprises thousands of samples from several hundred finely labelled brain structures (Hawrylycz et al., 2012; www.human.brain-map.org). While not at single cell resolution, this is a far higher spatial resolution than is typically available in microarray data from brain tissue. Practical considerations such as the smaller size of the mouse brain allow for even more comprehensive sampling using *in situ* hybridization. This method was used in a systematic, brain-wide survey resulting in the Allen Mouse Brain Atlas (AMBA), which has been cast as a three dimensional atlas at 200 micron resolution (Lein et al., 2007; www.mouse.brain-map.org/). In both the AHBA and the AMBA, the locations of samples have been labeled according to conventionally defined brain atlases at multiple levels of granularity, allowing analysis of both broad and fine neuroanatomical regions. Though based on different data collection techniques, the high resolution and multi-level structure labelling of these datasets make them well-suited for investigating molecular neuroanatomy. Additionally, Chapter 3 compares the AHBA to another publicly available microarray dataset from the adult human brain, previously described by Gibbs et al. (2010). This chapter provides a description of these datasets, details a procedure for standardization used in Chapters 3 and 4, and briefly discusses some of the methodological decisions made in processing these datasets.

2.2 Allen Human Brain Atlas (AHBA)

The AHBA includes six genome-wide microarray datasets from adult donors aged 24-57 with no history of neuropathology (Hawrylycz et al., 2012; www.human.brain-map.org/). These data were collected from post-mortem samples and analyzed using a custom Agilent 8x60k cDNA array chip. Microarray results were preprocessed (including normalization to account for array-specific biases, within-batch intensity differences, within-donor batch-level effects, and cross-brain effects) by the research team at the Allen Institute for Brain Science (AIBS). Between 300 and 450 anatomically distinct samples were available from the left hemisphere for each donor. Although samples were also available from the right hemisphere for two donors, only left-hemisphere samples were used here. The neuroanatomical region from which each sample was obtained is annotated according to a hierarchical classification scheme (using standardized nomenclature and based on gyral / sulcal landmarks in the cerebral cortex), and the ~60,000 probes were annotated with relevant gene symbols by the AIBS group.

2.2.1 Probe selection

For genes represented by two probes, the probe with the highest mean intensity across all samples for all donors was selected. For genes represented by more than two probes, each probe's Pearson Product-Moment Correlation Coefficient (PCC) with each of the other probes for this gene across all samples for all donors was calculated, and the probe with the highest mean PCC (the "most average" probe) was selected. The “probe-per-gene” data includes ~32,000 genes. The subset of genes from this dataset that are treated here varies by analysis, and is specified in each case.

2.3 Allen Mouse Brain Atlas (AMBA)

The AMBA provides cellular resolution expression profiles based on non-isotopic *in situ* hybridization (ISH) for ~20,000 genes in the 56-day-old male C57BL/6J mouse. Volumetric datasets were used here, in which sections from each experiment were registered to a common three-dimensional template and binned into voxels with 200μm resolution (Lein et al., 2007; Ng et al., 2007). In this volumetric space, the expression of a gene is summarized by a measure of expression energy (i.e., the average intensity of pixels in the ISH image intersecting a given voxel). For this analysis, we used data from image series in the coronal plane, comprising 4,108 genes with restricted expression patterns, which had been selected by the AIBS team for a brain-wide survey. For consistency with the human data, only voxels from the left hemisphere were used in these analyses.

2.4 Gibbs dataset.

Microarray expression data from 4 distinct neuroanatomical regions - the frontal lobe (FL, Brodmann areas 9 and 46), temporal lobe (TL, Brodmann areas 21, 41, and 42), cerebellum (CB), and pons (PO) - were obtained from Gene Expression Omnibus (GEO), Accession Number GSE15745. These data were previously reported by Gibbs et al. (2010). Gene expression profiling was performed using Illumina humanRef-8 v2.0 Expression BeadChips as described in van der Brug et al. (2008). Expression data were processed and quality controlled using procedures similar to those described in Wolock et al (2013). The un-normalized mRNA microarray expression data were downloaded from

GEO and associated with the sample phenotype annotations and the mRNA microarray probe annotations for the platform used, GEO platform accession GPL6104. Low quality and technical probes as annotated using illuminaHumanv2.db from the Bioconductor package were removed from further analysis. Quantile normalization (Bolstad et al., 2003) was applied using normalized exponential quantile normalization (Shi et al., 2010) with the *neqc* function from Bioconductor (Ritchie et al., 2011). Categorical covariates were controlled using ComBat (Johnson et al., 2007), which corrected for effects of batch, tissue source and sex. ComBat has been previously shown to have the best overall performance, relative to other commonly used methods, in removing artifactual correlations induced by group effects (Chen et al., 2011), including when applied to Illumina platforms (Kitchen et al., 2010). The *correctBatchEffect* routine from *limma* (Smyth, 2005) was used to correct for the continuous covariates age, ancestry and post-mortem interval (PMI) using linear regression. After preprocessing, the *Gibbs dataset* consisted of 19,910 features (i.e., probes) and the number of samples that passed quality control varied by tissue source ($N^{FL} = 137$, $N^{TL} = 134$, $N^{CB} = 135$, $N^{PO} = 137$).

2.5 Probe standardization

Some form of normalization is necessary for meaningful comparisons across expression values from different probes in microarray datasets such as the AHBA and the Gibbs dataset. For consistency, normalization procedures described here were also applied to the AMBA.

A "weighted z-scoring" procedure (described in Ch. 2; see also Myers et al., 2015) was designed to give equal weight to brain structures regardless of the number of available samples. To some extent, this corrects for bias resulting from non-uniform sampling - in particular, the AHBA dataset included a greater number of samples from human cerebral cortex (1216 samples) than from any subcortical structure (1043 samples total). Because the Gibbs dataset includes one sample per region from each donor (i.e., uniform sampling across regions), weighted z-scoring amounts to equally weighting all samples in that dataset.

Details of the weighted z-scoring procedure are given below. Briefly, for each gene (probe), weighted means and standard deviations were calculated across available samples such that the set of all samples from a set of selected brain regions received equal overall weight, and these were used to normalize all expression values. Note, however, that the differential expression analysis in Chapter 3 required data that had not been standardized, and that all analyses in Chapter 5 used conventional z-scoring.

In this procedure, expression data are represented as a matrix E(*g, s*), indicating the expression level of gene *g* in sample *s*. This matrix is partitioned into *K* distinct sets, $R_1, R_2, \ldots R_k$, where R_j contains the indices of all samples in that set (i.e., brain region). For gene *g* and partitioning *S*, a weighted mean expression level is defined as:

$$\mu^S(g) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{|R_k|}\sum_{i\epsilon R_k} E\,(g,i)\right) \qquad (1)$$

and the weighted standard deviation is defined as:

$$\sigma^S(g)=\sqrt{\frac{\sum_{k=1}^{K}\frac{1}{|R_k|}\left(\sum_{i\epsilon R_k}\left(E(g,i)-\mu^S(g)\right)^2\right)}{K\left(\frac{\sum_{k=1}^{K}|R_k|-1}{\sum_{k=1}^{K}|R_k|}\right)}} \qquad (2)$$

The most basic case would be all samples in the dataset belonging to the same region, i.e. $S = \{R_1\}$ and $R_1 = \{1, 2, ..., N\}$, where N is the total number of samples. The weighted values would then reduce to the standard formulations for mean and standard deviation, and each sample would receive equal weight. With different partitions S (i.e., different groupings of the samples based on neuroanatomical labels), these procedures give equal overall weight to the sets of samples in each region R_k. The standardized expression level of gene g in sample i given partition S is then:

$$\tilde{E}^S(g,i) = \frac{\left(E(g,i)-\mu^S(g)\right)}{\sigma^S(g)} \qquad (3)$$

2.6 Data selection, processing, and labelling

Several methodological decisions contributed to the selection and processing of the data used in these analyses. This section includes a brief discussion of some of these decisions.

2.6.1 Weighted z-scoring

First, "weighted z-scoring" of probes (described above) was used as a means of addressing the presence of non-uniform sampling of brain regions in the AHBA dataset, and applied also to the AMBA dataset (as well as the Gibbs dataset, where it was equivalent to conventional z-scoring). To examine the effect of weighted z-scoring, the AHBA and AMBA datasets used in Chapter 3.3 and Chapter 4 (which used this

procedure) were compared to the same datasets when probes were standardized using conventional z-scoring. For the weighted z-scoring procedure, the following regions were used to partition the brain: the cerebral cortex, hippocampal formation, amygdala, striatum, globus pallidus, thalamus, hypothalamus, midbrain, pons, medulla, and cerebellum. (Chapter 4 gives the numerical identifiers associated with these regions in the AHBA and AMBA databases; see Table 4.1.)

Weighted z-scoring has the effect of de-emphasizing the influence of more heavily sampled brain regions on brain-wide expression values; therefore, its effect varied somewhat by brain region. Pearson's Product-Moment Correlation Coefficients (PCCs) were calculated between sample (or voxel) expression profiles within each region used to partition the brain after weighted z-scoring, and within each of the same regions after conventional z-scoring. Mean within-region PCCs changed by an average (across regions) of ~0.1 in the human data, ~0.05 in the mouse data, and ~0.03 across species, with the cortex (and to a lesser extent the hippocampus and amygdala) showing increased correlations and other regions showing decreases.

2.6.2 Standardization of probes in the AMBA

For consistency, the same standardization procedure was applied to probes in the mouse dataset as to probes in the human dataset, though it is not strictly necessary to cast the mouse *in situ* hybridization based datasets in relative terms as is necessary for the human microarray-based datasets. Within-probe standardization of the mouse data has the effect of emphasizing differences between voxel expression profiles. Mean within-region PCCs for broad mouse regions decrease by an average of ~0.5 when either

weighted or unweighted z-scoring is applied to the mouse expression data; however, the mean PCC between two expression profiles drawn randomly from anywhere in the mouse decreases by ~0.65, suggesting that mouse regions appear more distinct after standardization.

2.6.3 Restriction to the left hemisphere

Similarly, because right-hemisphere samples were discarded from the two whole-brain human datasets, right-hemisphere voxels in the mouse brain were also discarded to maintain consistency across species. Not surprisingly, results did not substantially change when the mouse data were restricted to left-hemisphere voxels due to the strong left-right symmetry in the AMBA: mean left-hemisphere expression profiles are highly correlated with mean right-hemisphere expression profiles for the 11 broad mouse regions (mean cross-hemisphere PCC ~0.92; ± ~0.06). This was also true of 9 finer sub-cortical regions (dentate gyrus, Ammon's horn, subiculum, caudoputamen, nucleus accumbens, external globus pallidus, internal globus pallidus, cerebellar cortex, and cerebellar nuclei; mean cross-hemisphere PCC ~0.91, standard deviation ~0.06). Sixteen cortical areas also showed high cross-hemisphere correlations (ectorhinal area, perirhinal area, temporal association areas, posterior parietal association areas, retrosplenial area, agranular insular area, orbital area, infralimbic area, prelimbic area, anterior cingulate area, visual areas, auditory areas, visceral area, gustatory areas, somatosensory areas, somatomotor areas, and frontal pole of the cerebral cortex; mean cross-hemisphere PCC ~0.87, standard deviation ~0.06). Additionally, within-region mean PCCs for the broad

mouse regions change by an average of only ~0.01 when right-hemisphere voxels were included.

2.6.4 Neuroanatomical labels

In addition to decisions regarding pre-processing of the data, this study depends on the assignment of samples and voxels to conventionally defined neuroanatomical structures, and on decisions about which neuroanatomical labels to include in the definition of a structure. For example, in the AHBA ontology, "Hippocampal formation" is a child (substructure) of "Cerebral cortex." In order to study the hippocampal formation separately from the rest of the cortex, labels corresponding to any part of the hippocampal formation were excluded from the definition of the cerebral cortex. More generally, conventional neuroanatomy sometimes struggles with conflicting opinions regarding the identities and borders of different structures and their homologs (Bota et al., 2003). Therefore, this work focuses on regions of the brain whose definitions are more or less well-established and agreed upon, but is subject to any bias introduced by the labeling schemes applied.

CHAPTER 3: LARGE-SCALE, WITHIN-SPECIES PATTERNS OF NEUROANATOMICAL GENE EXPRESSION

3.1 Introduction

Studies of gene expression across the brain have revealed a close relationship between molecular and conventional neuroanatomy in individual species. This relationship is demonstrated by preferential or exclusive expression of certain genes in a given brain region, and a number of studies have identified such genes for coarsely-defined regions throughout the mouse brain (e.g. Pavlidis and Noble, 2001; Zirlinger et al., 2001; Lein et al., 2007). The Anatomic Gene Expression Atlas (AGEA), which includes online tools designed for exploring the AMBA, shows molecular relationships between brain regions based on both preferential gene expression in certain brain regions, and correlations between gene expression profiles from throughout the mouse brain (Ng et al., 2009).

The molecular underpinnings of conventional region boundaries emerge in striking detail in the AMBA. Unsupervised clustering of voxels (based on multivariate patterns of gene expression at each location) yields clusters which correspond closely to mouse brain areas and cortical layers, with cluster boundaries appearing even between cortical areas (Bohland et al., 2010). A similar analysis limited to expression of neuron marker genes yields clusters corresponding to over fifty brain areas (Ko et al., 2013). Voxel clustering also shows the molecular heterogeneity of brain structures, where expression profiles from subcortical nuclei break into small clusters while those from

more homogenous structures such as the cerebral cortex form large individual clusters (Lein et al., 2007).

Other studies have used a variety of computational techniques to elucidate the relationship between gene expression and neuroanatomy in the mouse brain, and specifically the AMBA. Mahfouz and colleagues (2015) revealed distinctions between mouse brain regions by applying an optimized version of the data reduction technique t-Distributed Stochastic Neighbor Embedding (t-SNE, Van der Maaten and Hinton, 2008) to the AMBA (and the AHBA; see below). Grange et al. (2014) cast the AMBA dataset as a linear combination of cell types, using 64 previously measured expression profiles to model the spatial distribution of different transcriptomically defined cell types. French and Pavlidis (2011), also using the AMBA as well as rat data from the Brain Architecture Management System (BAMS; Bota et al., 2005), found that regional gene expression profiles are statistically related to regional connectivity.

Comparable organization has been found based on human brain gene expression, where one of the first comprehensive applications of this approach showed brain samples clustering by region of origin (Roth et al., 2006). Neocortical samples grouped together, and showed more similarity to the hippocampus and amygdala than to the rest of the subcortex, while the cerebellum stood out with a particularly unique molecular profile. Similar to the finding by Lein et al. (2007) in mouse, Hawrylycz and colleagues (2012) showed much higher internal homogeneity for gene expression profiles in the human neocortex than in many subcortical structures, whose varying cellular architecture is reflected by higher numbers of differentially expressed genes. Still, topographical

relationships between human cortical areas were preserved in their gene expression profiles, reflecting graded differences in cortical cell populations and possibly relative positions of progenitor cells in early development. Application of t-SNE to the AHBA also revealed few within-cortex distinctions, but did show separation between broader brain regions (Mahfouz et al., 2015).

Nearly all of the human studies mentioned above use the AHBA, which offers uniquely high spatial resolution (as compared to other gene expression datasets from the human brain) but low sample size (six donors). Similar sampling properties are shared by the datasets used in Roth (2006), Johnson (2009), and Khaitovich (2004), though without such high spatial resolution (19, 13, and 8 brain areas, respectively). In contrast, Oldham et al. (2008) and Gibbs et al. (2010) use datasets from over 100 donors each, but with samples from only 3 or 4 brain regions in each donor. Relatively fine neuroanatomical specificity and broad coverage (i.e. sampling from areas throughout the brain) comes at a price of low sample size, which limits the ability to distinguish individual variability from noise. These trade-offs emphasize the need for validation across datasets. Most of the work described in this dissertation focuses on the AHBA and AMBA; however, a comparison of the AHBA with the dataset from Gibbs et al. (2010) is presented below.

This chapter examines large-scale structure in gene expression data from the human and mouse brain, using a largely exploratory approach applied to the AHBA and the AMBA. Transcriptomic structure in the AHBA is cross-validated through comparison to a second dataset with larger sample size and coarser spatial resolution.

Visualizations revealing organization of samples based on gene expression and its correspondence to neuroanatomical labels serve to lay the groundwork for the studies presented in Chapters 4 and 5.

3.2 Modes of variability

The singular value decomposition (SVD, described briefly below) is useful for de-noising and reducing the dimensionality of high-dimensional datasets. Essentially, the singular value decomposition of a matrix yields (1) an orthonormal basis for its columns, (2) an orthonormal basis for its rows, and (3) "singular values" indicating how each vector in each orthonormal basis is scaled. In a sample-by-gene expression matrix, (1) defines a rotated coordinate system for genes and (2) for samples, while (3) indicates the spread of datapoints along each axis in the new system.

To understand why this is useful, compare the new space defined by (2) to the coordinate space of the original sample data. This original space is defined by thousands of axes, one per gene, and each sample is represented in the space by as many coordinates. Because genes do not vary *independently*, the way samples are spread out in the space may be effectively described by a smaller number of orthogonal axes, each of which captures as much variability across samples as possible. If we know how much variability each axis captures, we can assess how many sources of variability there are in the data. Singular values give us this information. (This is also true if we reverse "samples" and "genes" in the above example; singular values describe the variability captured by each axis in either of the orthonormal vector sets).

Below, the first use made of SVD is to see how much of the original variability in the data is captured as more axes (dimensions) in the new coordinate space are included, beginning with those which capture the most variability (i.e. those with the highest corresponding singular values). If most of the variability in the data can be captured by just a few orthogonal axes, then the structure of the data (i.e., its correlations across anatomical space and across genes) is relatively simple. The more dimensions are required to describe variability in the data, the more rich the structure of that data. The second use of SVD made below is to project the sample data down to the three-dimensional space defined by the three axes in the new space with the highest singular values; i.e., those that capture the most variability. The resulting visualization of sample locations in this three-dimensional space gives a sense of whether and how samples separate based on region of origin.

Bohland et al. (2010) applied SVD to the AMBA, finding rich structure in the data as well as region-based separation of voxels in the three-dimensional space capturing the most variability. Here, SVD is applied to both a the AMBA, and also to the AHBA. Note that since Bohland et al.'s (2010) study, a new algorithm has been applied to the AMBA to register the expression data to the reference space; therefore these analyses use a different version of the AMBA (see documentation for the AMBA dataset at www.mouse.brain-map.org/).

3.2.1 Gene selection and standardization

Human orthologs were identified for 3,792 of the genes available in the mouse dataset using NCBI HomoloGene (NCBI Resource Coordinators, 2015;

www.ncbi.nlm.nih.gov/homologene). Expression data for these genes (in both species) constituted a common dataset, which was used in both this section and Section 3.3. Expression data for each gene was standardized using the weighted z-scoring procedure described in Chapter 2, with the following 11 broad brain regions used to partition the data: cerebral cortex, hippocampus, amygdala, striatum, globus pallidus, thalamus, hypothalamus, midbrain, pons, medulla, and cerebellum. For the human data, this standardization was performed separately for each donor.

3.2.2 Singular value decomposition

Formally, the singular value decomposition of a $p \times q$ matrix M is defined as $M = USV^T$, where U is a $p \times q$ matrix, and both S and V are $q \times q$ matrices. The columns of U and S are known as *left* and *right singular vectors*, respectively. The matrix S is only non-zero along the diagonal, which contains *singular values*. Each singular value corresponds to both a right and a left singular vector.

When M is centered such that row and column means are 0, each right singular vector is an eigenvector of the covariance matrix $M'M$ (the covariance between rows of M, without standardizing by number of columns, since singular values will scale the eigenvector), while the square of the corresponding singular value is the variance described by that eigenvector. In other words, if M is a sample-by-gene matrix (as here), the first right singular vector indicates the direction of the greatest "spread" among the samples (and the first singular value indicates its extent); the second right singular vector indicates the direction of second-greatest spread that is orthogonal to the first (and the second singular value indicates its extent); and so on. The same is true of left singular

vectors, the covariance matrix of MM', and the variance / covariance of genes. Right singular vectors are referred to here as *spatial modes*, and left singular vectors as *gene modes* (borrowing from Bohland et al., 2010). These are equivalent to the orthogonal axes of the rotated coordinate spaces described at the beginning of this section.

The first use of SVD in this section (mentioned above) is to see how much of the original variability in the data is captured as more modes are included. As the variance captured by a mode is the square of the corresponding singular value, the proportion of the total variance captured by that mode is that variance divided by the sum of squared singular values. The cumulative sum of this proportion, moving from the first singular value onward, shows how many modes are required to capture a given proportion of variance in the data.

The second use of SVD in this section is to project the samples to a 3-dimensional subspace that captures the largest possible proportion of their variance. Because $MV = US$, the matrix US is computed while retaining only the first 3 singular values in S. This is the projection of the sample data in M onto the first spatial modes contained in V, but is easier to compute than using M and V directly, due to S having non-zero values only along the diagonal. (This projection is performed only for samples; however, since M has been double-centered, the same procedure could be applied to project genes onto the first three gene modes.)

SVD was applied to the AMBA, to each donor in the AHBA separately, and to a single expression matrix combining samples from all human donors (the latter for visualization of samples projected into a 3D space). In each case, the expression matrices

included data from the 3,792 probes described in above in "Gene selection and standardization", and from left-hemisphere samples / voxels from the 11 broad regions listed in that section. All expression matrices were double-centered (i.e., row and column means were subtracted from each value) before SVD was performed. For comparability to Bohland et al. (2010), mouse probes were not converted to z-scores for this analysis. Note that probe and voxel selection differ between the two studies: Bohland et al. (2010) used 3,041 probes chosen for consistency with a second dataset (also part of the AMBA, but not used here), and used data from the whole brain rather than the left hemisphere only.

3.2.3 Results

The proportion of variance explained by each singular value (i.e. that value squared divided by the sum of squares of all singular values in *S*) is shown as a cumulative sum in Fig 3.1 for the AMBA and the six donors of the AHBA. Over 100 modes are required to explain at least 90% of the variance in the human data, while 222 are required to explain the same proportion in the mouse data. The higher spatial resolution of the mouse data may explain its comparatively rich structure; different voxels can reflect different cell type distributions that would be averaged together in a sample from the AHBA. The spread of values across human donors is relatively small, with donor H0351.2001 requiring more modes than the others to account for the same proportion of the variance. Donor sample counts, in ascending number of modes required to explain 90% of the variance, were: 363, 292, 392, 374, 434, 404.

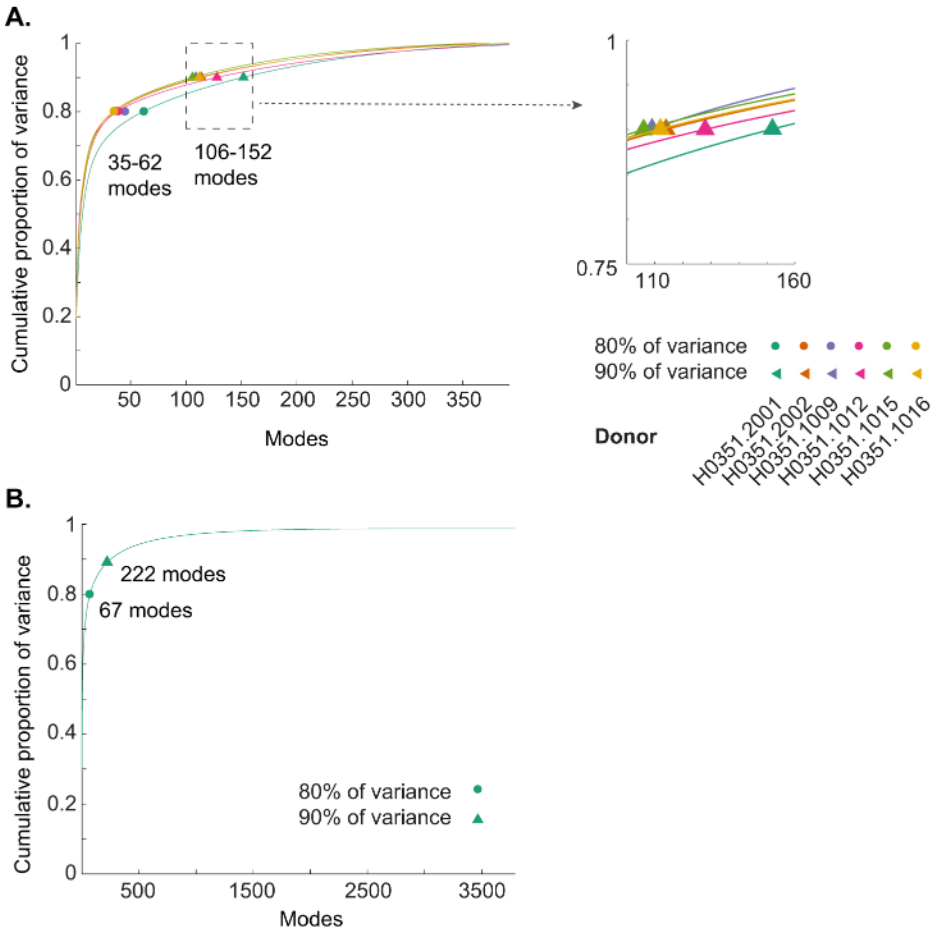


Fig 3.1. Cumulative proportion of variance captured by modes in the human (A) and mouse (B) expression matrices. Circles and triangles indicate number of modes required to capture at least 80% and 90% of the variance in the data, respectively. SVD was performed separately in each human donor. In A, colors correspond to donors, and inset shows the spread of values across human donors for number of modes required to capture at least 90% of the variance.

When human samples and mouse voxels were projected onto the first three modes, clusters appeared corresponding roughly to several broad regions (Fig 3.3). The cerebellum was particularly distinct from other parts of the brain in both species, particularly the human; however, the hippocampal formation, amygdala, striatum, and thalamus also show some separation.

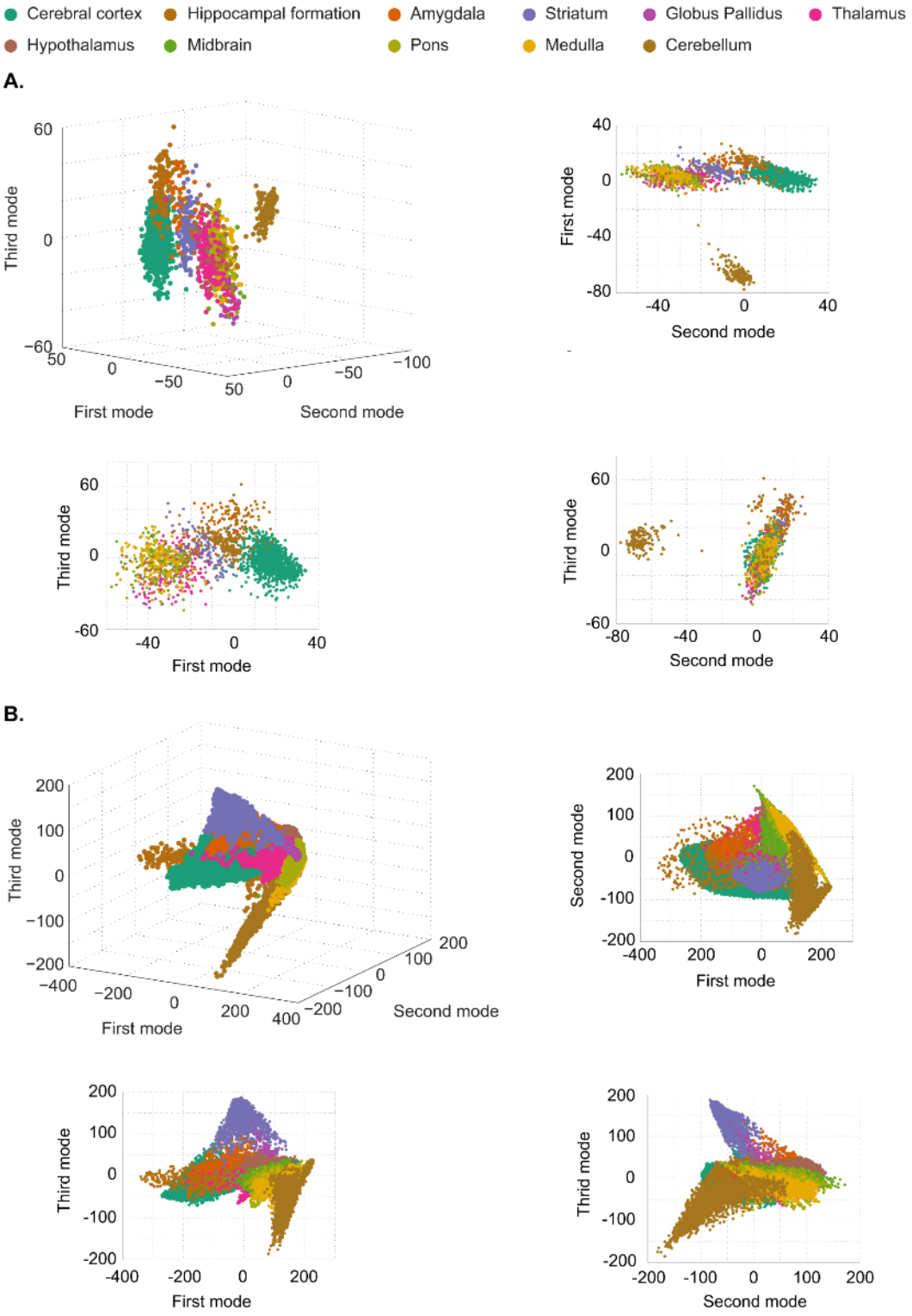


Fig 3.2. Projection of human samples (A) and mouse voxels (B) onto the first three spatial modes (i.e. right singular vectors). Additional plots show projection onto each pair of those modes. Dot color indicates region of origin.

3.3 Correlations between expression profiles from different locations in the brain

3.3.1 Calculation and comparison to empirical null distribution

To assess the regional organization of transcriptomic relationships, tissues throughout the brain were compared in each species using Pearson's Product-Moment Correlation Coefficient (PCC) calculated between pairs of profiles. Genes were selected and standardized as described in Section 3.2. PCCs were calculated (i) between all pairs of samples from the human data, and (ii) between all pairs of voxels from the mouse data.

Two additional calculations were made to assess the strength of within-region similarity. First, the quartiles of all within-region PCCs were calculated for each of 11 broad regions and a number of finer regions (varying by species). Second, PCCs for each region were compared to a distribution of 1000 PCCs calculated between sample pairs (or, in the mouse data, voxel pairs) for which one sample originated within the region, and the other did not. For each region, the mean of the resulting percentile ranks was calculated to assess specificity of within-region correlations.

3.3.2 Results

Heatmaps of PCCs between expression profiles show organization based on brain region in both species (Fig 3.3). Overall, PCC values are higher in the human than mouse. This probably results from the lower spatial resolution of the AHBA. Samples average across more cells than voxels, making them both less prone to noise and less revealing of distinct cell type populations.

In both species, blocks of high values (in comparison to the rest of the matrix) correspond to brain regions including the cerebral cortex, thalamus, hypothalamus, striatum, amygdala, and cerebellum. Striatal expression profiles show particularly high PCCs with each other in comparison to other within-region PCCs. In some cases, expression profiles from different regions show slightly elevated PCCs. This is true of the cerebral cortex and the hippocampal formation, and to a lesser extent of the cerebral cortex and amygdala. The midbrain, pons, and medulla are represented by a single block of positive-valued PCCs. Fig 3.3E shows elevated PCCs within the mouse caudal pallidum and globus pallidus, with the internal segment showing even higher PCCs, while PCCs between samples from the human globus pallidus do not appear to distinguish between the internal and external segments (Fig 3.3B). However, it is possible that coarse sampling of the human globus pallidus has obscured distinctions between cell type distributions in the GPi and GPe. In both species, the cerebellum panel (Fig 3.3C, F) shows higher correlations within the cerebellar cortex and cerebellar nuclei than across those sub-structures.

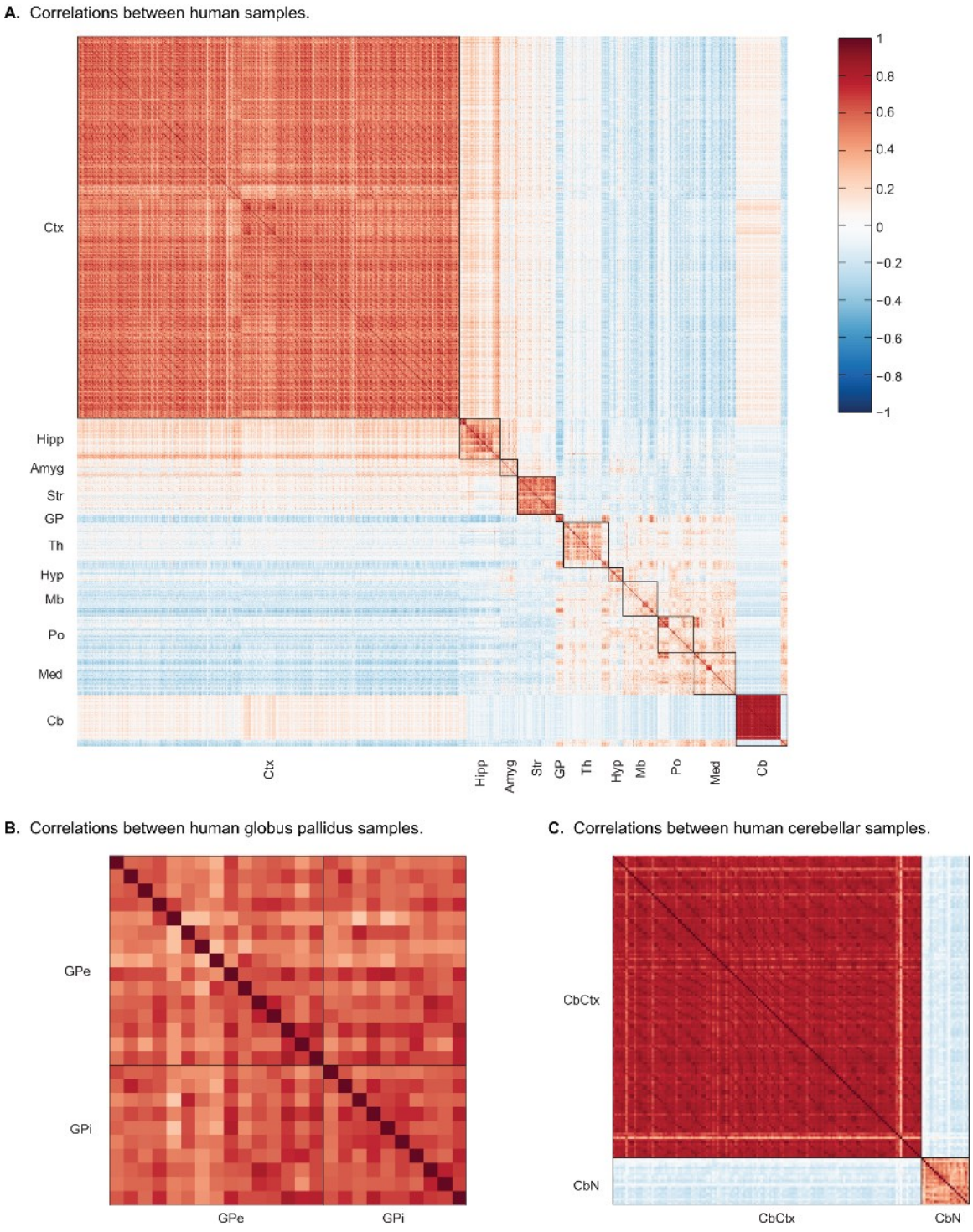


Fig 3.3 (cont. on next page). Correlations between expression profiles of human samples and mouse voxels, within-species. Color bar applies to all plots.

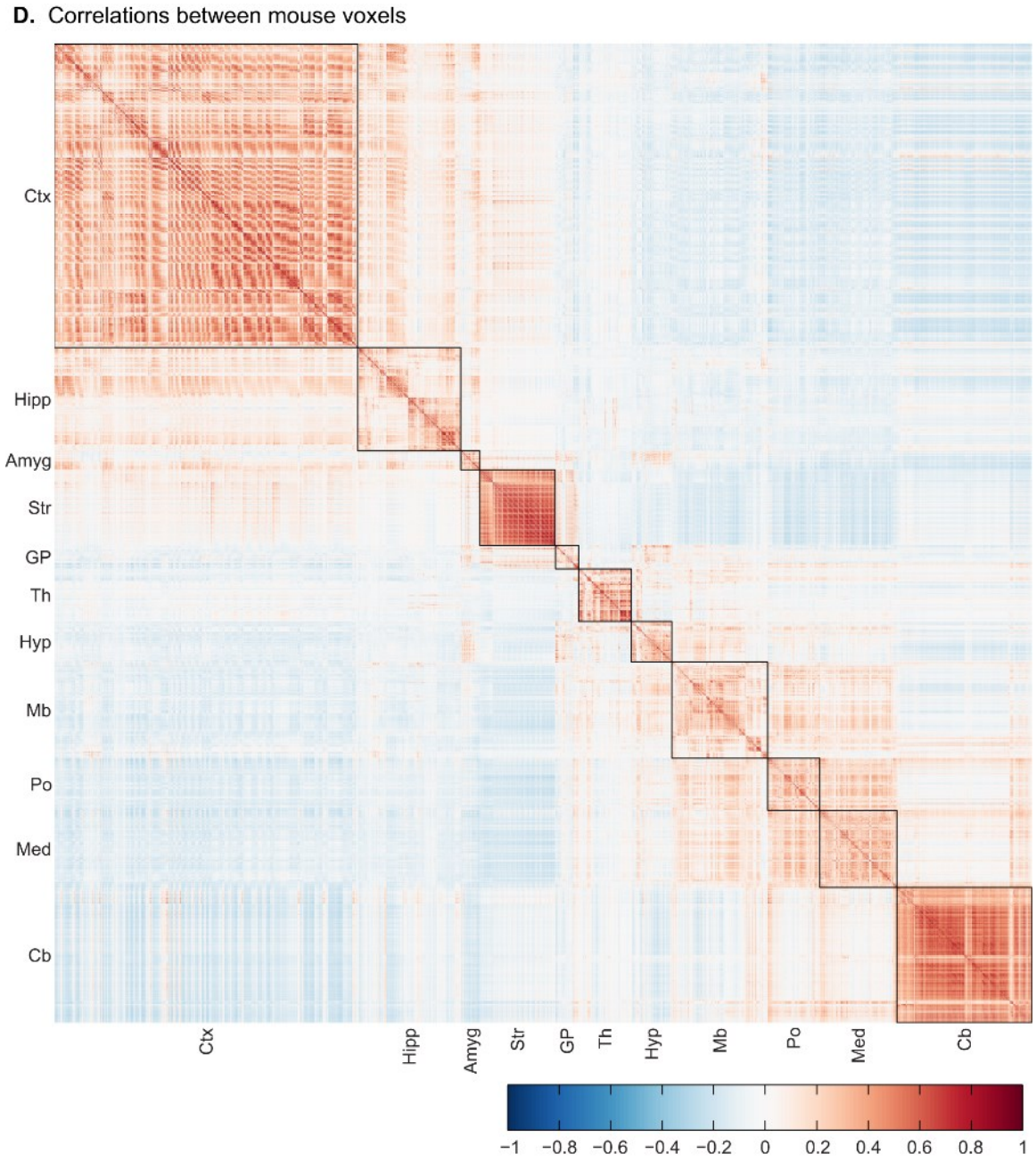


Fig 3.3 (cont. on next page). Correlations between expression profiles of human samples and mouse voxels, within-species. Color bar applies to all plots. Voxels from the cerebellum without a more specific label are grouped at the end.

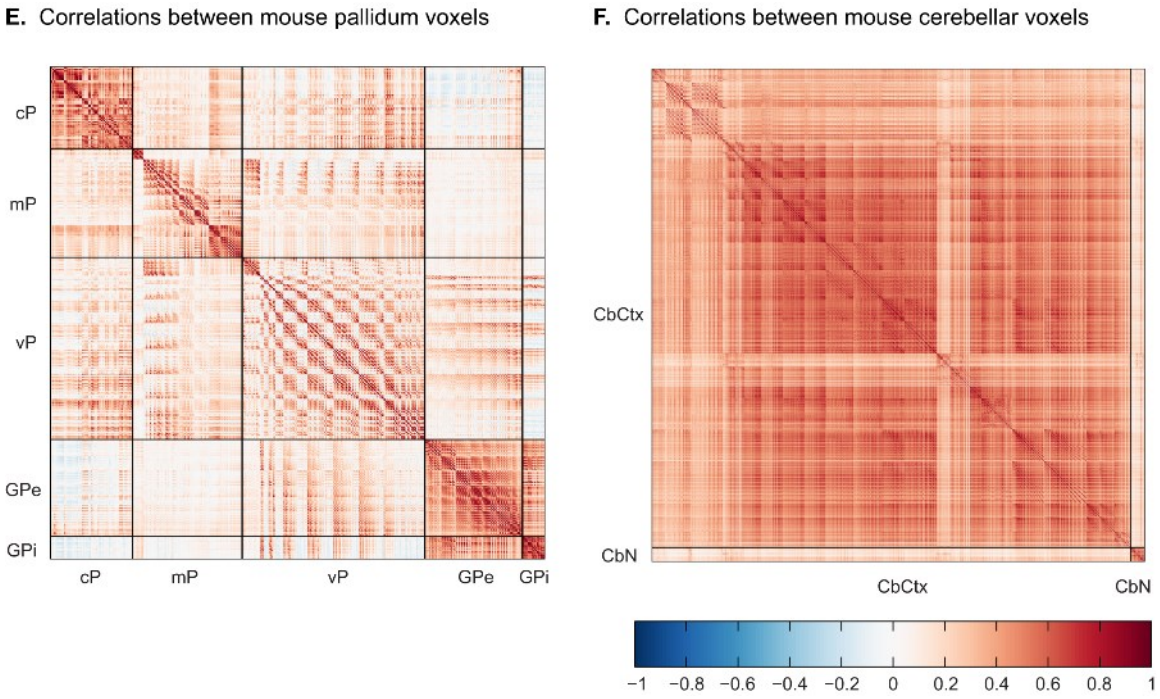


Fig 3.3. Correlations between expression profiles of human samples and mouse voxels, within-species. Color bar applies to all plots. In F, voxels without a more specific label than "Cerebellum" are not included.

Fig 3.4 summarizes within-region distributions of PCCs using "box plots." For the most part, the median PCC within a fine region was higher than for the parent region (though not always to a great extent), pointing to the distinct molecular compositions of sub-structures within broad regions. Exceptions included human cerebellar nuclei, with no PCCs reaching the mean value for the cerebellum as a whole (Fig 3.4A), and human cerebral cortex, where cortical areas showed mean PCCs very similar to the mean for the parent lobe (Fig 3.4B).

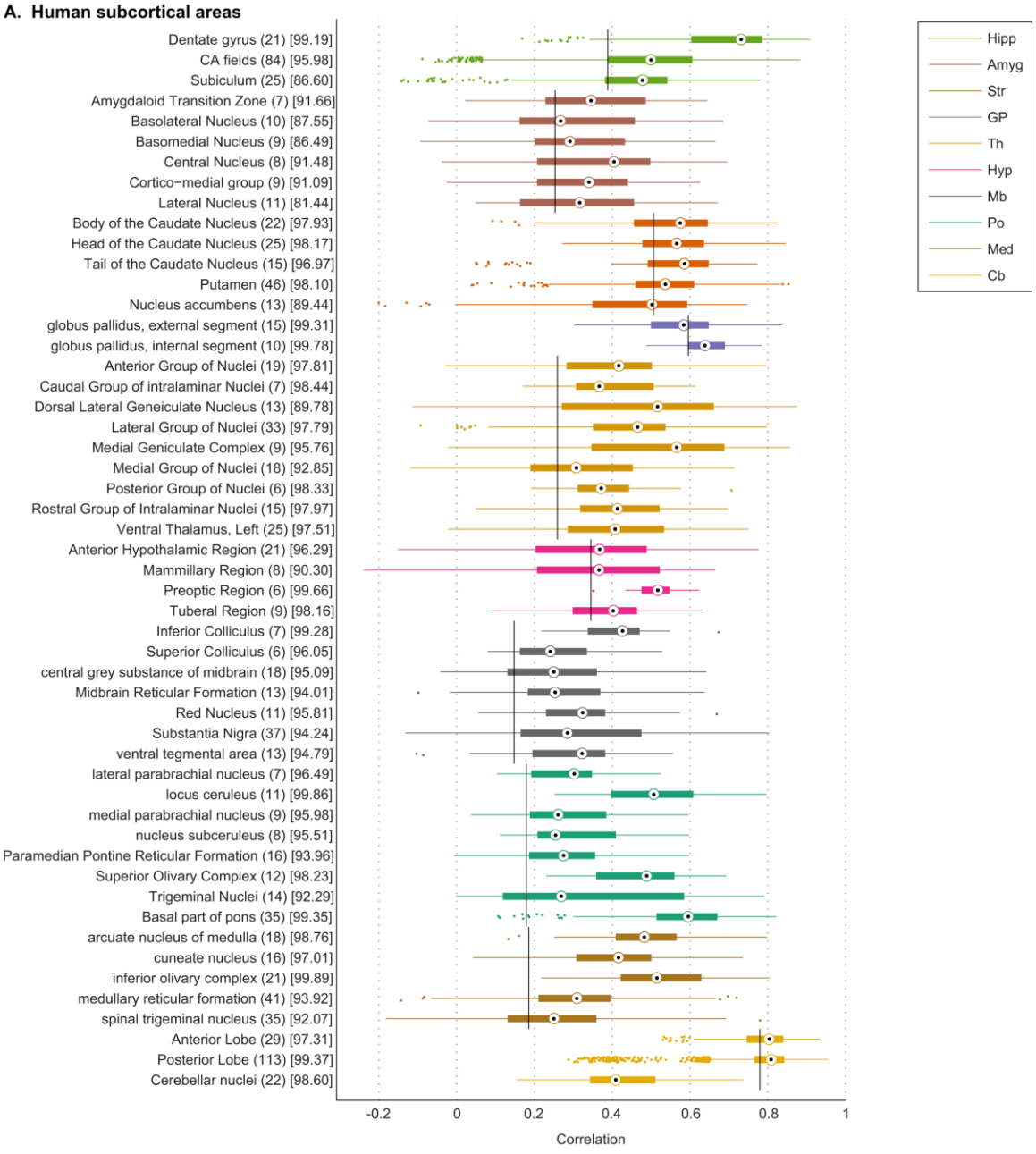


Fig 3.4 (cont. on next page). Median correlation between expression profiles from within the same brain region. Box edges represent the 25th and 75th percentile values and whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlation of parent structure. Numbers in parentheses are sample / voxel counts. Numbers in square brackets are mean percentile rank of within-region correlations in distributions of correlations where one expression profile belongs to the region in question and the other is from elsewhere in the brain.

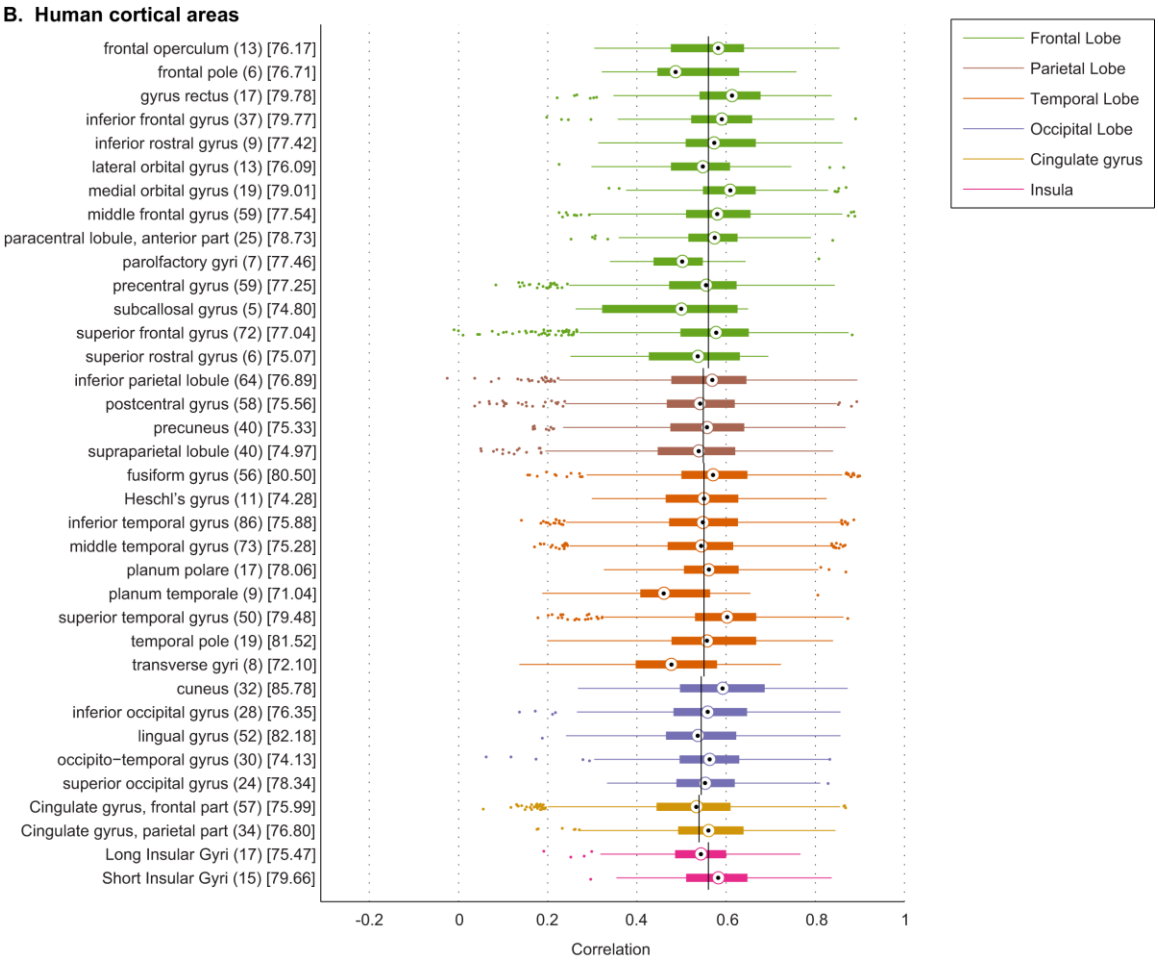


Fig 3.4 (cont. on next page). Median correlation between expression profiles from within the same brain region. Box edges represent the 25th and 75th percentile values and whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlation of parent structure. Numbers in parentheses are sample / voxel counts. Numbers in square brackets are mean percentile rank of within-region correlations in distributions of correlations where one expression profile belongs to the region in question and the other is from elsewhere in the brain.

Fig 3.4 (cont. on next page). Median correlation between expression profiles from within the same brain region. Box edges represent the 25th and 75th percentile values and whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlation of parent structure. Numbers in parentheses are sample / voxel counts. Numbers in square brackets are mean percentile rank of within-region correlations in distributions of correlations where one expression profile belongs to the region in question and the other is from elsewhere in the brain.

Fig 3.4 (cont. on next page). Median correlation between expression profiles from within the same brain region. Box edges represent the 25th and 75th percentile values and whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlation of parent structure. Numbers in parentheses are sample / voxel counts. Numbers in square brackets are mean percentile rank of within-region correlations in distributions of correlations where one expression profile belongs to the region in question and the other is from elsewhere in the brain.

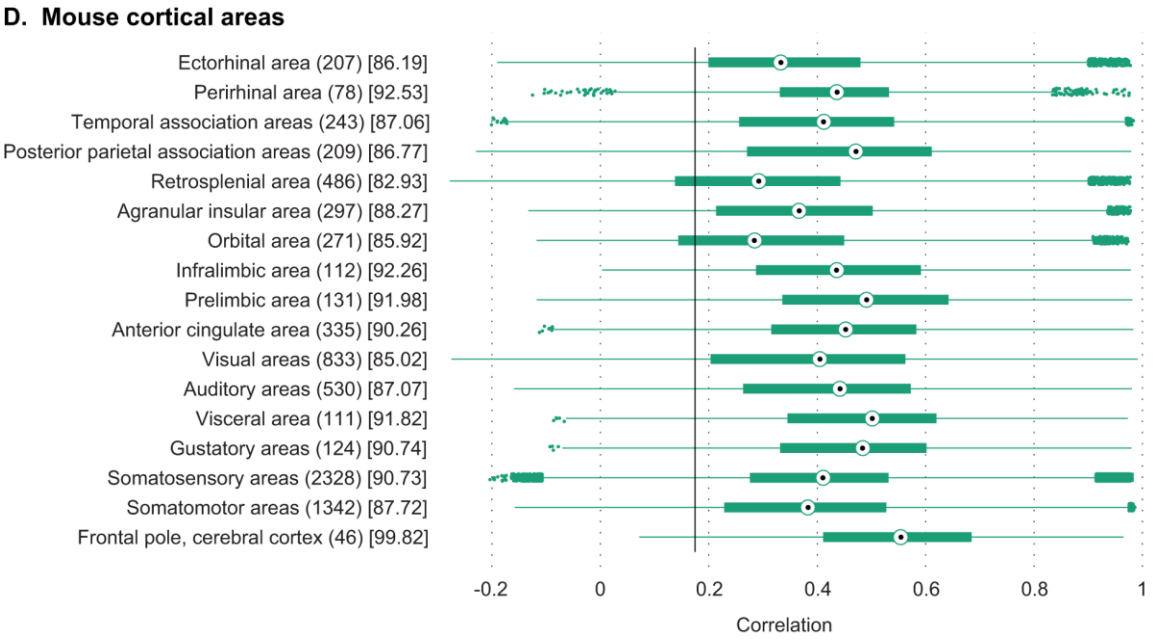


Fig 3.4. Median correlation between expression profiles from within the same brain region. Box edges represent the 25th and 75th percentile values and whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlation of parent structure. Numbers in parentheses are sample / voxel counts. Numbers in square brackets are mean percentile rank of within-region correlations in distributions of correlations where one expression profile belongs to the region in question and the other is from elsewhere in the brain.

For each fine brain region, the mean percentile rank of within-region PCCs against an empirical null distribution is given in Fig 3.4 (bracketed value following region name and sample count). High percentile ranks indicate that samples within the region resemble each other more strongly, overall, than they resemble samples from elsewhere in the brain. Except for the mean ranks of cortical areas, most exceed the 90th percentile in both species. The distribution of within-region mean percentile ranks (grouping cortical and subcortical areas) for both the human and mouse is significantly different from (with higher values than) a normal distribution with a mean at the 50th

percentile ($p < 0.03x10^{-8}$ and $p < 0.06x10^{-12}$ for the human and mouse, respectively; one-tailed Kolmogorov-Smirnov test).

The distribution of correlations from expression profiles within the same region tends to have a smaller range in the human than mouse data, and a greater tendency toward negative skew (Fig 3.4). This is probably also a result of lower spatial resolution in the AHBA, where larger samples can be expected to smooth variation in cell type distributions within a region, so that sample profiles tend to be similar.

3.4. Cross-dataset validation of anatomical relationships (in human brain)

3.4.1 Selection of probes for common gene set

Each of the 19,910 probes available in the preprocessed and quality controlled Gibbs dataset (see Section 2.4 for description) was mapped, where possible, to a unique gene symbol using the Illumina annotation file Human-Ref-8_V3_0_R3_11282963_A.txt. This resulted in 12,202 unique gene symbols, and 1,041 duplicate gene symbols (i.e., multiple probes mapping to the same gene). For the majority of genes, two or more relevant probes were available in the AHBA dataset. To create a common gene set across datasets, a simple algorithm was used to choose a single probe per gene from each dataset. For each gene, the probe with the highest mean correlation (across samples) with other probes for the same gene was chosen; in cases with exactly two probes, the probe with highest mean intensity was chosen. This resulted in a set of 11,841 genes represented by individual probes in both the Gibbs and AHBA datasets, which was used in the below analyses.

3.4.2 Differential expression of individual genes

Differential expression (DEX) of a gene was quantified by contrasting the expression levels in all samples from one brain region with expression levels in all samples from another region. In the Gibbs dataset, paired t-tests (pairing samples from the same donors) were performed, excluding any donors for whom samples were not available from both brain regions. In the AHBA dataset, 2-sample t-tests were performed across all samples (pooled across donors) within each of the two regions of interest, allowing for different population variances (i.e., Welch's t-test). P-values were corrected for multiple comparisons using the linear step-up false discovery rate (Benjamini and Hochberg, 1995). A minimum fold-change parameter was used, with its value varied to assess the sensitivity of results to the particular parameter choice.

3.4.3 Comparing region-specific expression profiles

From either dataset, a set of brain region specific expression profiles, each of which is a vector of length 11,841 encoding the average (over samples from that region) relative expression level of each of the genes in the common gene set, was calculated. To compensate for differences in probe efficacies and, especially, to account for the non-uniformities in sampling across brain regions, we standardized expression values based on a partition of samples into the 4 gross neuroanatomical areas represented in the Gibbs datasets (frontal cortex - FCTX, temporal cortex - TCTX, cerebellum - CB, and pons). For the Gibbs dataset, this amounted to equally weighting all samples. For the AHBA, this meant that means and standard deviations were computed over only the samples from

these four brain areas (or their sub-regions), and that each region was given the same total weight regardless of numbers of samples. In the AHBA, these standardization procedures were performed independently for each donor. Standardizing the AHBA expression profiles based on expression values in only the regions represented in the Gibbs sample enabled a ‘‘fair’’ comparison between datasets. PCCs were computed between the 4 region-specific expression profiles derived from the Gibbs dataset and the expression profiles from a set of AHBA regions in each of the 6 donor brains.

3.4.4 Results

Fig 3.5 shows three comparisons of DEX across regions in the AHBA and Gibbs datasets. In both datasets, over 20% of genes in the common gene set showed significant DEX between all pairs of the four brain regions except the frontal and temporal cortex (paired *t*-tests, $p < 0.05$ after correction using FDR and a minimum $\log_2$ fold change of 0.5). A volcano plot (showing $\log_2$ fold change against $-\log_{10}$ corrected p-value) for all genes in the FCTX / TCTX comparison based on the Gibbs data set is shown in Fig 3.5B. Only 26 genes (represented by dark blue circles) were differentially expressed between the two cortical areas in both datasets (Table 3.1). However, in general, there is greater agreement in the set of DEX genes across datasets (i.e., >63% overlap relative to the smaller gene set for all other region pairs). Fig 3.5C shows that the fraction of genes that show DEX in both datasets for each region pair remains relatively stable even as the fold change threshold is varied, again with the exception of FCTX / TCTX. Even as fewer genes meet the FC threshold, the number of overlapping genes is far greater than would be expected by chance for all tested FC values (hypergeometric tests, $p < 1 \times 10^{-275}$ for all

tests). The overlap in DEX genes in cortex is less stable; however, for fold changes below 1.0, the size of the overlapping DEX gene set is larger than expected by chance (hypergeometric tests, $p < 0.025$ for all tests). Because the FCTX / TCTX comparisons have fewer DEX genes overall, the estimated proportion is likely to be somewhat imprecise even below a FC of 1.0. Therefore, larger datasets may ultimately be necessary to provide an improved estimate of the number of genes with differential expression across cortical regions.

Symbol	Name	Location
TGFBI	transforming growth factor, beta-induced, 68kDa	5q31
TNNT2	troponin T type 2 (cardiac)	1q32
ZBBX	zinc finger, B-box domain containing	3q26.1
CNTN6	contactin 6	3p26-p25
COL5A2	collagen, type V, alpha 2	2q14-q32
DUSP13	dual specificity phosphatase 13	10q23.1
GAL	galanin/GMAP prepropeptide	11q13.2
GPX3	glutathione peroxidase 3 (plasma)	5q23
GSTM5	glutathione S-transferase mu 5	1p13.3
APOC1	apolipoprotein C-I	19q13.2
KCTD4	potassium channel tetramerization domain containing 4	13q14.12-q14.13
KNG1	kininogen 1	3q27.3
LAMP5	lysosomal-associated membrane protein family, member 5	20p12
LGR6	leucine-rich repeat containing G protein-coupled receptor 6	1q32.1
LXN	latexin	3q25.32
MET	MET proto-oncogene, receptor tyrosine kinase	7q31
ARL9	ADP-ribosylation factor-like 9	4q12
NECAB2	N-terminal EF-hand calcium binding protein 2	16q23.3-q24.1
NEFH	neurofilament, heavy polypeptide	22q12.2
ASGR2	asialoglycoprotein receptor 2	17p
PCP4	Purkinje cell protein 4	21q22.2
PDGFD	platelet derived growth factor D	11q22.3
PDYN	prodynorphin	20p13
PRRX1	paired related homeobox 1	1q24.3
BHLHE22	basic helix-loop-helix family, member e22	8q12.1
STC2	stanniocalcin 2	5q35.2

Table 3.1. Common DEX genes between Gibbs and AHBA datasets for TCTX vs. FCTX. DEX genes have $\log_2$ fold change > 0.5, FDR-corrected p-value < 0.05.

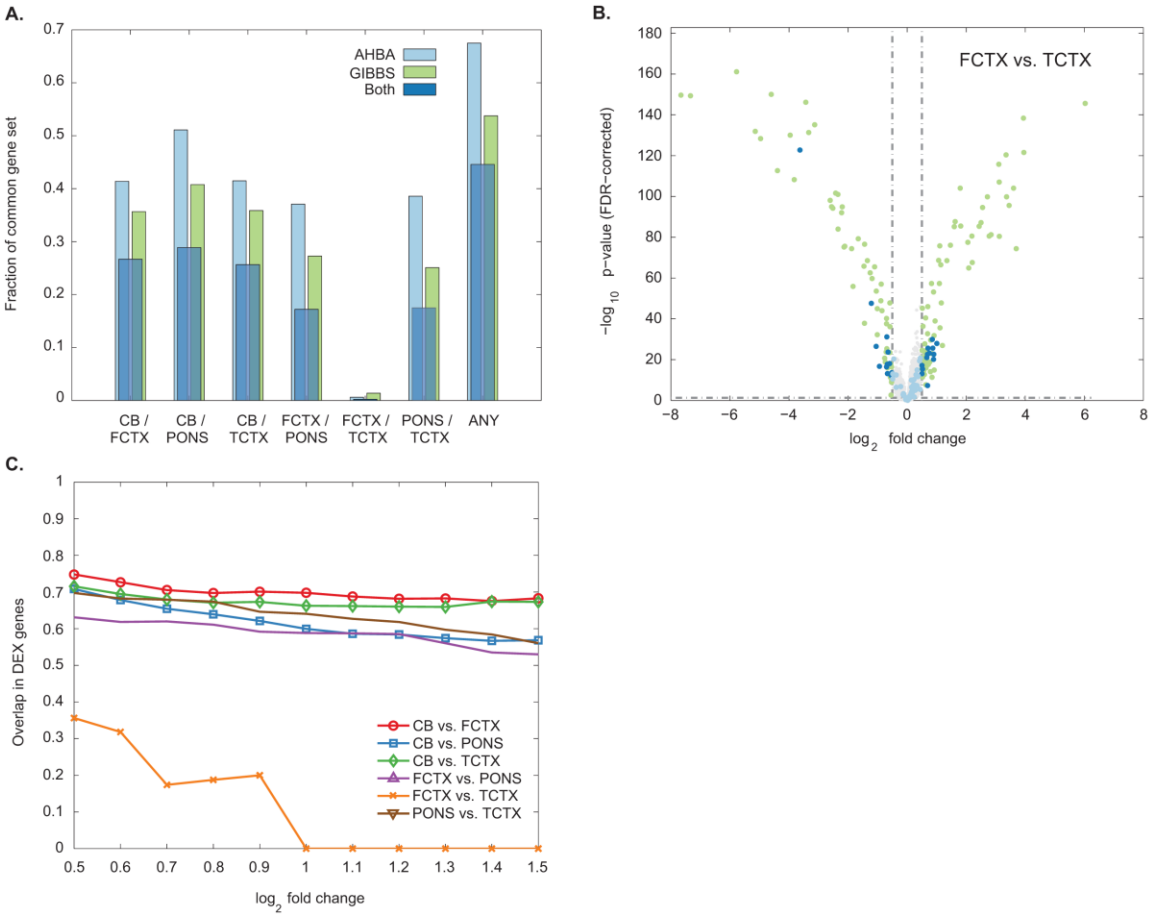


Fig 3.5. Cross-dataset comparison of differential expression across regions. A. Proportion of genes in the common gene set that are differentially expressed between pairs of structures in each dataset (minimum $\log_2$ fold change of 0.5 and $p < 0.05$, FDR-corrected). Light blue bars indicate proportions of DEX genes in the AHBA dataset (2-sample *t*-test), green bars in the Gibbs dataset (paired t-test), and dark blue bars show the fraction of overall genes showing DEX in both datasets. B. Volcano plot showing DEX in FCTX vs. TCTX for all genes in Gibbs dataset. Gray dots are not significant, green dots show DEX in Gibbs only, light blue dots in AHBA only, and dark blue dots show DEX in both datasets. C. Percentage overlap in DEX genes vs. $\log_2$ fold change threshold (in all cases $p < 0.05$, FDR corrected).

Fig 3.6A shows correlations between each of the Gibbs regions and several regions from the AHBA. These correlations are preferentially high between each Gibbs region and the profile from the corresponding region in the AHBA, consistently across donors. The cerebellum profile in each dataset is negatively correlated with all non-cerebellar profiles in the other dataset. FCTX and TCTX profiles are positively correlated

with all AHBA cortical structures, although each shows a slightly but not significantly higher correlation with the corresponding cortical lobe in the AHBA. Gibbs cortical profiles are also positively correlated with AHBA hippocampal profiles. Figs 3.6B and 3.6C show finer anatomical resolution for cortical, cerebellar and pons regions. From Fig 3.6C it is clear that the Gibbs CB profile compares with the AHBA cerebellar cortex and not the cerebellar nuclei, which are actually more molecularly similar to the Gibbs PONS samples.

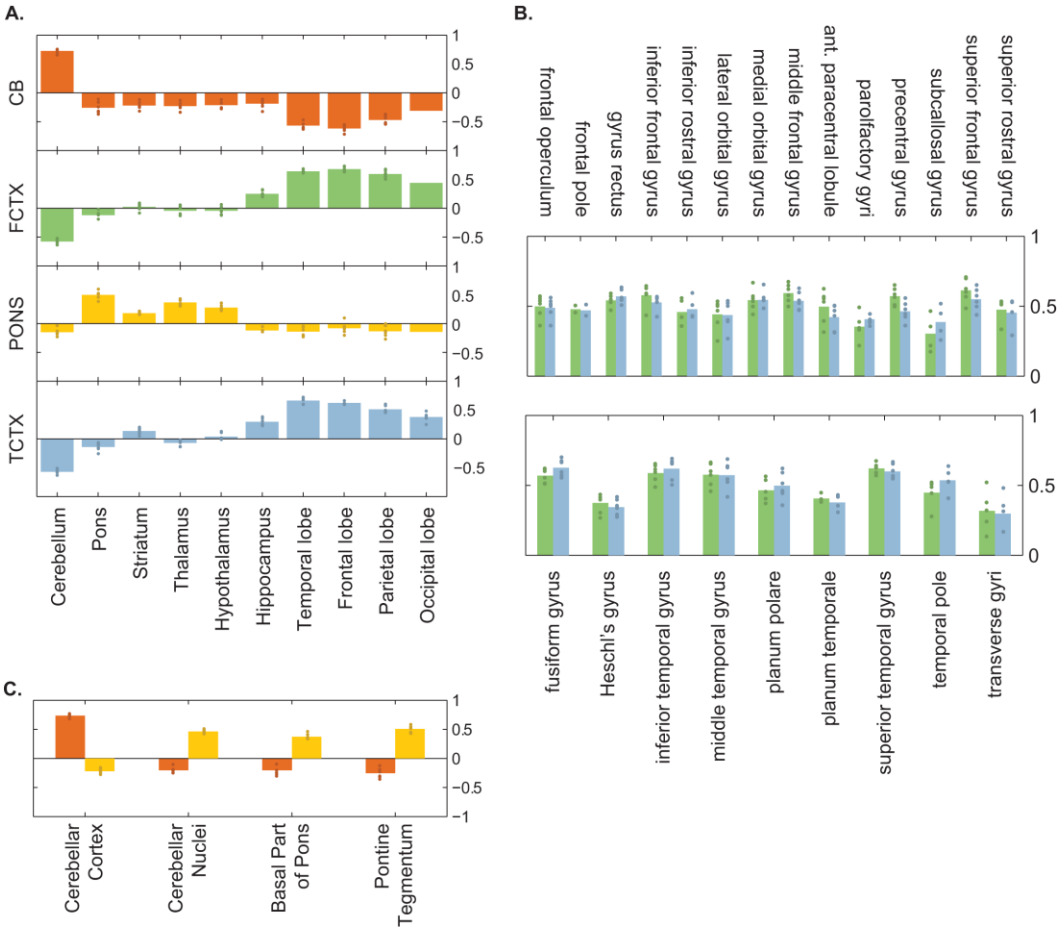


Fig 3.6. Cross-dataset comparison of region expression profiles. Correlations between Gibbs cerebellum (CB; orange bars), frontal cortex (FCTX; green bars), pons (PONS; yellow bars), and temporal cortex (TCTX; blue bars) samples and expression profiles derived from AHBA regions. Bar heights represent average PCC across AHBA donor brains, and individual dots indicate PCC for each donor. A. Correlations with broad regions. B. FCTX and TCTX correlations with individual gyri in the frontal lobe (top) and temporal lobe (bottom). C. CB and PONS correlations with cerebellar and pontine sub-structures.

3.5 Discussion

The visualizations and summary statistics presented above consistently show correspondence between transcriptomic and neuroanatomical organization in both the human and mouse brain. The number of modes required to account for 90% of the variance suggests rich structure in each of these datasets, and is similar across the six human donors. Even when sample and voxel expression profiles were projected onto a three-dimensional subspace (a reduction in dimensionality by at least two orders of magnitude in each data set), however, groups emerged corresponding to coarse brain regions, suggesting a dominant regional structure present in the expression patterns of sets of genes. These results generally agree with Bohland et al. (2010)'s findings in the AMBA, where 67 and 271 modes were required to explain 80 and 90% of the variability in the data, respectively, and where the cerebral cortex, striatum and cerebellum were particularly distinct in the subspace defined by the first three spatial modes.

Subsequent analyses showed this transcriptomic organization at multiple neuroanatomical scales. The broadest of these scales is apparent in Fig 3.3, where most telencephalic structures (the cerebral cortex, hippocampal formation, amygdala, and to a lesser extent the striatum) show a modest degree of transcriptomic similarity that is not shared with the rest of the brain. This result echoed the clustering of human cortex, hippocampus, and amygdala samples observed by Roth et al. (2006) and of mouse cortex, hippocampus, and striatal voxels observed by Lein et al. (2007). Similarly, Zapala et al. (2005) found that adult mouse brain samples group transcriptomically according to

position along the neural tube during development, with the included telencephalic structures (neocortex, hippocampus and olfactory bulbs) forming one cluster.

The globus pallidus was an exception to the relative similarity of telencephalic structures, sharing this weak similarity only with the striatum and tending to group with brainstem structures and the cerebellar nuclei (Fig 3.3). In early development, neurons of the globus pallidus and striatum originate primarily from the medial and lateral ganglionic eminence (Olsson et al., 1998). These transient, raised areas of the developing telencephalon do generate cells for other telencephalic structures; for example, most cortical interneurons may originate there (Kriegstein and Noctor, 2004). However, the striatum and pallidum derive most of their neurons from these structures, and develop into the two main components of the basal ganglia. Notably, in addition to its projections to the thalamus in the main "loop" of basal ganglia motor control, the mature internal globus pallidus projects to motor control centers in the brainstem (Hikosaka, 2007). It is possible that the development of this circuitry influenced the transcriptomic profile of the globus pallidus such that traces of its earlier origins are obscured.

The substructures of the diencephalon did not, on the other hand, show this broad-level similarity, with the thalamus and hypothalamus having near-zero or slightly negative correlations with each other within each species (Fig 3.3). This result differs from Lein et al.'s study (Lein et al., 2007), where the mouse thalamus and hypothalamus both showed positive correlations with samples from the brainstem. This difference may be explained by the fact that those authors used a list of over 5000 genes selected to contain many genes with expression patterns restricted to certain regions. It is possible

that some of the genes included in their study (but not ours) distinguished the diencephalon and brainstem from the telencephalon and cerebellum. Note also that the status of the hypothalamus as part of the diencephalon, based on diencephalic development, is subject to debate (e.g. Larsen et al., 2001; Puelles and Rubenstein, 2003; see also Lim and Golden, 2007).

The brainstem also showed a slight but consistent tendency towards large-scale similarity between samples / voxels from across the midbrain, pons, and medulla. Finally, in our analyses, the cerebellum (Fig 3.2) and specifically cerebellar cortex (Fig 3.3) did not consistently group with other regions in either the mouse or the human brain. This distinct pattern of gene expression reflects the distinct cellular composition of cerebellar cortex, with the Purkinje and granular layers in particular each dominated by one cell type. These results are consistent with Zapala et al. (2005; in the mouse) and Roth et al. (2006; in the human), though Roth et al.'s (2006) unsupervised clustering of brain samples found that sub-structures of the midbrain, pons and medulla clustered with samples from the thalamus and hypothalamus as well as with each other. Note that (as Zapala et al., 2005 point out) the midbrain, pons and medulla develop primarily from three adjacent vesicles of the neural tube, while the cerebellum originates from the rhombic lip, a distinct, transient structure at the boundary between midbrain and hindbrain (Wingate 2001, Fink et al. 2006). However, cerebellar nuclei (which do show elevated correlations with the brainstem; Fig 3.3) as well as some brainstem nuclei also originate from the rhombic lip. The correspondence between mature transcriptomic relationships and position along the neural tube may still hold, but if so must be traced to

more specific locations of origin, and timing of neurogenesis must also be considered (see e.g. Ray and Dymecki, 2009; Landsberg et al., 2005; Wang et al., 2005).

At a finer scale, most neuroanatomical regions of greater specificity showed somewhat stronger within-region similarity than their parent regions (Fig 3.4). The most striking exception, the human cerebellar nuclei, reinforces that a key takeaway of this analysis is the importance of neuroanatomical resolution. There are too few samples to examine individual cerebellar nuclei in the human dataset, but the increase in mean within-region correlation for individual cerebellar nuclei in the mouse data suggests that these nuclei have distinct patterns of gene expression (Fig 3.4C). Combining samples from different substructures may help to increase the signal-to-noise ratio of a characteristic profile of gene expression; however, in cases like the cerebellar nuclei, it also means combining substructures that may have quite heterogeneous cellular composition, and thus distinct molecular profiles. Hence, our analysis faces a trade-off between signal-to-noise ratio (which increases with additional samples from a brain region) and specificity of expression profiles (which requires fine resolution of samples taken from consistent cellular environments). The different trade-off made by the AHBA and the AMBA is also likely to be responsible for the lower correlations between mouse voxels than human samples (Fig 3.3), the lack of genome-scale molecular distinctions between sub-structures of the human globus pallidus (Fig 3.3B), and the tendency toward within-region correlations between human samples to have a narrower range and more negative skew than mouse voxels (Fig 3.4). Future studies might help to address the optimal partitioning of the dataset to balance these two considerations. Additionally, it

would be informative to make cross-dataset comparisons with some form of normalization for tissue sample size.

The AHBA makes it possible to examine the human brain at different scales, while a second human dataset including only four coarsely defined regions offers greater confidence in representing the population due to the relatively large number of donors (the "Gibbs dataset"; Gibbs et al., 2010). Genes that were differentially expressed across those four regions (the frontal lobe, temporal lobe, pons, and cerebellum) in the second dataset overlapped significantly with genes differentially expressed across the same regions in the AHBA (Fig 3.5). The small number of differentially expressed genes across frontal and temporal lobe in both datasets is consistent with previous studies, showing relatively (though not entirely) uniform gene expression across cortical areas (e.g. Bohland et al., 2010; Hawrylycz et al., 2012; Mahfouz et al., 2015). Genome-scale expression profiles of those four regions within the AHBA dataset showed correspondence to those within the Gibbs dataset, consistently across the six AHBA donors (Fig 3.6). This correspondence also showed some specificity at a finer level. Gibbs frontal lobe samples were taken from BA9/46, which cover portions of the middle and superior frontal gyri; these are among the AHBA regions that are most correlated with the Gibbs frontal cortex profile (along with the adjacent precentral gyrus). Additionally, the Gibbs cerebellum profile is strongly correlated with the AHBA cerebellar cortex, but negatively correlated with cerebellar nuclei profiles. This suggests that Gibbs cerebellar samples were selectively taken from the cerebellar cortex. Gene expression profiles sampled from the cerebellum exhibit broad negative correlations with

most other brain regions, confirming the very distinct mode of cerebellar gene expression observed in Figs 3.2 and 3.3.

It is important to note that with the exception of this comparison to the Gibbs dataset, all human analyses presented here and throughout this dissertation were performed using the AHBA. Comparison across AHBA donors in assessing sources of variability in the AHBA (Fig 3.1) and tissue correlations between the AHBA and Gibbs datasets (Fig 3.6) provide some degree of validation. However, confidence in these results will require eventual replication in other datasets. Similarly, not only the analyses presented but most of the works cited regarding the mouse brain use the AMBA. The Allen Brain atlases offer an extremely high level of spatial resolution, and systematic comparisons to other datasets will need to account for different sampling properties, as in the comparison to the Gibbs data discussed above.

One difference that this work does not directly address, particularly in comparisons between the AHBA and the AMBA, is in the size of tissue samples (or voxels). Lower correlations between mouse voxels, as well as greater transcriptomic distinctions between fine regions in the mouse, probably result from the higher spatial resolution of the AMBA. Normalization for tissue sample size might be difficult, given that in a dataset such as the AHBA this can vary by brain structure. Nevertheless, given the importance of cross-dataset comparisons, it would be well worth an attempt in future work.

CHAPTER 4: A COMPARATIVE STUDY OF MOLECULAR ORGANIZATION IN THE MOUSE AND HUMAN BRAIN

4.1 Introduction

Anatomical homologies between the human and mouse brain are relatively well-understood. Despite substantial divergence--most notably in the folds and elaborate areal patterning of the human cerebral cortex as opposed to that of the mouse--neuroanatomists are able to identify homologs of broad human brain structures in the mouse brain, and of many finer regions as well. Cytoarchitecture, myeloarchitecture, and inter-areal neuronal connectivity, observable by such long-established techniques as light microscopy and histochemical assays, form the conventional basis for these definitions of brain regions. Such conventional markers are, however, dependent upon local molecular phenomena operating at a much smaller scale, either directly resulting in the signal of interest or giving rise to the observable structure through a developmental program. With the development in recent years of technology to measure gene expression with high throughput, brain regions can now be characterized and delineated from the perspective of gene products and their interactions. The molecular mechanisms effected by these interactions form the cellular environment underlying conventional markers of region identity and support the functions associated with the region. In contrast to homologies defined at the macro/structural level, regional correspondences between molecular environments of the human and mouse brain remain largely unexplored. The elucidation of molecular-level homologies has substantial implications for the use of mouse models of human neuropathologies, which often rely on the implicit assumption of conserved

molecular mechanisms that underlie the homologous neuroanatomical features. This chapter describes the development and application of tools for the evaluation of molecular correspondences across species using gene expression profiles, and for the identification of groups of genes which may preferentially drive those molecular correspondences for specific regions. These analyses use the Allen Human Brain Atlas (AHBA; Hawrylycz et al., 2012) and the Allen Mouse Brain Atlas (AMBA; Lein et al., 2007), which are described in greater detail in Chapter 2.

4.1.1 Surveying regional gene expression

Comparative studies confirm that the relationships between conventional and molecular neuroanatomy extend across species, with structural homologies carrying molecular-level correspondences. Such correspondences between the human and chimpanzee (one of the closest living relatives to humans) have been identified by cross-species comparisons of (i) differential expression across regions, (ii) similarity relationships between gene expression profiles from different brain regions, and (iii) networks based on co-expression relationships between genes (Khaitovich, 2004; Oldham et al., 2006). These analyses revealed strong conservation of gene networks across species, but also specializations thought to be introduced by evolution, particularly impacting the cerebral cortex. In a study spanning phylogenetic classes, Pfenning et al. (2014) found molecular specialization specific to regions responsible for vocal motor function, which appeared in the human and zebra finch brain but not in other species that do not engage in vocal learning. Strand et al. (2007) made several comparisons between the human and C57BL/6 mouse brain. Using samples from the caudate, cerebellum, and

motor cortex, they found that genes which show preferential expression in a human region tend to show the same preference in the homologous mouse region. Additionally, unsupervised clustering of samples from both species (based on genes with high variance of expression across samples) yielded clusters corresponding perfectly to region labelings.

This evidence for a strong cross-species molecular relationship between these three brain regions points to the possibility for a direct, high-resolution comparative study of gene expression in regions throughout the brain, which the AHBA and AMBA together make possible. One such comparison was recently performed by Hawrylycz et al. (2015). In this study, a brain-wide co-expression consensus network (Langfelder et al., 2011) was defined based on the AHBA, and 32 "core transcriptional modules" were identified using weighted gene co-expression network analysis (WGCNA; Zhang and Horvath, 2005). Several of these modules were well preserved in the AMBA (based on an aggregate preservation score across all genes in the module; Langfelder et al., 2011), including modules with strong expression in the striatum, thalamus, and cerebellum. Even in well-preserved modules, however, a small proportion of individual genes in the mouse were poorly correlated with the human module eigengene and, in some cases, genes showed strong associations with a different module. Furthermore, a number of modules, particularly those not enriched for neuronal expression, were poorly preserved across species. These results suggest the conservation of co-expression relationships that are relevant to specific brain regions, but also indicate points of divergence within the larger co-expression networks.

4.1.2 Challenges of molecular neuroanatomy and mouse models

The neocortex is of special interest in comparative studies of humans and other mammals, as the seat of profound cognitive differences as well as a key structure in many of the human neuropathologies modelled in other mammals. However, identifying molecular-level similarities and differences that are relevant to cortical function is made exceptionally difficult by the fact that both human and mouse neocortex show relatively uniform gene expression across cortical areas (see e.g., Hawrylycz et al., 2012; Bohland et al., 2010; Mahfouz et al., 2015). Variation of gene expression across the cortex is present, and shows some consistency between the human and the mouse, with genes that group together in the one tending to group together in the other (Oldham et al., 2008; Miller et al., 2010). These two studies, though, defined gene clusters based on co-expression across cortical samples regardless of the area of origin. In the rhesus macaque, Bernard et al. (2012) did take area of origin into account and identified groups of genes with cortical area preferences, with striking differences between V1 and the rest of the neocortex appearing largely consistent with the human brain but less so with the mouse. Many expression preferences were heavily influenced by proximity between cortical areas, reflecting the rostrocaudal gradients of gene expression that are a part of brain development (see Sansom and Livesey, 2009 for a review of gradients in human cortical development). Genes may also be expressed in different types of patterns (e.g., laminar, widespread, or sparse) in the cortex. Such patterns show high conservation across human cortical areas--more so than across species when compared with mouse cortex (Zeng et al., 2012). Laminar variation in cell type densities results in neocortical

layers showing different gene expression patterns (Ng et al., 2009; Belgard et al., 2011; Bernard et al., 2012); however, the AHBA lacks laminar specificity because cortical samples were taken across layers rather than tangentially through the cortex (Hawrylycz et al., 2012). In addition to the relatively homogenous molecular nature of the neocortex, many neocortical areas in the human brain lack clear neuroanatomical homologs in the mouse, making molecular-level homologies doubly challenging to identify.

Mouse models used to investigate learning and memory, and sometimes Alzheimer's Disease specifically, make the hippocampal formation another structure of particular interest. The broad structure of the hippocampal formation is consistent between the human and mouse, each including the dentate gyrus, Ammon's horn, and subicular complex (see Ding, 2013 for a detailed comparison of the subicular complex in human, monkey, and rodent). Additionally, the role of the striatum in Parkinson's Disease (PD) has made it the focus of a large number of studies in mice (see Le et al., 2014 for a review of mouse models of PD). The most substantial divergence of the human and mouse striatum lies in the separation of the human caudate nucleus from the putamen by the internal capsule, where the mouse caudoputamen is a single structure. Unlike the neocortex, the hippocampal formation and striatum are composed of substructures with identifiable homologs in the mouse and human, and which are more molecularly distinct from each other than are neocortical areas (Bohland et al., 2010; Hawrylycz et al., 2012).

Nevertheless, developing effective and reproducible mouse models of AD and PD, in which these brain structures play central roles, has proved difficult (Duff, 2004;

Hardy, 2006; Le et al., 2014). A probable source of some of this difficulty has been identified by Burns et al. (2015) through analysis of the expression profiles of disease-implicated genes. This study found striking differences between the up- and down-regulation of genes of interest in several human diseases (as well as aging in the human brain) and the up- and down-regulation of these genes in the various mouse models they examined. Burns et al.'s findings suggest that mouse models of human disease may benefit from better knowledge of similarities and differences in the transcriptomic environments present in relevant brain structures and systems. This reinforces an argument made by Strand et al. (2007), suggesting that understanding the initial conditions (i.e., healthy state) of the transcriptome in the human and mouse brain is highly relevant to comparisons of disease conditions in the two species.

4.1.3 Overview of the current approach

Rather than focusing on specific genes of interest, the current study deals primarily with groups of genes and their co-expression relationships, which underlie the molecular environment that influences any given gene. We evaluated the overall similarity of gene expression in regions throughout the human and mouse brain by correlating expression profiles, both within- and across-species. We also sought to identify groups of genes which preferentially "drive" cross-species similarity for different regions. To do this, we developed a quantitative measure to assess the extent to which a given set of genes provides a region-specific molecular signature that is consistent across species. Because it is not computationally feasible to assess all possible subsets of genes, we defined candidate gene sets in two ways. The first was data-driven, where genes were divided

into sets based on their co-expression relationships. Co-expression is an indicator of functional relationships between genes (Eisen et al., 1998; Lee, 2004; Wei et al., 2006); therefore, this approach is influenced by relationships – both known and unknown – between genes. In the second approach, gene sets were defined based on common annotations; for example, a gene set might include all genes known to distinguish a certain cell type or to be involved in a certain metabolic pathway. In addition to these region-specific analyses, gene-gene similarity was evaluated across species by comparing each gene's brain-wide expression pattern in the human with its expression pattern in the mouse, and examined genes that showed very high or very low cross-species similarity. Brain-wide co-expression relationships between cell-type marker genes (for neurons, oligodendrocytes, and astrocytes) were also assessed, both within- and between-species.

The overall aim of this chapter is to better quantify molecular-level correspondences between the human and mouse brain at a large scale, and to provide inroads for further study of neural homologies. The efficacy of mouse models depends upon homology at all scales, including preservation of the molecular mechanisms underlying the function of individual brain regions, making this area of study important in biomedical research and in the development of drugs or other neurotherapeutic interventions. While homologies can be considered in terms of individual genes, the expression of one gene always occurs under the influence of others. This raises the question of which aspects of the molecular environment as a whole are conserved across species. Like conventional neuroanatomical markers, this environment varies across the brain, reflecting the localization of brain function and dysfunction to specific structures

and circuits. To our knowledge, this study is the first to comprehensively quantify the similarity between local transcriptomic environments in the human brain and their equivalents in one of the most common model organisms.

4.2 Methods

See Chapter 2 for formal descriptions of the Allen Human Brain Atlas (AHBA) and the Allen Mouse Brain Atlas (AMBA).

4.2.1 Gene selection

Human orthologs were identified for 3,792 of the genes available in the mouse dataset using NCBI HomoloGene (NCBI Resource Coordinators, 2015; www.ncbi.nlm.nih.gov/homologene). Expression data for these genes constituted a common dataset, which was used in all comparisons of the AMBA and AHBA reported here.

4.2.2 Regions of the brain

In the AHBA and AMBA, human samples and mouse voxels are assigned neuroanatomical labels at multiple levels of granularity based on detailed reference atlases. For effective interspecies comparison, the subsequent analyses treat eleven coarsely defined brain structures (Table 4.1). Only samples and voxels belonging to one of these eleven broad brain regions were used in any part of these analyses. Analyses which examine finer subdivisions of these regions focus primarily on those regions in the second column of Table 4.1. In the AHBA ontology, the "cerebral cortex" includes the

hippocampal formation. Here, the hippocampal formation was excluded from the definition of cerebral cortex in order to study it as a separate structure. Human cerebral cortex was therefore defined as the neocortex (the frontal, temporal, parietal, and occipital lobes, and the insula), cingulate cortex, and parahippocampal gyrus. In the AMBA ontology, "cerebral cortex" also includes the hippocampal formation as well as amygdalar nuclei; to study these structures separately, mouse cerebral cortex was defined here using the AMBA term "isocortex."

Broad regions (11)	**Fine regions (9)**
Cerebral cortex (4008 [excluding 4249] / 315)	
Hippocampal formation (4249 / 1089)	Dentate gyrus (12891 / 726)
	Ammon's horn (12892-12895 / 375)
	Subiculum (12896 / 502)
Amygdala (4327 / 278, 131, 295, 319, 780)	
Striatum (4277 / 672, 56)	Caudoputamen (4278, 4287 / 672)
	Nucleus accumbens (4290 / 56)
Globus pallidus / Pallidum or dorsal pallidum[2] (4293 / 803 or 818)	External segment of globus pallidus (12897 / 1022)
	Internal segment of globus pallidus (12898 / 1031)
Thalamus (4392 / 549)	
Hypothalamus (4540 / 1097)	
Midbrain (9001 / 313)	
Pons (9131 / 771)	
Medulla (9512 / 354)	
Cerebellum (4696 / 512)	Cerebellar cortex (4697 / 528)
	Cerebellar nuclei (4780 / 519)

Table 4.1. List of brain regions. Regions were defined using the neuroanatomical labels associated with the given numeric identifiers (human / mouse) in the AHBA and AMBA ontologies. [1]

1. The relevant reference atlases can be viewed at http://atlas.brain-map.org/atlas?atlas=265297125 (human) and http://atlas.brain-map.org/atlas?atlas=1 (mouse), or downloaded directly (with associated IDs and other metadata) from http://api.brain-map.org/api/v2/structure_graph_download/10.xml (human) or http://api.brain-map.org/api/v2/structure_graph_download/1.xml (mouse).

2. Neuroanatomical labels for the ventral, caudal, and medial as well as the dorsal pallidum (globus pallidus) are included in the AMBA, but not the AHBA. In comparisons of homologous brain structures, the human globus pallidus was compared only to the mouse dorsal pallidum.

4.2.3 Probe standardization

Probes in both the AHBA and AMBA were standardized using the weighted z-scoring procedure described in Chapter 2, with the broad regions given in Table 4.1 given equal weight. Weighted z-scoring in the AHBA was performed within-donor.

4.2.4 Correlations between individual samples / voxels

To assess the regional organization of transcriptomic relationships, tissues throughout the brain were compared across species using Pearson's Product-Moment Correlation Coefficient (PCC) calculated between pairs of profiles. These coefficients were calculated between each human sample and each mouse voxel. The cross-species sample/voxel PCCs were pooled into distributions corresponding to the 121 pairings of broad regions in the human and mouse brain (Table 4.1). The same procedures were followed using the 9 fine brain regions listed in Table 4.1. Cross-species PCC distributions for some structure pairs were compared using two-sample Kolmogorov-Smirnov tests, with p-values corrected for multiple comparisons using the Bonferroni method.

For each putatively homologous cross-species region pair, the quartiles of all sample-voxel PCCs was calculated to assess the strength of cross-species similarity for each region at a global scale. Additionally, sample-voxel PCCs for homologous regions were compared to a distribution of 1000 PCCs calculated between sample/voxel pairs for which either the sample or the voxel originated within the region, and the other member of the pair did not. Means of the resulting percentile ranks were calculated to assess specificity of regional correlations.

4.2.5 Region-specific homology score

A "homology score" was defined to quantify the extent to which a given gene set provides a molecular signature that is specific to a given brain region, and which is consistent across species (see below for selection of gene sets). Homology scores were calculated separately for each human donor brain. A gene set's homology score for a given brain region was based on correlations between its human expression profile for that region (averaged across all samples available for a given donor brain) and the orthologous expression profile at each voxel in the mouse data. This yields a map of correlations across the entire mouse brain, as schematized in Fig 4.1. Before using specific subsets of genes, correlation maps and homology scores were calculated for each of the 11 broad seed regions using the full list of 3,792 orthologous genes.

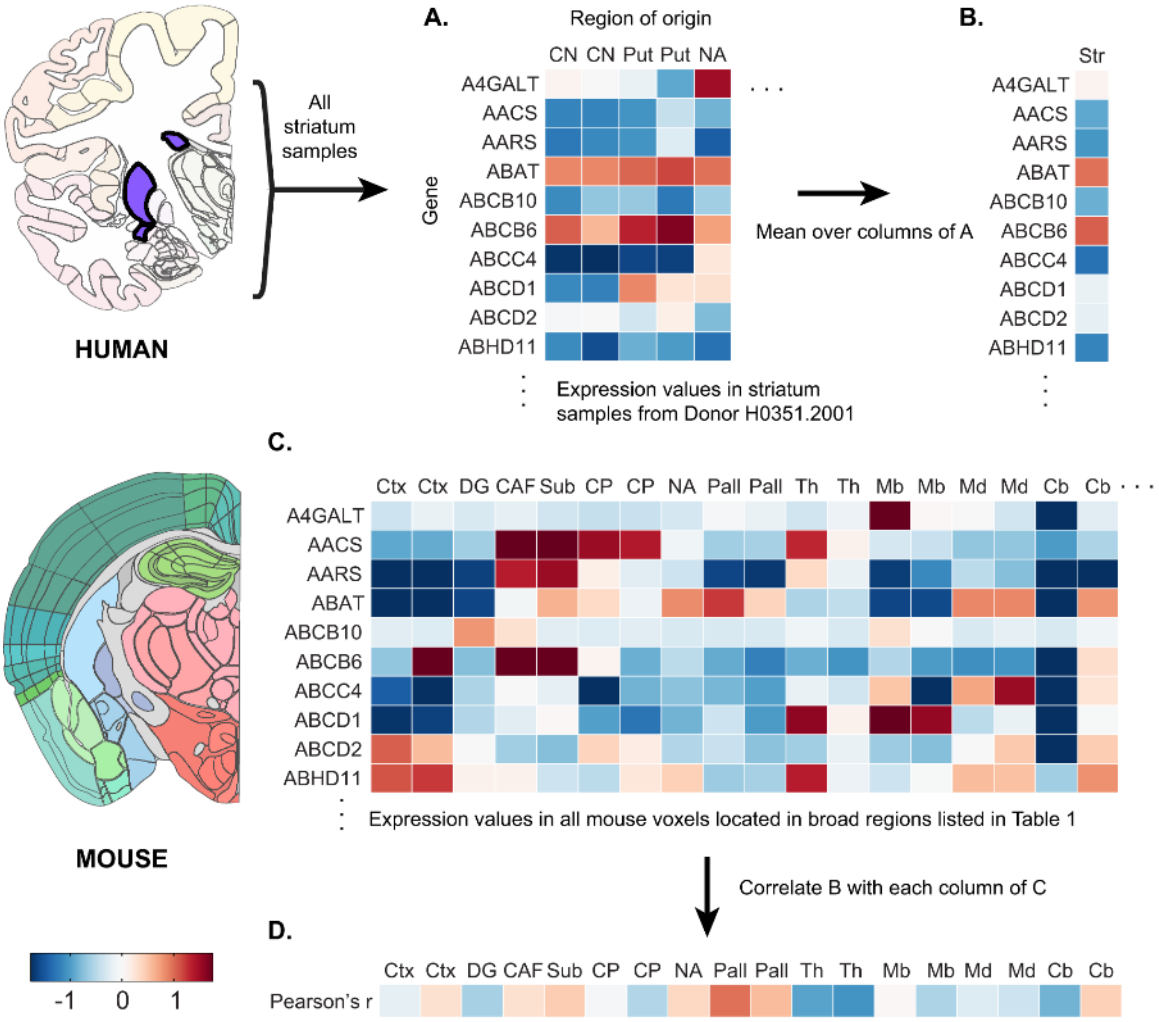


Fig 4.1. Schematic of method for calculating correlation maps. Color bar at bottom left applies to all panels in this figure. A. All left-hemisphere samples from a "seed" brain region are selected from one human donor. B. The standardized profiles of those samples are averaged into a mean profile for the seed region. C. Standardized expression profiles for all mouse voxels located within any of the broad brain regions listed in Table 1 were used for comparison. D. The correlations between the expression profile of the human seed region and each mouse voxel are computed to generate a map of correlations across the mouse brain. Reference atlas images adapted from files downloaded from the Allen Institute for Brain Science. Image credit: Allen Institute.

The procedure for calculating a homology score based on a correlation map is formalized below. We compare the mean correlation across voxels falling within a mouse target region (usually the putative homolog to the human seed region) to the mean correlation across voxels falling elsewhere in the mouse brain. If the former value is

larger than the latter, then the expression of the gene set carries preferential molecular similarity between the human seed region and the mouse target region. Because this score is intended to measure region-specific similarity, a penalty for non-specificity is then applied to ensure that a gene set does not receive a high homology score if its expression profile for the human seed region is highly similar to a non-target mouse region in addition to the target mouse region.

4.2.6 Definition (Correlation map)

We define $E^G(i)$ as a vector representing the standardized expression levels for human sample *i* across gene set *G*. E_R^G then represents the average expression profile (across *N* available samples) for a given human seed region *R*:

$$E_R^G = \frac{1}{N}\sum_{i=1}^{N} E^G(i) \qquad (1)$$

For each mouse voxel, we then calculate the PCC between the seed region profile E_R^G and the mouse expression vector defined across orthologs of gene set *G* at each voxel *v*, denoted as $PCC_R^G(v)$. If any gene had missing data for certain voxels, that gene was excluded from the calculation of $PCC_R^G(v)$ for those voxels. This procedure yields a map of PCCs across the mouse brain.

4.2.7 Definition (Homology score)

Given the map of voxelwise PCCs with E_R^G, we calculate the average of the PCC values within the mouse target region R':

$$IN_{R'}^{G} = \frac{1}{V_{R'}} \sum_{v \in R'} PCC_{R}^{G}(v) \qquad (2)$$

where $V_{R'}$ is the number of voxels in *R'*. We then calculate the average of the PCC values outside the target region *R'*:

$$OUT_{R'}^{G} = \frac{1}{(V_{\text{total}} - V_{R'})} \sum_{v \notin R'} PCC_{R}^{G}(v) \qquad (3)$$

where V_{total} is the total number of voxels analyzed in the mouse brain. Negative mean PCCs (i.e., patterns of gene expression that are inversely related across species) would affect the homology score without having a clear biological interpretation; therefore, we threshold $IN_{R'}^{G}$ and $OUT_{R'}^{G}$ at zero.

The difference of these quantities, $\left[IN_{R'}^{G}\right]^{+} - \left[OUT_{R'}^{G}\right]^{+}$, provides information about cross-species homology. However, while this difference is sensitive to a concentration of high PCCs in the target region, it is not necessarily specific to such a concentration. It may yield a high value even if some non-target region also has a concentration of high PCCs, provided that the overall average ($OUT_{R'}^{G}$) remains low. Therefore, we introduced a penalty for non-specificity. This penalty is based on the highest mean PCC with the seed region for any single non-target mouse region:

$$r'_{max} = \underset{\mathrm{r} \neq \mathrm{R}'}{\operatorname{argmax}}(IN_{r}^{G}) \qquad (4)$$

The penalty term λ is defined as the mean of the differences between the mean PCC for r'_{max} and each mean PCC yielded by another non-target region:

$$\lambda_{R'}^{G} = \frac{1}{M-2} \sum_{r \neq R', r \neq r'_{max}} \left([IN_{r'_{max}}^{G}]^{+} - [IN_{r}^{G}]^{+} \right) \qquad (5)$$

where *M* is the total number of mouse regions analyzed (i.e., those which are homologs to the eleven broad seed regions; see "Human seed regions and mouse target regions," below). This penalty will have a large value if one mouse region outside the target has a much larger mean PCC than the others, showing that gene set *G* drives cross-species similarity between the human seed region and a mouse region that is not the target. It will have no impact (value of zero) if the PCCs are uniform outside the target region.

The homology score H_R^G of gene set *G* for human seed region *R* is then defined as:

$$H_R^G = \left(\left[IN_{R'}^G \right]^+ - \left[OUT_{R'}^G \right]^+ \right) - \lambda_{R'}^G \qquad (6)$$

Thus, the homology score quantifies the extent to which higher PCCs are concentrated within the mouse homolog to the human seed region (relative to the rest of the mouse brain), and are not specifically concentrated in any other mouse region.

The difference of thresholded mean PCCs falls in the interval $[-1,1]$, where a value of one would indicate a PCC of 1 with each voxel in the mouse target region and a maximum PCC of zero with any non-target voxel. Conversely, a value of -1 one would indicate a PCC of 1 with each non-target voxel, and a maximum average PCC of zero inside the target region. The value of H_R^G, however, may fall below -1 due to the penalty term $\lambda_{R'}^G$. This asymmetrical range reflects the fact that a seed region may be correlated only with the target region (H_R^G = 1), only with voxels outside the target region (H_R^G = -1), or only with a specific non-target region (H_R^G < -1). In practice, however, values below -1 did not occur.

Each homology score was converted to a percentile rank in an empirical chance distribution. For each candidate gene set, 1,000 sets of randomly selected genes were

generated, each of the same size as the original gene set. Homology scores were computed for these random sets, and the resulting distribution was used to calculate the percentile rank of H_R^G for the original gene set.

4.2.8 Human seed regions and mouse target regions

Candidate gene sets (see below) were first scored using each of the 11 broad human brain regions (Table 4.1) as seeds. Target regions in the mouse (i.e., homologous regions) were determined based on common nomenclature and through surveys of the anatomical literature. Gene sets were then scored using the fine human regions of Table 4.1 as seeds, which also have relatively well-established homologs in the mouse. In the case of the caudate nucleus and putamen, which are distinct structures in the human but not the mouse, the mouse caudoputamen was used as the target region for each.

Our purpose in using the additional set of fine seed regions was to determine whether the specificity of a gene set's cross-species correspondence relative to the rest of the brain was affected by increasing the specificity of the neuroanatomical region of interest. Therefore, computation of homology scores for a given fine seed region excluded the part of the PCC map falling within the parent structure but outside the target region. For example, a gene set's homology score for the dentate gyrus was not affected by PCCs falling within Ammon's horn, the subiculum, or any of the mouse retrohippocampal regions.

4.2.9 Candidate gene set identification: Data-driven gene sets

Initial candidate gene sets were identified based on co-expression relationships by applying weighted gene co-expression network analysis (WGCNA; Zhang and Horvath, 2005) to the mouse dataset using a publicly available R package (Langfelder and Horvath, 2008). A network of genes was defined in which edge weights encoded the absolute value of the correlation between pairs of genes' spatial expression patterns. Co-expression similarity was measured by topological overlap (TO; Zhang and Horvath, 2005)). Average linkage hierarchical clustering was then applied, and the resulting dendrogram was cut using a dynamic tree-cutting algorithm (Langfelder and Horvath, 2008).

WGCNA was applied to the full-brain, unweighted, standardized mouse data. In order to emphasize relationships across the brain structures that would be used as broad seed regions when calculating homology scores, the mouse expression dataset was averaged (across voxels) into the homologs of those regions, with the exception that the striatum and pallidum were treated as a single structure (in later analyses they would be treated separately, as in Table 4.1). WGCNA was performed on the resulting 10 x 3,792 matrix with default parameters, which sets the minimum gene set size to 20. Each of the resulting gene modules was used as a candidate gene set and assessed for region-specific molecular similarity across species.

4.2.10 Candidate gene set identification: Annotation-based gene sets

A second group of candidate gene sets was defined, consisting of all genes associated with a series of annotations representing some function, cellular mechanism,

or other association. Three of these sets were composed of genes which show cell-type specific expression for neurons, astrocytes, and oligodendrocytes in the postnatal mouse brain (Cahoy et al., 2008); these genes (those intersecting our common gene set) had significant differential expression in each of these three cell types compared to the others and a fold change of at least 20.

Additional gene sets were defined using annotations which were statistically over-represented in the data-driven gene sets. We examined over-represented annotations for data-driven gene sets whose homology scores ranked at the 80th percentile or higher for at least one broad seed region. Such annotations were expected to more specifically correspond to groups of genes which had common, conserved roles across the two species.

The Molecular Signatures Database (MSigDB; Subramanian et al., 2005) includes over 10,000 gene annotations from many sources, including publications and online neuroinformatics resources such as the Gene Ontology (Ashburner et al., 2000) and the KEGG pathways database (Kanehisa and Goto, 2000; Kanehisa et al., 2014). All available annotation terms and their associated gene lists were downloaded from MSigDB. The genes associated with each annotation were limited to those appearing in our common set of 3,792 genes. Over-representation in the selected data-driven sets was assessed using the hypergeometric test, with p-values corrected for multiple comparisons using the linear step-up false-discovery rate (FDR; Benjamini and Hochberg, 1995). Additionally, annotations had to be represented in a gene set by at least 3 genes to be considered over-represented in that set.

Annotations that were over-represented with FDR-corrected p-value less than 0.01 in any of the selected data-driven sets were then manually curated to select only annotations that could be specifically related to brain structure or function. Annotations were selected from this list to define gene sets, each of which was composed of all genes in our common gene set that were associated with the annotation. Homology scores were then calculated and converted to percentile ranks for both the cell-type marker gene sets and these additional annotation-based gene sets, using both the broad and fine seed regions.

4.2.11 Brain-wide similarity of orthologous gene expression profiles across species

We measured the similarity between each gene's pattern of expression across the human brain and the pattern of expression of its ortholog across the mouse brain. The expression matrix for each human donor was averaged into 16 brain regions by starting with the 11 broad regions listed in Table 4.1 and replacing the hippocampal formation, globus pallidus, striatum, and cerebellum with their corresponding fine regions. To allow a one-to-one correspondence with brain regions in the mouse, the human caudate nucleus and putamen were treated as a single structure for this analysis. The region-averaged AHBA data was then averaged across donors, yielding a single 3,792 x 16 expression matrix. The AMBA dataset was averaged into the same 16 structures, and PCCs between expression vectors for orthologous genes in mouse and human were calculated.

Genes with PCCs in the top 5% (190 genes) were clustered based on their expression patterns across the 16 human brain regions using average linkage hierarchical clustering. Leaf order of the resulting dendrogram was used to order genes in heatmap

representations of their expression patterns across the 16 regions of interest in both the human and mouse brain. Genes with PCCs in both the top and bottom 5% were examined for over-represented annotations, using the MSigDB as described above (see "Annotation-based gene sets").

4.2.12 Effects of penalty score and regional expression

The effect of the penalty term λ on the homology scores was evaluated by converting un-penalized scores to percentile ranks in a distribution of un-penalized scores from randomly selected gene sets. These ranks were compared to those obtained using the penalty term.

To assess the relationship between a gene set's homology score for a brain region and its expression in that region, percentile ranks of H_R^G were compared to percentile ranks of mean expression values in each broad region, for all gene sets. The same randomly selected gene sets used to calculate percentile ranks for H_R^G were used to calculate percentile ranks for mean expression values in a region. In separate analyses, the mean regional expression was calculated in mouse and human brain by averaging entries in the expression sub-matrix that indexes all genes in the set and all samples or voxels in the region. For human data, this mean value and its percentile rank were calculated for each donor individually before being averaged across donors.

4.2.13 Cell-type markers

An additional analysis was performed on the three cell-type marker gene sets (Cahoy et al., 2008; see “Annotation-based sets”, above). To assess the similarity of expression patterns related to cell types across the human and mouse brain, cross-species

PCCs between gene expression patterns (calculated as above) were averaged within each cell-type marker gene set. This average PCC was compared to a distribution of average PCCs calculated from 10,000 randomly selected gene sets of the same size as the original, and its percentile rank in this empirical chance distribution was calculated.

4.3 Results

4.3.1 Correlation heatmaps

Heatmaps of cross-species PCCs between expression profiles for individual AHBA samples and AMBA voxels show global organization based on brain region (Fig 4.2). Blocks of high values (in comparison to the rest of the matrix) correspond to brain regions including the cerebral cortex, thalamus, hypothalamus, amygdala, and cerebellum, with striatal expression profiles showing especially high PCCs with each other. In some cases, expression profiles between non-homologous regions show somewhat elevated PCCs. The midbrain, pons, and medulla are represented by a single block of positive-valued PCCs, in addition to showing elevated PCCs with the cerebellar nuclei. To a lesser extent, this is also true of the cerebral cortex, hippocampal formation, amygdala, and striatum. These features also appear in the within-species correlation heatmaps (see Chapter 3, Fig 3.3), which show higher within-region PCCs overall than those in Fig 4.2. Fig 4.2 also shows this organization at a finer level for the pallidum and cerebellum, where the dorsal pallidum, cerebellar cortex, and cerebellar nuclei show higher PCCs with their homologs than with non-homologous sub-structures.

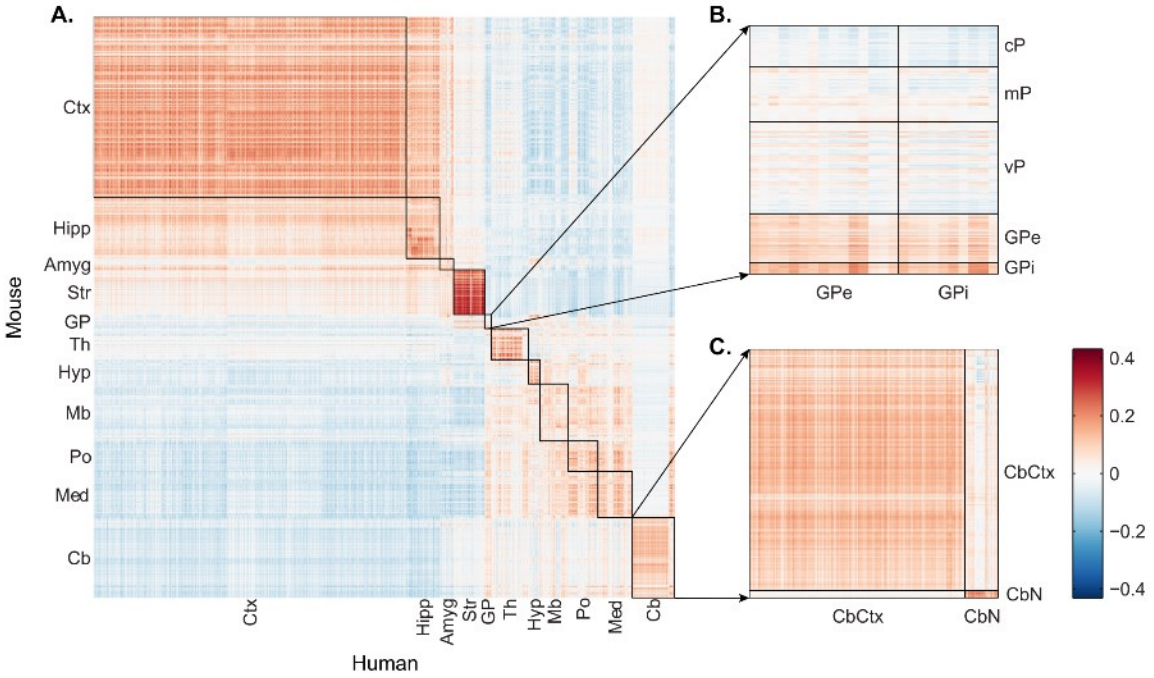


Fig 4.2. Correlations between gene expression profiles of human samples and mouse voxels. A. Correlations between each human sample (horizontal axis) and each mouse voxel (vertical axis). B. Correlations between the human and mouse pallidum. C. Correlations between the human and mouse cerebellum. Of the voxels from the mouse cerebellum included in this figure, 473 were labelled only with the term "Cerebellum," with no finer structure specified. Correlations between these voxels and human samples are along the bottom of A, and are not included in C.

4.3.2 Cross-species correlation distributions

Each distribution in Fig 4.3 represents a histogram of all values within a submatrix of the correlation matrix shown in Fig 4.2. For the broad regions (Fig 4.3A), these distributions highlight patterns described above in the correlation matrix. Distributions of PCCs between homologous regions (along the main diagonal) generally showed a rightward shift relative to those between most non-homologous regions. Slightly right-shifted distributions also appeared between the groups of regions noted above (the cerebral cortex, hippocampal formation, amygdala, and striatum as opposed to the midbrain, pons, and medulla). The thalamus had a bimodal distribution, as did

several non-homologous region pairs. Closer examination revealed that the human ventral and mouse dorsal thalamus have a median correlation of only 0.002, while the medians for the human dorsal / mouse dorsal, human ventral / mouse ventral, and human dorsal / mouse ventral thalamus distributions fall between 0.068 and 0.081. The shapes of the distributions between the human cerebellum and all mouse homologous regions includes a high peak near 0, which reflect human cerebellar cortex (Figs 4.2 and 4.3B). In most cases, human cerebellum also showed a smaller number of more positive PCCs (with mouse brainstem structures and pallidum), or negative PCCs (with other mouse telencephalic structures); these are due to human cerebellar nuclei (Figs 4.2 and 4.3B). Mouse cerebellar correlations with non-homologous human brain structures do not show this shape. In the AMBA, layers of cerebellar cortex are distinct; in the AHBA, cerebellar cortical samples may include more than one layer. This combination of cell type distributions may obscure consistent similarities and differences between human cerebellar cortical profiles and most of the mouse brain, although it does not appear to obscure similarity with the mouse cerebellar cortex.

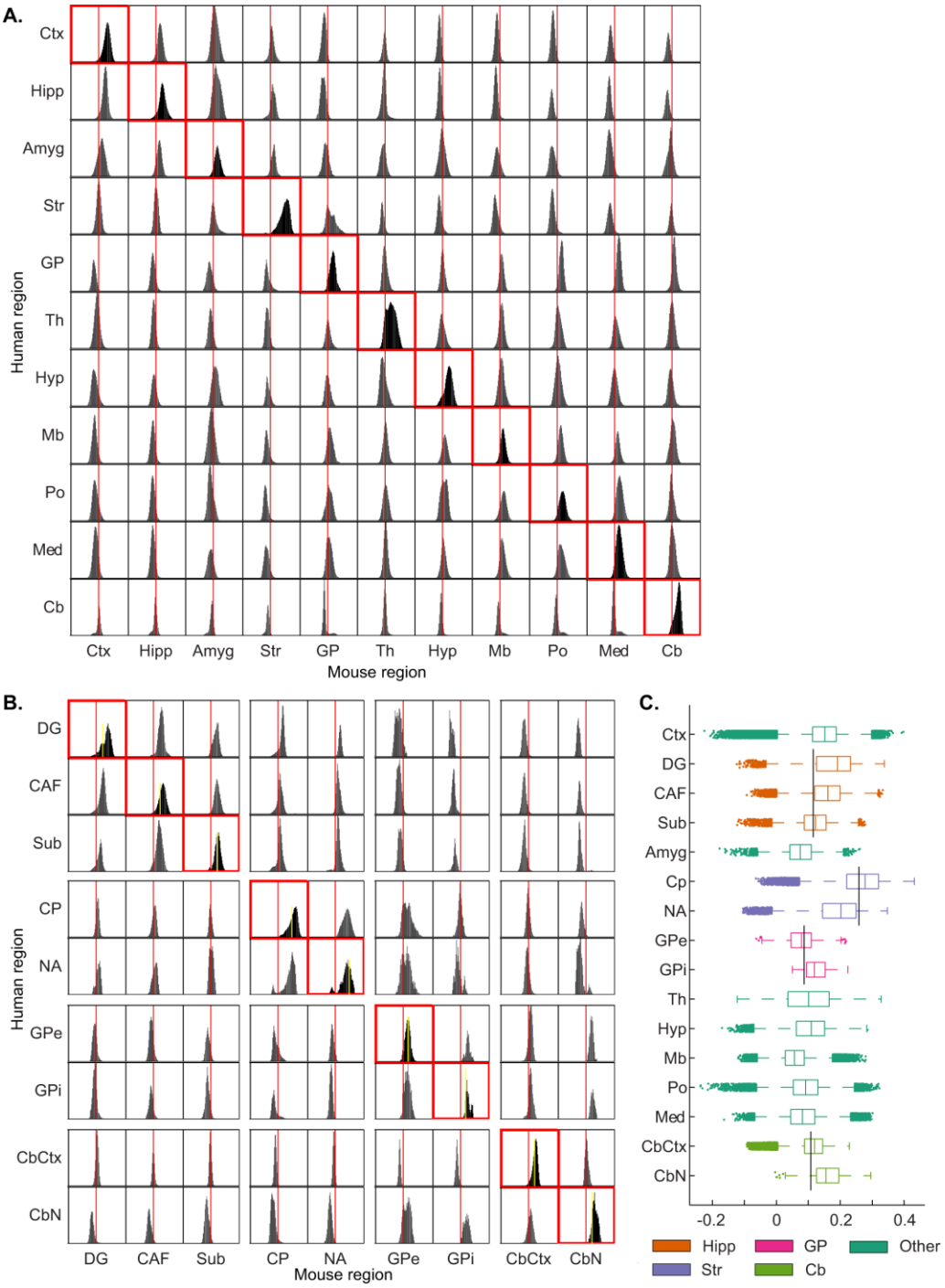


Fig 4.3. Distributions of correlations between expression profiles from a human region and a mouse region. Homologous region pairs are outlined in red. Horizontal axis limits are -0.5 and 0.5; red vertical lines indicate zero. A. Broad brain regions. B. Fine brain regions. Dashed yellow vertical line indicates median correlation for parent brain region. C. Median correlation between all sample-voxel pairs where sample and voxel are from homologous regions. Box edges represent the 25th and 75th percentile values; whiskers extend to any data points within 1.5 times the interquartile range from the rendered boxes. Box colors correspond to parent brain regions. Black vertical lines show median cross-species correlations of parent structures.

For the most part, the PCC distributions for fine regions shown in Fig 4.3B echoed the organization shown for broad regions, with a tendency towards higher PCCs between profiles of homologous regions than between profiles from regions thought to be anatomically and functionally distinct. This was largely true even for sub-structures of the same broad region. The difference between the distribution of correlations for any homologous fine region pair and the pooled distributions of correlations for non-homologous pairs within the same broad region (those not outlined in red) was significant for all sub-structures after Bonferroni correction (one-tailed two sample Kolmogorov-Smirnov test, $p < 0.01$, $N = 9$ homologous pairs) except for the nucleus accumbens ($p > 0.6$) and GPe ($p > 0.3$). Surprisingly, the cerebellar nuclei showed elevated PCCs with the GPe and GPi, and several bimodal distributions were observed for fine structure pairings, suggesting further anatomical specificity of gene expression relationships that go beyond the labeling available.

Fig 4.3C summarizes the above data, showing medians of the PCC distributions for homologous structures (outlined in red in Fig 4.3A and 4.3B). As Fig 4.2 indicated, the striatum showed the highest median cross-species PCC; this primarily reflects transcriptomic similarity between the human and mouse caudoputamen. These values are consistently lower than within-species PCCs between expression profiles of a given region (compare Fig 4.3C to Fig 3.2). However, they did tend to be higher than cross-species PCCs where the sample and the voxel were from non-homologous regions (Fig 4.4A). At a finer scale, this tendency held only within the cerebellum; expression

profiles within the other broad regions were not more similar when originating from the same sub-region than when originating from different sub-region (Fig 4.4B).

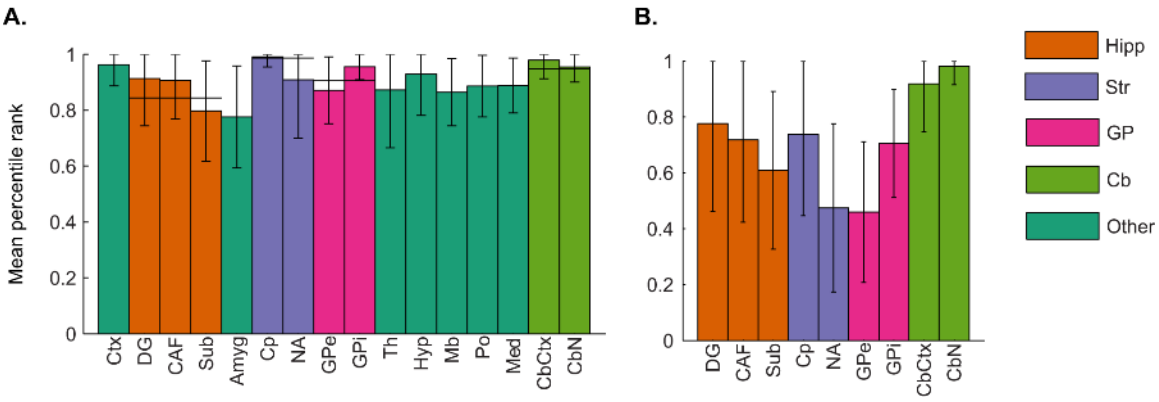


Fig 4.4. Mean percentile rank of within-region correlations in empirical null distributions. Bars correspond to brain regions and bar length to mean percentile rank of correlations within the region, based on the appropriate simulated distribution. Error bars show standard deviation. Bar colors correspond to parent brain regions. A. Correlations in null distribution are between expression profiles within the region and expression profiles from elsewhere in the brain. Solid horizontal lines show the mean percentile rank within the parent region. B. Correlations in null distribution are between expression profiles within the region and expression profiles from elsewhere within the parent region.

4.3.3 Genome-scale correlation maps

The initial analysis of region-specific expression used the full list of 3,792 orthologous genes to generate correlation maps (by the procedure schematized in Fig 4.1). Maximum intensity projections (Fig 4.5) show the similarity between an average expression profile from a human brain "seed" region (from one donor brain) and all voxels in the mouse brain. These correlation maps show a varying degree of anatomical specificity, as anticipated from sample-based correlation results above (e.g., the human striatum profile has particularly high correlations with mouse striatal voxels).

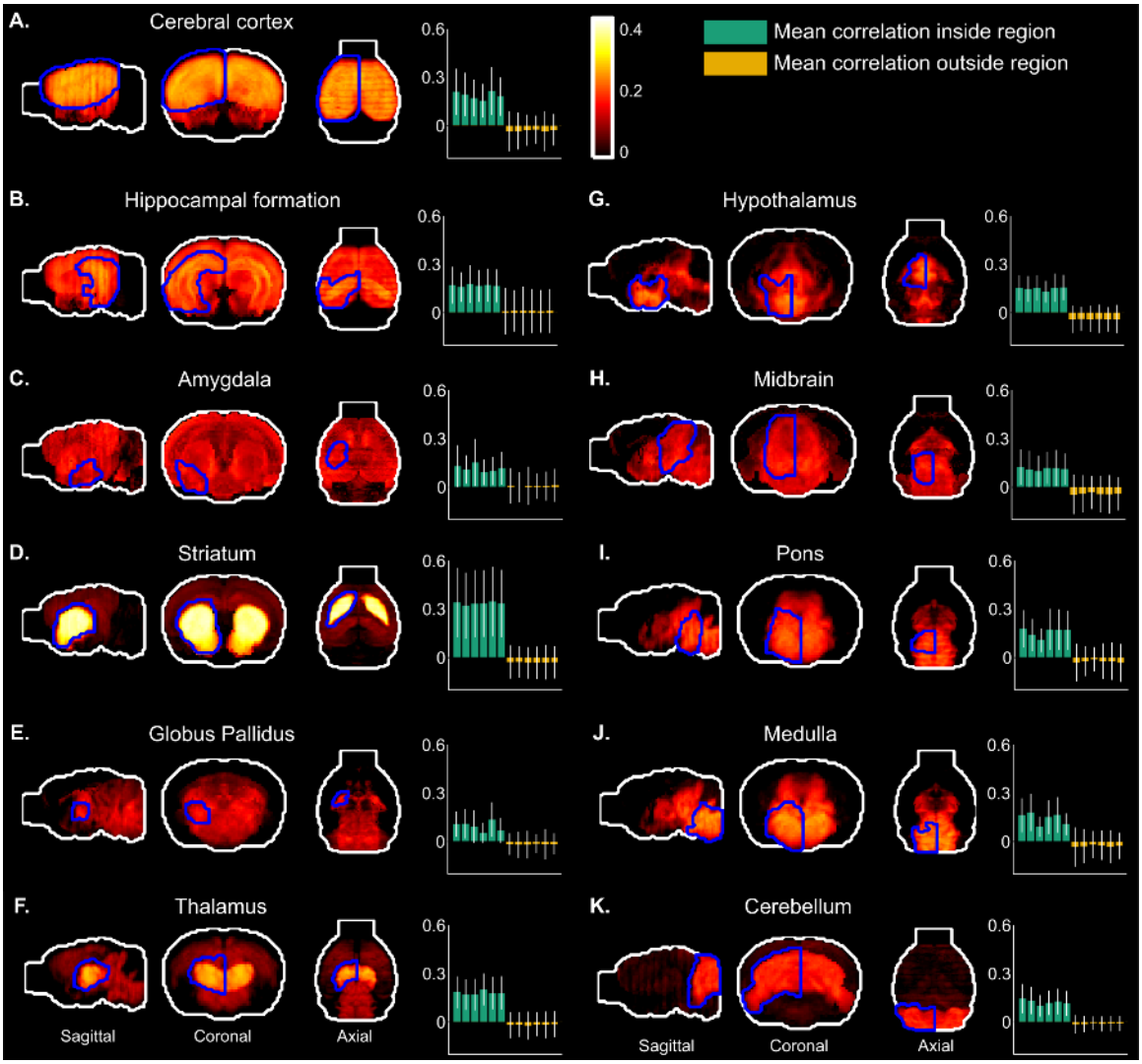


Fig 4.5. Maps of mouse voxel correlations with human seed regions. Correlation maps for 11 broad brain regions, generated using left-hemisphere samples from a single donor (H0351.2002). Correlation maps are shown as maximum intensity projections in each of the cardinal planes, in which the value at a given location in the two-dimensional plane is the maximum value found along the perpendicular axis at that location. Mouse homologs to human seed regions are outlined in blue in the left hemisphere. Bar plots show mean correlations across voxels inside and outside the mouse homolog to the seed region (green and yellow bars, respectively), for all donors. Individual bars correspond to human donors. Error bars show standard deviation across voxels.

The PCCs inside the homologous mouse brain region were then compared to those outside the homolog in each human donor, yielding extremely consistent results (Fig 4.5, bar plots). All broad seed regions showed modest but positive mean PCCs inside their homologs (for each donor brain), with the striatum yielding the highest mean

across donors ($r = 0.34$). Most showed near-zero (and usually slightly negative) mean PCC values for voxels outside their homologs. The human hippocampal formation profile yielded small positive values outside its homolog due to a tendency towards positive PCCs with cortical, amygdalar, and striatal voxels (mean PCCs of 0.16, 0.10, and 0.07, respectively). Similarly, the human amygdala showed a small positive value due to positive mean PCCs for cortical, hippocampal, and striatal voxels (mean PCCs of 0.08, 0.09, and 0.09).

4.3.4 Homology scores of candidate gene sets

Next we sought to quantify the degree to which specific sets of genes provided region-specific, cross-species homologies, based on the anatomical specificity of the calculated correlation maps. Sets of genes were determined in either a data-driven manner using WGCNA or based on curated annotations.

4.3.5 Data-driven sets

WGCNA applied to the region-averaged mouse expression data resulted in 31 gene sets; see Fig 4.6 for dendrogram. Un-penalized homology scores and penalties (see Equations 5 and 6) for the full list of 3,792 genes and for each of the 31 data-driven gene sets are shown in Fig 4.7 for each of the 11 broad seed regions. For all seed regions, there were one or more gene sets that provide a higher un-penalized homology score than the full gene set. Penalties were then subtracted from the un-penalized scores for each gene set. After enforcing the penalty, at least one gene set per seed region increased the homology score relative to the full gene set, with the exception of the cerebral cortex.

Homology scores were then converted into percentile ranks, with average ranks across donors (in comparison to an empirical null distribution) shown in Fig 4.8A.

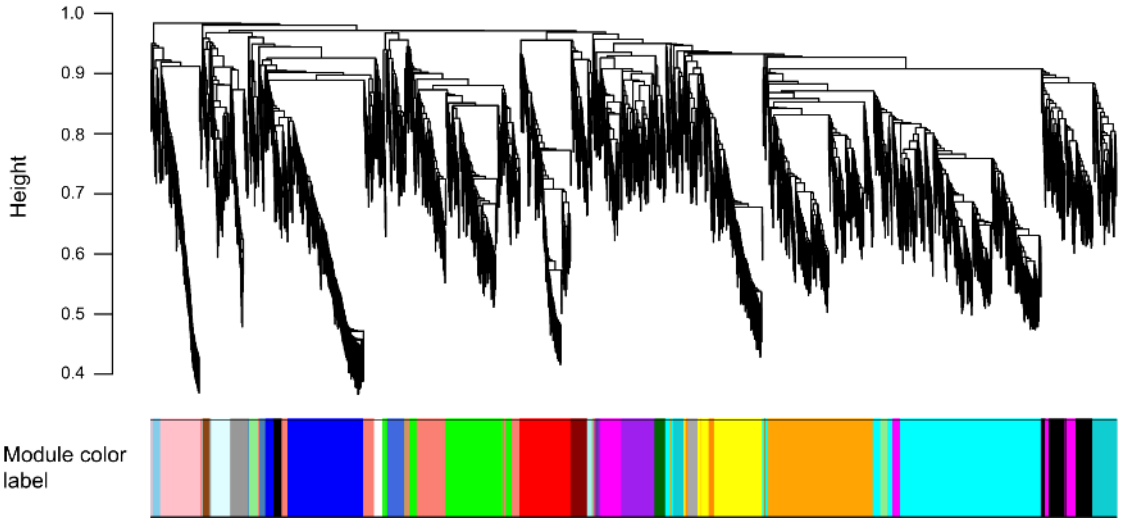


Fig 4.6. Dendrogram resulting from hierarchical clustering of genes using WGCNA on mouse data after averaging into the 10 broad regions shown in Table 1, with striatum and pallidum treated as a single structure. Dissimilarity is based on topological overlap. Colors of horizontal band correspond to gene sets.

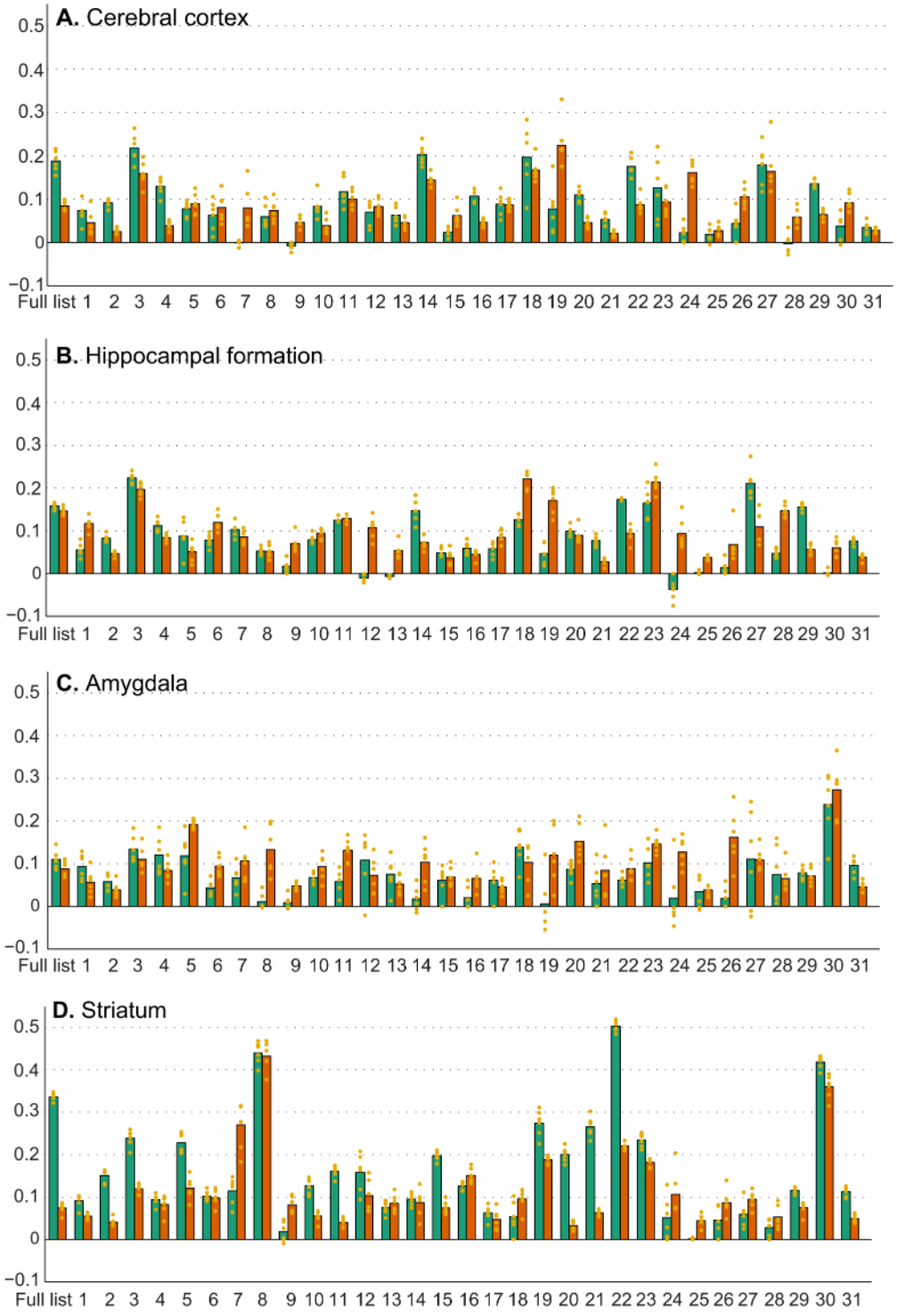


Fig 4.7 (cont. on next page). Un-penalized homology scores and penalties of full gene list and data-driven subsets. Green bar height indicates average un-penalized score across the six human donors. Orange bar height indicates average penalty. Yellow dots represent individual donors.

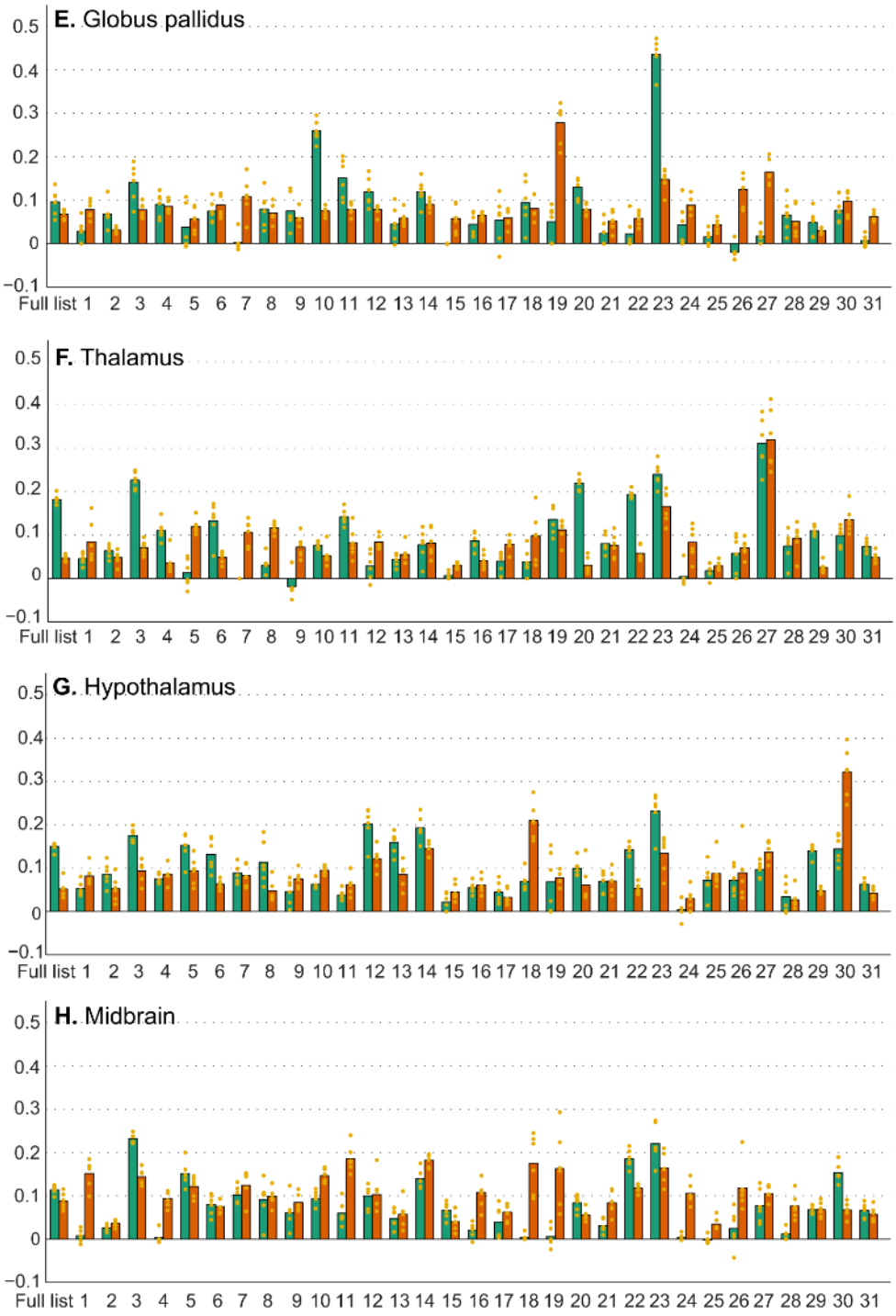


Fig 4.7 (cont. on next page). Un-penalized homology scores and penalties of full gene list and data-driven subsets. Green bar height indicates average un-penalized score across the six human donors. Orange bar height indicates average penalty. Yellow dots represent individual donors.

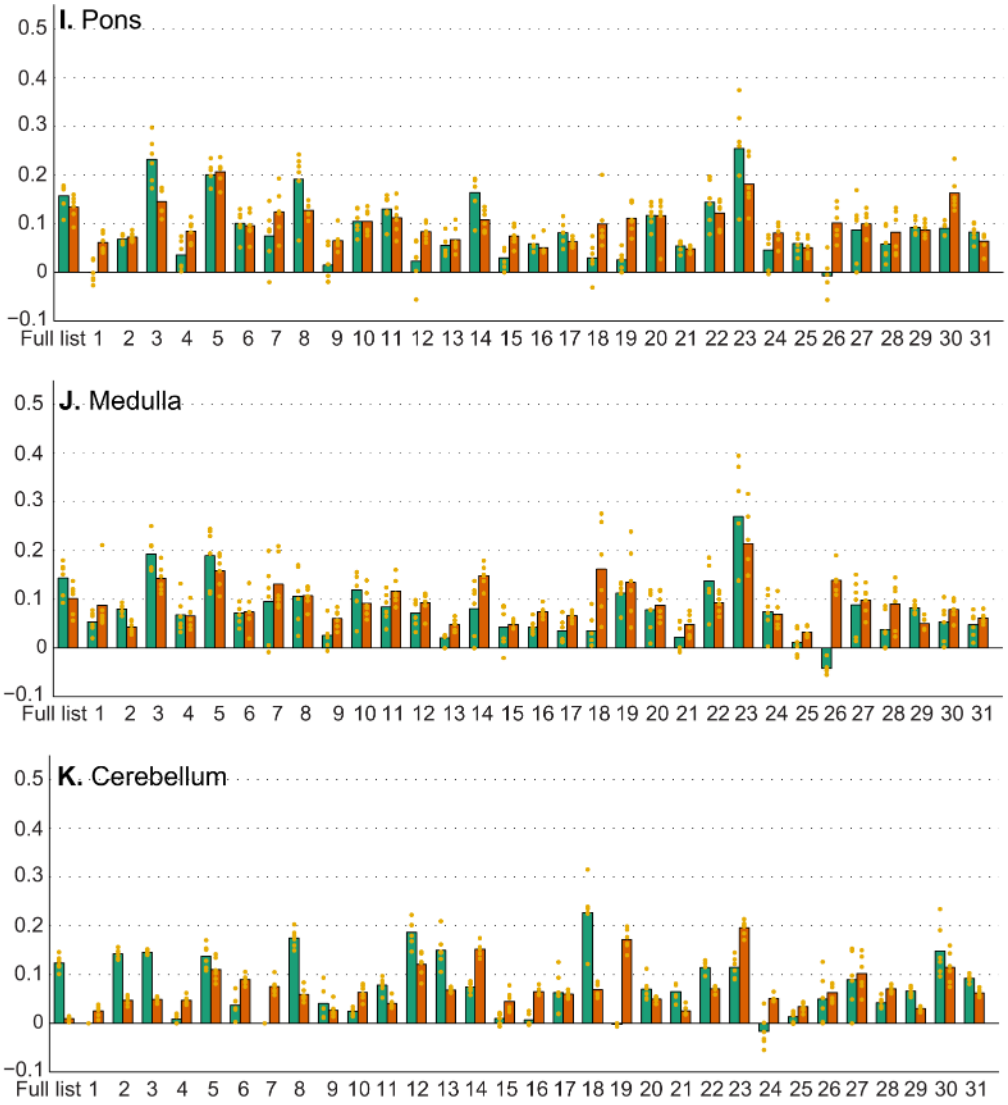


Fig 4.7. Un-penalized homology scores and penalties of full gene list and data-driven subsets. Green bar height indicates average un-penalized score across the six human donors. Orange bar height indicates average penalty. Yellow dots represent individual donors.

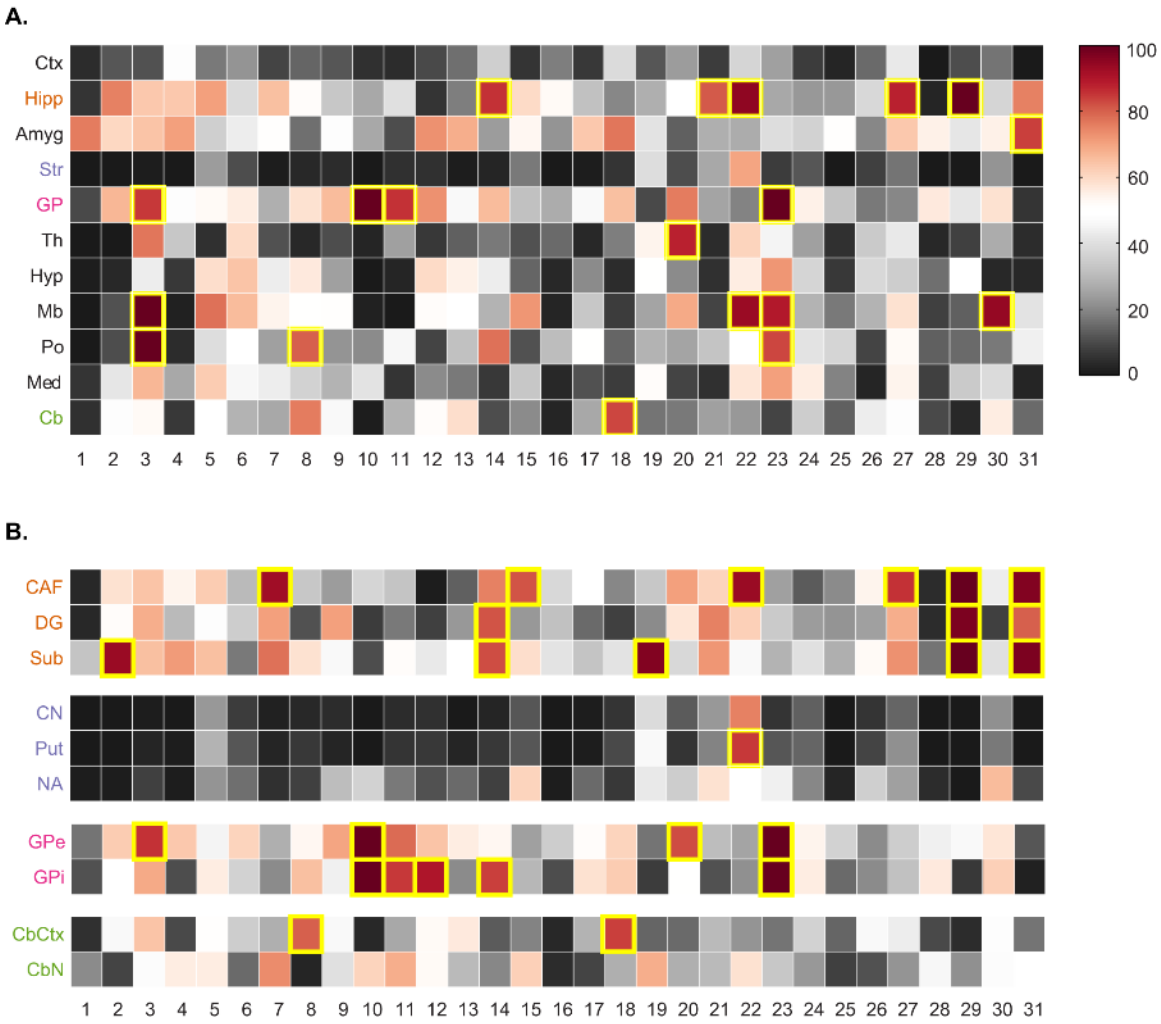


Fig 4.8. Homology score percentile ranks (in empirical null distributions) of data-driven gene sets. Color bar applies to both panels. White indicates chance (50th percentile) performance. Yellow outlines indicate values of 80th percentile or above. A. Broad regions. B. Sub-structures of the hippocampal formation, striatum, globus pallidus, and cerebellum.

In general, there was a sparse pattern of data-driven gene sets that provided strong homology scores for individual brain regions when compared to random gene sets. We established a threshold for identifying possible sets of interest for further investigation at the 80th percentile, with sets surpassing this threshold for any brain region highlighted in yellow in Fig 4.8. Seven seed regions showed between one and five gene sets which

scored at or above this threshold. For the most part, different gene sets yielded high percentile ranks for different brain regions (though there were exceptions; e.g., Sets 3 and 23 in the globus pallidus, midbrain, and pons). No data-driven gene sets scored at or above the 80th percentile for the cerebral cortex, hypothalamus, striatum, or medulla. This was due in part to the penalty term λ (Equation 5), which decreased the score when a relatively high concentration of correlations occurred in a specific non-homolog region of the mouse brain. When un-penalized homology scores were compared to distributions of un-penalized scores from the randomly selected gene sets, at least one data-driven gene set met the cutoff for each of these seed regions except the cerebral cortex, for which the highest-ranking gene set was at the 77th percentile. In one extreme example, Set 22's percentile rank for the striatum increased from 70 to 99.88 when it was no longer penalized for similarity to the pattern of expression in the globus pallidus.

For a subset of brain regions, homology scores were additionally computed for each of their sub-structures according to the Allen Reference Atlas hierarchy (Fig 4.8B). In some cases, a gene set showed a high rank for only some portion of the larger structure (e.g., Set 2 for the subiculum only); in others, a gene set maintained high ranks throughout the larger structure (e.g., Set 29 for the CA fields, dentate gyrus, and subiculum). Fig 4.8B also shows that results for the striatum and cerebellum as a whole were largely determined by the caudoputamen and cerebellar cortex, respectively.

4.3.6 Annotation-based sets

Three candidate gene sets were defined using neuron, oligodendrocyte, and astrocyte markers of the postnatal mouse brain (Cahoy et al., 2008). Eleven additional

gene sets were defined by identifying over-represented annotations in the 14 data-driven gene sets which had a mean percentile rank of at least 80 for at least one broad seed region (Fig 4.8A). Of the 255 annotations that were over-represented in these sets (FDR-adjusted p-value < 0.01), 11 were selected to form candidate gene sets (Table 4.2).

Name in MSigDB (Abbreviation)	Description	Enriched data-driven set
REACTOME_OPIOID_SIGNALLING (OP)[1, 2]	48 genes	Set 23 (5 occurrences)
KEGG_LONG_TERM_POTENTIATION (LTP)[3, 4]	32 genes	Set 3 (9 occurrences)
CIRCADIAN_RHYTHM (Circ)[5]	8 genes	Set 3 (5 occurrences)
BLALOCK_ALZHEIMERS_DISEASE_DN (Alzdn)[6]	461 genes down-regulated in human hippocampus with Alzheimer's Disease	Set 3 (50 occurrences); Set 18 (14 occurrences)
MCCLUNG_COCAINE_REWARD_5D (Coc)[7]	41 genes up-regulated in mouse nucleus accumbens after 5 days cocaine treatment	Set 22 (11 occurrences)
LU_AGING_BRAIN_DN (FLdn)[8]	77 genes down-regulated in human frontal lobe with age	Set 18 (4 occurrences)
LEE_AGING_CEREBELLUM_DN (CBdn)[9]	27 genes down-regulated in mouse cerebellum with age	Set 18 (3 occurrences)
LEIN_PONS_MARKERS (Po)[10]	63 genes (those appearing in the current data) out of the 100 most specific to mouse pons	Set 10 (31 occurrences)
LEIN_MIDBRAIN_MARKERS (Mb)[10]	57 genes out of the 100 most specific to mouse midbrain	Set 23 (8 occurrences)
LEIN_MEDULLA_MARKERS (Med)[10]	53 genes out of the 100 most specific to mouse medulla	Set 10 (23 occurrences)
MODY_HIPPOCAMPUS_POSTNATAL (HFpost)[11]	31 genes up-regulated in mouse postnatal hippocampus	Set 18 (3 occurrences)

Table 4.2. Annotation terms selected to define new candidate gene sets. Description field includes the number of genes with this annotation found in the list of 3,792 genes used here. All annotations were over-represented with FDR-corrected p < 0.01 in the specified data-driven set (i.e., WGCNA module).

1. Croft et al. (2014). 2. Milacic et al. (2012). 3. Kanehisa and Goto (2000). 4. Kanehisa et al. (2014). 5. Ashburner et al. (2000). 6. Blalock et al. (2004). 7. McClung and Nestler (2003). 8. Lu et al. (2004). 9. Lee et al. (2000). 10. Lein et al. (2007). 11. Mody et al. (2001).

Percentile ranks of homology scores for all annotation-based gene sets are shown in Fig 4.9. Neuron markers met or exceeded the 80th percentile for all regions except the hippocampal formation and the thalamus. Analysis of hippocampal sub-structures shows that the neuron markers in fact yielded high region-specific correspondence for the dentate gyrus (99th percentile), but lower ranks for the CA fields and subiculum (Fig 4.10). This was due to a penalty for similarity to the cortex, causing the mean rank to drop from 91st to 66th for the CA fields, 92nd to 32nd for the subiculum, and 98th to 74th for the hippocampal formation as a whole. In the thalamus, even the un-penalized scores for neuron markers ranked, on average, only at the 42nd percentile.

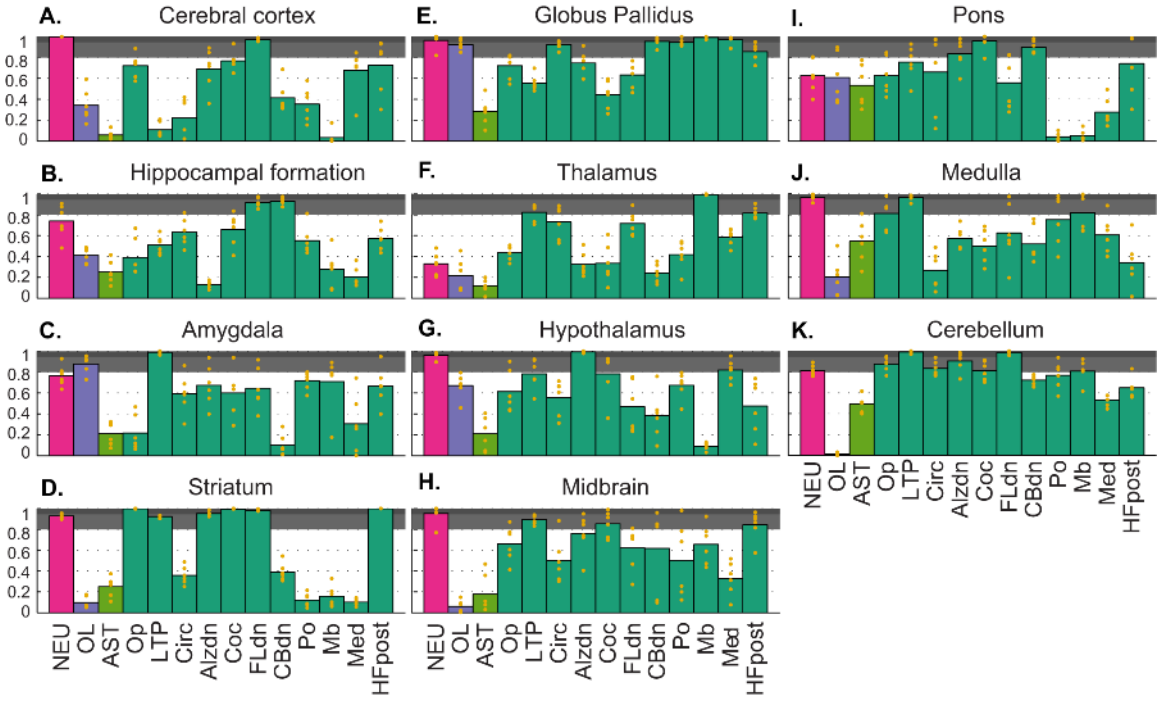


Fig 4.9. Homology score percentile ranks of annotation-based gene sets for broad seed regions. Bar height indicates average percentile rank across the six human donors; yellow dots are values for individual donors. Bar colors for cell-type marker sets correspond to Fig 4.13. Grey shaded area denotes the 80th to the 95th percentile, and darker grey the 95th to 100th percentile values for random gene sets of the same size. See Table 4.2 for full names and descriptions of gene sets.

Fig 4.10. Homology score percentile ranks of annotation-based gene sets for sub-structures of four broad regions. Bar height indicates average percentile rank across the six human donors; yellow dots are values for individual donors. Grey shaded area denotes the 80th to the 95th percentile, and darker grey the 95th to 100th percentile values for random gene sets of the same size. See Table 4.2 for full names and descriptions of gene sets.

Homology scores for oligodendrocyte and astrocyte markers not only had low percentile ranks for most regions, but were almost always negative before being converted to percentile ranks. Exceptions included oligodendrocyte markers in the amygdala and globus pallidus, the only two regions where either glial gene set showed homology scores with relatively high ranks compared to random gene sets (mean score

and rank 0.09 and 88th percentile for amygdala; 0.16 and 92nd percentile for globus pallidus).

For the most part, the other 11 annotation-based sets showed more variable results across the broad regions. However, a striking number of annotation-based sets showed high region-specific correspondence for the striatum, globus pallidus, and cerebellum. This tendency appeared to be common to the three human striatal sub-structures and to the external and internal globus pallidus, but within the cerebellum it was specific to the cerebellar cortex (Fig 4.10).

For each of the other broad regions, at least one of these 11 annotation-based sets yielded a percentile rank of at least 80 (Fig 4.9). In the cerebral cortex, for example, genes that have been found to be down-regulated in the aging human frontal lobe ("FLdn"; Lu et al., 2004) had a percentile rank of 97, notably higher than any of the data-driven gene sets. For the hippocampal formation, both FLdn and a set of genes down-regulated in aging mouse cerebellum ("CBdn"; Lee et al., 2000) exceeded the 90th percentile. The thalamus showed the strongest region-specific correspondence with mouse midbrain marker genes ("MB"; Lein et al., 2007) and the hypothalamus with genes downregulated in the human hippocampus with Alzheimer's disease ("AlzDn"; Blalock et al., 2004). Both gene sets ranked at the 99th percentile.

Genes involved in the long-term potentiation pathway ("LTP"; Kanehisa and Goto, 2000; Kanehisa et al., 2014) showed high percentile ranks for most regions, with a mean percentile rank of ~75 across the 11 seed regions. The most striking exceptions, however, were the cerebral cortex, hippocampal formation, and globus pallidus (11th,

51st, and 56th percentile). For the cerebral cortex and hippocampal formation, this was due to high penalties for similarity to the striatum (un-penalized scores ranked at the 99th and 86th percentile, respectively).

4.3.7 Effects of penalty term and regional expression

Fig 4.11 shows homology score percentile ranks obtained using the penalty term (Equation 5) against percentile ranks obtained without it. The penalty term sometimes had the effect of increasing percentile rank, and sometimes of decreasing it.

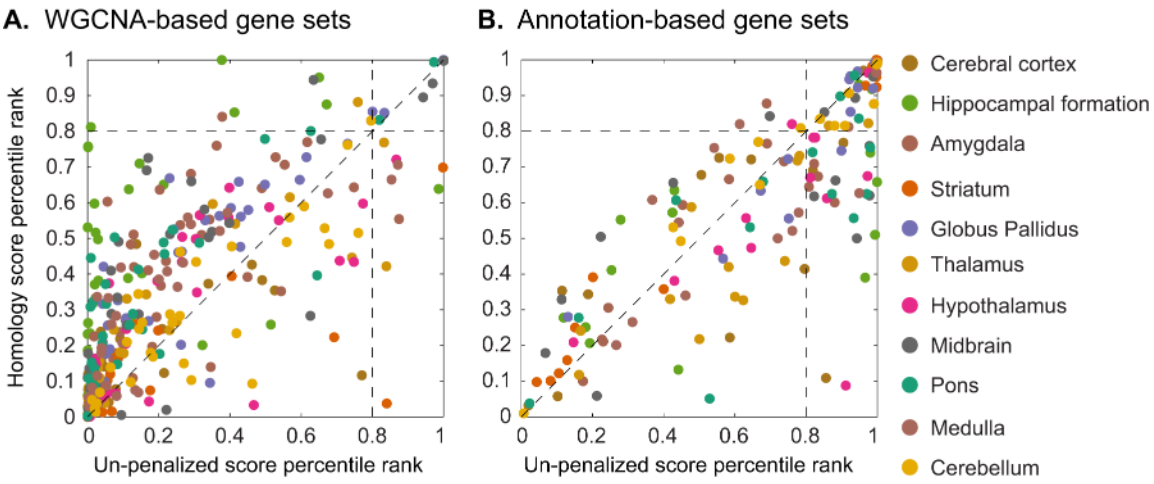


Fig 4.11. Homology score percentile rank against un-penalized score percentile rank. Each datapoint results from one gene set for one brain regions. Dashed horizontal and vertical lines mark the 80th percentile. Dashed diagonal line runs from (0, 0) to (1, 1).

Fig 4.12 compares gene sets' homology scores for each broad region to their mean expression in that region (after converting both values to percentile ranks). Most homology scores that rank over the 80th percentile occur in a brain region where the gene set showed either much lower or much higher expression than in other regions.

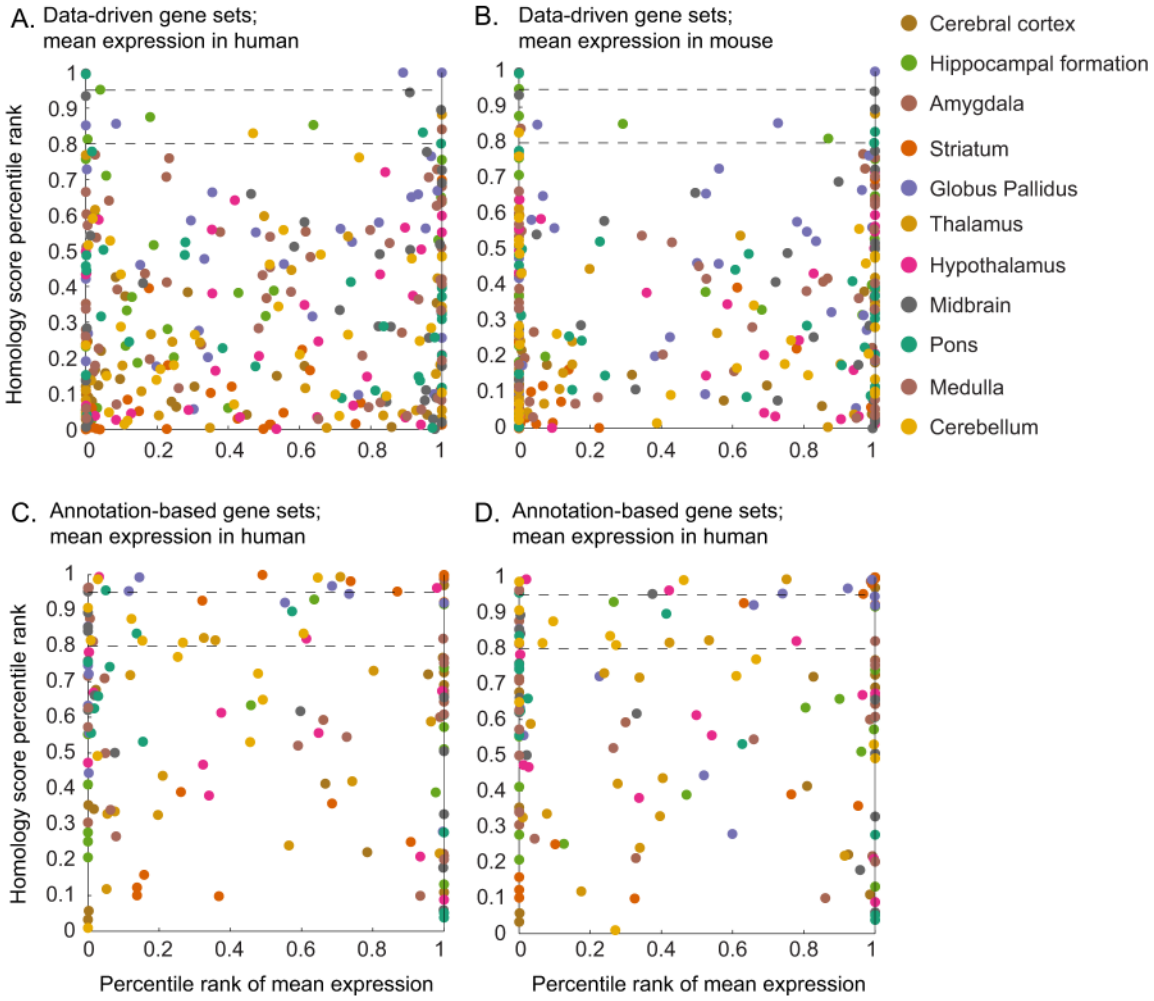


Fig 4.12. Homology score percentile rank against mean expression percentile rank. Each datapoint represents results from one gene set for one brain region. Mean expression percentile ranks that use human data are averaged across human donors (as are homology score percentile ranks). Dashed horizontal lines show the 80th and 95th percentiles for homology scores.

4.3.8 Brain-wide similarity across species

Finally, we analyzed the similarity of cross-species, brain-wide gene expression profiles for individual genes and for sets of cell-type marker genes. Fig 13A shows the distribution of PCCs between each gene's expression pattern (summarized into 16 regions) across the human brain and its corresponding expression pattern across the

mouse brain (quartiles at -0.09, 0.24, 0.56). For cell-type markers, the median PCC between a gene's expression profile across the human brain and its profile across the mouse brain was 0.65 for neuron markers, 0.63 for oligodendrocyte markers, and -0.18 for astrocyte marker genes. When compared to an empirical chance distribution of average PCCs of 10,000 randomly selected gene sets of the same size, the percentile ranks of these values were 100 for neuron markers (higher than all random gene sets), ~99 for oligodendrocyte markers, and 0.0001 for astrocyte markers.

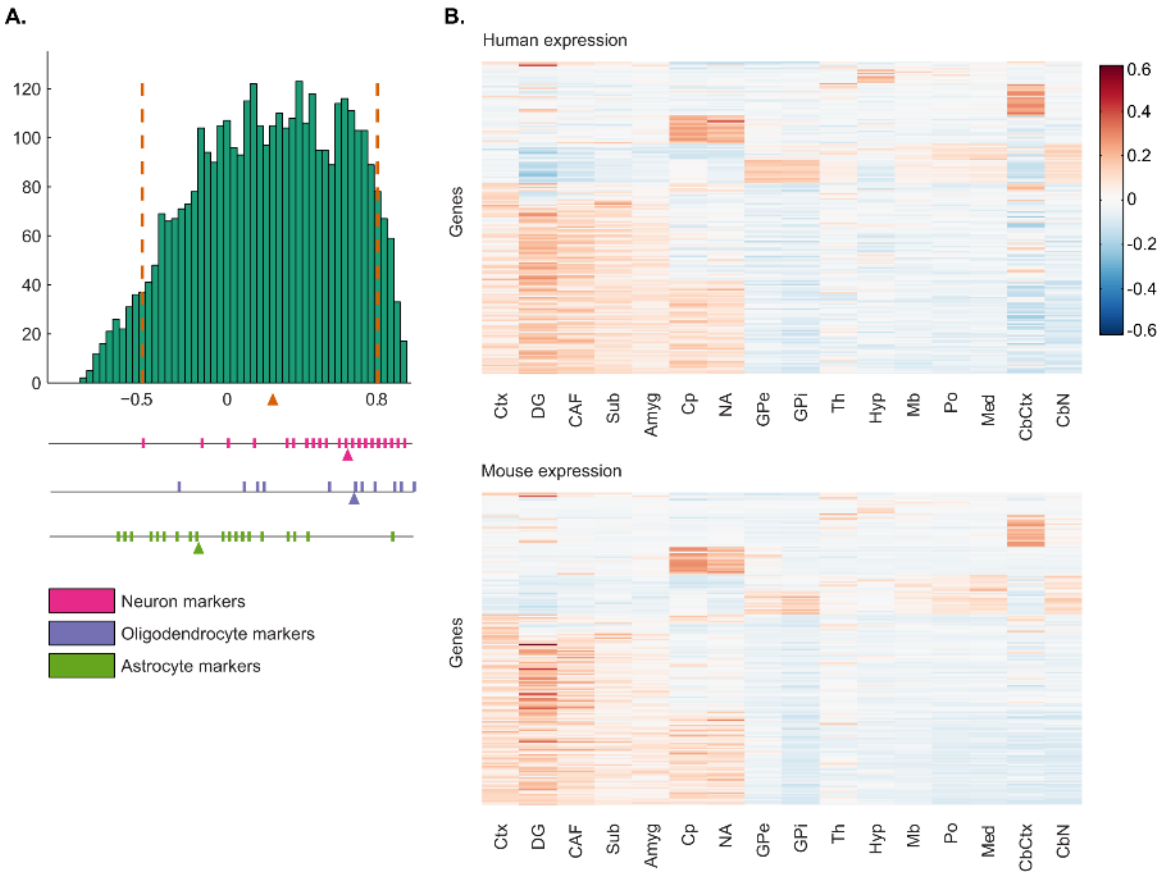


Fig 4.13. Cross-species correlations and expression heatmaps of orthologous genes. A. Distribution of correlations between brain-wide expression profiles for orthologous genes. Dashed orange lines mark the 5th and 95th percentiles. Orange triangle indicates distribution median. Cross-species correlations for cell-type markers are represented below, with median correlation for each group of cell-type markers indicated by a triangle. B. Heatmap of human brain-wide expression patterns of genes with cross-species correlations in the top 5%. Genes are ordered by clusters obtained by hierarchical clustering of their human expression patterns. C. Mouse brain-wide patterns of the same genes shown in B, in the same order.

The top 5% of gene-gene PCCs ranged from 0.83 to 0.98. Heatmaps of these 190 genes' expression patterns across the human and mouse brain show clusters of genes with higher expression values for the striatum, or hypothalamus, or cerebellar cortex than for the other regions (Fig 4.13). Other genes were most strongly expressed in multiple regions, such as the cerebral cortex, hippocampal formation, and to a lesser extent the amygdala, or for these regions together with the striatum. Finally, one cluster of genes showed slightly higher expression in the globus pallidus, thalamus, midbrain, pons, medulla, and cerebellar nuclei than in the rest of the human brain, though this cluster's preferences for those structures were less apparent in the mouse brain.

Table 4.3 shows the median within-species PCC for each group of cell-type marker genes. Although these values were low, they were higher than for randomly selected genes, with the exception of astrocyte markers in the mouse. Brain-wide expression patterns of neuron and oligodendrocyte markers are shown in Fig 4.14. Overall, neuron markers had their strongest expression in the cortex and hippocampal formation, and to a lesser extent the amygdala, of both species. Oligodendrocyte markers showed higher expression in the brainstem and cerebellar nuclei.

Cell type	Human (samples)	Human (region averages)	Mouse (voxels)	Mouse (region averages)
Neuron	0.20 (100)	0.28 (100)	0.19 (100)	0.28 (99.9)
Oligodendrocytes	0.34 (100)	0.46 (100)	0.19 (99.0)	0.32 (96.9)
Astrocytes	0.14 (99.9)	0.26 (99.9)	0.05 (41.2)	0.16 (62.9)

Table 4.3. Within-species correlations of mouse cell-type markers. The median of all correlations between the markers for a cell type is followed in parentheses by the percentile rank of that median correlation in a distribution of 10,000 randomly selected gene sets.

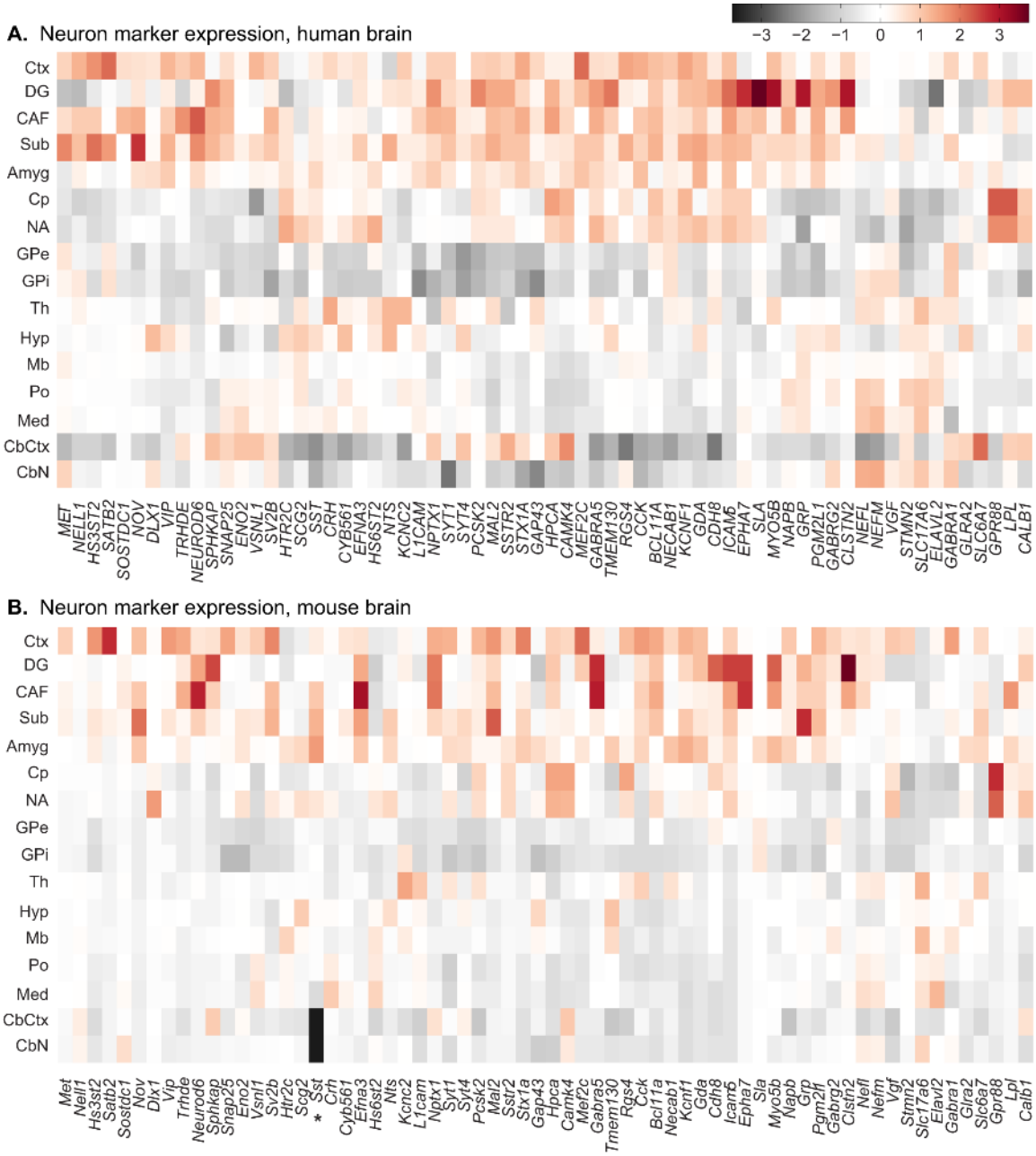


Fig 4.14 (cont. on next page). Heatmaps of human and mouse brain-wide expression patterns of mouse cell-type marker genes. Color scale is consistent across panels. Expression in human brain regions was averaged across donors. *In panel B, note that *Sst* has NaN values for all voxels of CbCtx and CbN in the AMBA.

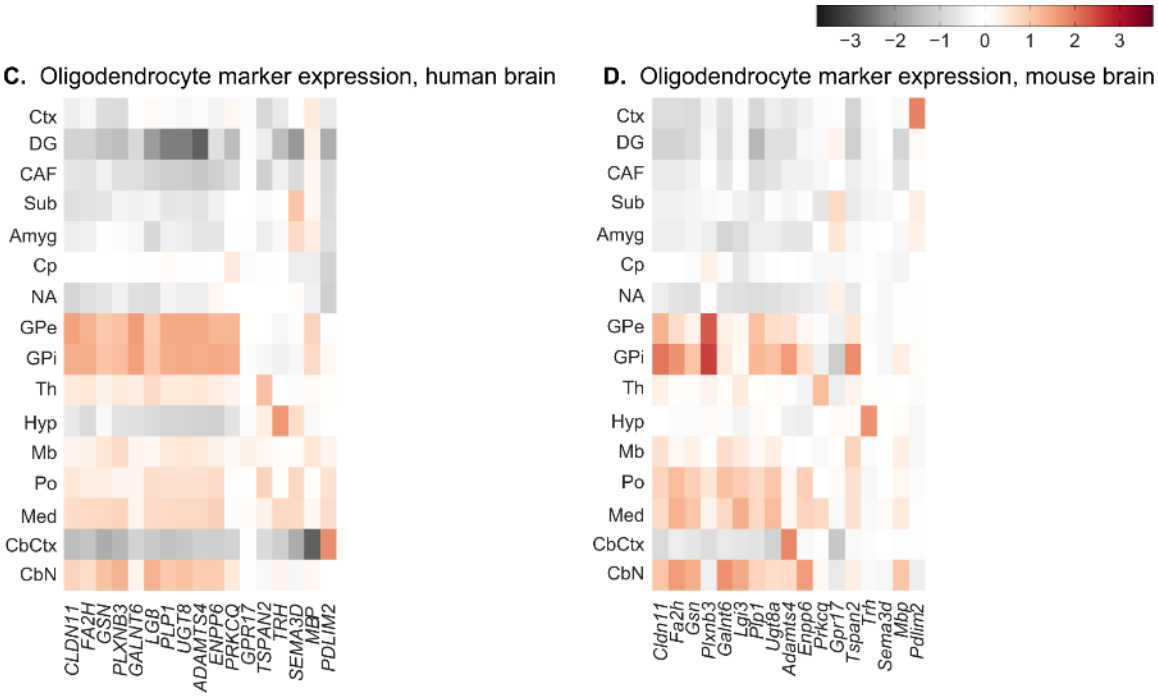


Fig 4.14. Heatmaps of human and mouse brain-wide expression patterns of mouse cell-type marker genes. Color scale is consistent across panels. Expression in human brain regions was averaged across donors.

4.3.9 Annotations of genes with high correlations across species

45 annotations were over-represented in the 5% most-correlated genes (FDR-corrected p-value < 0.01; Table 4.4). These included long-term potentiation, long-term depression, calcium signaling (Kanehisa and Goto, 2000; Kanehisa et al., 2014), glutamate signaling, nerve impulse transmission, and synaptic transmission (Ashburner et al., 2000). Highly-correlated genes also showed over-representation of several annotations curated from publications, including some that were used previously to define annotation-based gene sets such as genes down-regulated in human frontal cortex with age (Lu et al., 2004), genes involved in the mouse nucleus accumbens' response to

cocaine treatment (McClung and Nestler, 2003), and genes down-regulated in human hippocampus with Alzheimer's Disease (Blalock et al., 2004).

Annotation term in MSigDB	Associated genes in dataset	Occurrences in gene set
KEGG_LONG_TERM_POTENTIATION	32	13
GLUTAMATE_RECEPTOR_ACTIVITY	17	8
LU_AGING_BRAIN_DN	77	17
REACTOME_NEURONAL_SYSTEM	175	27
CAHOY_NEURONAL	71	16
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	130	21
REACTOME_UNBLOCKING_OF_NMDA_RECEPTOR_GLUTAMATE_BINDING_AND_ACTIVATION	12	6
V$RFX1_02	86	16
REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL	97	17
IONOTROPIC_GLUTAMATE_RECEPTOR_ACTIVITY	10	5
BLALOCK_ALZHEIMERS_DISEASE_DN	461	44
KEGG_CALCIUM_SIGNALING_PATHWAY	98	16
GLUTAMATE_SIGNALING_PATHWAY	16	6
KEGG_LONG_TERM_DEPRESSION	29	8
MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	197	24
REACTOME_CREB_PHOSPHORYLATION_THROUGH_THE_ACTIVATION_OF_CAMKII	8	4
REACTOME_TRAFFICKING_OF_AMPA_RECEPTORS	18	6
MIKKELSEN_IPS_WITH_HCP_H3K27ME3	32	8
BIOCARTA_CK1_PATHWAY	13	5
GSE19825_NAIVE_VS_IL2RAHIGH_DAY3_EFF_CD8_TCELL_UP	67	12
STARK_PREFRONTAL_CORTEX_22Q11_DELETION_UP	78	13
BIOCARTA_NOS1_PATHWAY	14	5
MCCLUNG_DELTA_FOSB_TARGETS_8WK	27	7
YOSHIMURA_MAPK8_TARGETS_UP	453	41
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS	28	7
LEIN_OLIGODENDROCYTE_MARKERS	44	9
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	240	26

MODULE_26	55	10
REACTOME_ACTIVATION_OF_NMDA_RECEPTOR_UPON_GLUTAMATE_BINDING_AND_POSTSYNAPTIC_EVENTS	22	6
REACTOME_RAS_ACTIVATION_UOPN_CA2_INFUX_THROUGH_NMDA_RECEPTOR	10	4
TRANSMISSION_OF_NERVE_IMPULSE	107	15
PLASMA_MEMBRANE	497	43
MEISSNER_NPC_HCP_WITH_H3K4ME2	194	22
ST_G_ALPHA_I_PATHWAY	16	5
V$NRSF_01	58	10
MODULE_20	24	6
MCCLUNG_COCAINE_REWARD_5D	41	8
DOANE_BREAST_CANCER_ESR1_DN	18	5
KEGG_GAP_JUNCTION	42	8
METABOTROPIC_GLUTAMATEGABA_B_LIKE_RECEPTOR_ACTIVITY	7	3
MODULE_415	7	3
PID_IL8CXCR1_PATHWAY	7	3
REACTOME_TRAFFICKING_OF_GLUR2_CONTAINING_AMPA_RECEPTORS	12	4
SYNAPTIC_TRANSMISSION	104	14
WATANABE_COLON_CANCER_MSI_VS_MSS_UP	7	3

Table 4.4. Annotations over-represented in genes with cross-species correlations in the top 5%. These terms were over-represented with $p < 0.01$ after correction using FDR.

4.4 Discussion

Previous studies have used neuroanatomically-linked gene expression data to elucidate the tight relationship between molecular and conventional neuroanatomy and have shown that some aspects of this relationship (i.e., specific genes that are differentially expressed between pairs of homologous brain regions) persist across species as closely related as human and chimpanzees (Khaitovich, 2004; Oldham et al., 2006) and as distantly related as humans and birds (Pfenning et al., 2014). Here, we have advanced understanding of the correspondences between the molecular neuroanatomical

architecture of the human and mouse brain in several ways. The present study used high-resolution datasets from the AIBS to enable a direct, systematic comparison of adult human and mouse gene expression in a set of 11 broadly defined regions and several finer sub-regions throughout the brain. We developed and applied methods which support three approaches to quantifying these comparisons. The first was global, comparing high-dimensional gene expression profiles from each location in the brain of one species to expression profiles from all locations in the brain of the other. The second approach was designed to evaluate similar expression patterns between the human and mouse brain that are specific to a brain region. We identified distinct subsets of genes which preferentially drive this molecular similarity for different regions, while other subsets show negligible similarity for the same region. Third, we took a gene-centered approach, examining the similarity between a gene's spatial expression pattern across the human brain with its orthologous pattern across the mouse brain. These three approaches in sum reveal a highly structured relationship between gene expression profiles throughout the human and mouse brains. They also demonstrate that this relationship – and in turn the similarity of the local molecular "environment" – varies considerably, not only in its overall strength, but in which genes are co-expressed in similar ways in each brain region.

4.4.1 Brain-wide comparisons of expression profiles

Similarities between gene expression profiles from throughout the human and mouse demonstrate how the local molecular architecture corresponds with conventional neuroanatomical boundaries. We found this correspondence, based on our full set of

3,792 orthologous genes, to vary by neuroanatomical region and by species, but its broad outlines appeared in cross-species comparisons as well as the within-species study described in Chapter 3. Correlations between profiles from the two different species were overall lower than correlations within-species (compare Fig 4.2 to Fig3.3). This may be a result of comparing microarray data (the AHBA) to ISH data (the AMBA). Lee et al. (2008) found relatively low correlations between the AMBA and microarray data from adult mouse brains of similar strains (~0.4-0.5 for most brain regions). The authors suggested that this was due to differences in dynamic range and signal detection thresholds for the two types of data.

The neuroanatomical organization shared by the two species was, however, similar to that appearing within-species in Chapter 3, at both broad and fine scales. The cerebral cortex, hippocampal formation, amygdala, and striatum showed some transcriptomic similarity across species that is not shared with the rest of the brain, as did the midbrain, pons, and medulla (Figs 4.2 and 4.3A). The striatum stands out with the highest within-region correlations of the broad brain regions, reflecting its consistent cellular architecture, while the amygdala and brainstem structures have lower within-region correlations (Fig 4.3C; note that these lower correlations were still higher than expected by chance as shown in Fig 4.4A). The lack of enhanced transcriptional similarity between the thalamus and hypothalamus seen in Chapter 3 also appears in Figs 4.2 and 4.3A. The cerebellum remained distinct from other structures, showing a highly positively skewed distribution of cross-species correlations only when samples and voxels were both chosen from the cerebellum (Fig 4.3A), consistent with the findings of

Strand et al. (2007). Altogether, the relationships that emerge in both within- and across-species correlation heatmaps point to the developmental plan discussed in Chapter 1.1, the basic divisions of which are common to vertebrate brains (Sanes et al., 2012, Ch. 2).

Transcriptomic organization seen within both the mouse and human brain in Chapter 3 showed cross-species correspondence at a slightly finer scale as well. Finer neuroanatomical regions tended to show somewhat stronger cross-species similarity than their parent regions (Figs 4.2 and 4.3); additionally, cross-species similarity within a parent region was nearly always significantly greater between homologous than non-homologous sub-regions.

4.4.2 Region-specific homology scores

We next considered the question of region-specific homology not only using the full list of 3,792 orthologous genes but also focusing on groups of genes that might show enhanced similarity for different neuroanatomical entities. Candidate gene sets were evaluated in each brain region using a homology score, which provided a continuous measure of the extent to which a group of genes encoded a cross-species consistent, region-specific "fingerprint." The full gene list provides such a fingerprint, to some extent, for most broad brain regions in that the human region's mean expression profile tends to have slightly higher correlations with voxels inside its mouse homolog than voxels in other regions (Fig 4.5). We began with the assumption that these enhanced correlations between homologous samples may be driven by different genes for different brain regions since each region may require a distinct, though likely overlapping, set of gene products that specify its structure and function. Our approach was not designed to

identify a unique gene set responsible for enhanced cross-species similarity for each brain region. Instead, gene sets and brain regions may have a many-to-many relationship in which (a) multiple gene sets provide strong signatures for the same region, and (b) the same gene set may provide enhanced cross-species similarity for multiple brain regions.

The latter possibility can occur if the same gene set provides uncorrelated expression profiles for distinct brain regions (i.e., the genes are expressed in a different pattern in the two areas). The penalty term λ (see Equation 5) enforces the idea that the "signature" provided by a given gene set must be uncorrelated with its signatures for other regions by decreasing the homology score to the extent that it shows enhanced similarity between a human region and a non-homologous mouse region. For example, data-driven Gene Set 23 provides relatively high homology scores for the globus pallidus, midbrain, and pons (Fig 4.8A); these homology scores would be reduced if, for example, the pattern in which these genes are expressed in the human globus pallidus was highly correlated with their expression pattern in the mouse midbrain or pons. Instead, the gene set simply provides a "basis" for comparison, with each of the human brain signatures defined over that basis both sufficiently distinct from one another and sufficiently similar with expression vectors defined over that basis in the homologous mouse brain regions. The penalty term also had the unexpected result of sharply increasing percentile ranks for certain gene sets primarily because they reduce this kind of "cross-region similarity" as compared to chance, particularly for several data-driven sets (i.e., those grouped by within-mouse brain-wide co-expression relationships), which showed strong homology

scores for the hippocampus only when their unusually low similarity with mouse cortex was taken into account by the application of the penalty (Fig 4.11).

This region-specific approach is not oriented toward individual genes which mark a given brain region in both species, but towards groups of genes whose products may have consistent region-specific interactions. The gene groups tested often showed a tendency towards particularly high or low expression values for the brain region in question (Fig 4.12), but did not simply contain anatomical marker genes. In fact, a set of genes which all have similar standardized expression values for that region (i.e., a relatively "flat" profile) in either species will result in cross-species correlations that may be dominated by noise, even if those expression values are large. The use of correlation as the measure of similarity is best suited to identifying gene sets with diverse expression values within a given region and its homolog. These diverse expression values reflect the complex, local biological environment within the region, including cell type populations of varying densities. Because expression of a gene depends on many such complex local interactions, the identification of consistent local molecular environments defined over a set of genes may help to improve our understanding of the appropriateness of knockdown / knockout mouse models of human brain disorders and the effectiveness of pharmacological agents that implicitly target specific brain regions.

4.4.3 Data-driven candidate gene sets

Gene sets were defined based on the organization of their brain-wide expression patterns in the mouse using WGCNA (Fig. 4.6; Langfelder and Horvath, 2008). WGCNA, as applied here, groups genes whose expression patterns show spatial

correlations at a broad scale, suggesting potential functional relationships between genes assigned to the same module or set. This approach has the advantage of being a discovery method, open to influence by gene-gene relationships that have not yet been identified, but which may be important in a local neuronal environment. However, a disadvantage of this approach is limited biological interpretability: that is, to the extent that we do not know specific functions or pathways underlying gene groupings, we likewise cannot identify conserved functions that underlie a high homology score.

Using this approach, specific gene sets were found that enhanced molecular similarity (relative to random gene sets) very strongly for some regions, moderately for others, and not at all for the cerebral cortex (Figs 4.7, 4.8). This variability does not appear to correspond with higher or lower variability across the six individual human donors. For example, neither the cerebral cortex nor the striatum showed especially high homology scores for any gene set, yet neither had particularly high variability across donors (their highest cross-donor interquartile rank for any data-driven set was 0.35 and 0.14, respectively, which was lower than most other broad regions). Because there was a tendency toward reduced variation in standardized expression values across the cortex relative to subcortical regions, co-expression relationships in the latter structures may have played a relatively dominant role in the creation of modules in WGCNA. Overall, it is not clear why using WGCNA in this way yielded gene sets with region-specific similarity for some brain regions and not others, but it is possible that other approaches to clustering genes might reveal different gene sets with stronger homology scores for some regions.

For further interpretation of the WGCNA-based gene sets, we used the g:Profiler online tool (Reimand et al., 2016) to assess enrichment of Gene Ontology annotation terms including "IEAs", or "Inferred from Electronic Annotations" (which are automatically assigned and have not been reviewed by a curator). Using this option and a background set consisting of the full list of 3,792 genes, a majority of data-driven sets were enriched for several brain-related annotations from the Gene Ontology. In particular, Set 3 over-represented a substantial number of annotations related to functions that occur in many areas of the brain (e.g., dendrite development, synaptic transmission, and postsynaptic density; note that Set 3 was also enriched for LTP-related genes in the previous analysis using MSigDB), and very few GO annotations that are not brain-related. Interestingly, Set 3 had higher-ranking homology scores for regions throughout the brain than most other data-driven sets (Fig 4.8). A few GO terms that were over-represented in data-driven sets have associations with specific regions; for example, learning and memory (Set 3), diencephalon development (Set 20) , and response to cocaine (Set 23). However, the only one of these with a clear relationship to a region where the gene set showed strong correspondence was diencephalon development (Set 20 had a percentile rank of 88 for the thalamus).

4.4.4 Cell-type markers as candidate gene sets

Varying distributions of cell types are central to the differentiation of brain structures. Genes whose expression marks certain cell types are thus natural candidates for assessment using our homology score. Neuron markers showed strong region-specific correspondence in nearly all regions (Fig 4.9). The relatively weak

correspondence of neuron markers for the hippocampal formation as a whole results from reduced similarity of the CA fields and subiculum rather than the dentate gyrus (Fig 4.10), and is due to a strong penalty for similarity to the cerebral cortex (i.e., neuron markers are expressed similarly in hippocampus and cortex). The low correspondence of neuron markers for the thalamus is notable, reflecting surprisingly weak correlations in cross-species expression profiles defined across neuron marker genes (the mean correlation within the mouse thalamus was 0.17, which ranked at the 40th percentile relative to random gene sets.) It is possible that thalamic nuclei with different neuron populations have varying degrees of cross-species correspondence, resulting in relatively weak correspondence for the thalamus as a whole. It is also possible that thalamic sub-structures were sampled differently in the two species, to a greater extent than other brain structures which include many nuclei.

Astrocyte markers, on the other hand, provided homology scores that were near or below chance (random gene sets) for all regions (Fig 4.9). These results are consistent with a comparison of co-expression relationships from Hawrylycz et al. (2015), in which modules of genes that were co-expressed in the Allen Human Brain Atlas were assessed for preservation in the Allen Mouse Brain Atlas. In that analysis, better-preserved modules showed higher proportions of neuron markers, while the module containing predominantly astrocyte markers was poorly preserved. Astrocytes in the human and rodent show structural, functional and molecular differences (Oberheim et al., 2006, 2009), and evidence of these differences has been found previously in large gene expression datasets (Miller et al., 2010). It is possible that astrocyte markers identified in

the mouse, as these were, may simply not provide strong cell type specific markers in the human brain. Our results support a much stronger conservation of neuron-specific markers, which are expressed in relatively unique, but cross-species consistent, patterns across brain areas.

4.4.5 Other annotation-based candidate gene sets

We additionally identified over-represented annotations across the high-scoring data-driven gene sets. We hypothesized that such annotations would reveal some of the highly conserved functions and pathways that drove high homology scores. If over-represented attributes are important to cross-species homologies (i.e., due to conserved local functions), then the more complete list of genes associated with an annotation might serve as an even stronger basis for molecular similarity. Over-represented annotations from the MSigDB for the data-driven gene sets, however, included many general terms that could not be clearly related to the brain or nervous system function. This was expected, of course, as many groups of genes work together throughout the body, and genes may take on different roles in different tissues. Therefore, the process of selecting which annotations (and resulting gene sets) to examine further was necessarily somewhat subjective. However, the selected annotations included many of those with a clear relationship to brain function.

Homology scores for these gene groups showed a different pattern from those for the data-driven gene sets or cell-type markers (Fig 4.9). Here, most gene sets showed variable performance across the eleven broad regions (unlike cell-type marker sets), while nearly all broad regions examined showed multiple gene sets ranking at or above

the 80th percentile (unlike our data-driven sets). One possible reason for the general trend toward increased homology scores for the annotation-based gene sets is that genes chosen were, in effect, required to have established brain functions. It is important to note that gene sets with unexpectedly low homology scores may also provide important insights; for example, a group of genes previously shown to be down-regulated in the human hippocampus in Alzheimer's Disease ("Alzdn"; Blalock et al., 2004) showed reduced cross-species similarity specific to the hippocampus in comparison with randomly selected gene lists of the same size. This suggests that products of these genes may function in distinct molecular environments in the mouse and human hippocampus, which should be considered in preclinical research that may target these genes.

In some cases, our results revealed enhanced similarity of gene expression profiles specific to certain brain regions with known functions that are associated with the annotation common to the genes in the set. For example, a set of genes down-regulated with age in the human frontal lobe ("FLdn"; Lu et al., 2004) yielded a higher homology score than more than 96 percent of random gene sets of the same size in the cerebral cortex (Fig 4.9A). This set includes genes involved in synaptic plasticity and neuronal survival, which, while functionally relevant for the cerebral cortex, are clearly relevant across all brain structures. Indeed, the "FLdn" set also scored highly for the hippocampal formation, striatum, and cerebellum.

A group of 41 genes whose expression in the mouse nucleus accumbens was shown to change in response to cocaine treatment ("Coc"; McClung and Nestler, 2003) showed strong region-specific similarity between the mouse and human striatum (Fig

4.9D). This cross-species similarity was maintained for both the nucleus accumbens and the caudoputamen (Fig 4.10B). Interestingly though, the human nucleus accumbens profile for this gene set has a similar average correlation with both mouse nucleus accumbens profiles ($r \sim 0.43$) and mouse caudoputamen profiles ($r \sim 0.49$). The nucleus accumbens is a major reward center in the brain, implicated in addiction (Carlezon and Thomas, 2009). McClung et al.'s finding that cocaine treatment regulates expression of these genes (McClung and Nestler, 2003) suggests that they are functionally important in the mouse nucleus accumbens. Here, we have shown that these genes show similar co-expression patterns that are unique to the striatum and are consistent in each of its sub-structures, and which are conserved between the mouse and human brain. Thus, these genes form the basis of a similar, functionally relevant molecular environment in the striatum of the two species, and directly suggest their relevance for mouse models of addiction. Genes associated with this annotation were also over-represented in the only data-driven set whose homology score for the striatum ranked above the 50th percentile (Fig 8A).

Genes involved in the opioid signaling pathway ("Op") provided high homology scores for both the dorsal and ventral striatum. Opioid receptors in the dorsal striatum have been implicated in ethanol consumption in rats (Nielsen et al., 2012), and more specifically, down-regulation in the dorsal striatum of the opioid peptides *PDYN* (in the gene set Op) and *PENK* (not in Op) has been implicated in alcoholism (Sarkisyan et al., 2015). In the ventral striatum, μ-opioid stimulation affects food intake, most likely through an effect on pleasurability of tastes (Kelley et al., 2002). This gene set also

provided a relatively high homology score in the medulla, which includes centers where opioids are involved in pain modulation (Lovick, 1985; Fields, 2004), respiratory suppression (Lovick, 1985; White and Irvine, 1999; Montandon et al., 2011), and cardiovascular function (Lovick, 1985; Tjen-A-Looi et al., 2007). However, without enough samples to analyze individual nuclei in the medulla, it is unclear which function or functions affected by the opioid signaling pathway may have contributed to this result.

On the other hand, three gene sets originally defined as marking the mouse midbrain, pons and medulla ("Mb", "Po", and "Med"; Lein et al., 2007) showed weak similarity for those specific regions (Fig 9). In the source paper, the authors identified the 100 genes with expression patterns most specific to a given brain region in the AMBA, based on the ratio of voxels expressing the gene that were inside and outside the region. The Mb, Po and Med gene sets were subsets of the top 100 genes identified for each structure (i.e., those which appear in our common gene set). Lein et al. note, however, that the midbrain, pons, and medulla do not show strongly enriched expression even for their "top 100" genes; rather, the most-specific genes for each of these regions show brain-wide expression patterns that extend beyond the brain region in question. In this study, we found that these gene sets were expressed in different patterns across the corresponding regions in mouse and human, suggesting that strong expression in these brain regions in mouse does not necessarily predict strong expression in the homologous regions in human.

There were many cases in which the common annotation used to define a gene set had no obvious relationship to brain regions for which it received a high homology score.

The cerebellum showed a tendency towards high percentile ranks for nearly all the annotation-based gene sets. FLdn, Coc, and Op are mentioned above for yielding high homology scores for brain regions with related functions; however, these sets also yielded high scores for other regions for which there is no obvious explanation. We do not know what functions these genes may have in these regions, or why they show enhanced cross-species similarity for regions not clearly associated with the annotation. In these cases, the implicated molecular mechanisms remain as obscure as with the data-driven gene sets.

In sum, the use of a region-specific homology score makes it clear that (i) there exists region-specific molecular similarity across species, and (ii) different sets of genes drive or enhance this specificity for different brain regions. The interpretation of the results for any given subset of genes is, at this point, somewhat less clear, and will require further examination of known gene functions, gene-gene interactions, and the expected composition of each region's underlying cell types.

4.4.6 Gene-gene comparisons

The AHBA and AMBA together also enabled an analysis of the similarity of brain-wide expression profiles for orthologous genes across species. While the median correlation between orthologous gene expression patterns (summarized to a set of 16 neuroanatomical regions spanning the brain) was slightly positive (~0.24), many genes had strongly positive or indeed negative correlations across species. The overall distribution of cross-species PCCs was heavily skewed, with the bulk of the density focused on positive values (Fig 4.13A). Highly correlated genes clustered into groups

that were preferentially expressed in different regions throughout the brain, suggesting their importance to brain-specific and region-specific functions (Fig 4.13B). Many of the most correlated genes are known to be associated with evolutionarily ancient mechanisms important for brain function such as long-term potentiation (Sacktor, 2012), use of glutamate as a neurotransmitter (Tikhonov and Magazanik, 2009), and use of calcium as a signal transductor (Cai et al., 2015; Table 4.4). Other genes with high cross-species similarity have shown changed expression associated with Alzheimer's Disease (Blalock et al., 2004), or aging in the human cortex (Lu et al., 2004). This result suggests that normal or pathological processes can impact even the most conserved molecular environments, and that mouse models are an appropriate tool for studying how these processes impact such genes in the human brain. Markers of mouse neurons and oligodendrocytes (Cahoy et al., 2008) showed strongly conserved brain-wide patterns as well, reflecting the similarly varying cell-type populations across the brains of the two species (Figs 4.13A, 4.14). The brain-wide expression patterns of mouse astrocyte markers, however, differed between the human and mouse, echoing their weak region-specific relationships (see "Cell-type markers as candidate gene sets", above). While the region-specific analyses performed were designed to suggest conserved local mechanisms using the co-expression relationships between different genes, gene-centric studies offer a broader view of the conservation of expression patterns for specific genes of interest. Our results suggest cross-species similarities in the brain-wide expression of many genes implicated in a variety of brain functions, but also demonstrate an

exceptional amount of variability in the degree of conservation of anatomical expression patterns across the genome.

4.4.7 Limitations and future directions

More complete interpretation of the present results, particularly those regarding region-specific molecular signatures, depends on knowledge of the functions of and relationships between both brain regions and genes. There are many protein-coding genes whose functions remain unknown, and many more (indeed a majority) for which our knowledge about their roles in the brain is sparse. Annotating the thousands of genes in the mammalian genome is a slow and arduous process; however, as these annotations accumulate, information regarding the functions of gene sets which provide consistent, cross-species molecular signatures for individual brain regions may enable new interpretations of our results. Because it is not computationally feasible to assess all possible gene subsets for region-specific homology, other approaches to defining candidate gene sets may also be informative. In particular, for each pair of homologous regions, one might apply optimization techniques, which would iteratively eliminate genes in order to optimize a region's homology score (see, for example, a related application to find genes whose co-expression patterns correlate with anatomical connectivity; French and Pavlidis, 2011).

We have focused exclusively on two large, publicly accessible gene expression datasets, the Allen Human Brain Atlas and the Allen Mouse Brain Atlas. Use of other gene expression datasets may offer opportunities to (i) assess generalization of the results to a larger human population (though we often observed a high degree of consistency in

our results across the six available donor brains) and to other strains of the lab mouse (Geurts et al., 2011; Sigmund, 2000), (ii) expand analyses of the neocortex by incorporating layer-specific information, which is not available in the AHBA , and (iii) ensure that findings are robust across expression data collected using different techniques, given that microarray data depend on the somewhat inconsistent correspondence between abundance of mRNA and protein level (Greenbaum et al., 2003). Further, gene expression is a dynamic process, and the available datasets require us to assume a meaningful "snapshot" of this process in the adult animal. The expression of many genes changes rapidly and/or systematically during brain development as they regulate, among other things, the differentiation of brain areas (Johnson et al., 2009; Thompson et al., 2014). Some local molecular environments may be similar or different across species for reasons that can only be understood by using data from earlier developmental stages to study the processes that created them. Even in later stages, while the expression of some genes may remain stable in the brain through adulthood, others undergo transient changes in expression which no one gene expression dataset can illuminate (for example, the transcription factor *c-FOS* is expressed following neuronal activity; Kovács, 1998). Other important considerations can only be addressed by looking beyond the present technologies. Gene expression is influenced by alternative splicing events, which are ubiquitous in the brain (Grabowski, 1998). Additionally, transcriptional regulatory networks (Vaquerizas et al., 2009; Ravasi et al., 2010) and post-transcriptional mechanisms (Day and Tuite, 1998) may also differ across species.

CHAPTER 5: THE TRANSCRIPTIONAL LANDSCAPE OF GENES IMPLICATED IN SPEECH AND LANGUAGE DISORDERS

5.1 Introduction

Developmental disorders of speech and language are highly heritable, and over two dozen genes have been implicated in one or more of these pathologies. There is an extensive body of work indicating that certain measures of speech and/or language ability are associated with DNA variants at loci within a given gene, or with structural variations of the chromosome which affect the gene (for two reviews, see Fisher et al., 2003; Graham et al., 2015). However, the causal links between genotypic variations and phenotypic measures that capture speech and language function remain largely obscure. This chapter, which is an extension of previously published work (Bohland et al., 2014), reviews the putative associations between genes and speech / language abilities and uses gene expression data to suggest relationships between the neuroanatomical localization of implicated genes, and how these may be used to suggest hypotheses regarding the roles those genes play in speech / language function. "Candidate" genes, curated from the literature on the genetics of speech and language and entered into a database described in the next section, are examined to determine where (if anywhere) in the brain they are preferentially expressed, and to characterize their co-expression relationships throughout the brain.

Speech and language disorders are nearly always polygenic: loci within many genes each exert a small degree of influence on variation of the phenotype (Fisher et al., 2003). These influences are characterized as "quantitative trait loci" (QTLs) and stand in

contrast to rare situations in which a variant in a single gene explains a disruption of normal speech or language function (e.g., a single *FOXP2* mutation is causally implicated in developmental verbal dyspraxia, or DVD, in ~2% of cases; MacDermot et al., 2005). The predominance of statistically weak QTLs among genetic influences on speech and language disorders (and neurological and neuropsychiatric disorders more broadly) poses a challenge, since the identification of multiple loci with small effects requires more studies with higher statistical power than do monogenic influences. Indeed, nearly all the genotype-phenotype associations reviewed in this chapter were found in only a small percentage of the cohorts studied (e.g., mutations in *NAGPA* and *GNPTG* were found in only 2% and 4% of subjects with persistent developmental stuttering, or PDS, respectively; Kang et al., 2010). Just as a given phenotype may have multiple genetic influences, a given gene may influence multiple phenotypes, and these may range from cognitive abilities to basic cellular functions. For example, *NAGPA* and *GNPTG*, with another PDS candidate (*GNPTAB*; Kang et al., 2010), are involved in directing enzymes to the lysosome, which affects cellular processes such as waste disposal (Settembre et al., 2013). Thus, although the phenotypes of interest here are all related to speech and language, genes which impact them are not "speech genes" or "language genes", but may influence many very different processes, some of which remain unknown.

The most obvious mechanism for genetic variants to impact behavior is by making specific changes in brain areas where those genes are expressed. A single gene, expressed in multiple functionally relevant brain systems, could therefore impact multiple disorders. Similarly, genes that are co-expressed might therefore impact the same

disorder. This suggests that, though many disorders are polygenic, different disorders may often be impacted by the same genes. In an influential paper proposing what has become known as the Generalist Genes Hypothesis, Plomin and Kovas (2005) argued that this is the case for most learning disorders (in language, reading and mathematics). The hypothesis is based primarily on high genetic correlations between different learning disorders, and between aspects of a single disorder. Briefly, the genetic correlation between two traits is the likelihood that a gene associated with one trait will also be associated with the other. Any observed covariation, however, does not guarantee that both associations are causal: for example, if one trait directly affects the other trait, this will increase their genetic correlation regardless of shared genetic influences. However, the consistently high genetic correlations between several measures of language, reading, and mathematical ability are at least suggestive of overlap between the genes influencing these abilities (Plomin and Kovas, 2005). While the Generalist Genes Hypothesis is relevant to learning disorders, it should be noted that not all the disorders of speech and language function discussed here are always considered learning disorders: DVD and PDS, for example, are usually thought of as speech motor disorders, which concern the inability to effectively produce (but not necessarily failure to learn) speech sounds. Most genetic studies in this field focus on associations with a single speech / language disorder. However, a few of the genes discussed below were originally associated with one such disorder, but were then implicated in another as well. For example, *ATP13A4* has been implicated in specific language impairment (SLI) and DVD (Kwasnicka-Crawford et al.,

2005; Worthey et al., 2013), and *CMIP* in both SLI and dyslexia (Newbury et al., 2009, 2011; Scerri et al., 2011).

Plomin and Kovas also hypothesized that the same genes which influence learning disabilities also influence normal variation in learning ability (Plomin and Kovas, 2005). In other words, learning disabilities are not etiologically distinct from learning ability, but are the tail end of the overall population distribution. Of the genes analyzed here, only a few have been tested for association with normal variation (i.e., using a typical, large population sample). *DYX1C1*, *DCDC2*, *KIAA0319* and *TTRAP* were associated with reading and spelling skill (Paracchini et al., 2011; Lind et al., 2010; Paracchini et al., 2008; Luciano et al., 2007), and *ROBO1* with nonword repetition and short-term storage of verbal sequences (Bates et al., 2011). However, one study failed to replicate the association for *DCDC2*, or to find association between three other genes associated with dyslexia and dyslexia-related behavioral measures in the same large, typical cohort (*MRPL19*, *C2ORF3*, and *KIAA0319*; Paracchini et al., 2011). Thus, the extent to which speech / language ability and disability share a genetic etiology is not yet clear. Note also that two genes are included here entirely due to associations found in subjects with no neurological disorders: *CACNA1C* with performance on a lexical access task, and *GRM3* with neural responses to unexpected phonemes (Krug et al., 2010; Harrison et al., 2008).

The genes treated in this chapter (see Table 5.1 for a complete list) are curated from publications using a range of methods to identify genetic causes of speech / language disorders, including genome-wide association studies, whole-exome

sequencing, and sequencing of just a few individuals with a disorder or a single family with members who have a disorder. This list of speech / language candidate genes, or "SL genes", errs on the side of inclusion: in many cases, the evidence for association is a single study that has not been replicated. Most of these genes are associated with one or more of four disorders: dyslexia (also sometimes known as reading disorder), specific language impairment, developmental verbal dyspraxia (also known as childhood apraxia of speech in the United States), and persistent developmental stuttering. These are discussed in turn below.

Symbol	**Aliases**	**Entrez ID**	**Associated phenotype(s)**	**Sources**
ADARB2	*RED2*	105	PDS	Kraft (2010)
AP4E1	*CPSQ4*, *SPG51*	23431	PDS(2015)	Raza et al. (2015)
ARNT2		9915	PDS	Kraft (2010)
ATP13A4		84239	SLI DVD	Kwasnicka-Crawford et al. (2005) - SLI Worthey et al. (2013) - DVD
ATP2C2	*hSPCA2*	9914	SLI	Newbury et al. (2009)
BCL11A		53335	DVD	Peter et al. (2014)
BDNF		627	SLI	Simmons et al. (2010)
CACNA1C		775	SVF	Krug et al. (2010)
CEP63	*SCKL6*	80254	DYX	Einarsdottir et al. (2015)
CFTR		1080	SLI(O'Brien et al., 2003)	O'Brien et al. (2003)
CMIP		80790	SLI DYX	Newbury et al. (2009, 2011) - SLI Scerri et al. (2011) - DYX
CNTNAP2		26047	SLI DYX	Vernes et al. (2008) - SLI Peter et al. (2011) - DYX Newbury et al., (2011) - SLI, DYX
CTNNA3	*VR22*, *ARVD13*	29119	PDS	Kraft (2010)
CYP19A1		1588	DYX DVD	Anthoni et al. (2012) - DYX, DVD
DCDC2		51473	DYX	Deffenbacher et al. (2004); Meng et al. (2005a); Schumacher et al. (2006)
DGKI		9162	DYX	Matsson et al. (2011)
DIP2A	*DIP2*, *C21orf106*	23181	DYX	Poelmans et al. (2009); Kong et al. (2016)
DOCK4		9732	DYX	Pagnamenta et al. (2010)
DRD2		1813	PDS[1]	Lan et al. (2009)
DYX1C1	*EKN1*	161582	DYX[2]	Taipale et al. (2003); Scerri et al. (2004); Wigg et al. (2004); Brkanac et al. (2007); Marino et

				al. (2007); Dahdouh et al. (2009); Lim et al. (2011); Newbury et al. (2011); Paracchini et al. (2011); Zhang et al. (2012); Mascheretti et al. (2013)
ERC1	*ELKS*	23085	DVD	Thevenon et al. (2013)
EYA2		2139	PDS	Kraft (2010)
FADS2		9415	PDS	Kraft (2010)
FMN1		342184	PDS	Kraft (2010)
FOXP1		27086	ELS	Hamdan et al. (2010)
FOXP2		93986	DVD DYX SLI POV	Lai et al. (2001); MacDermot et al. (2005); Feuk et al. (2006); Lennon et al. (2007) - DVD Peter et al. (2011) - DYX Rice et al. (2009) - SLI Tolosa et al. (2010) - POV
GNPTAB		79158	PDS	Kang et al. (2010)
GNPTG		84572	PDS	Kang et al. (2010)
GPLD1		2822	DYX	Meng et al. (2005a)
GRM3		2913	MMN	Harrison et al. (2008)
KIAA0319		9856	DYX[3] SLI	Francks et al. (2004); Cope et al. (2005); Harold et al. (2006); Paracchini et al. (2008); Venkatesh et al. (2013a) - DYX Rice et al. (2009); Newbury et al. (2011) - SLI
NAGPA		51172	PDS	Kang et al. (2010)
NFXL1		152518	SLI	Villanueva et al. (2015)
NRSN1	*VMP*	140767	DYX	Deffenbacher et al. (2004)
PCSK5		5125	PDS	Kraft (2010)
PLXNA4		91584	PDS	Kraft (2010)
ROBO1		6091	DYX[4] SLI	Hannula-Jouppi et al. (2005) - DYX Bates et al. (2011) - SLI
SETBP1		26040	ELS	Filges et al. (2011); Marseglia et al. (2012)
SLC24A3		57419	PDS	Kraft (2010)
SRPX2		27286	DVD	Roll (2006)
THEM2		55856	DYX	Francks et al. (2004); Cope et al. (2005); Harold et al. (2006); Paracchini et al. (2008); Venkatesh et al. (2013a)
TTRAP		51567	DYX	Francks et al. (2004); Cope et al. (2005); Harold et al. (2006); Paracchini et al. (2008); Venkatesh et al. (2013a)

Table 5.1. Candidate speech / language genes included in these analyses ("SL genes"). DVD = developmental verbal dyspraxia, DYX = dyslexia, ELS = expressive language skills, MMN = mismatched negativity response to unexpected phonemes, PDS = persistent developmental stuttering, POV = poverty of speech, SLI = specific language impairment, SVF = semantic verbal fluency. Note that some of the sources use gene symbols from the "Alias" column. Best-supported / first-found association is listed first in "Associated phenotypes".

1. Kang et al. (2011a) found no association between *DRD2* and PDS. 2. Bellini et al. (2005), Marino et al. (2005), Meng et al. (2005b), Ramachandra et al. (2008), and Venkatesh et al. (2011) found no association between *DYX1C1* and dyslexia. 3. Paracchini et al. (2011) found no association between *KIAA0319* and dyslexia. 4. Venkatesh et al (2013b) found no association between *ROBO1* and dyslexia.

Before delving into the genes and phenotypes listed in Table 5.1, several important caveats should be noted. First, speech / language disorders are complex phenotypes. Defining a phenotype of interest as a disorder simplifies analysis, but does obscure the impact of genes on specific aspects of the disorder (*endophenotypes*). Similarly, it is not truly the gene as a whole that impacts the phenotype of interest, but one or more variants within the gene (possibly working in concert with variants in other genes). This discussion, and the subsequent analyses, deal with the relatively coarse level of genes and (primarily) disorders. The studies cited in Table 5.1 include more detailed information on both implicated genotype and, in many cases, impacted phenotype. Second, an association between genotype and phenotype need not be causal. Variants in two different genes may be statistically associated, and if one variant influences a given phenotype, the other will show a relationship with that phenotype as well. Third, certain genetic variants known as *expression quantitative trait loci* (eQTLs) impact the expression of other genes. In other words, the gene indicated by genetic research may not always be the same as the gene whose expression is relevant to the speech / language phenotype. The identification of eQTLs requires both genotype and gene expression data, and may be key in pointing to genes whose expression profiles are of interest in this research.

5.1.1 Dyslexia candidates

The ability to read and spell rests on typical development of cognitive functions including orthographic processing, phonemic awareness, and phonological short-term memory. Skill in these areas can be measured with a variety of well-established

assessments including reading rate and accuracy measures, non-word repetition, and rapid automatized naming. Several genomic regions have been implicated in susceptibility to dyslexia, in particular on chromosomes 15 and 6. The first region (DYX1) includes the gene ***DYX1C1***, which was associated with dyslexia in several cohorts (Taipale et al., 2003; Scerri, 2004; Wigg et al., 2004; Brkanac et al., 2007; Marino et al., 2007; Dahdouh et al., 2009; Newbury et al., 2011; Paracchini et al., 2011; Lim et al., 2011; Zhang et al., 2012; Mascheretti et al., 2013). *DYX1C1* is arguably the most-studied gene on this list, and some studies have failed to find an association between this gene and dyslexia (Bellini et al., 2005; Marino et al., 2005; Meng et al., 2005b; Ramachandra et al., 2008; Venkatesh et al., 2011). ***CYP19A1*** is also located in DYX1 and has been associated with dyslexia (Anthoni et al., 2012). The second region (DYX2) includes a haplotype (a set of genomic markers that are usually inherited together) which has been associated with dyslexia and which spans some of ***KIAA0319***, all of ***TTRAP***, and regulatory regions of ***THEM2*** (Cope et al., 2005; Francks et al., 2004; Harold et al., 2006; Paracchini et al., 2008; Venkatesh et al., 2013a), though one study found no association between dyslexia and *THEM2* (Venkatesh et al., 2013b). DYX2 also includes three other genes which have been associated with dyslexia. ***DCDC2*** has received the most attention (Deffenbacher et al., 2004; Meng et al., 2005a; Schumacher et al., 2006; Wilcke et al., 2009; Newbury et al., 2011); however, single nucleotide polymorphisms (SNPs) in ***GPLD1*** and ***NRSN1*** have each shown association with performance on several tests used in the diagnosis of dyslexia (Meng et al., 2005a; Deffenbacher et al., 2004).

Additionally, ***CEP63*** (Einarsdottir et al., 2015), ***DIP2A*** (Poelmans et al., 2009; Kong et al., 2016), ***DGKI*** (Matsson et al., 2011), ***DOCK4*** (Pagnamenta et al., 2010), and ***ROBO1*** (Hannula-Jouppi et al., 2005) have all been implicated in dyslexia, with one non-replication of the association with *ROBO1* (Venkatesh et al., 2013b). Finally, there is some evidence for involvement of ***FOXP2*** (Peter et al., 2011), ***CNTNAP2*** (Newbury et al., 2011; Peter et al., 2011), and ***CMIP***. (Scerri et al., 2011) Those two, however, are primarily and most commonly associated with other disorders, *FOXP2* with DVD and *CNTNAP2* and *CMIP* with SLI (see sections on DVD and SLI, below).

Most of these genes are known or hypothesized to have roles in brain-specific functions, including cortical neuron migration (*DYX1C1*, Wang et al., 2006; *KIAA0319*, Paracchini, 2006; *DCDC2*, Meng et al., 2005a), neurite development (*NRSN1*, Araki and Taketani, 2009; *DIP2A*, Poelmans et al., 2011; *DOCK4*, Ueda et al., 2008), axon guidance (*ROBO1*; Seeger et al., 1993; Kidd et al., 1998), cerebral cortex growth (*CEP63*; Sir et al., 2011), cortical neuron proliferation, sexual differentiation of brain areas, development of the neural circuitry underlying vocalizations in songbirds (*CYP19A1*, Morris et al., 2004; Forlano et al., 2006; Martínez-Cerdeño et al., 2006; Diotel et al., 2010), and presynaptic signaling (*DGKI*; Yang et al., 2011). *CYP19A1* codes for aromatase, which is also essential to the development of vocal and auditory circuits in songbirds and vocalizing fish (Forlano et al., 2006). The involvement of some of these genes in cortical development may be a clue to the anomalies of cortical neuron migration found in brains of dyslexic donors (Galaburda et al., 1985), and reading deficits in subjects with a cortical migration disorder (Chang et al., 2005). For a review of

possible connections between *DYX1C1*, *ROBO1*, *KIAA0319*, and *DCDC2* and issues in cortical development, see Galaburda et al. (2006).

Interestingly, *DOCK4*, *DGKI*, *FOXP2*, and *CNTNAP2* are all located on chromosome 7q31-35. This is within a genomic region that has been linked to autism, a disorder that also has a language component (Abrahams and Geschwind, 2008). *CNTNAP2* and *DOCK4* have both been associated with autism (Peñagarikano and Geschwind, 2012; Pagnamenta et al., 2010), as has *FOXP2* in some populations but not others (e.g., Gauthier et al., 2003; Gong et al., 2004).

Anthoni et al (2006) have suggested *MRPL19* and the adjacent *C2ORF3* as dyslexia candidates. *MRPL19* is also adjacent to an intergenic region containing a risk haplotype, and the two genes are in strong linkage disequilibrium (i.e., variants in these genes are statistically associated). The authors also found co-expression between these genes and other dyslexia candidates (*DYX1C1*, *ROBO1*, *DCDC2* and *KIAA0319*), and attenuated expression of both genes in carriers of the risk haplotype. However, three later studies failed to identify any association between these genes and dyslexia (Paracchini et al., 2011; Scerri et al., 2011; Venkatesh et al., 2013b). Therefore, *MRPL19* and *C2ORF3* have been excluded from the list of candidate genes treated here.

5.1.2 Specific language impairment (SLI) candidates

SLI is a deficit in the normal development of expressive and / or receptive language skills with no more general explanation (such as hearing impairment or intellectual disability). SLI is therefore a heterogeneous disorder, although children with SLI characteristically speak in short, simplified sentences (Newbury et al., 2005).

Although the classification of language disorders has changed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V, American Psychiatric Association, 2013), all research discussed here uses the term SLI as described. Phonemic awareness and phonological short-term memory are impaired in SLI as in dyslexia (Bishop et al., 1996; Shaywitz and Shaywitz, 2005), and the relationship between the two disorders is a subject of much debate, with some arguing that the underlying deficits are the same and that dyslexia should be considered a mild form of SLI (see Bishop and Snowling, 2004, for a review). It is not surprising, therefore, that several genes are associated with both disorders.

CNTNAP2 and ***CFTR*** are both located near ***FOXP2*** (at the autism susceptibility locus on chromosome 7q, discussed above), and all three genes have been associated with SLI (Vernes et al., 2008; Newbury et al., 2011; O'Brien et al., 2003; Rice et al., 2009). In fact, *FOXP2* down-regulates *CNTNAP2* (Vernes et al., 2008), which codes for a neurexin protein that may have a role in human cortical development (Abrahams et al., 2007). ***ATP2C2*** and ***CMIP***, located within a risk locus on chromosome 16q (SLI Consortium, 2002), have also been associated with SLI (Newbury et al., 2009, 2011). Implicated genes at other loci include ***NFXL1*** (Villanueva et al., 2015), ***ATP13A4*** (Kwasnicka-Crawford et al., 2005), and ***BDNF*** (Simmons et al., 2010). *ATP13A4* has been hypothesized to have a role in early neural development (Vallipuram et al., 2010). *BDNF*, or "brain-derived neurotrophic factor", has several important roles especially in central nervous system development, including regulation of dendrite growth and

synaptic plasticity, and is implicated in many other neurological and neuropsychiatric disorders (McAllister et al., 1999; Autry and Monteggia, 2012).

KIAA0319, a candidate gene for dyslexia, shows association with SLI as well (Rice et al., 2009; Newbury et al., 2011). It has also been argued that ***ROBO1***'s association with dyslexia results from an impact on phonological short-term memory, which is equally central to SLI (Bates et al., 2011).

5.1.3 Developmental verbal dyspraxia (DVD) candidates

DVD (also known as childhood apraxia of speech, or CAS) is not always consistently defined, but is generally considered to be characterized by an impaired ability to produce (and possibly sequence) the orofacial movements required for speech, in the absence of muscle weakness or paralysis (e.g., Lai et al., 2001; Ferry et al., 2008). In rare cases of DVD, mutations in ***FOXP2*** cause or are associated with the disorder (Lai et al., 2001; MacDermot et al., 2005; Feuk et al., 2006; Lennon et al., 2007). *FOXP2* is a transcription factor that primarily down-regulates the expression of other genes, and its hundreds of putative targets have roles in neural transmission, synaptic plasticity, and axon guidance (and other cellular functions not specific to the CNS; Vernes et al., 2007).

Another DVD candidate, ***BCL11A***, is located in a region linked to dyslexia (Peter et al., 2014) and is involved in cortical neuron migration (Wiegreffe et al., 2015). A deletion specific to *BCL11A* was found in a patient with DVD as well as more generalized apraxia, hypotonia, and motor delays. Although the report acknowledges the requirement that difficulty producing speech sounds is not accounted for by muscle weakness, it is unclear how the diagnosis of DVD was made given the generalized

apraxia and hypotonia (Peter et al., 2014). ***SRPX2*** is also associated with DVD accompanied by other traits, primarily seizures (Roll, 2006). This gene is down-regulated by *FOXP2* (Roll et al., 2010), and its over-expression in mice affects synapse density and interferes with ultrasonic vocalization (Sia et al., 2013). ***ERC1*** is located in a region of overlap between deletions in several people with DVD (Thevenon et al., 2013). There is also some evidence that ***ATP13A4***, previously mentioned for its potential role in SLI, may be involved in some cases of DVD (Worthey et al., 2013).

Finally, in addition to its association with dyslexia, ***CYP19A1*** has been associated with speech sound disorder (SSD), which is also characterized by difficulty in producing intelligible speech sounds (Anthoni et al., 2012). It is important to note that the distinction, if any, between DVD / CAS and SSD in the view of these researchers and those who diagnosed their subjects is unclear; the DSM-V has since defined DVD as a sub-type of SSD.

5.1.4 Persistent developmental stuttering (PDS) candidates

PDS is characterized by syllable repetitions and prolongations, and interruptions of speech flow, typically developing gradually between ages 3 and 8. In many children who initially develop these dysfluencies, they resolve spontaneously within a few years; in PDS, they may persist into adulthood (Ashurst and Wasson, 2011). While the mechanisms underlying PDS are unclear, multiple brain areas and circuits have been implicated, in particular the basal ganglia (see Craig-McQuaide et al., 2014 for a review).

Mutations in ***GNPTAB***, ***GNPTG***, ***NAGPA***, and ***AP4E1*** have been associated with PDS in several populations (Kang et al., 2010; Raza et al., 2015). The first three are part

of a signaling pathway that directs enzymes to the lysosome, an organelle with roles in a wide range of processes including waste disposal, nutrient sensing, and membrane repair (Settembre et al., 2013). Mutations in *GNPTAB* and *GNPTG* are known to cause lysosomal storage disorders. Kang and Drayna (2012) point out that these disorders do sometimes have surprisingly specific effects in which only certain organs show defects, and they can include neurological deficits. However, the mechanistic connection between this pathway and the rather specific PDS phenotype is unknown. *AP4E1* is also involved in sorting proteins, and in neurons it may mediate the transport of AMPA glutamate receptors to the postsynaptic domain (Matsuda et al., 2008). Cases of microcephaly and intellectual disability have also been linked to mutations in *AP4E1* (Moreno-De-Luca et al., 2011; Kong et al., 2013).

One unpublished genome-wide association study of PDS, though smaller than most genome-wide studies (84 people with PDS and 107 controls), suggested nine candidate genes based on statistically significant variants (Kraft, 2010). Several of these genes have known relationships with a variety of brain functions and neurological disorders. ***ARNT2***, ***PLXNA4***, and ***CTNNA3*** have roles in neuroendocrinological cell development (Hosoya et al., 2001), axon guidance (Suto et al., 2005), and cell-cell adhesion (Smith et al., 2011), respectively. All three are associated with autism (Maestrini et al., 2010; Bacchelli et al., 2014; Di Napoli et al., 2015). Additionally, *ARNT2* regulates *BDNF*, an SLI candidate (Pruunsild et al., 2011). *PLXNA4* is also associated with Alzheimer's Disease and Parkinson's Disease (Jun et al., 2014; Schulte et al., 2013), and there is conflicting evidence for *CTNNA3*'s involvement in Alzheimer's

Disease (Smith et al., 2011). ***ADARB2*** affects glutamate levels in white matter (Kawahara et al., 2003), while differences have been found in white matter of people with PDS (Connally et al., 2014). ***FADS2*** modulates the effect of fatty acid intake during development, and it has been argued that this affects cognition (Rizzi et al., 2013). The other four genes indicated by this GWAS are ***EYA2***, ***FMN1***, ***SLC24A3***, and ***PCSK5***, which are not, based on a survey of current literature, implicated in any neurological disorders or brain-specific functions (Kraft, 2010).

Finally, the hypothesis that PDS involves excessive dopamine (and in particular D2 receptors in the striatum; Alm, 2004) resulted in a study showing an association between ***DRD2*** and PDS (Lan et al., 2009). This association, found in a Chinese cohort, did not appear in either a Brazilian or a European cohort (Kang et al., 2011a). It is unknown whether the initial finding in a Chinese cohort was a false positive, or whether the association is present in the first population but not the latter two; however, the authors of the non-replication point out that the allele associated with PDS in the Chinese cohort was far less common in the latter two cohorts, and that this difference may have obscured the association in the second study.

5.1.5 Other phenotypes related to speech and language ability

Genes that are examined in this chapter also include ***FOXP1*** and ***SETBP1***. Both are implicated in expressive language impairments, accompanied by broader effects (aggressiveness and obsessive-compulsive behavior for *FOXP1*, Hamdan et al., 2010; delayed / impaired motor skills and distinctive facial features for *SETBP1*, Filges et al., 2011; Marseglia et al., 2012).

Schizophrenia has a language component, and two genes implicated in schizophrenia also show association with language measures in healthy controls. The first is ***GRM3*** (Harrison et al., 2008), which predicted neural responses to unexpected phonemes as measured by mismatch negativity using magnetoencephalography. (Kawakubo et al., 2011). The second is ***CACNA1C*** (Green et al., 2010), which was associated with performance on a lexical access task (Krug et al., 2010). There is evidence that performance on this task is impaired in dyslexia and SLI; therefore, *CACNA1C* may be a candidate gene for these disorders (Cohen et al., 1999; Weckerly et al., 2001).

Finally, in addition to connections with multiple disorders discussed above, ***FOXP2*** has been associated with poverty of speech in people with schizophrenia (Tolosa et al., 2010) and ***CNTNAP2*** with language acquisition (Whitehouse et al., 2011; Al-Murrani et al., 2012) and response to syntactic violations as measured by event-related brain potentials (Kos et al., 2012).

For many of these genes, the mechanisms relating genotype to speech and language abilities are completely unknown. For some genes, little or nothing is known of their functions. Given our sparse knowledge of these genes, their roles and their relationships, it is not easy to move beyond making lists of candidates that have shown statistical associations. Neuroanatomically-specific gene expression data suggest a first step. As mentioned earlier, the most obvious mechanism for genetic variants to influence these phenotypes is through effects on brain areas where those genes are expressed. The expression profiles of candidate genes are therefore of interest, not only across the brain

as a whole, but across samples from brain structures of particular relevance to speech and language functions.

In addition to cerebral cortex, the basal ganglia and cerebellum support many aspects of these functions. Cortico-basal ganglia and cortico-cerebellar circuitry are involved not only in speech-related motor control (Kent, 2000; Wildgruber et al., 2001; Riecker et al., 2005) but in speech processing, including at the phoneme level (Booth et al., 2007; Peeva et al., 2010; see Mariën et al., 2013; Kotz et al., 2009 for reviews of CB and BG involvement, respectively). In these language-related circuits, cortical projections from the basal ganglia and cerebellum have different but overlapping distributions at the thalamus, suggesting an interplay between their roles (Barbas et al., 2013).

The cerebral cortex, basal ganglia and cerebellum have all been implicated in dysfunctions of speech and language as well as healthy functioning. Dyslexia has long been tied to disturbances of cortical neuronal migration, particularly in the temporal lobe (Galaburda, 2005; Giraud and Ramus, 2013). Areas from all four cortical lobes (particularly parietal and inferior frontal) show differences between people with dyslexia and controls, both in structure and in functional activations during relevant language tasks, as do the anterior and posterior lobes of the cerebellum (see Eckert, 2004 for a review). Ullman and Pierpont (2005) have proposed that SLI results from deficits in the procedural memory system, including the cerebral cortex, basal ganglia and cerebellum. All three structures have also been implicated in PDS (Alm, 2004; Brown et al., 2005) and in DVD (Belton et al., 2003; Vargha-Khadem et al., 1998). Nicolson and Fawcett

(2007) hypothesized more generally that many developmental disorders might arise from disturbances of cortico-basal ganglia and cortico-cerebellar circuitry.

By determining where in the brain candidate genes are expressed, and whether / where they are co-expressed, we can consider our knowledge of these genes in light of our incomplete, but more substantial, knowledge of larger-scale functional neuroanatomy and pathophysiology of the disorders. The study described in this chapter identifies brain areas where these genes are preferentially expressed, and shows their co-expression relationships using multi-dimensional scaling and co-expression network analysis approaches.

5.2 Speech and Language Disorders Database

A major gap exists between studies that aim to identify candidate genes for developmental language disorders and brain imaging studies of the neural bases for linguistic functions, which may be impacted by those same disorders. This section outlines the construction of a novel web-accessible database that aims to narrow this gap by bringing both data types into a common framework. This database also, in large part, provides the list of candidate genes shown in Table 5.1. The overall database (http://neurospeech.org/sldb) contains a series of manually curated results from the published literature describing (i) genes or chromosomal regions (genetic loci) for which there is evidence of association to speech and language-related phenotypes, and (ii) results from the brain imaging literature describing localized structural or functional abnormalities observed in populations of individuals with heritable speech or language

disorders compared to typically developing control subjects. Thus these efforts bring results from two fundamentally different levels of investigation together into a common database system. This section focuses on the portions of the database concerned with candidate genes rather than with brain imaging studies. The overall database structure is further described in Bohland et al. (2014).

5.2.1 Candidate genes

The database currently contains records related to 27 individual genes that have been implicated, with varying degrees of evidence, in speech- or language-related phenotypes. The current gene list includes genes for which some association has been found to DYX, SLI, DVD, and PDS; it also includes genes that have been specifically linked to quantitative measures of speech and language processing.

For each of the genotype-phenotype relationships curated from the literature and stored in the database, our web-based interface provides users simple access to the curated details, and also provides links to the Entrez Gene database (http://www.ncbi.nlm.nih.gov/gene), the original publication via Pubmed (http://www.ncbi.nlm.nih.gov/pubmed/), and − using simple URL based mapping − to the Allen Human Brain Atlas (AHBA) website (http://human.brain-map.org) depicting brain expression profiles for the gene of interest. The Allen Brain Atlas Application Programming Interface (API) is used to provide direct links to download a complete set of pre-processed, normalized human gene expression data from the AHBA for each gene of interest, in either JSON or XML format. A screen capture demonstrating the primary entry point in the database (http://neurospeech.org/sldb), which summarizes database

contents for these genes, is shown in Fig 5.1. The table shown in the screen capture shows basic gene metadata, and provides links to AHBA data for each gene in the database. It also gives a summary of how many studies with positive association results for this gene have been curated and included in the database to date, and the total number of gene-phenotype associations reported for each gene. This allows the novice user to quickly ascertain which genes are of highest interest (reflected in many studies), and also places responsibility on the curators to ensure that sampling of studies is as free from bias as possible. Clicking the row corresponding to an individual gene leads to a page summarizing the reports of associations, as well as any replication failures (i.e., studies that tested examined the same gene and did not find association with a similar phenotype) entered into the database, for that gene. This more detailed view is depicted in Fig 5.2 for the gene *ROBO1*.

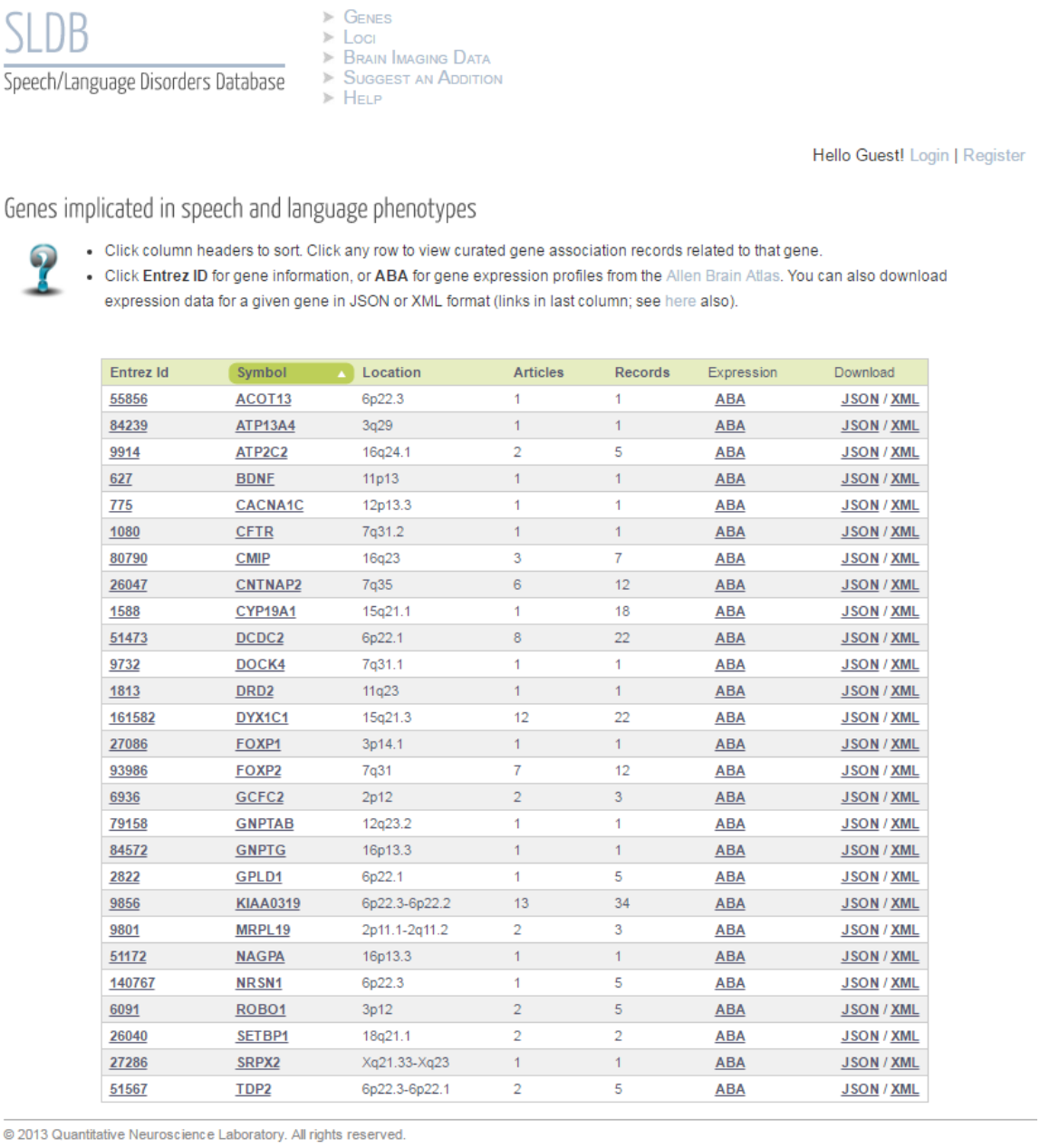


Fig 5.1. Screen capture of a (partial) view into the database showing a sortable table of genes implicated in speech / language phenotypes, sorted by gene symbol. From this primary view, the user is able to quickly survey the list of genes implicated in speech and language disorders that are currently in the database, as well as the number of reports curated by gene, and hyperlink to a number of relevant resources for each record, including the Entrez Gene page for the gene of interest, and the Allen Human Brain Atlas page for the gene of interest. The right-most links (labeled "JSON/XML) provide a mechanism to download gene expression data for this gene using the Allen Brain Atlas API.

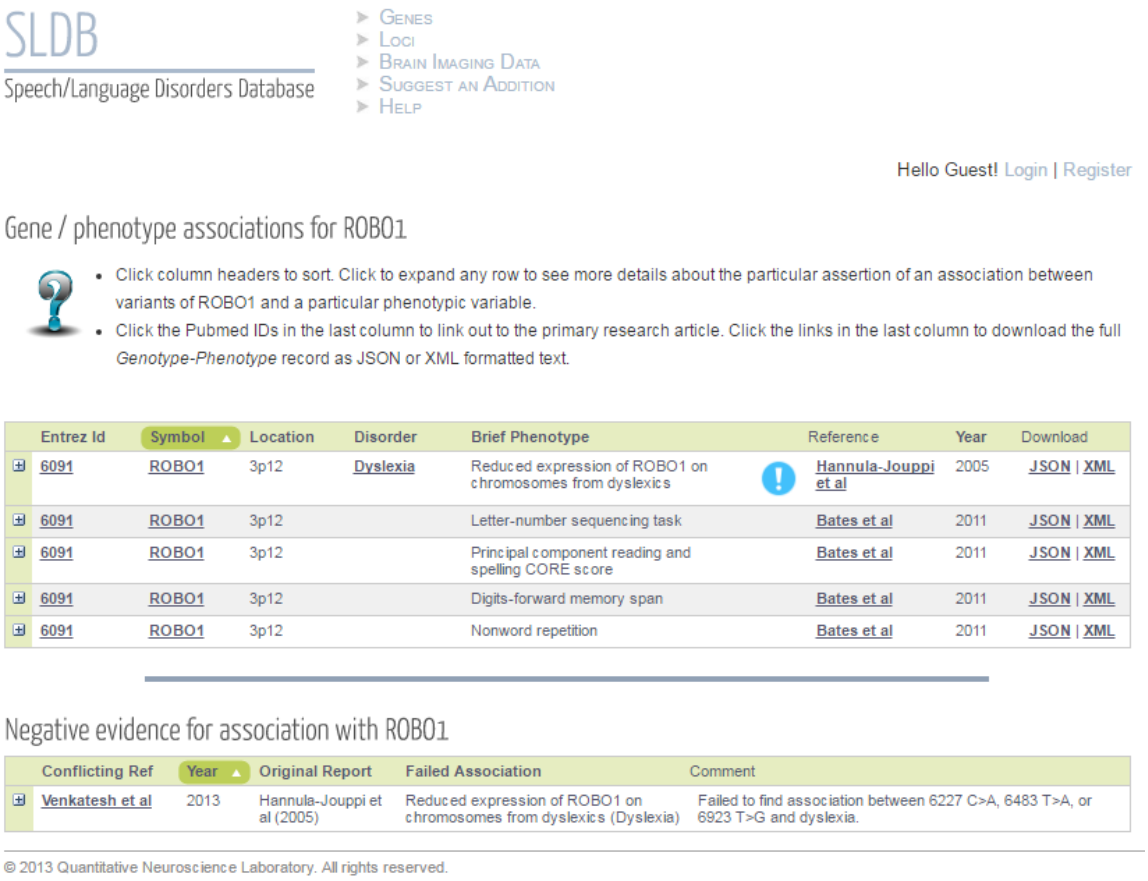


Fig 5.2. Screen capture of a (partial) view into the database detailing reports for the gene *ROBO1*. This view provides two sortable tables, the first of which shows positive evidence for association with any speech or language phenotype, and the second of which shows negative results. The study at bottom failed to replicate the original association of SNPs in *ROBO1* to susceptibility for dyslexia, and the corresponding record in the table at top is flagged with a blue icon. Users can expand the rows in either table to provide more detailed information about study methods. Finally, users can download each record as a JSON or XML structured text file.

5.2.2 Inclusion criteria

Genes were selected for entry into the database by a process intended to minimize potentially subjective interpretation of results from the literature. A series of search terms form the basis of RSS (Really Simple Syndication) feeds, current versions of which are available through the website (see http://neurospeech.org/sldb/help). These search terms take the following forms:

1. *gene <phenotype>*
2. *genetics <phenotype>*
3. *linkage <phenotype>*
4. *SNP <phenotype>*

For example, searches include *gene dyslexia* and *linkage stuttering*. Phenotypes with multi-word names are encased in quotations (e.g., *gene "specific language impairment"*). Initially, an additional search expression beginning with the term *locus* was used; this was dropped because the relevant results were a subset of those turned up by the term *linkage*. Strings currently used as phenotype names include: stuttering, dyslexia, reading disability, "verbal dyspraxia", "specific language impairment", "SLI", "childhood apraxia of speech", "language delay", and "fluency."

In addition, once a gene of interest is established in the literature, follow-up studies can be found by explicitly searching for the gene symbol[1] appearing concurrently with certain search terms. Therefore, an additional search expression, and corresponding RSS feed, takes the form:

(<gene1> OR <gene2> OR ... <geneN>) AND (speech OR language OR <phenotype1> OR <phenotype2> ... OR <phenotypeN>)

[1] Note that Pubmed will automatically explode search terms with synonymous gene symbols as well as synonyms from Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS) ontologies.

where <gene> terms are replaced by official gene symbols for genes in our database, and <phenotype> terms are the same as noted above.

These searches return hundreds of results, many of which do not fulfill the above criteria. Pubmed's Really Simple Syndication (RSS) feeds, coupled with an RSS reader, however, provide an efficient method for sifting through new literature as it appears, and for flagging new studies for inclusion. As new studies are located, they are added to an electronic queue for inclusion in the database. (Some genes included in this chapter have not yet been entered into the database.)

5.2.3 Concluding remarks

The approach used in this chapter, like any work attempting to bridge the molecular and neuroanatomical levels, relies in part upon careful and thorough curation of the literature to establish the best candidate gene list according to current knowledge, and the ability to refine and expand that list to reflect continuing research. The database described above facilitates this curation, brings information about genotype-phenotype relations specific to speech and language disorders together with results from neuroimaging research, and provides the ability to link to and / or download spatial gene expression data. Thus, these efforts represent a first step toward bringing molecular level information into cognitive and computational theories of speech and language function.

5.3 Preferential expression of genes implicated in speech and language disorders

This section focuses on the expression patterns of individual speech / language candidate genes across the brain, and particularly on where those genes show unusually

high expression levels, or *preferential expression*. Intuitively, preferential expression (as used here) means that in a given brain area, the gene's expression level "stands out" (relative to its expression elsewhere) *more* than other genes' expression levels stand out (relative to their own expression elsewhere). In other words, we are testing for areas where the deviation of a gene's expression from its brain-wide average (or average across some other "parent" structure) is large compared to other genes.

Importantly, strong expression does not necessarily imply preferential expression. A gene showing uniformly strong expression throughout the brain is not considered preferentially expressed in a given brain area. In fact, because of differences in probe efficacy, we cannot distinguish between a gene with uniformly strong expression or uniformly weak expression throughout the brain in these data. Conversely, preferential expression does not imply strong expression. A gene may show only slightly higher expression levels in an area than it does elsewhere, but if few other genes show even that much of an increase, the gene may meet the criteria for preferential expression (see Methods). In practice, higher expression levels in a particular area frequently co-occurs with preferential expression.

5.3.1 Methods

Expression values were standardized within-probe across all left-hemisphere samples using conventional z-scoring (i.e., not the weighted z-scoring described in Chapter 2). For each of the ~32,000 genes in the AHBA, probes were z-scored across all left-hemisphere samples and a mean expression value was calculated for each of 10 broad regions by averaging across samples. For each of 88 sub-regions, the un-standardized

probe data was z-scored across left-hemisphere samples *within the "parent" region only*, and values for those samples were then averaged to calculate the sub-region mean expression profile. The above steps were performed within each donor brain (though not all donors had left-hemisphere samples available for all 88 fine regions).

Each speech / language (SL) candidate gene's mean expression value for a region was converted to a percentile rank in the distribution of mean values for all genes in that region. This was also done separately for each donor. Preferential expression was defined as a percentile rank of 95 or greater in all donors with samples available for the brain region. This is equivalent to an uncorrected p-value under 0.05. The results shown here do not survive Bonferroni correction for multiple comparisons (for N = 42 genes x 10 regions, or N = 42 genes x 88 regions).

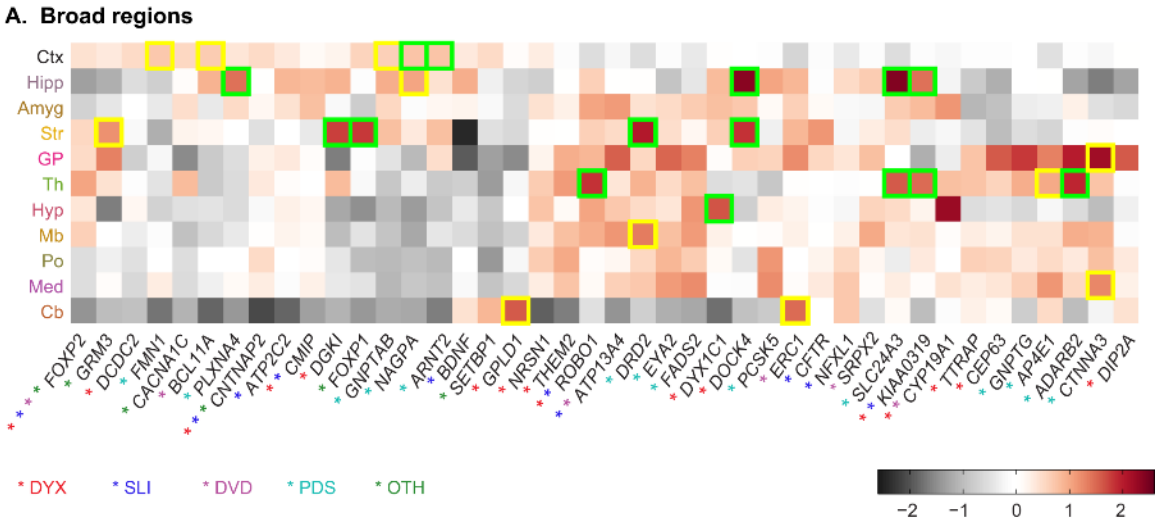
Of the full set of ~32,000 genes, those that were preferentially expressed in at least one of the 11 broad area were identified, and the hypergeometric test was used to assess whether the SL genes were overrepresented in this group. The same test was then performed using the 88 fine areas.

5.3.2 Results

Of the 42 SL genes, 12 were among those showing preferential expression in at least one broad region (3,041 genes total, $p < 1 \times 10^4$, hypergeometric test) and 16 in at least one fine region (3,498 genes total, $p < 1 \times 10^6$).

Mean expression values of the SL genes appear in Fig 5.3. In addition to the green outlines which flag preferentially expressed genes (those with a percentile rank of at least 95 in all donors), yellow outlines indicate a percentile rank of at least 90 in all

donors. Note also that although preferential expression is based on a minimum value across donors, values indicated in the heatmaps are *mean* expression values across all donors.



Ctx = Cerebral cortex; Hipp = Hippocampal formation; Amyg = Amygdala; Str = Striatum; GP = Globus pallidus; Th = Thalamus; Hyp = Hypothalamus; Mb = Midbrain; Po = Pons; Med = Medulla; Cb = Cerebellum.

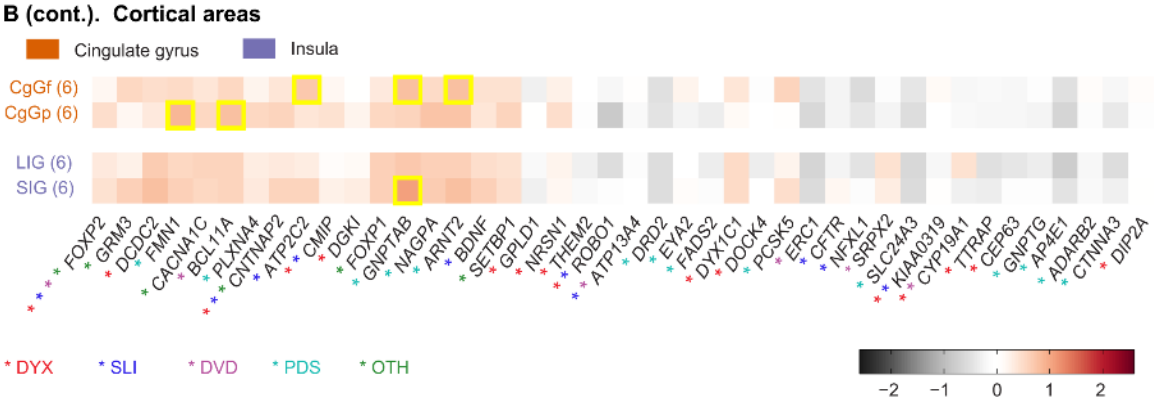
Fig 5.3 (cont. on next page). Expression heatmaps of speech / language candidate genes. Values are averaged across donors. Genes are ordered by clusters obtained by hierarchical clustering of values across broad brain regions. Asterisks next to each gene symbol indicate phenotypes the gene has been associated with. Colored outlines indicate expression at or above the 95th percentile (green boxes) or 90th percentile (yellow boxes) for that region within each donor. Parenthetical values following region names indicate the number of donors with left-hemisphere samples available.

Few genes were preferentially expressed in more than one broad region, with the exceptions of DOCK4 (HF, STR), SLC24A3 (HF, TH) and KIAA0319 (HF, TH) (Fig 5.3). SL genes that showed preferential expression tended to do so in the hippocampal formation, striatum, or thalamus (4 genes in each case); also note that cerebral cortex ranked five SL genes at the 90th percentile or above for all donors. Although *FOXP2* was not preferentially expressed in any broad region, its highest percentile ranks were in

the cerebral cortex, striatum, and thalamus (minimum across donors of 80th, 81st, and 87th percentile, respectively).

SFG = superior frontal gyrus; MFG = middle frontal gyrus; IFG = inferior frontal gyrus; SRoG = superior rostral gyrus; IRoG = inferior rostral gyrus; LOrG = lateral orbital gyrus; MOrG = medial orbital gyrus; PCLa = paracentral lobule, anterior part; PaOG = parolfactory gyri; FP = frontal pole; fro = frontal operculum; GRe = gyrus rectus; PrG = precentral gyrus; SCG = subcallosal gyrus; PoG = postcentral gyrus; IPL = inferior parietal lobule; SPL = supraparietal lobule; PCu = precuneus; FuG = fusiform gyrus; HG = Heschl's gyrus; STG = superior temporal gyrus; MTG = middle temporal gyrus; ITG = inferior temporal gyrus; PLP = planum polare; PLT = planum temporale; TP = temporal pole; TG = transverse gyri; Cun = cuneus; SOG = superior occipital gyrus; IOG = inferior occipital gyrus; LiG = lingual gyrus; OTG = occipito-temporal gyrus.

Fig 5.3 (cont. on next page). Expression heatmaps of speech / language candidate genes. Values are averaged across donors. Genes are ordered by clusters obtained by hierarchical clustering of values across broad brain regions. Asterisks next to each gene symbol indicate phenotypes the gene has been associated with. Colored outlines indicate expression at or above the 95th percentile (green boxes) or 90th percentile (yellow boxes) for that region within each donor. Parenthetical values following region names indicate the number of donors with left-hemisphere samples available.

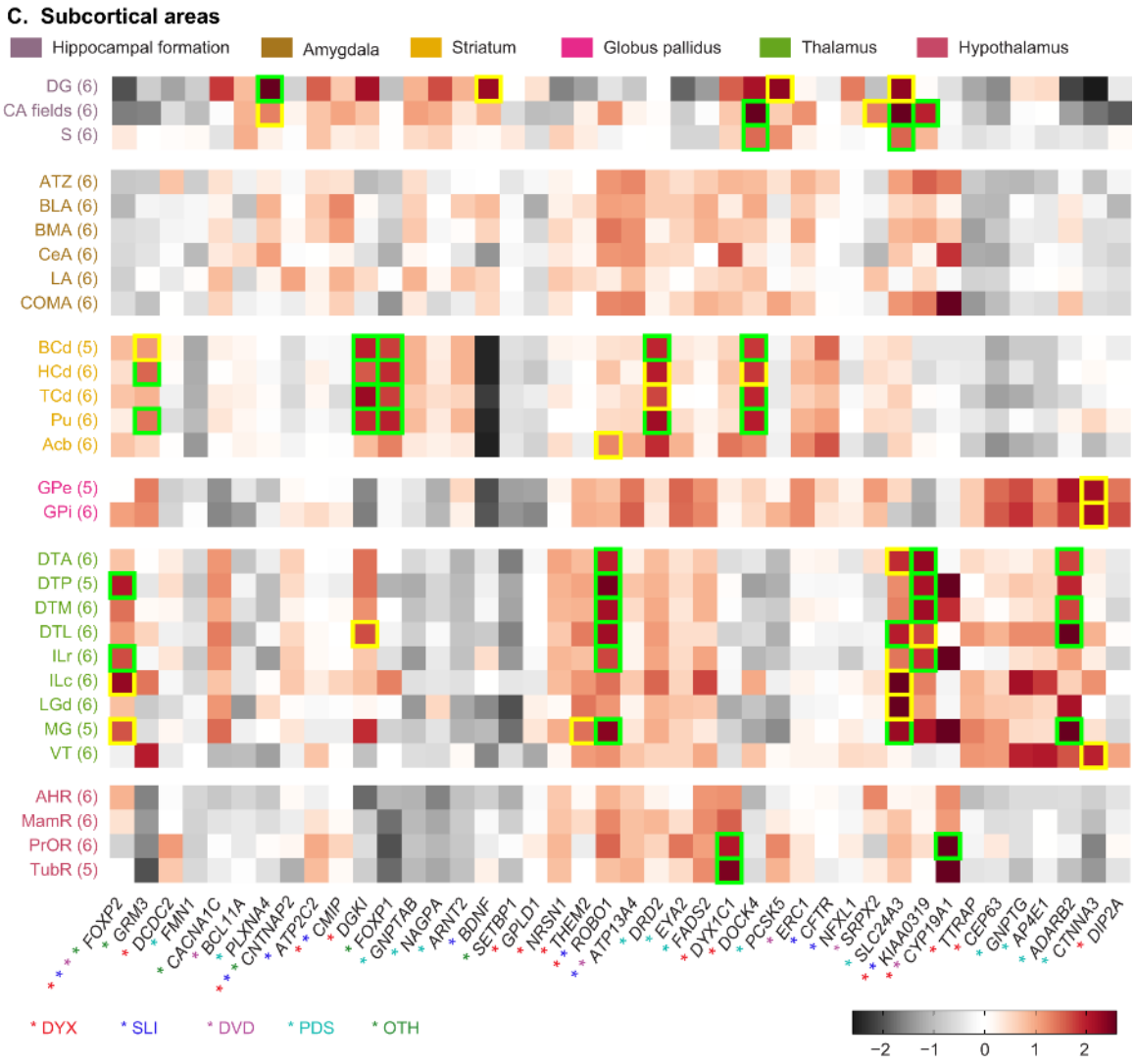


CgGf = Cingulate gyrus, frontal part; CgGp = Cingulate gyrus, parietal part; LIG = Long Insular Gyri; SIG = Short Insular Gyri.

Fig 5.3 (cont. on next page). Expression heatmaps of speech / language candidate genes. Values are averaged across donors. Genes are ordered by clusters obtained by hierarchical clustering of values across broad brain regions. Asterisks next to each gene symbol indicate phenotypes the gene has been associated with. Colored outlines indicate expression at or above the 95th percentile (green boxes) or 90th percentile (yellow boxes) for that region within each donor. Parenthetical values following region names indicate the number of donors with left-hemisphere samples available.

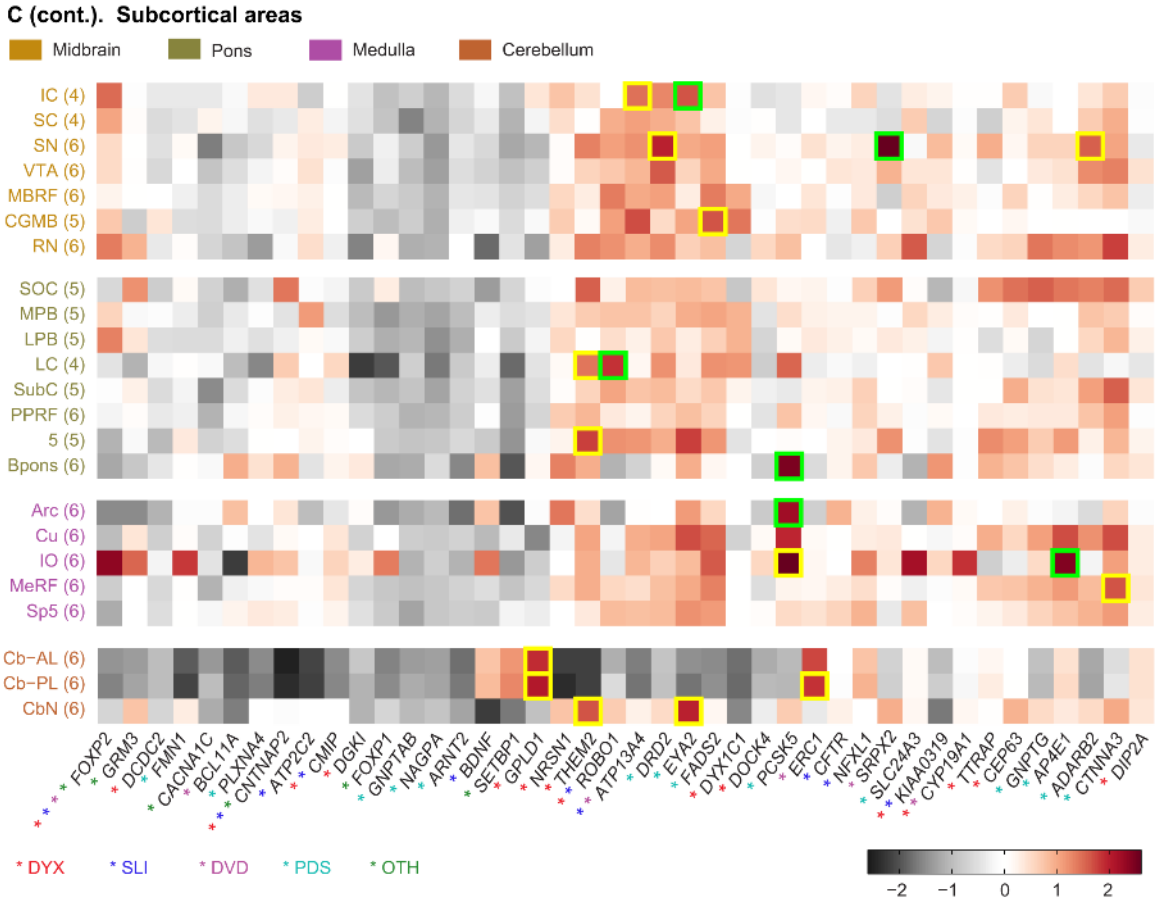
These expression patterns did not group by associated disorders (Fig 5.3). However, most PDS candidates did have higher expression values (and sometimes preferential expression) in either the telencephalon (*GNPTAB*, *NAGPA*, *ARNT2*, *PLXNA4*), the amygdala, striatum, diencephalon, and brainstem (*DRD2*, *EYA2*, *FADS2*), or the globus pallidus, thalamus, and brainstem (*GNPTG*, *AP4E1*, *ADARB2*, *CTNNA3*) than in other brain areas.

A set of 93 finer brain regions revealed a few more genes with preferential expression in areas throughout the brain (though still none in any amygdalar nuclei). Most SL genes showed relatively modest variation in either expression percentile rank within any given telencephalic or diencephalic broad region, with the exceptions of differences between hippocampal sub-structures, lower values in the nucleus accumbens than the rest of the striatum, and some distinction between the dorsal and ventral thalamus.



DG = Dentate gyrus; CA fields = CA fields; S = Subiculum; ATZ = Amygdaloid Transition Zone; BLA = Basolateral Nucleus; BMA = Basomedial Nucleus; CeA = Central Nucleus; LA = Lateral Nucleus; COMA = Cortico-medial group; BCd = Body of the Caudate Nucleus; HCd = Head of the Caudate Nucleus; TCd = Tail of the Caudate Nucleus; Pu = Putamen; Acb = Nucleus accumbens; GPe = globus pallidus, external segment; GPi = globus pallidus, internal segment; DTA = Anterior Group of Nuclei; DTP = Posterior Group of Nuclei; DTM = Medial Group of Nuclei; DTL = Lateral Group of Nuclei; ILr = Rostral Group of Intralaminar Nuclei; ILc = Caudal Group of intralaminar Nuclei; LGd = Dorsal Lateral Geneiculate Nucleus; MG = Medial Geniculate Complex; VT = Ventral Thalamus, Left; AHR = Anterior Hypothalamic Region; MamR = Mammillary Region; PrOR = Preoptic Region; TubR = Tuberal Region.

Fig 5.3 (cont. on next page). Expression heatmaps of speech / language candidate genes. Values are averaged across donors. Genes are ordered by clusters obtained by hierarchical clustering of values across broad brain regions. Asterisks next to each gene symbol indicate phenotypes the gene has been associated with. Colored outlines indicate expression at or above the 95th percentile (green boxes) or 90th percentile (yellow boxes) for that region within each donor. Parenthetical values following region names indicate the number of donors with left-hemisphere samples available.



IC = Inferior Colliculus; SC = Superior Colliculus; SN = Substantia Nigra; VTA = ventral tegmental area; MBRF = Midbrain Reticular Formation; CGMB = central grey substance of midbrain; RN = Red Nucleus; SOC = Superior Olivary Complex; MPB = medial parabrachial nucleus; LPB = lateral parabrachial nucleus; LC = locus ceruleus; SubC = nucleus subceruleus; PPRF = Paramedian Pontine Reticular Formation; 5 = Trigeminal Nuclei; Bpons = Basal part of pons; Arc = arcuate nucleus of medulla; Cu = cuneate nucleus; IO = inferior olivary complex; MeRF = medullary reticular formation; Sp5 = spinal trigeminal nucleus; Cb-AL = Anterior Lobe; Cb-PL = Posterior Lobe; CbN = Cerebellar nuclei.

Fig 5.3. Expression heatmaps of speech / language candidate genes. Values are averaged across donors. Genes are ordered by clusters obtained by hierarchical clustering of values across broad brain regions. Asterisks next to each gene symbol indicate phenotypes the gene has been associated with. Colored outlines indicate expression at or above the 95th percentile (green boxes) or 90th percentile (yellow boxes) for that region within each donor. Parenthetical values following region names indicate the number of donors with left-hemisphere samples available.

In contrast, several genes had high expression values in only one or two nuclei of the midbrain, pons, and / or medulla (e.g. *FOXP2*, *BCL11A, DYX1C1*, *PCSK5*, and *SLC24A3*). However, few of these met the criteria for preferential expression. In the

midbrain, three genes were preferentially expressed (*SRPX2*) or had minimum ranks of at least 90 in the substantia nigra (*DRD2*, *ADARB2*). In the medulla, the inferior olivary complex ranks *PCSK5* and *AP4E1* at or above the 90th and 95th percentile for all donors, respectively. Several other genes show stronger expression there than elsewhere in medulla or, for the most part, the rest of the brainstem (e.g. *FOXP2*, *FMN1*). This is maintained when expression values are converted to percentile ranks: five genes have higher minimum percentile ranks (across donors) in the inferior olivary complex than in any other brainstem area (*FOXP2*, *GRM3, FMN1*, *FOXP1, AP4E1*).

In the cerebellum, most genes showed higher expression in either the cerebellar cortex or cerebellar nuclei. In particular, *GPLD1* and *ERC1* are preferentially expressed in the cerebellar cortex, and *THEM2* and *EYA2* in the cerebellar nuclei.

5.4 Co-expression modularity

This section examines potential enrichment of co-expression for a given gene set within specific brain regions. Enriched co-expression may point to particular brain regions through which candidate genes could influence behavioral phenotypes, and ultimately may suggest mechanisms for that influence. Here, a "modularity score" is used to measure the extent to which a group of genes is co-expressed across samples within a given region (i) relative to other genes within the region, and (ii) relative to their own co-expression across randomly selected samples.

5.4.1 Methods

AHBA gene expression profiles were converted to z-scores across samples. Gene-gene co-expression networks were then generated using the PCC computed across expression profiles for pairs of genes. All PCC matrices were of size 32,536 x 32,536 (the number of unique genes in the dataset). Each co-expression network used samples from a particular brain region *and its sub-regions*. To account for tissue sampling biases (i.e., an overrepresentation of samples from one brain region), the PCC approach was adjusted to allow weighting of individual samples given a partition set S containing K regions, R_1, R_2,. . .R_K, where R_j contains the integer indices of all samples in that brain region. The resulting equation for weighted correlation between gene m and gene n is given by:

$$r_{mn}^{\mathrm{S}} = \frac{\sum_{k=1}^{K}\left(\sum_{i\in R_k}\frac{1}{|R_k|}\left(E^{\mathrm{AHBA}}(m,i)-\mu^{S}(m)\right)\left(E^{\mathrm{AHBA}}(n,i)-\mu^{S}(n)\right)\right)/K}{\sigma^{S}(m)\sigma^{S}(n)} \qquad (1)$$

For each AHBA region analyzed, the partition function S was set to represent the set of its child regions in the reference hierarchy. For example, the children of the region *Cerebral Cortex* in the hierarchy are *Frontal Lobe, Insula, Limbic Lobe, Occipital Lobe, Parietal Lobe,* and *Temporal Lobe*, each of which receives equal weight (despite non-uniform sampling) in computing the correlation coefficient for the cortical co-expression network. In order to obtain reasonable correlation estimates, only regions with at least 30 samples available were included.

Enrichment of co-expression within a brain region was characterized by comparing the modularity of a provided gene set to values expected by chance in each

AHBA region-specific co-expression network. The modularity score for a given gene set *G* was defined as:

$$M_G = \frac{1}{|G|(|G|-1)} \sum_{i \in G} \sum_{j \in G, j \neq i} \text{abs}\left(F(r_{ij})\right) - \frac{1}{|G|(|\bar{G}|)} \sum_{i \in G} \sum_{j \in \bar{G}} \text{abs}\left(F(r_{ij})\right) \quad (2)$$

where $F(\ )$ denotes the Fisher *r* to *z* transformation, and $\bar{G}$ is the complement of gene set *G* (i.e., the genes in the dataset that are *not* in set G). The value M_G then indicates the difference between the average co-expression score between pairs of genes in G and the average co-expression score between pairs of genes where one gene is in G and one gene is not in G.

Two methods were used to assess the extent to which a given gene set had high modularity in a given brain region (and its subregions), both using randomization approaches. In the first method (gene permutation), a series of 1000 randomly selected gene sets of the same cardinality as the gene set of interest are used to generate a distribution of expected modularity scores, for each region-specific network. Then, the modularity score for set G is standardized by subtracting the mean and dividing by the standard deviation of scores obtained in these random draws. Thus the final, region-specific standardized modularity score reflects how modular a given gene set is in a given brain area relative to other gene sets in units of standard deviation (i.e., a score of 5 is five standard deviations higher than average). P-values are also obtained by calculating the percentile rank of M_G in the relevant empirical chance distribution (i.e., without assuming chance distributions were Gaussian). In results presented here, the p-values are Bonferroni corrected for multiple comparisons across brain regions (N = 60).

A second randomization procedure (sample permutation) compared the modularity of a gene set in a given brain region to the modularity of the same gene set in randomly selected samples. Specifically, a null distribution was generated by recomputing M_G across randomly selected groups of left-hemisphere samples from the AHBA, with the number of samples matched to the number of samples available in the brain region of interest. It should be noted that the sample permutation method is more computationally intensive than the gene permutation approach above because it requires recalculating a weighted correlation network for every permutation using a new subset of available samples. As above, a set of 1000 random selections of samples was used to calculate each empirical null distribution. Based on these chance distributions, standardized modularity scores and Bonferroni-corrected p-values were calculated as described above. Scores based on the sample permutation approach describe how modular the gene set is in samples from a specific brain area – i.e., the anatomical specificity for this gene set – whereas scores based on gene permutations reflect how modular the gene set is in a brain region relative to other gene sets. These complementary pieces of information are presented separately for each gene set of interest.

5.4.3 Results

Fig 5.4 shows co-expression modularity based on the gene permutation approach (left column) and the sample permutation approach (right column) of the five groups of genes defined by associated disorders. The first row corresponds to the complete set of 42 SL genes. These genes showed higher co-expression modularity than average for

random genes (i.e., positive modularity z-scores) in all 60 regions. The genes' modularity was significant ($p < 0.05$ after correction for multiple comparisons) in the cerebral cortex, as well as specifically in the frontal, temporal and limbic lobes. Within those lobes, the inferior frontal, inferior temporal, fusiform, frontal cingulate, and parahippocampal gyri also showed enhanced modularity for these genes. Other regions showing significant modularity included the diencephalon, pons, and medulla.

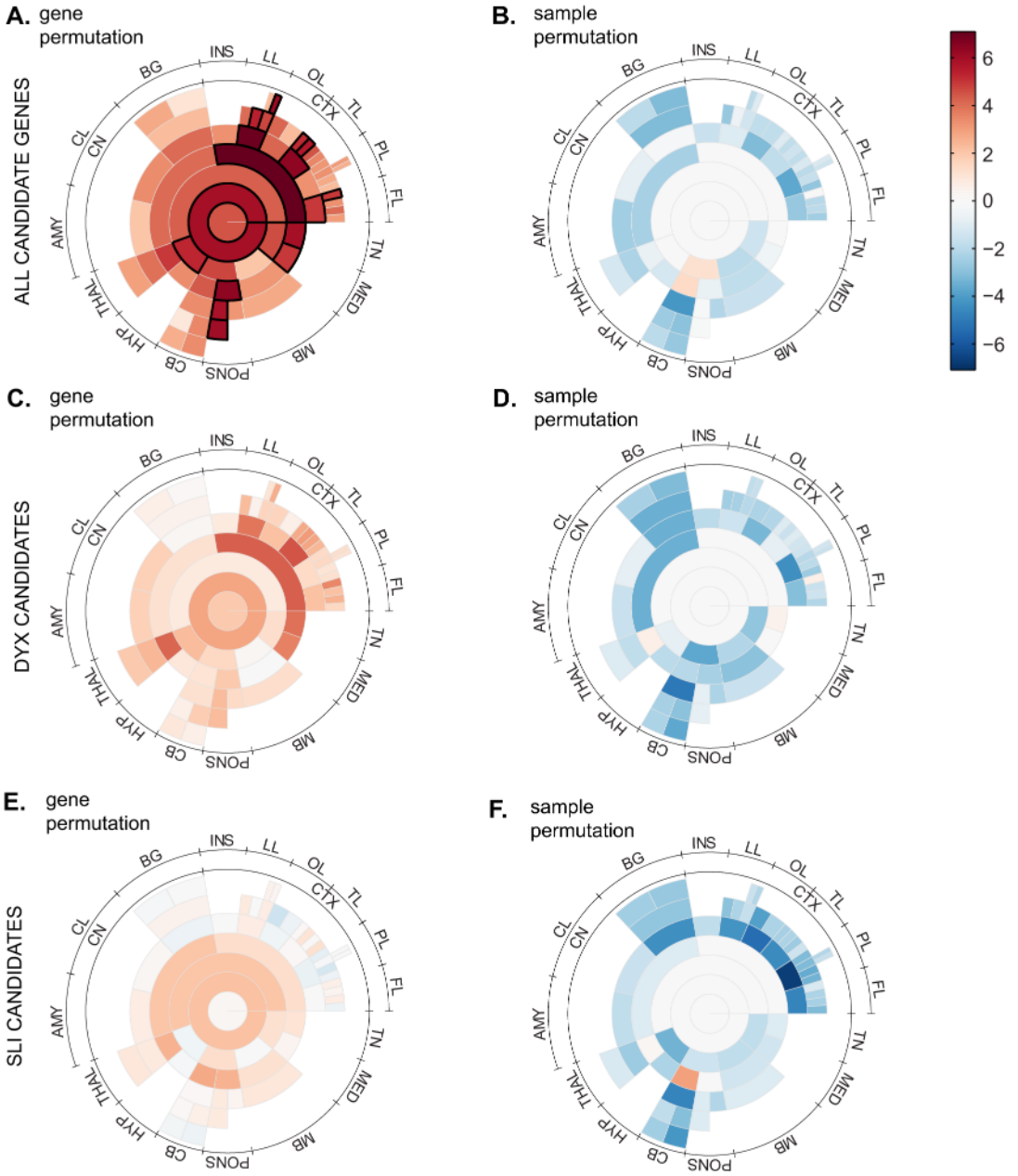


Fig 5.4 (cont. on next page). Enrichment of speech / language gene candidates in region-specific AHBA networks. Each row represents regional enrichment of a gene set based on co-expression modularity relative to other gene sets (gene permutation, left column) and relative to random samples (sample permutation, right column). Rows illustrate results for the full set of 42 candidates (A, B), and for subsets consisting of genes implicated in dyslexia (C, D), and specific language impairment (E, F). See panel K for a legend indicating the brain regions corresponding to each individual wedge. Bold outlined wedges indicate significant results ($p < 0.05$, Bonferroni-corrected). For sample permutations, values for the brain and grey matter are set to 0 (since it is impossible to select random samples outside of those structures), as well as the telencephalon and cerebral cortex (because slightly fewer than half the samples are outside these structures).

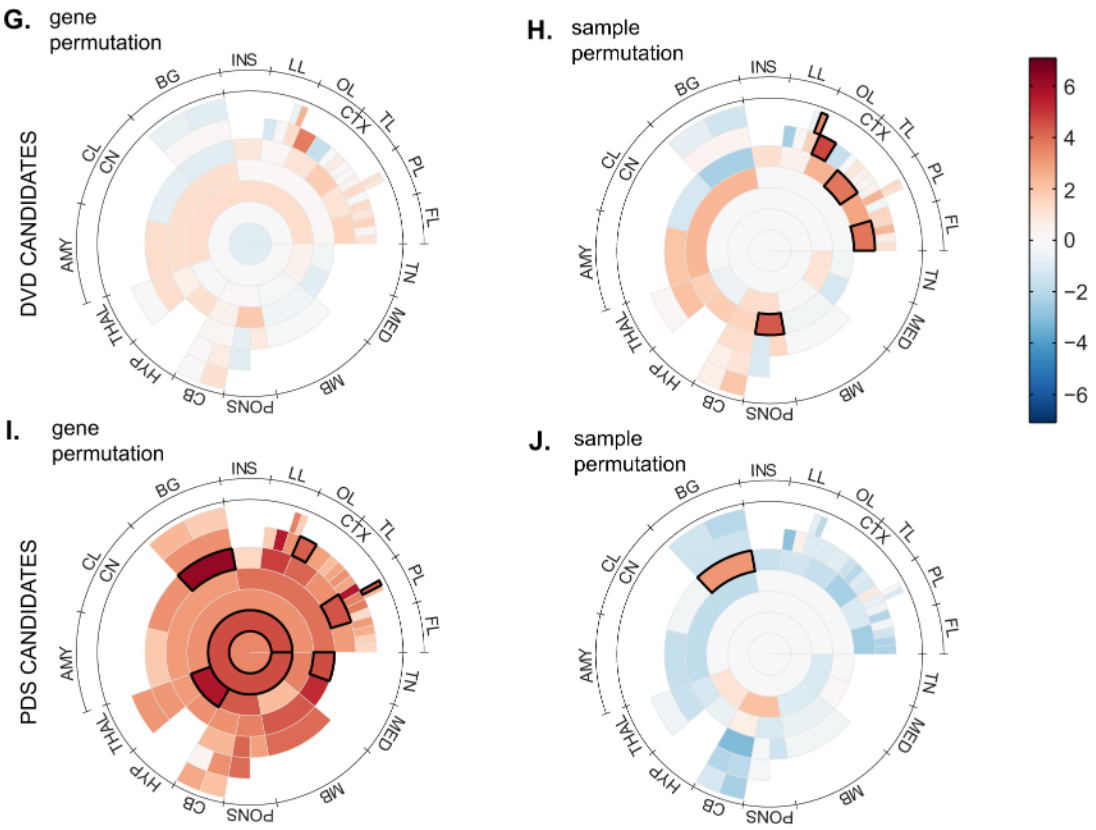


Fig 5.4 (legend on next page). Enrichment of speech / language gene candidates in region-specific AHBA networks. Each row represents regional enrichment of a gene set based on co-expression modularity relative to other gene sets (gene permutation, left column) and relative to random samples (sample permutation, right column). Rows illustrate results for developmental verbal dyspraxia (G, H), and persistent developmental stuttering (I, J). See panel K for a legend indicating the brain regions corresponding to each individual wedge. Bold outlined wedges indicate significant results ($p < 0.05$, Bonferroni-corrected). For sample permutations, values for the brain and grey matter are set to 0 (since it is impossible to select random samples outside of those structures), as well as the telencephalon and cerebral cortex (because slightly fewer than half the samples are outside these structures).

K.

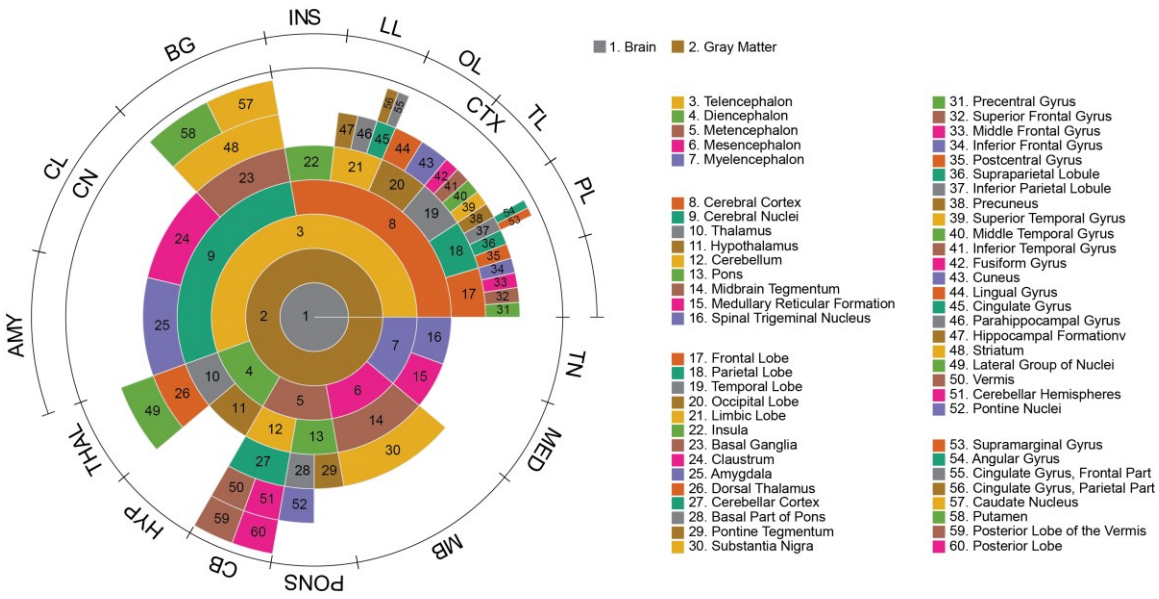


Fig 5.4. Enrichment of speech / language gene candidates in region-specific AHBA networks. Panel K is a legend indicating the brain regions corresponding to each individual wedge in panels A-J.

Neither dyslexia, SLI nor DVD candidate genes (Fig 5.4, second through fourth rows) reached significant modularity in any brain region. However, the dyslexia candidate group showed largely positive values that were highest (over 3) in the temporal lobe and inferior temporal gyrus, as well as the limbic lobe, thalamus and both sub-structures of the medulla. PDS candidate genes (Fig 5.4, fifth row) showed enhanced modularity in the basal ganglia and diencephalon. The parietal lobe, and specifically the angular gyrus, also reached significance, as well as the lingual gyrus and spinal trigeminal nuclei.

The SL candidates as a group, dyslexia candidates, and SLI candidates showed similar or slightly lower modularity in most brain regions than across randomly selected samples, though the full SL gene list and the SLI candidates both yielded modest positive

values in the cerebellum. In contrast, the DVD candidates showed enhanced modularity relative to random samples in the frontal and temporal lobes, lingual and frontal cingulate gyri, and the pons, and the PDS candidates in the basal ganglia.

5.5 Co-expression landscape

The following analysis uses the similarity / dissimilarity between expression profiles of the SL genes to visually analyze their overall *co-expression landscape*, and to examine whether genes implicated in certain subclasses of these disorders cluster in this co-expression space. In the two-dimensional landscape, calculated using non-metric multi-dimensional scaling (MDS), distances between genes are based on their co-expression. Genes with either positively or negatively correlated expression values across samples (either of suggests a regulatory relationship) are represented as nearby locations, while genes whose expression profiles have no clear relationship are represented as distant locations. This technique does not create a perfect representation of the distance relationships, but does provide a useful visualization for exploration of the anatomical expression patterns of genes implicated in related phenotypes. In this way, the co-expression landscape of candidate genes is examined here both within the brain as a whole and separately within the cerebral cortex, basal ganglia, and cerebellum.

5.5.1 Methods

Pearson's Product-moment Correlation Coefficients (PCCs) were calculated between each pair of SL genes, based on their expression profiles across an anatomically relevant set of samples. Correlation distance between two genes was then defined as one

minus the absolute value of this PCC. Using the magnitude of the PCCs allows the distance measure to treat inverse relationships, such as one gene down-regulating another, as constituting "similar" expression.

Distance matrices were calculated by pooling samples from all donors for each of four sample subsets: (i) the entire left hemisphere, (ii) left cerebral cortex only, (iii) left basal ganglia only and (iv) left cerebellum only. In each case, probes were z-scored across the samples in the subset (using conventional, i.e. un-weighted z-scoring). Cerebral cortex was defined to include the frontal, parietal, temporal, and occipital lobes, the cingulate gyrus, and the insula. Basal ganglia included the striatum, pallidum, subthalamic nucleus, and substantia nigra.

For each disorder (dyslexia, SLI, DVD, and PDS), pair-wise distances "within-group" were compared to those "across-group". Genes within a group included all those associated with the disorder (including those also associated with one or more additional disorders). Across-group distances were calculated between all possible pairs of a gene in the group with a gene outside the group. For each group, a one-tailed, two-sample t-test was performed on the two sets of distances to test for significantly smaller within-group than across-group differences. The Bonferroni method was used to correct for multiple comparisons (i.e., 4 groups x 4 sample sets = 16 comparisons).

Donor variability was assessed by calculating within-donor distance matrices and performing a Mantel test (Mantel, 1967) for each pair of donors. In the Mantel test, the PCC between two distance matrices is compared to a distribution of PCCs between the first matrix and randomly permuted versions of the second. The proportion of the

distribution that is greater than the original PCC determines the PCC's statistical significance. This corrects for the dependence between values in a distance matrix (i.e., a change in one value of the matrix entails a change in the other values, since the gene's "location" relative to all the other genes has been altered). 10,000 random permutations per donor pair were used.

To visualize the co-expression distance relationships between genes, non-metric multi-dimensional scaling was applied to calculate the genes' coordinates in a two-dimensional space. This method solves an optimization problem such that the distances between pairs of points (genes) in the 2D embedding space approximate a monotonic transformation of the input distance matrix. The method specifically minimizes "stress", or the squared difference between the input and output distance matrices, normalized by the sum of all squared input distances.

5.5.2 Results

PDS candidate genes were the only group to show a smaller mean within-group than across-group distance in all sample subsets except the cerebellum (Fig 5.5). Within-group distances f or PDS candidates were significantly smaller than across-group distances ($p < 0.05$ after correction for multiple comparisons) only in the basal ganglia ($p = 0.0005$).

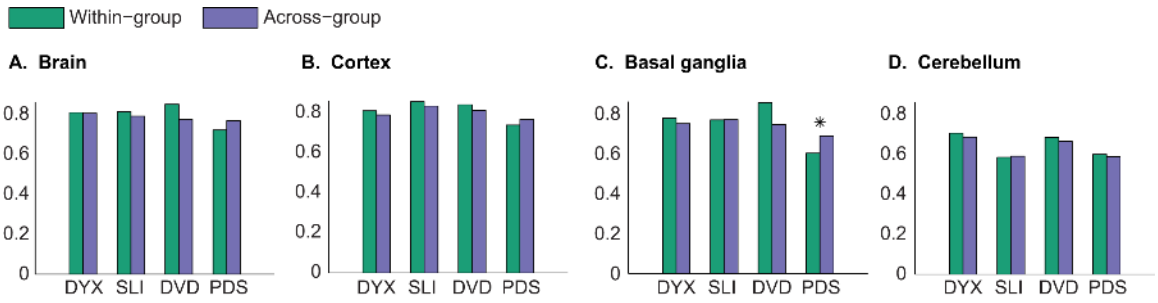


Fig 5.5. Within- and across-group correlation distances. Bar height is mean distance across pairs of genes. Asterisks indicate differences with $p < 0.05$ after multiple comparisons correction.

The mean pair-wise PCC between donor distance matrices was 0.63 for the brain, 0.43 for cerebral cortex, 0.45 for the basal ganglia, and 0.44 for the cerebellum. For each sample subset, each of the fifteen donor pairs had a higher PCC than the maximum of the empirical null distribution (i.e., $p = 0$, Mantel test).

The two-dimensional "landscapes" based on samples from all donors are shown in Fig 5.6. While there is some loose grouping of genes in the cortex and a single tight cluster in the cerebellum, for the most part genes do not appear to separate by phenotype. Genes with high proximity in the cerebellum (Fig 5.6D) include *GPLD1*, *ERC1*, *THEM2*, and *EYA2*, which all showed preferential expression for either the cerebellar cortex or cerebellar nuclei (Fig 5.3B). Other genes shown in the expanded area of Fig 5.6D (e.g., *BDNF*, *SETBP1*) had percentile ranks in a cerebellar structure that did not meet the criteria for preferential expression, but nevertheless did a preference for one or the other. Most genes outside of that densely populated area of Fig 5.6D did not show a preference (e.g., *AP4E1*, *BCL11A*).

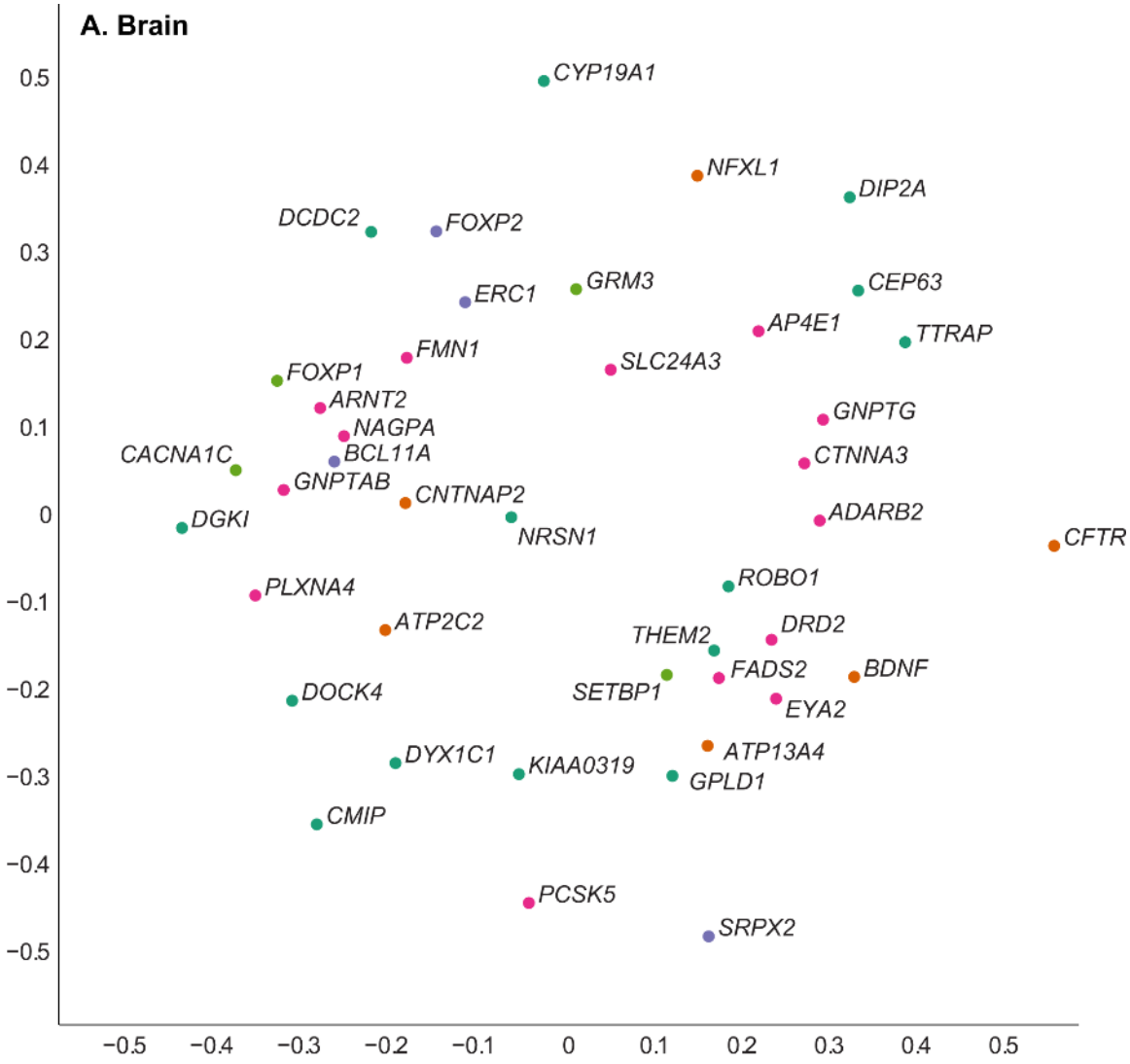


Fig 5.6 (cont. on next page). Two-dimensional representation of gene expression pattern relationships, using multi-dimensional scaling. For genes associated with multiple disorders, datapoint color is based on the best-supported or first association.

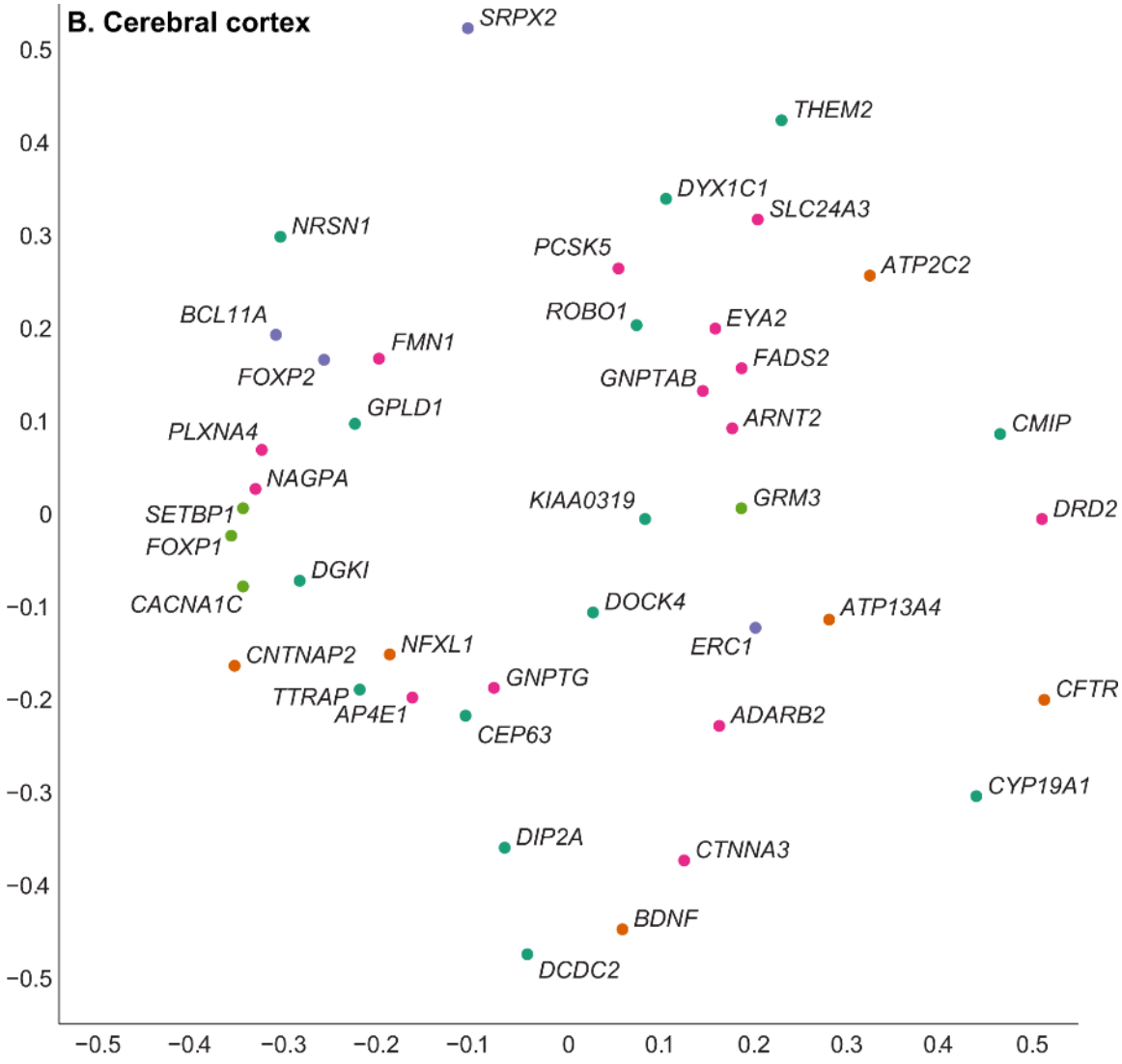


Fig 5.6 (cont. on next page). Two-dimensional representation of gene expression pattern relationships, using multi-dimensional scaling. For genes associated with multiple disorders, datapoint color is based on the best-supported or first association.

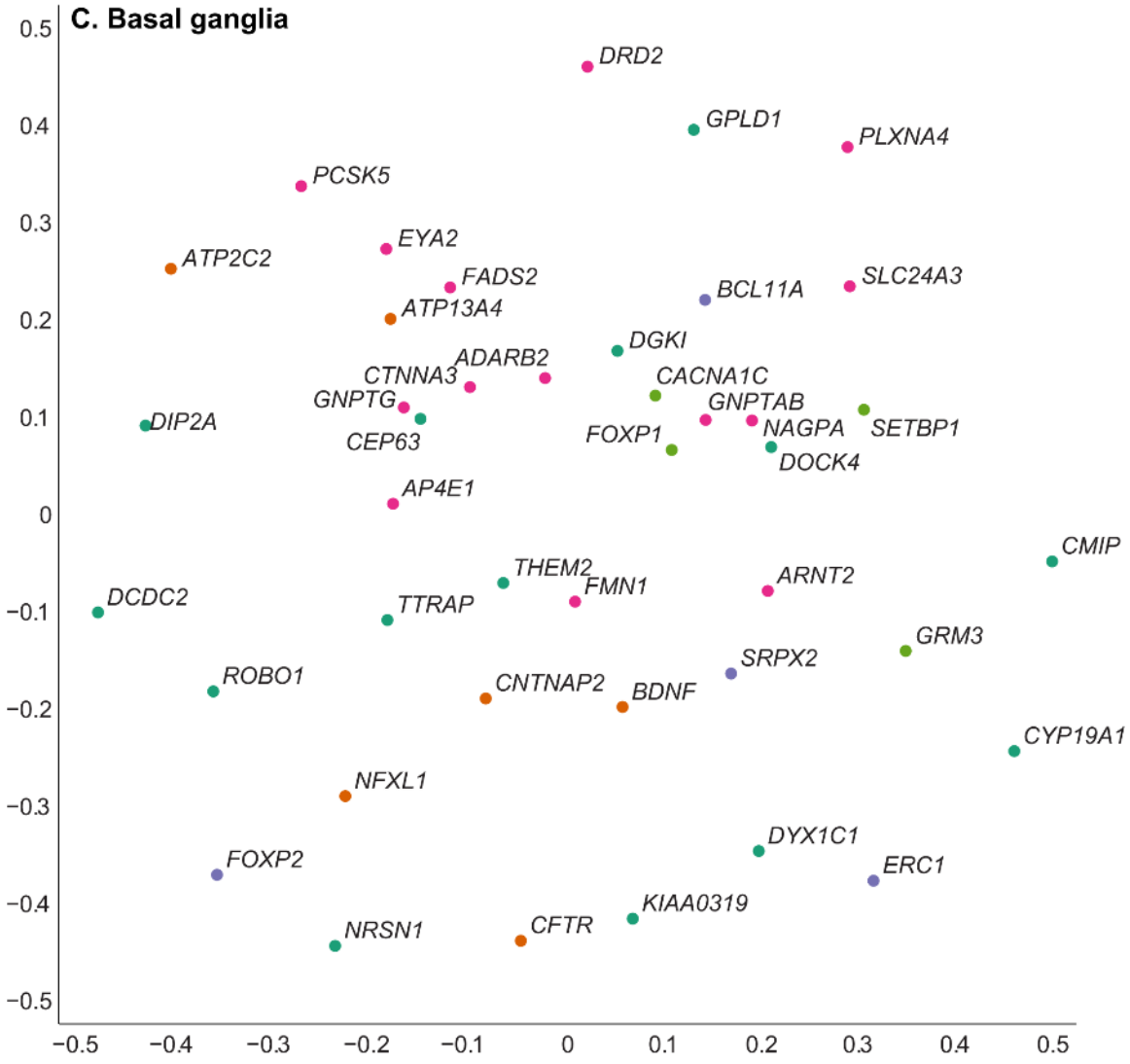


Fig 5.6 (cont. on next page). Two-dimensional representation of gene expression pattern relationships, using multi-dimensional scaling. For genes associated with multiple disorders, datapoint color is based on the best-supported or first association.

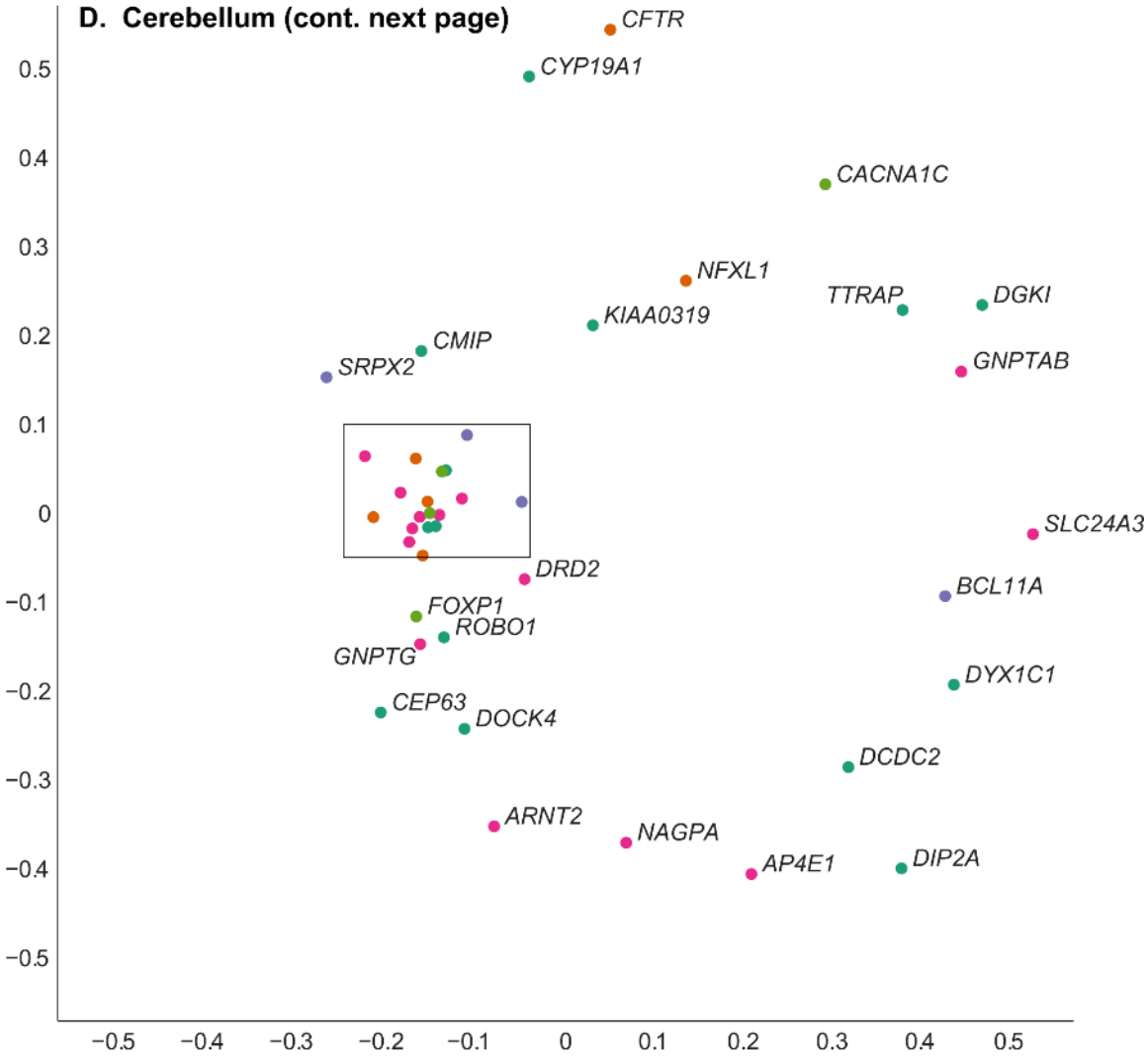


Fig 5.6 (cont. on next page). An expanded view of the boxed area is shown in the next plot. Two-dimensional representation of gene expression pattern relationships, using multi-dimensional scaling. For genes associated with multiple disorders, datapoint color is based on the best-supported or first association.

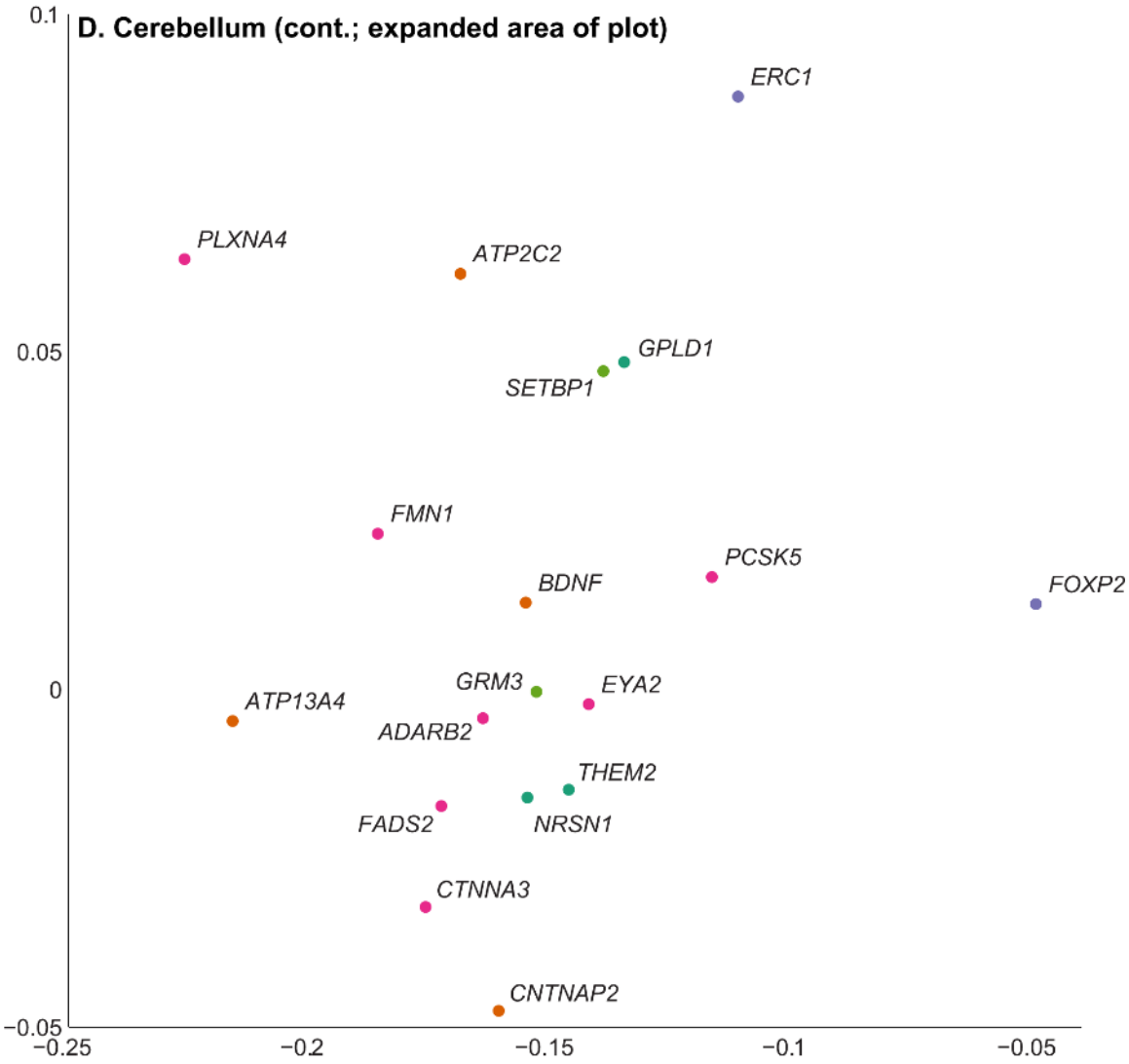


Fig 5.6. Expanded image of tightly clustered datapoints from the previous plot. Two-dimensional representation of gene expression pattern relationships, using multi-dimensional scaling. For genes associated with multiple disorders, datapoint color is based on the best-supported or first association.

5.6 Co-expression networks using topological overlap

The previous analyses deal with comparisons of gene spatial expression profiles. An alternative approach to examining relationships between genes is to instead compare patterns of co-expression with the other genes. In other words, rather than asking only if *FOXP2* and *CNTNAP2* are co-expressed, we might ask whether *FOXP2*'s co-expression

relationships to the other SL genes resemble those of *CNTNAP2*. One distance measure that is based on similarity of co-expression relationships is *topological overlap* (Zhang and Horvath, 2005; Yip and Horvath, 2007). Briefly, each gene is treated as a node in a co-expression network. Genes are then hierarchically clustered based on similarity between their patterns of co-expression, or topological overlap. The resulting dendrograms can reveal the overall structure in the relationships between gene expression patterns.

5.6.1 Methods

Weighted gene co-expression analysis (WGCNA; Zhang and Horvath, 2005) was applied to the 42 SL genes using a publicly available R package (Langfelder and Horvath, 2008). A network of genes was defined in which edge weights encoded the absolute value of the correlation between pairs of genes' spatial expression patterns. Co-expression similarity was measured by topological overlap (Zhang and Horvath, 2005; Yip and Horvath, 2007), a measure which quantifies the degree to which two nodes' neighborhoods overlap. Because of the small number of genes, resulting in a dendrogram with easily distinguishable branches, the "dynamic tree-cutting algorithm" included in the R package was not used (Langfelder et al., 2008).

WGCNA was applied separately across each of the four sample subsets used in the previous section: (i) the entire left hemisphere, (ii) left cerebral cortex only, (iii) left basal ganglia only and (iv) left cerebellum only. Samples from all donors were used. Similarity matrices (topological overlap matrices) were also calculated for each donor individually. As in the previous section, the Mantel test was applied to each pair of donor

matrices by computing the PCC between vectors composed of the upper-triangular parts of the two matrices, and comparing it to PCCs resulting from 10,000 random permutations of one of the matrices.

5.6.2 Results

The dendrograms in Fig 5.7 show small groups of genes at shorter distances than others (i.e., with more similar patterns of co-expression relationships with the other SL genes). These groups vary somewhat by structure, but for the most part do not reflect associated disorders. The cerebellum alone does not show this tendency to cluster (Fig 5.7D). Genes that appeared very close to each other in the cerebellum's co-expression landscape have similar co-expression relationships with other genes (compare genes in the expanded area of Fig 5.6D to genes at the far right of Fig 5.7D). After that, genes are merged into the dendrogram with fairly regular spacing, reflecting the circle of datapoints around the tight cluster in Fig 5.6D.

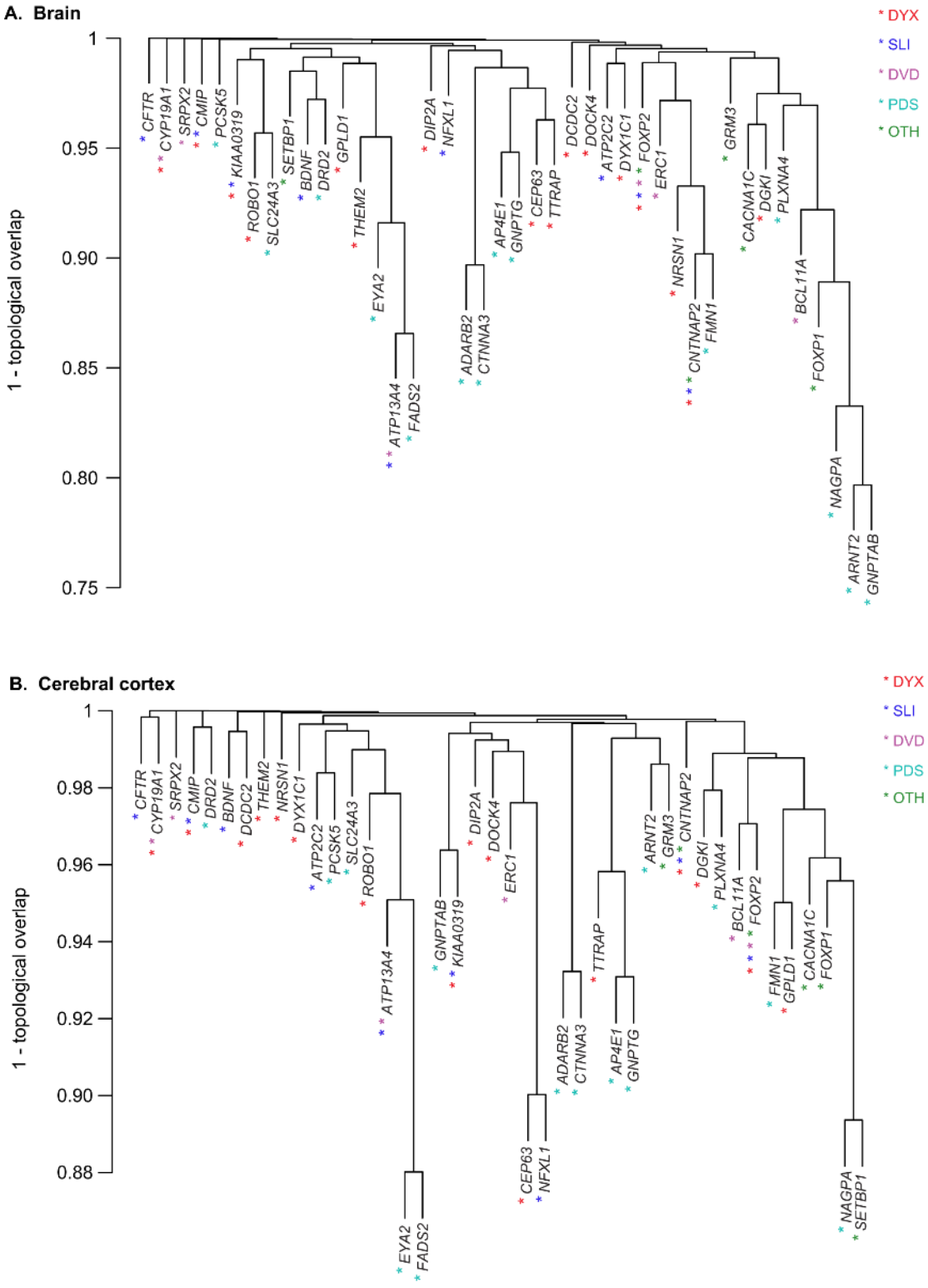


Fig 5.7 (cont. on next page). Hierarchical clustering of genes in a co-expression network. Asterisks next to each gene symbol indicate disorders the gene has been associated with.

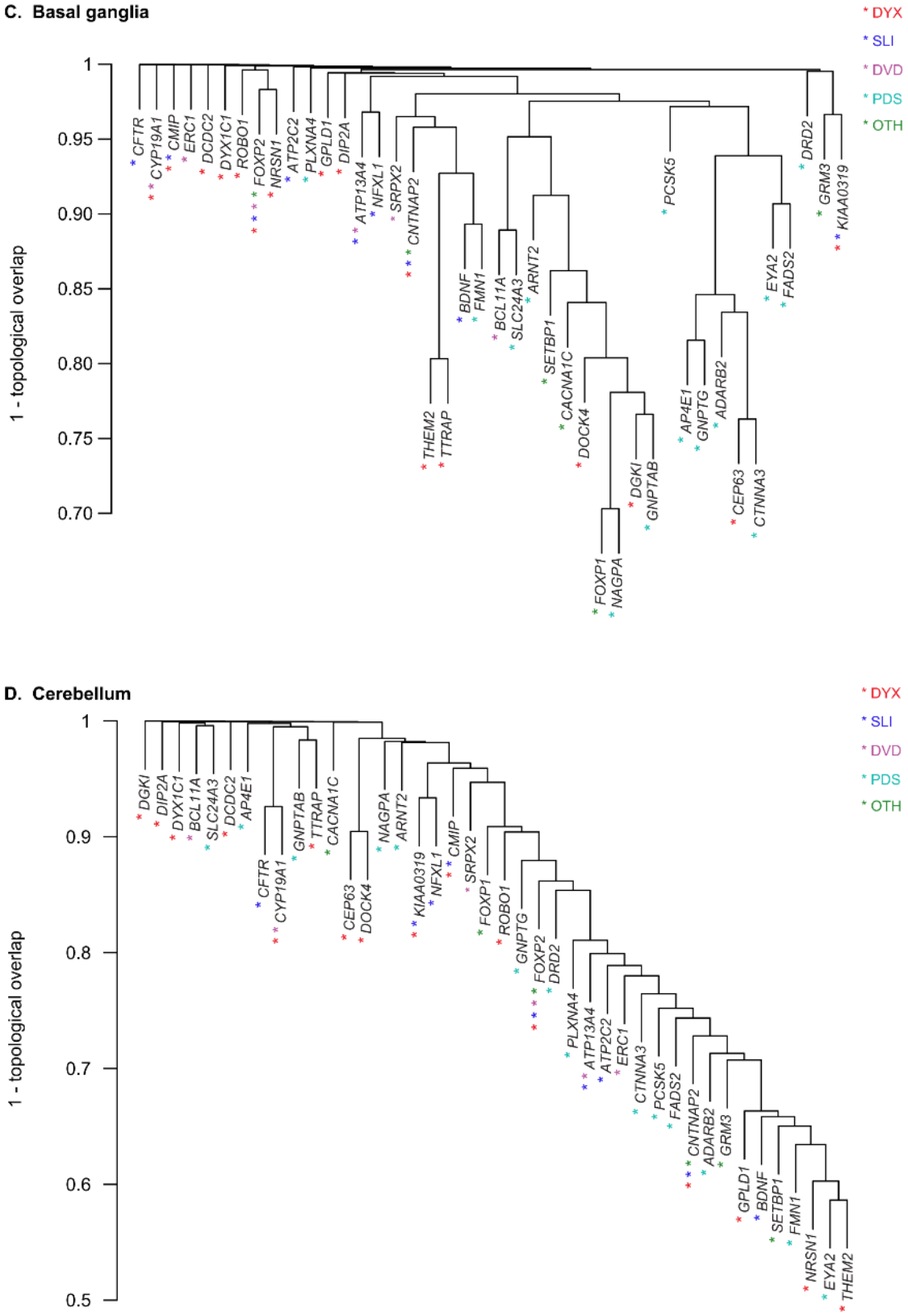

C. Basal ganglia
* DYX
* SLI
* DVD
* PDS
* OTH
1 - topological overlap
1
0.95
0.90
0.85
0.80
0.75
0.70
CFTR
CYP19A1
CMIP
ERC1
DCDC2
DYX1C1
ROBO1
FOXP2
NRSN1
ATP2C2
PLXNA4
GPLD1
DIP2A
ATP13A4
NFXL1
SRPX2
CNTNAP2
THEM2
TTRAP
BDNF
FMN1
BCL11A
SLC24A3
ARNT2
SETBP1
CACNA1C
DOCK4
DGKI
GNPTAB
FOXP1
NAGPA
PCSK5
AP4E1
GNPTG
ADARB2
CEP63
CTNNA3
EYA2
FADS2
DRD2
GRM3
KIAA0319
D. Cerebellum
* DYX
* SLI
* DVD
* PDS
* OTH
1 - topological overlap
1
0.9
0.8
0.7
0.6
0.5
DGKI
DIP2A
DYX1C1
BCL11A
SLC24A3
DCDC2
AP4E1
CFTR
CYP19A1
GNPTAB
TTRAP
CACNA1C
CEP63
DOCK4
NAGPA
ARNT2
KIAA0319
NFXL1
CMIP
SRPX2
FOXP1
ROBO1
GNPTG
FOXP2
DRD2
PLXNA4
ATP13A4
ATP2C2
ERC1
CTNNA3
PCSK5
FADS2
CNTNAP2
ADARB2
GRM3
GPLD1
BDNF
SETBP1
FMN1
NRSN1
EYA2
THEM2

Fig 5.7. Hierarchical clustering of genes in a co-expression network. Asterisks next to each gene symbol indicate disorders the gene has been associated with.

The mean pair-wise PCC between donor topological overlap matrices was 0.68 for the brain, 0.34 for cerebral cortex, 0.55 for the basal ganglia, and 0.57 for the cerebellum. Using the Mantel test, the largest p-value for any pair of donors in any of the sample subsets was $p = 0.0095$ (i.e., the lowest PCC between donors was higher than 99.05% of PCCs in the empirical null distribution).

5.7 Persistent developmental stuttering candidate genes in the basal ganglia

Both the regional networks and co-expression landscape analyses pointed to co-expression relationships between PDS candidates specifically in the basal ganglia. These relationships were stronger than expected by chance relative not only to randomly selected genes (Fig 5.4I), but also to the other SL genes (Fig 5.5C). These observations motivated a closer examination of PDS candidate gene expression within the basal ganglia.

5.7.1 Co-expression in the basal ganglia

Correlations between PDS candidate gene expression profiles, defined across samples from the basal ganglia are shown in Fig 5.8. In the basal ganglia, these genes fell into two groups which had positively correlated expression profiles within-group, but almost entirely negative profile correlations across-group. These groups were designated "Group A" (8 genes) and "Group B" (6 genes). To a lesser extent, these relationships appear to be maintained in the striatum and its sub-regions as well, but not the globus

pallidus. Group A genes showed significantly higher within-group than across-group correlations in the basal ganglia, striatum, caudate nucleus, putamen, and nucleus accumbens, and Group B genes only in the basal ganglia (one-tailed two sample Kolmogorov-Smirnov test, $p < 0.05$ after Bonferroni correction for 5 regions x 2 gene groups = 10 comparisons). With the exception of the nucleus accumbens, the effect size decreased with greater neuroanatomical specificity: the difference between mean correlation with Group A and mean correlation across the two groups for the basal ganglia was 0.68, for the striatum 0.39, and for the caudate, putamen, and nucleus accumbens 0.25, 0.37, and 0.60 respectively. The difference between Group B's mean correlation and the mean cross-group correlation in the basal ganglia was 0.47. (Note that only 25 left-hemisphere samples were available for the globus pallidus and 13 for the nucleus accumbens; hence their exclusion from Section 5.4's co-expression modularity analysis, where only brain regions with at least 30 available samples were included.)

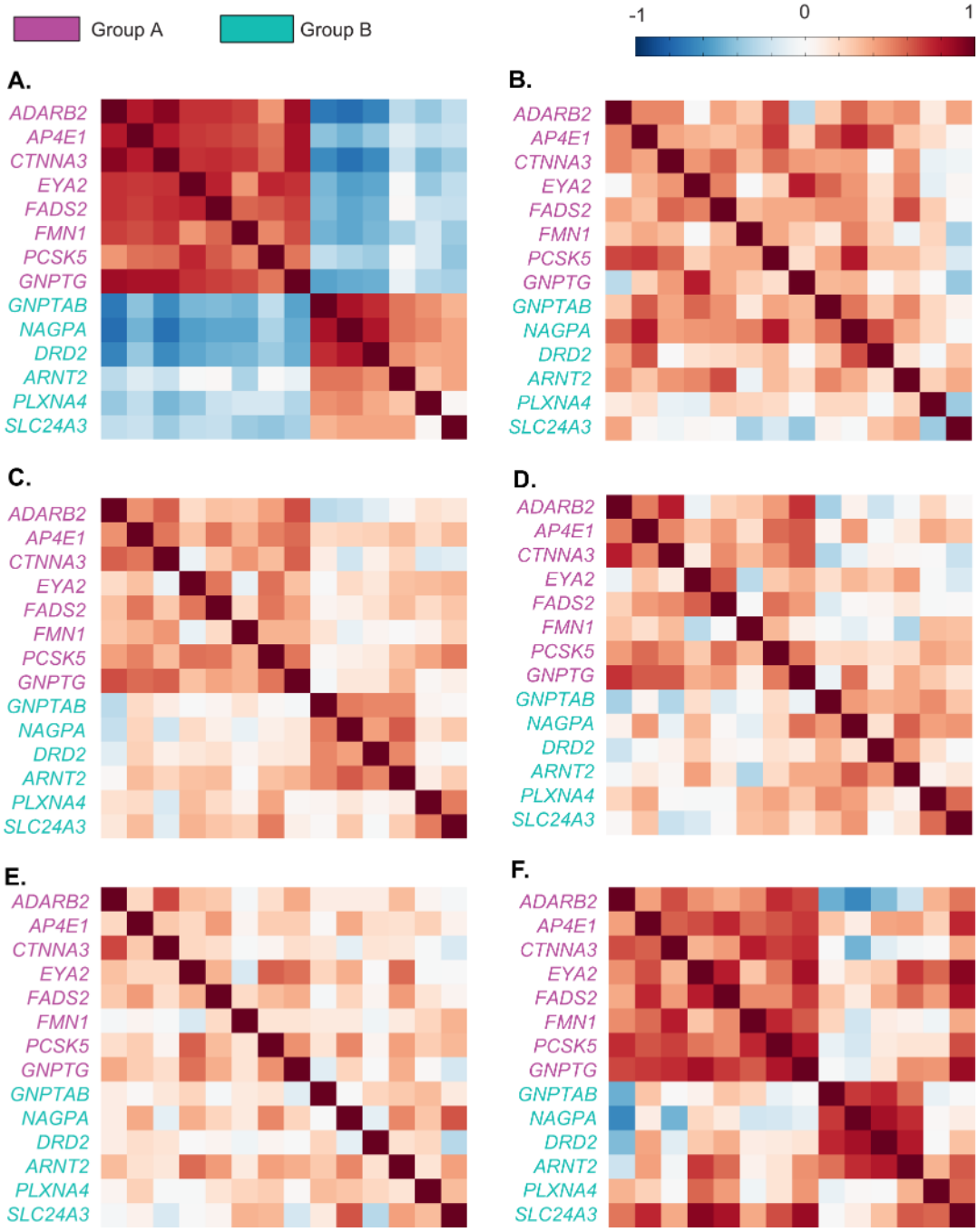


Fig 5.8. Correlation heatmaps of persistent developmental stuttering candidate genes in the basal ganglia and sub-structures. Genes are ordered to group those with high correlations. A. Basal ganglia (146 samples). B. Globus pallidus (25 samples). C. Striatum (121 samples). D. Putamen (46 samples). E. Caudate nucleus (62 samples). F. Nucleus accumbens (13 samples).

5.7.2 Differential expression across sub-structures of the basal ganglia

As shown in Chapter 3, *sample expression profiles* from the striatum and pallidum tend to be distinct from each other, but relatively consistent within-region (Fig 3.2A). This suggests that many genes may have consistently higher expression values across striatal than pallidal samples, and others consistently higher expression across pallidal than striatal samples: i.e., many genes may be differentially expressed across the two structures. Strong positive correlations between *gene expression profiles* within the basal ganglia may therefore result from genes that are consistently higher in either the striatum or the pallidum.

Fig 5.9A confirms that the eight "Group A" genes consistently (though sometimes weakly) showed stronger expression in the globus pallidus than the striatum, while the six "Group B" genes showed the opposite tendency. This may partially explain the high within-group correlations shown in Fig 5.8A, and the decrease of this effect within the striatum. Group A genes (which had shown significantly enhanced correlations in the striatum as well) also showed consistently (though weakly) higher expression in the putamen than either the caudate or nucleus accumbens, with the exception of *EYA2*. However, none of these genes had a log2 fold change of at least 0.5 between the striatum and globus pallidus or between any two sub-regions of the striatum, so they were not differentially expressed by the criteria used in Chapter 3.

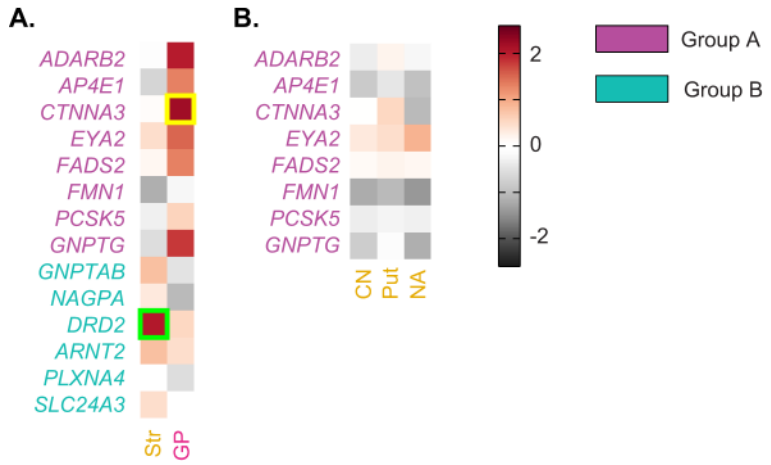


Fig 5.9. Stuttering candidate gene expression in basal ganglia substructures. A. Expression in the striatum and globus pallidus. Sub-set of heatmap shown in Fig 5.3A. B. Expression of "Group A" genes in striatal sub-structures. This differs from Fig 5.3C in that expression values for the head, body and tail of the caudate nucleus have been averaged together.

5.8 Discussion

The analyses discussed here constitute an early step in using the transcriptome to investigate genes implicated in speech and/or language disorders ("SL genes") in a neuroanatomical context, which may provide clues about their association with those higher-level functions. Because each of these genes probably accounts for only a small part of the associated disorder's prevalence, identifying some form of convergence or consistency among the different candidate genes is necessary for moving beyond a list of genes with few known relationships and, ultimately, understanding their shared influence on speech and language. The current approach seeks this convergence in common elements of expression profiles across the brain, which could help to illuminate how these genes have similar impacts on complex behavioral phenotypes.

The Speech and Language Disorders Database described above is intended in part to aid the growth and refinement of this gene list (Figs 5.1 and 5.2). The rest of the

chapter, which examined the expression and co-expression of genes associated with speech and language functions in the adult human brain, may also be helpful in focusing on the most relevant genes by revealing relationships (or lack thereof) between those with little evidence supporting their inclusion, and those with relatively well-established relevance. For example, the genes implicated by Kang et al. (2010) in persistent developmental stuttering, while surprising in their general role in encoding proteins within the lysosomal enzyme targeting pathway, have strong evidence for involvement with stuttering in three cohorts. Here we showed that an additional set of genes – those with strongest evidence based on a single genome-wide association study (Kraft, 2010) – showed strong co-expression relationships with the four lysosomal pathway genes specifically in the basal ganglia, a set of brain structures with relevance to PDS. This result lends support to the possibility that alterations in the genes from Kraft (2010) may have at least some similar impacts on the brain, and encourages attempts to validate these genes, such as future association studies with larger sample sizes or more detailed case-control studies using established cohorts of people who stutter.

5.8.1 Preferential expression

The importance of a thorough examination of the overall set of candidate genes in neuroanatomically-specific datasets is confirmed by the fact that many of them showed preferential expression in one or more brain regions (a statistically significant number, at two different anatomical scales; Fig 5.3). That is, these genes have expression patterns that are more anatomically specific than observed for randomly selected genes. The broad-scale regions most often preferred--the cerebral cortex, hippocampal formation,

striatum, and thalamus--all have important roles in speech and / or language functions (e.g. Gabrieli et al., 1998; Kotz et al., 2009; Duff and Brown-Schmidt, 2012; Barbas et al., 2013). Four genes showed high percentile ranks in the cerebellar cortex (*GPLD1*, *ERC1*) or cerebellar nuclei (*THEM2*, *EYA2*), also important for speech and language (Mariën et al., 2013). *FOXP2*, though not preferentially expressed by the current criterion in any broad region, had minimum percentile ranks (across all donors) of at least 80 in the cerebral cortex, striatum, and thalamus. These results, as well as *FOXP2*'s high expression and minimum percentile rank of 89 in the inferior olivary complex, are consistent with Lai et al. (2003). That study showed restricted *FOXP2* expression in the developing human brain, particularly in the cortex, striatum, thalamus, inferior olivary complex, and cerebellum, suggesting that associations between *FOXP2* and speech / language phenotypes may be due to an important role for this gene in the development of structures related to speech and language, and particularly motor control.

Notably, the inferior olivary complex showed increased relative expression and percentile ranks compared to other areas for several other SL genes as well (Fig 5.3C). This structure is involved in motor learning and timing, and projects to both the cerebellar cortex and cerebellar nuclei (e.g. Martin et al., 1996; De Zeeuw, 1998). It is possible that some of these genes could, when disrupted, in turn disrupt speech- or language-related functions due to changes of expression in the inferior olive. The superior olivary complex is involved in both ascending and descending auditory pathways; therefore, it is interesting that four PDS candidate genes as well as *CNTNAP2* (implicated in SLI and dyslexia) and *THEM2* (implicated in dyslexia) all showed

relatively strong percentile ranks there (Fig 5.3C; *ADARB2*, *AP4E1*, *CTNNA3*, and *GNPTG* had mean ranks over 80, but in each case one donor fell under 80). Finally, the substantia nigra, which yielded high percentile ranks of two PDS candidates (*DRD2* and *ADARB2*) and a DVD candidate (*SRPX2*), is integral to basal ganglia function (Graybiel, 2000).

Many SL genes showed (not always preferential) expression across multiple brain areas, usually either cortical or subcortical (Fig 5.3). As noted in Section 5.1, the functions of SL genes are not limited to their potential importance to speech and language. The impact of variants in these genes may depend on the molecular environment of a given brain areas, which makes points of convergence between expression profiles of multiple gene candidates particularly interesting. The expression patterns of SL genes as a group were indeed often restricted to brain areas with known roles in speech and language. Despite this, genes associated with a specific disorder did not, for the most part, have common preferences for certain brain areas.

5.8.2 Regional networks

Co-expression relationships often imply common functions or pathways (Eisen et al., 1998; Lee, 2004; Wei et al., 2006), through which different genes might influence the same processes. Strong co-expression of SL genes in a given brain structure, therefore, suggest that their shared influence on speech and language might be effected through changes in that structure. Here, we looked in different brain structures for unusual *co-expression modularity* of the SL genes: i.e., an exaggerated difference between the genes' co-expression with each other and their co-expression with other genes.

The SL genes as a group showed significantly enhanced co-expression modularity in several brain areas already known to play important roles in speech and language, including the inferior frontal gyrus and the temporal lobe. This enhancement was observed when compared to randomly selected genes in each brain area; however, when the same values were compared to the co-expression modularity scores for these genes but random neuroanatomical samples, no results reached significance (Figure 5.4A, B). This suggests that, while the overall set of genes has some modular network structure, it does not appear to have strong neuroanatomical specificity. It is more likely that meaningful neuroanatomical results should arise for smaller gene sets related to more specific phenotypes.

The set of PDS candidates showed their highest co-expression modularity (compared to random gene sets) in the basal ganglia, as well as significant values in the thalamus, parietal lobe, and trigeminal nucleus of the medulla (Fig 5.4I). Importantly, when the co-expression scores were compared to those obtained for the same genes and random brain samples, only the basal ganglia showed a significant result (Fig 5.4J). Thus, this set of subcortical nuclei shows enhanced co-expression modularity across the set of PDS candidate genes (including nine genes with only suggestive evidence from Kraft, 2010) both compared to other gene sets of the same size, and compared to different anatomical areas.

Neither dyslexia nor SLI candidates showed significant co-expression modularity in any of the brain regions examined. Dyslexia genes did show their highest co-expression modularity in the thalamus, cerebral cortex as a whole, and temporal lobe (Fig

5.4C). Though the significance of these values did not survive multiple comparisons correction and the modularity is not neuroanatomically specific (Fig 5.4D), it is worth noting given that the thalamus is an important structure in speech / language circuits, that phonological processing deficits are a central feature of dyslexia (Shaywitz and Shaywitz, 2005), and that the disruption of cortical neuron migration associated with dyslexia occurs in the temporal lobe (Galaburda et al., 1985). Finally, DVD candidates showed enriched co-expression relative to random samples in the frontal and temporal lobes, frontal cingulate and lingual gyri, and pons (Fig 5.4H); however, their co-expression was not significant in these brain areas relative to other genes (Fig 5.4G).

In a few cases, a group of genes showed high co-expression modularity in a brain area with no clear relationship to the associated phenotype; in particular, the fusiform, frontal cingulate, and parahippocampal gyri for SL genes as a whole, and the parietal lobe and lingual gyrus for PDS candidates (these also stood out for SLI candidates, but were not significant; Fig 5.4A, I, E). However, none of these showed anatomical specificity (Fig 5.4B, J, F).

Overall, though co-expression modularity was higher in several areas known to support speech and language functions, only co-expression of PDS candidate genes in the basal ganglia was significantly stronger than chance relative to both other genes and random brain samples.

5.8.3 Stuttering and the basal ganglia

Neither correlation distances nor topological overlap networks (Figs 5.5, 5.6, and 5.7) indicated clear correspondence between the transcriptomic relationships of the SL

genes and the phenotypes with which they are associated, with the sole exception of PDS candidate genes in the basal ganglia (Fig 5.5C). This result reinforces the co-expression modularity analysis discussed above (Fig 5.4I, J). Within the basal ganglia, the PDS candidates appear to separate into two groups of genes which are positively correlated within-group and negatively correlated with across-group (Fig 5.8). This grouping might be due to (often slightly, and never significantly) higher expression in either the pallidum (*ADARB2*, *AP4E1*, *CTNNA3*, *EYA2*, *FADS2*, *FMN1*, *PCSK5*, and *GNPTG*) or the striatum (*GNPTAB*, *NAGPA*, *DRD2*, *ARNT2*, *PLXNA4*, and *SLC24A3*; Fig 5.9).

The basal ganglia have long been thought to play a key role in stuttering (see Alm, 2004 for a review). Because the basal ganglia subnuclei are small and contain large populations of inhibitory neurons, differences between people with PDS and controls are difficult to identify and interpret through neuroimaging (Civier et al., 2013). However, lesions of the basal ganglia have been associated with acquired stuttering (Ludlow et al., 1987; Tani and Sakai, 2011; Theys et al., 2013). The implication of the basal ganglia in stuttering, a disorder that interferes with proceeding from one motor action to the next, is consistent with their proposed role in the selection of actions (see e.g. Redgrave et al., 1999). Pharmacological evidence also supports basal ganglia involvement, as blocking type D2 dopamine receptors (D2Rs) has been shown to reduce stuttering (Stager et al., 2005). D2Rs are heavily expressed in the striatum, and necessary for its dense dopaminergic innervation. Civier et al. (2013) have developed a computational model suggesting that dopaminergic excess in the striatum, as well as abnormalities of cortico-striatal projections from ventral primary motor cortex, could result in the dysfluencies

characteristic of stuttering. Briefly, excessive dopaminergic activity in the striatum is hypothesized to put a "ceiling effect" on the outgoing signal to produce the next syllable, preventing this signal from gaining sufficient strength relative to signals representing competing syllables and leading to its re-selection.

The findings of the current study suggest that the 14 PDS candidate genes may be working in concert, perhaps with a similar impact on basal ganglia circuitry. This is particularly interesting given the varying genetic evidence supporting the inclusion of these genes as candidates. Although the pharmacological evidence and model discussed above suggests a connection between PDS and *DRD2*, which codes for type 2 dopamine receptors, genetic evidence for the association is weak (Lan et al., 2009 found an association in a Han Chinese cohort; Kang et al., 2011a failed to replicate the finding in a Brazilian and a European cohort). Similarly, 9 genes are implicated by a single genome-wide association study with a relatively small sample size (Kraft, 2010). In contrast, larger studies in multiple populations provide support for *GNPTG*, *GNPTAB*, and *NAGPA* (Pakistani, British, and North American cohorts; Kang et al., 2010) as well as *AP4E1* (Pakistani, Cameroonian, and North American cohorts; Raza et al., 2015). The strong co-expression relationships between these four genes and the less well-investigated candidates, which are strongest within a structure implicated in PDS, lends support to the inclusion of the genes from Lan et al (2009) and Kraft (2010) as candidate genes. Further, these results suggest that the basal ganglia may be of particular importance in attempting to establish the mechanisms by which alterations to these genes, and possibly others yet to be identified, impact fluent speech.

5.8.4 Limitations and future directions

Evidence implicating the genes treated here in disorders of speech and language varies substantially. The inclusion of *CNTNAP2*, for example, is supported by multiple studies as well as the fact that it is a regulatory target of the extremely well-established candidate transcription factor *FOXP2* (Vernes et al., 2008; Peter et al., 2011; Newbury et al., 2011). *DRD2*'s association with PDS, on the other hand, was found in only one study of a Han Chinese cohort (Lan et al., 2009), while a study of a European and a Brazilian cohort failed to replicate the finding (Kang et al., 2011a). This raises the possibility of a false positive, or perhaps (as the authors of the second study suggest) that the association is not causal but is due, for example, to linkage disequilibrium with variants in a different gene in some populations. The current list of candidate genes undoubtedly includes some implicated by false positives or non-causal associations. Many genes with important roles in these complex phenotypes are also undoubtedly missing from the list. Continuing genetic research into these phenotypes will probably both refine and expand the current set of gene-disorder associations, as well as associations between specific variants and endophenotypes. This will create an increasingly solid basis for future analyses of their transcriptomic profiles using the general approach defined here.

This attempt to relate genotype to phenotype through gene expression makes the assumption that variants impact expression of the genes that contain them. As a result, the approach in its current form will not yield useful results in situations where a variant influences a behavioral phenotype by altering the expression of a different gene (unless that gene is co-expressed with the gene in which the variant resides). Some variants may

indirectly change the expression of another gene through a co-expression network (in which genes up- or down-regulate each others' expression). In other cases, a variant may directly change the expression of a different gene; these are found in expression quantitative trait loci (eQTLs; see Section 5.1). Efforts to map eQTLs may allow future transcriptomic analyses to address some of these cases.

Transcriptomic analyses of implicated genes will also benefit from a comprehensive use of current knowledge regarding the brain structures and circuits underlying language functions. The fact that certain structures (such as the cerebral cortex, basal ganglia, and cerebellum) have major roles in language functions was used to focus the co-expression landscape and co-expression network analyses presented here, and to interpret the regional co-expression of genes associated with different phenotypes. A more detailed and nuanced use of the neuroimaging results represented in the Speech and Language Disorders Database and the vast surrounding literature could provide further insight into the transcriptomic results presented here, and might suggest further avenues of investigation. In particular, systematic use of this literature could focus the analyses on particular cortical areas supporting different speech and language functions, rather than treating the cerebral cortex as a whole.

Speech and language production and comprehension are, of course, dynamic processes, and some research has directly related gene expression to functional activity in the brain. As early as 1991, expression of the gene *c-fos* was shown to reflect tonotopic maps in the mouse dorsal cochlear nucleus and inferior colliculus (Ehret and Fischer, 1991). In human subjects with Fragile X syndrome, expression of the gene *FMR1*

(implicated in intellectual disability) in lymphocytes was correlated with activity in the middle frontal and supramarginal gyri (Menon et al., 2000). Richiardi et al. (2015) related functional activity to gene expression across the brain (as opposed to expression in blood cells) by defining resting state functional networks in human subjects and comparing these to gene expression in the AHBA. If such a relationship exists between speech / language networks and gene expression, it would be worth identifying genes responsible for that relationship. These may become candidate speech / language genes (or, if already candidates, such a finding would provide additional support for their inclusion).

A major limitation of the current study is its exclusive focus on the adult brain. Dyslexia, SLI, DVD, and PDS are all developmental disorders, manifesting as children learn to speak or to comprehend spoken or written language (and in the case of SLI, signed language; see Marshall et al., 2006; Mason et al., 2010). A small influence at an early stage could have a serious impact on the developmental trajectory of brain structures and circuits, and many transcriptomic events through which an individual genotype may impact a behaviorally-defined phenotype occur transiently in the rapidly-changing molecular environment of the developing brain.

Genome-scale transcriptomic data for donor brains from 8 post-conceptional weeks (pcw) to 40 years of age are available as part of the publicly available dataset BrainSpan: Atlas of the Developing Brain (http://brainspan.org/). The BrainSpan Atlas includes RNA sequencing (RNA-Seq) expression data for each of 41 neurologically normal donors from 8 post-conceptional weeks to 40 years of age. RNA-Seq allows

direct quantification of expression by counting the number of transcripts in a sample (rather than quantifying image intensity based on a fluorescent label, as with microarray and ISH data), including measurement of alternative transcripts. This dataset has a low sample count and low spatial resolution; for most donors, only one sample per brain structure is available from 8-16 structures. There are also high spatial resolution microarray data (including ~300 structures) from four donor brains, from 15-21 post-conceptional weeks. The BrainSpan Atlas offers the possibility of examining the transcriptomic profiles of speech and language candidate genes at early stages corresponding to the onset of the disordered phenotype, and tracing those profiles over time. If some of these genes impact the normal development of brain systems supporting language function without leaving lasting evidence in the adult brain, then such data will be invaluable for quantifying transient preferential expression or co-expression in key brain structures.

The restriction to left-hemisphere samples in the current study is also worth considering, given asymmetry in the human brain. Human frontal cortex typically shows both functional and structural asymmetry of language-related areas. Abnormalities of cortical asymmetry have also been associated with dyslexia (e.g. Galaburda et al., 1985; Hynd et al., 1990; note however that some studies found normal asymmetry in people with dyslexia, e.g. Best and Demb, 1999; Rumsey et al., 1997), SLI (Gauger et al., 1997; De Fossé et al., 2004), and PDS (Chang et al., 2008; Foundas et al., 2004). Watkins et al. (2002) initially found reduced grey matter in the left inferior frontal gyrus associated with DVD, but using a more specific and selective model, concluded that the difference was in

fact bilateral (Belton et al., 2003). There is evidence that the basal ganglia also show structural asymmetry as well as hemispheric dominance related to motor control (e.g. Kooistra and Heilman, 1988; Scholz et al., 2000), and that the cerebellum's role in language as well as other cognitive and motor tasks is lateralized (Mariën et al., 2013; Stoodley, 2012). These asymmetries indicate the desirability of cross-hemispheric comparison of the expression profiles of genes implicated in those processes. In this study, such a comparison was precluded by the exclusion of right-hemisphere expression data, due to the lack of this data in four of the six donor brains in the AHBA (see Chapter 2). The expression and co-expression of the candidate genes could be compared either in the two donors with data available from both hemispheres, and in other datasets.

Thus far, however, studies of molecular neuroanatomy have revealed little difference in the expression of individual genes between hemispheres in adulthood (Hawrylycz et al., 2012; Pletikos et al., 2014) or even mid-fetal stages (Sun, 2005; Johnson et al., 2009; Lambert et al., 2011; Pletikos et al., 2014). One study found global transcriptomic symmetry from early fetal stages on (beginning 10 weeks post-conception; Pletikos et al., 2014). However, another identified 27 genes showing differential expression across cortical hemispheres from about 12-19 weeks post-conception (Sun, 2005), after which most (not all) of cortical neuron proliferation and migration is complete (de Graaf-Peters and Hadders-Algra, 2006). Therefore, hemispheric comparison may be most productive in developmental data from fetal stages before 19 weeks in datasets that include samples from both hemispheres and label the hemisphere

of origin (in the BrainSpan Atlas, the high-resolution microarray data is restricted to the left hemisphere and the RNA-Seq data does not indicate hemisphere of origin).

The connection suggested by these results between the PDS candidates, the basal ganglia, and the disorder itself could be validated by postmortem studies from people with PDS, should such data become available. If the expression or co-expression of these genes varied significantly between people who stutter and neurologically normal controls, the location and nature of the differences could provide a starting point for a model of the mechanisms by which the products of these genes impact fluent speech. Nine of the PDS candidate genes studied here are implicated by a single, small genome-wide association study, and little is known of most of them. Some are implicated in another neuropathology or brain-specific function by a single study (see Section 5.1.4), and four by none at all (*EYA2*, *FMN1*, *SLC24A3*, and *PCSK5*). As more is learned of the roles these genes play in the brain, this information might suggest more specific hypotheses (beyond anatomical localization) regarding their influence on fluent speech.

Some, but not all, of these analyses include cross-donor comparisons. Distance and topological overlap matrices based on individual donors were all strongly (and significantly) correlated, and our definition of preferential expression required that all donors show a minimum percentile rank of 95. However, the analyses of co-expression modularity and of PDS candidates in the basal ganglia would benefit from comparison of donors, where this is reasonable given the within-donor sample counts for a given brain area.

Finally, as with nearly all results based on human data in this dissertation, those presented in this chapter use only the AHBA. Application of these methods to other adult datasets is necessary to increase confidence in the results. Expansion of this work to datasets that distinguish between cortical layers would also make examinations of gene expression in the cerebral cortex potentially more fruitful, as laminar variation in cell type densities results in neocortical layers showing amplified differences in gene expression patterns (Belgard et al., 2011; Bernard et al., 2012).

5.8.5 Conclusion

The transcriptomic profiles of genes that impact speech and language ability offer the possibility of relating our knowledge of these disorders across very different levels of organization. This study is, to our knowledge, the first systematic examination of the brain-wide expression and co-expression relationships of speech and language candidate genes. It represents an important first step towards illuminating the roles of these genes and defining points of neuroanatomical convergence in the potential impact of a variety of DNA alterations driving similar, sometimes overlapping phenotypes.

CHAPTER 6: CONCLUSION

Though features such as cytoarchitecture and myeloarchitecture are more easily observable than are multivariate profiles at a molecular scale, the latter are no less integral to the structural and functional organization of the brain. Large gene expression datasets, while sparse in annotation and somewhat overwhelming in scope, contain extensive information about that scale. The work presented here examines the correspondence between transcriptomic organization and conventional neuroanatomy within and across species, and attempts to lay groundwork for the use of transcriptomic data to relate genotype and phenotype in behaviorally-defined disorders.

The approach taken in this work often blurs the distinction between exploratory data analysis (EDA; Tukey, 1977) and more traditional hypothesis-driven research. Even the first part of dissertation (Chapter 3), which makes heavy use of the sort of simple statistical summaries and visualizations that are central to EDA, is informed by prior knowledge of neuroanatomical labels, and tests simple hypotheses such as, "samples from within the cerebral cortex are more transcriptomically similar to each other than to other samples." The mouse-human comparative study (Chapter 4) selects genes with known common functions and poses the hypothesis that they will show conserved expression across species. In Chapter 5, genes were curated from relevant literature, and knowledge of the roles of different brain regions in speech and language functions informed both design and interpretation. These hypotheses, however, are very general and not always subjected to strict significance tests. This is because the nature of this approach is largely data-driven (as in EDA), using data from “experiments” not

specifically designed to address the questions posed but of sufficient scale to provide new insight. A data-driven approach can reduce bias, which is advantageous when faced with large, sparsely annotated datasets where it is difficult to know what will be relevant. However, in order to move towards more specific hypotheses, prior knowledge of genes, brain structures and complex phenotypes were used to inform these studies.

6.1 Summary of contributions

Molecular and conventional neuroanatomy have a close correspondence in both the mouse and human brain (e.g. Bohland et al., 2010; Roth et al., 2006). Chapter 3 took advantage of two high-throughput gene expression datasets with unusually high spatial resolution to reveal this relationship. These analyses showed the transcriptomic similarities and distinctions between brain structures at multiple levels. Additionally, this chapter compared two human datasets and identified consistencies between the expression profiles of different brain regions, in spite of the very different sampling properties of the datasets.

Homological relationships between the mouse and human brain have fundamental importance for the use of mouse models in both basic and clinical research. Conventional neuroanatomy (i.e., basic histochemical stains and tract tracing studies of connectivity) has been used extensively to understand these putative homologies, but less is known about the extent to which they are verifiable at the molecular level. Chapter 4 described the development and application of tools for identifying similarities and differences in brain-wide and regional molecular environments (defined by expression

levels across a pre-defined gene set) in the mouse and human brain, using two gene expression datasets with high anatomical resolution. The results revealed conserved molecular organization at multiple scales, as well as particular groups of genes whose expression patterns form unique regional fingerprints that are consistent between the two species. The diverse patterns of conservation of gene expression across the mouse and human brain is further reflected by the similarity of individual genes' brain-wide profiles, which was highly variable across the genes studied here. By applying these and related analyses to additional gene expression datasets and interpreting results in light of the larger context for gene expression (e.g. alternative splicing and post-transcriptional mechanisms), it may ultimately be possible to quantify homologies in the molecular architecture of the human and mouse brain, helping to bridge a seemingly vast divide between genomics and systems neuroscience. Such directions are of particular importance for understanding the mechanisms of heritable diseases of the nervous system and for improving and understanding the efficacy of drugs targeting the brain.

Genetic, neuroimaging and behavioral lines of research into speech and language disorders tend to be conducted in relative isolation. Chapter 5 provides the first detailed analysis of the expression patterns of genes implicated in these disorders throughout the human brain, a move in the direction of bridging the gap between genotype and phenotype through intervening brain regions and neural systems. The preferential expression and co-expression of many of these genes in regions already known to be important to speech and language supports their proposed roles in such functions. The most salient point identified in this study was the strong co-expression relationships

between genes associated with persistent developmental stuttering (PDS), which were not only preferentially expressed and co-expressed in the basal ganglia, but also showed stronger co-expression relationships with each other than with the other speech / language gene candidates specifically in that structure. These genes are differentially expressed across the striatum and pallidum, with one group more strongly expressed in the former and another in the latter, suggesting potentially different targets and mechanisms for impacting the same overall system. Finally, the co-expression relationships (especially within the basal ganglia) between genes identified in a genome-wide association study with relatively few subjects (Kraft, 2010) and genes with stronger evidence in PDS (Kang et al., 2010; Raza et al., 2015) provides a new source of support for roles for some of these genes in PDS.

6.2 Future directions

Chapters 3-5 include more detailed discussions of future work; however, there are common threads to those discussions that are worth reiterating here. First, this work has focused primarily on two high-throughput gene expression datasets out of many that have been made available since their advent in the 1990s. Chapter 3 implemented one cross-dataset comparison, showing relatively high consistency between anatomical region profiles in two expression datasets from the adult human brain. Further validation studies, and the application of similar methods to other datasets, will be necessary to identify most robust results.

Second, molecular neuroanatomy in the adult is the outcome of complex developmental processes intricately regulated by gene expression. Expanding this approach to transcriptomic data from a range of developmental stages may elucidate not only developmental mechanisms themselves, but also the latent causes of some features of the mature transcriptome. Most particularly, the study of speech and language disorders calls for attention to early development, when the processes forming key structures and circuitry are so rapid and precise that small differences could have profound effects, as well as to later developmental stages (i.e., childhood) when these disorders first appear.

Finally, functional annotations of genes are vital for both focusing analyses of transcriptomic data and interpreting their results. With prior knowledge of gene functions, computational resources can be allocated to the expression profiles of genes of interest, and hypotheses may be suggested regarding the biological significance and interpretation of some results. Such prior knowledge is still sparse relative to the number of protein-coding genes in the human and mouse genomes, and relative to the number of processes some of those genes impact. Furthermore, knowledge of their function within brain tissue specifically is even more sparse. Continuing genetic research, from knockout studies in model organisms to association studies in humans, may support or otherwise illuminate some of the results shown here, and suggest further avenues for future study.

BIBLIOGRAPHY

Abrahams, B.S., and Geschwind, D.H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. Nature Reviews Genetics *9*, 341–355.

Abrahams, B.S., Tentler, D., Perederiy, J.V., Oldham, M.C., Coppola, G., and Geschwind, D.H. (2007). Genome-wide analyses of human perisylvian cerebral cortical patterning. Proceedings of the National Academy of Sciences *104*, 17849–17854.

Alm, P. a (2004). Stuttering and the basal ganglia circuits: a critical review of possible relations. Journal of Communication Disorders *37*, 325–369.

Al-Murrani, A., Ashton, F., Aftimos, S., George, A.M., and Love, D.R. (2012). Amino-Terminal Microdeletion within the *CNTNAP2* Gene Associated with Variable Expressivity of Speech Delay. Case Reports in Genetics *2012*, 1–4.

Anthoni, H., Zucchelli, M., Matsson, H., Muller-Myhsok, B., Fransson, I., Schumacher, J., Massinen, S., Onkamo, P., Warnke, A., Griesemann, H., et al. (2006). A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. Human Molecular Genetics *16*, 667–677.

Anthoni, H., Sucheston, L.E., Lewis, B.A., Tapia-Páez, I., Fan, X., Zucchelli, M., Taipale, M., Stein, C.M., Hokkanen, M.-E., Castrén, E., et al. (2012). The Aromatase Gene CYP19A1: Several Genetic and Functional Lines of Evidence Supporting a Role in Reading, Speech and Language. Behavior Genetics *42*, 509–527.

Araki, M., and Taketani, S. (2009). Neurensin: A novel neuron-specific gene and its role in membrane trafficking and neurite outgrowth. In Recent Research Developments in Neuroscience 3, pp. 111–136.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25–29.

Ashurst, J.V., and Wasson, M.N. (2011). Developmental and persistent developmental stuttering: an overview for primary care physicians. Journal of the American Osteopathic Association *111*, 576–580.

Autry, A.E., and Monteggia, L.M. (2012). Brain-derived neurotrophic factor and neuropsychiatric disorders. Pharmacological Reviews *64*, 238–258.

Bacchelli, E., Ceroni, F., Pinto, D., Lomartire, S., Giannandrea, M., D'Adamo, P., Bonora, E., Parchi, P., Tancredi, R., Battaglia, A., et al. (2014). A CTNNA3 compound heterozygous deletion implicates a role for αT-catenin in susceptibility to autism spectrum disorder. Journal of Neurodevelopmental Disorders *6*, 17.

Barbas, H., García-Cabezas, M.Á., and Zikopoulos, B. (2013). Frontal-thalamic circuits associated with language. Brain and Language *126*, 49–61.

Bates, T.C., Luciano, M., Medland, S.E., Montgomery, G.W., Wright, M.J., and Martin, N.G. (2011). Genetic Variance in a Component of the Language Acquisition Device: ROBO1 Polymorphisms Associated with Phonological Buffer Deficits. Behavior Genetics *41*, 50–57.

Belgard, T.G., Marques, A.C., Oliver, P.L., Abaan, H.O., Sirey, T.M., Hoerder-Suabedissen, A., García-Moreno, F., Molnár, Z., Margulies, E.H., and Ponting, C.P. (2011). A Transcriptomic Atlas of Mouse Neocortical Layers. Neuron *71*, 605–616.

Bellini, G., Bravaccio, C., Calamoneri, F., Donatella Cocuzza, M., Fiorillo, P., Gagliano, A., Mazzone, D., del Giudice, E.M., Scuccimarra, G., Militerni, R., et al. (2005). No evidence for association between dyslexia and DYX1C1 functional variants in a group of children and adolescents from Southern Italy. Journal of Molecular Neuroscience *27*, 311–314.

Belton, E., Salmond, C.H., Watkins, K.E., Vargha-Khadem, F., and Gadian, D.G. (2003). Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. Human Brain Mapping *18*, 194–200.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Statistical Methodology *57*, 289–300.

Berchtold, N.C., Cribbs, D.H., Coleman, P.D., Rogers, J., Head, E., Kim, R., Beach, T., Miller, C., Troncoso, J., Trojanowski, J.Q., et al. (2008). Gene expression changes in the course of normal brain aging are sexually dimorphic. Proceedings of the National Academy of Sciences *105*, 15605–15610.

Bernard, A., Lubbers, L.S., Tanis, K.Q., Luo, R., Podtelezhnikov, A.A., Finney, E.M., McWhorter, M.M.E., Serikawa, K., Lemon, T., Morgan, R., et al. (2012). Transcriptional Architecture of the Primate Neocortex. Neuron *73*, 1083–1099.

Best, M., and Demb, J.B. (1999). Normal planum temporale asymmetry in dyslexics with a magnocellular pathway deficit. Neuroreport *10*, 607–612.

Bishop, D.V.M., and Snowling, M.J. (2004). Developmental Dyslexia and Specific Language Impairment: Same or Different? Psychological Bulletin *130*, 858–886.

Bishop, D.V.M., North, T., and Donlan, C. (1996). Nonword Repetition as a Behavioural Marker for Inherited Language Impairment: Evidence From a Twin Study. Journal of Child Psychology and Psychiatry *37*, 391–403.

Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., and Landfield, P.W. (2004). Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proceedings of the National Academy of Sciences *101*, 2173–2178.

Bohland, J.W., Bokil, H., Pathak, S.D., Lee, C.-K., Ng, L., Lau, C., Kuan, C., Hawrylycz, M., and Mitra, P.P. (2010). Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. Methods *50*, 105–112.

Bohland, J.W., Myers, E.M., and Kim, E. (2014). An Informatics Approach to Integrating Genetic and Neurological Data in Speech and Language Neuroscience. Neuroinformatics *12*, 39–62.

Bolstad, B.M., Irizarry, R.., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Booth, J.R., Wood, L., Lu, D., Houk, J.C., and Bitan, T. (2007). The role of the basal ganglia and cerebellum in language processing. Brain Research *1133*, 136–144.

Bota, M., Dong, H.-W., and Swanson, L.W. (2003). From gene networks to brain networks. Nature Neuroscience. *6*, 795–799.

Bota, M., Dong, H.-W., and Swanson, L.W. (2005). Brain Architecture Management System. Neuroinformatics *3*, 015–048.

BrainSpan: Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. © 2011. Available from: http://brainspan.org/.

Brkanac, Z., Chapman, N.H., Matsushita, M.M., Chun, L., Nielsen, K., Cochrane, E., Berninger, V.W., Wijsman, E.M., and Raskind, W.H. (2007). Evaluation of candidate genes for DYX1 and DYX2 in families with dyslexia. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics *144B*, 556–560.

Brown, S., Ingham, R.J., Ingham, J.C., Laird, A.R., and Fox, P.T. (2005). Stuttered and fluent speech production: An ALE meta-analysis of functional neuroimaging studies. Human Brain Mapping *25*, 105–117.

van der Brug, M.P., Blackinton, J., Chandran, J., Hao, L.-Y., Lal, A., Mazan-Mamczarz, K., Martindale, J., Xie, C., Ahmad, R., Thomas, K.J., et al. (2008). RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. Proceedings of the National Academy of Sciences *105*, 10244–10249.

Burns, T.C., Li, M.D., Mehta, S., Awad, A.J., and Morgan, A.A. (2015). Mouse models rarely mimic the transcriptome of human neurodegenerative diseases: A systematic bioinformatics-based critique of preclinical models. European Journal of Pharmacology *759*, 101–117.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., et al. (2008). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. Journal of Neuroscience *28*, 264–278.

Cai, X., Wang, X., Patel, S., and Clapham, D.E. (2015). Insights into the early evolution of animal calcium signaling machinery: A unicellular point of view. Cell Calcium *57*, 166–173.

Carlezon, W.A., and Thomas, M.J. (2009). Biological substrates of reward and aversion: A nucleus accumbens activity hypothesis. Neuropharmacology *56*, 122–132.

Chang, B.S., Ly, J., Appignani, B., Bodell, A., Apse, K.A., Ravenscroft, R.S., Sheen, V.L., Doherty, M.J., Hackney, D.B., O'Connor, M., et al. (2005). Reading impairment in the neuronal migration disorder of periventricular nodular heterotopia. Neurology *64*, 799–803.

Chang, S.-E., Erickson, K.I., Ambrose, N.G., Hasegawa-Johnson, M.A., and Ludlow, C.L. (2008). Brain anatomy differences in childhood stuttering. NeuroImage *39*, 1333–1344.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS ONE *6*, e17238.

Civier, O., Bullock, D., Max, L., and Guenther, F.H. (2013). Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. Brain and Language *126*, 263–278.

Cohen, M.J., Morgan, A.M., Vaughn, M., Riccio, C.A., and Hall, J. (1999). Verbal Fluency in Children: Developmental Issues and Differential Validity in Distinguishing Children with Attention-Deficit Hyperactivity Disorder and Two Subtypes of Dyslexia. Archives of Clinical Neuropsychology *14*, 433–443.

Colantuoni, C., Lipska, B.K., Ye, T., Hyde, T.M., Tao, R., Leek, J.T., Colantuoni, E.A., Elkahloun, A.G., Herman, M.M., Weinberger, D.R., et al. (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature *478*, 519–523.

Connally, E.L., Ward, D., Howell, P., and Watkins, K.E. (2014). Disrupted white matter in language and motor tracts in developmental stuttering. Brain and Language *131*, 25–35.

Cope, N., Harold, D., Hill, G., Moskvina, V., Stevenson, J., Holmans, P., Owen, M.J., O'Donovan, M.C., and Williams, J. (2005). Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. The American Journal of Human Genetics *76*, 581–591.

Craig-McQuaide, A., Akram, H., Zrinzo, L., and Tripoliti, E. (2014). A review of brain circuitries involved in stuttering. Frontiers in Human Neuroscience *8*.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Research *42*, D472-477.

Dahdouh, F., Anthoni, H., Tapia-Páez, I., Peyrard-Janvid, M., Schulte-Körne, G., Warnke, A., Remschmidt, H., Ziegler, A., Kere, J., Müller-Myhsok, B., et al. (2009). Further evidence for DYX1C1 as a susceptibility factor for dyslexia. Psychiatric Genetics *19*, 59–63.

Day, D.A., and Tuite, M.F. (1998). Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. Journal of Endocrinology. *157*, 361–371.

De Fossé, L., Hodge, S.M., Makris, N., Kennedy, D.N., Caviness, V.S., McGrath, L., Steele, S., Ziegler, D.A., Herbert, M.R., Frazier, J.A., et al. (2004). Language-association cortex asymmetry in autism and specific language impairment: Autism, SLI Language Asymmetry. Annals of Neurology *56*, 757–766.

De Zeeuw, C. (1998). Microcircuitry and function of the inferior olive. Trends in Neurosciences *21*, 391–400.

Deffenbacher, K., Kenyon, J., Hoover, D., Olson, R., Pennington, B., DeFries, J., and Smith, S. (2004). Refinement of the 6p21.3 quantitative trait locus influencing dyslexia: linkage and association analyses. Human Genetics *115*.

Di Napoli, A., Warrier, V., Baron-Cohen, S., and Chakrabarti, B. (2015). Genetic variant rs17225178 in the ARNT2 gene is associated with Asperger Syndrome. Molecular Autism *6*, 9.

Ding, S.-L. (2013). Comparative anatomy of the prosubiculum, subiculum, presubiculum, postsubiculum, and parasubiculum in human, monkey, and rodent: Comparative Neuroanatomy of the Subicular Cortices. Journal of Comparative Neurology *521*, 4145–4162.

Diotel, N., Page, Y.L., Mouriec, K., Tong, S.-K., Pellegrini, E., Vaillant, C., Anglade, I., Brion, F., Pakdel, F., Chung, B., et al. (2010). Aromatase in the brain of teleost fish: Expression, regulation and putative functions. Frontiers in Neuroendocrinology *31*, 172–192.

Duff, K. (2004). Transgenic mouse models of Alzheimer's disease: How useful have they been for therapeutic development? Briefings in Functional Genomics and Proteomics *3*, 47–59.

Duff, M.C., and Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. Frontiers in Human Neuroscience *6*.

Eckert, M. (2004). Neuroanatomical Markers for Dyslexia: A Review of Dyslexia Structural Imaging Studies. The Neuroscientist *10*, 362–371.

Ehret, G., and Fischer, R. (1991). Neuronal activity and tonotopy in the auditory system visualized by c-fos gene expression. Brain Research *567*, 350–354.

Einarsdottir, E., Svensson, I., Darki, F., Peyrard-Janvid, M., Lindvall, J.M., Ameur, A., Jacobsson, C., Klingberg, T., Kere, J., and Matsson, H. (2015). Mutation in CEP63 co-segregating with developmental dyslexia in a Swedish family. Human Genetics *134*, 1239–1248.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences *95*, 14863–14868.

Ferry, P.C., Hall, S.M., and Hicks, J.L. (2008). "Dilapidated" Speech: Developmental Verbal Dyspraxia. Developmental Medicine & Child Neurology *17*, 749–756.

Feuk, L., Kalervo, A., Lipsanen-Nyman, M., Skaug, J., Nakabayashi, K., Finucane, B., Hartung, D., Innes, M., Kerem, B., Nowaczyk, M.J., et al. (2006). Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. American Journal of Human Genetics *79*, 965–972.

Fields, H. (2004). State-dependent opioid control of pain. Nature Reviews Neuroscience *5*, 565–575.

Filges, I., Shimojima, K., Okamoto, N., Rothlisberger, B., Weber, P., Huber, A.R., Nishizawa, T., Datta, A.N., Miny, P., and Yamamoto, T. (2011). Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. Journal of Medical Genetics *48*, 117–122.

Fink, A.J. (2006). Development of the Deep Cerebellar Nuclei: Transcription Factors and Cell Migration from the Rhombic Lip. Journal of Neuroscience *26*, 3066–3076.

Fisher, S.E., Lai, C.S.L., and Monaco, A.P. (2003). DECIPHERING THE GENETIC BASIS OF SPEECH AND LANGUAGE DISORDERS. Annual Review of Neuroscience *26*, 57–80.

Forlano, P.M., Schlinger, B.A., and Bass, A.H. (2006). Brain aromatase: New lessons from non-mammalian model systems. Frontiers in Neuroendocrinology *27*, 247–274.

Foundas, A.L., Bollich, A.M., Feldman, J., Corey, D.M., Hurley, M., Lemen, L.C., and Heilman, K.M. (2004). Aberrant auditory processing and atypical planum temporale in developmental stuttering. Neurology *63*, 1640–1646.

Francks, C., Paracchini, S., Smith, S.D., Richardson, A.J., Scerri, T.S., Cardon, L.R., Marlow, A.J., MacPhie, I.L., Walter, J., Pennington, B.F., et al. (2004). A 77-Kilobase Region of Chromosome 6p22.2 Is Associated with Dyslexia in Families From the United Kingdom and From the United States. The American Journal of Human Genetics *75*, 1046–1058.

Fraser, H.B., Khaitovich, P., Plotkin, J.B., Pääbo, S., and Eisen, M.B. (2005). Aging and Gene Expression in the Primate Brain. PLoS Biology *3*, e274.

French, L., and Pavlidis, P. (2011). Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain. PLoS Computational Biology *7*, e1001049.

Gabrieli, J.D., Poldrack, R.A., and Desmond, J.E. (1998). The role of left prefrontal cortex in language and memory. Proceedings of the National Academy of Sciences *95*, 906–913.

Galaburda, A.M. (2005). Dyslexia--a molecular disorder of neuronal migration: the 2004 Norman Geschwind Memorial Lecture. Annals of Dyslexia *55*, 151–165.

Galaburda, A.M., Sherman, G.F., Rosen, G.D., Aboitiz, F., and Geschwind, N. (1985). Developmental dyslexia: Four consecutive patients with cortical anomalies. Annals of Neurology *18*, 222–233.

Galaburda, A.M., LoTurco, J., Ramus, F., Fitch, R.H., and Rosen, G.D. (2006). From genes to behavior in developmental dyslexia. Nature Neuroscience *9*, 1213–1217.

Gauger, L.M., Lombardino, L.J., and Leonard, C.M. (1997). Brain Morphology in Children With Specific Language Impairment. Journal of Speech Language and Hearing Research *40*, 1272.

Gauthier, J., Joober, R., Mottron, L., Laurent, S., Fuchs, M., De Kimpe, V., and Rouleau, G.A. (2003). Mutation screening ofFOXP2 in individuals diagnosed with autistic disorder. American Journal of Medical Genetics *118A*, 172–175.

Geurts, N., Martens, E., Verhenne, S., Lays, N., Thijs, G., Magez, S., Cauwe, B., Li, S., Heremans, H., Opdenakker, G., et al. (2011). Insufficiently Defined Genetic Background Confounds Phenotypes in Transgenic Studies As Exemplified by Malaria Infection in Tlr9 Knockout Mice. PLoS ONE *6*, e27131.

Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genetics *6*, e1000952.

Giraud, A.-L., and Ramus, F. (2013). Neurogenetics and auditory processing in developmental dyslexia. Current Opinion in Neurobiology *23*, 37–42.

Gong, X., Jia, M., Ruan, Y., Shuang, M., Liu, J., Wu, S., Guo, Y., Yang, J., Ling, Y., Yang, X., et al. (2004). Association between theFOXP2 gene and autistic disorder in Chinese population. American Journal of Medical Genetics *127B*, 113–116.

de Graaf-Peters, V.B., and Hadders-Algra, M. (2006). Ontogeny of the human central nervous system: What is happening when? Early Human Development *82*, 257–266.

Grabowski, P.J. (1998). Splicing Regulation in Neurons: Tinkering with Cell-Specific Control. Cell *92*, 709–712.

Graham, S.A., Deriziotis, P., and Fisher, S.E. (2015). Insights into the Genetic Foundations of Human Communication. Neuropsychology Review *25*, 3–26.

Grange, P., Bohland, J.W., Okaty, B.W., Sugino, K., Bokil, H., Nelson, S.B., Ng, L., Hawrylycz, M., and Mitra, P.P. (2014). Cell-type-based model explaining coexpression patterns of genes in the brain. Proceedings of the National Academy of Sciences *111*, 5397–5402.

Graybiel, A.M. (2000). The basal ganglia. Current Biology *10*, R509–R511.

Green, E.K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E., et al. (2010). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Molecular Psychiatry *15*, 1016–1022.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. *4*, 117.

Hamdan, F.F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.-O., Joober, R., et al. (2010). De Novo Mutations in FOXP1 in Cases with Intellectual Disability, Autism, and Language Impairment. The American Journal of Human Genetics *87*, 671–678.

Hannula-Jouppi, K., Kaminen-Ahola, N., Taipale, M., Eklund, R., Nopola-Hemmi, J., Kääriäinen, H., and Kere, J. (2005). The Axon Guidance Receptor Gene ROBO1 Is a Candidate Gene for Developmental Dyslexia. PLoS Genetics *1*, e50.

Hardy, J. (2006). A Hundred Years of Alzheimer's Disease Research. Neuron *52*, 3–13.

Harold, D., Paracchini, S., Scerri, T., Dennis, M., Cope, N., Hill, G., Moskvina, V., Walter, J., Richardson, A.J., Owen, M.J., et al. (2006). Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. Molecular Psychiatry *11*, 1085–1091, 1061.

Harrison, P., Lyon, L., Sartorius, L., Burnet, P., and Lane, T. (2008). Review: The group II metabotropic glutamate receptor 3 (mGluR3, mGlu3, GRM3): expression, function and involvement in schizophrenia. Journal of Psychopharmacology *22*, 308–322.

Hawrylycz, M., Miller, J.A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A.L., Jegga, A.G., Aronow, B.J., Lee, C.-K., Bernard, A., et al. (2015). Canonical genetic signatures of the adult human brain. Nature Neuroscience *18*, 1832–1844.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.

Hikosaka, O. (2007). GABAergic output of the basal ganglia. In Progress in Brain Research, (Elsevier), pp. 209–226.

Hosoya, T., Oda, Y., Takahashi, S., Morita, M., Kawauchi, S., Ema, M., Yamamoto, M., and Fujii-Kuriyama, Y. (2001). Defective development of secretory neurones in the hypothalamus of Arnt2-knockout mice. Genes Cells *6*, 361–374.

Hynd, G.W., Semrud-Clikeman, M., Lorys, A.R., Novey, E.S., and Eliopulos, D. (1990). Brain morphology in developmental dyslexia and attention deficit disorder/hyperactivity. Archives of Neurology *47*, 919–926.

Ji, S. (2011). Computational network analysis of the anatomical and genetic organizations in the mouse brain. Bioinformatics *27*, 3293–3299.

Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Ž., Coppola, G., Bogdanović, D., Geschwind, D.H., Mane, S.M., State, M.W., and Šestan, N. (2009). Functional and

Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron *62*, 494–509.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Jun, G., Asai, H., Zeldich, E., Drapeau, E., Chen, C., Chung, J., Park, J.-H., Kim, S., Haroutunian, V., Foroud, T., et al. (2014). *PLXNA 4* is associated with Alzheimer disease and modulates tau phosphorylation: *PLXNA4* : AD and Tau. Annals of Neurology *76*, 379–392.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research *28*, 27–30.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research *42*, D199–D205.

Kang, C., and Drayna, D. (2012). A role for inherited metabolic deficits in persistent developmental stuttering. Molecular Genetics and Metabolism *107*, 276–280.

Kang, C., Riazuddin, S., Mundorff, J., Krasnewich, D., Friedman, P., Mullikin, J.C., and Drayna, D. (2010). Mutations in the Lysosomal Enzyme–Targeting Pathway and Persistent Stuttering. New England Journal of Medicine *362*, 677–685.

Kang, C., Domingues, B.S., Sainz, E., Domingues, C.E.F., Drayna, D., and Moretti-Ferreira, D. (2011a). Evaluation of the association between polymorphisms at the DRD2 locus and stuttering. Journal of Human Genetics *56*, 472–473.

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011b). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–489.

Kawahara, Y., Ito, K., Sun, H., Kanazawa, I., and Kwak, S. (2003). Low editing efficiency of GluR2 mRNA is associated with a low relative abundance of ADAR2 mRNA in white matter of normal human brain. European Journal of Neuroscience *18*, 23–33.

Kawakubo, Y., Suga, M., Tochigi, M., Yumoto, M., Itoh, K., Sasaki, T., Kano, Y., and Kasai, K. (2011). Effects of Metabotropic Glutamate Receptor 3 Genotype on Phonetic Mismatch Negativity. PLoS ONE *6*, e24929.

Kelley, A.E., Bakshi, V.P., Haber, S.N., Steininger, T.L., Will, M.J., and Zhang, M. (2002). Opioid modulation of taste hedonics within the ventral striatum. Physiology & Behavior *76*, 365–377.

Kent, R.D. (2000). Research on speech motor control and its disorders. Journal of Communication Disorders *33*, 391–428.

Khaitovich, P. (2004). Regional Patterns of Gene Expression in Human and Chimpanzee Brains. Genome Research *14*, 1462–1473.

Kidd, T., Brose, K., Mitchell, K.J., Fetter, R.D., Tessier-Lavigne, M., Goodman, C.S., and Tear, G. (1998). Roundabout Controls Axon Crossing of the CNS Midline and Defines a Novel Subfamily of Evolutionarily Conserved Guidance Receptors. Cell *92*, 205–215.

Kitchen, R.R., Sabine, V.S., Sims, A.H., Macaskill, E.J., Renshaw, L., Thomas, J.S., van Hemert, J.I., Dixon, J.M., and Bartlett, J.M. (2010). Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. BMC Genomics *11*, 134.

Ko, Y., Ament, S.A., Eddy, J.A., Caballero, J., Earls, J.C., Hood, L., and Price, N.D. (2013). Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. Proceedings of the National Academy of Sciences *110*, 3095–3100.

Kong, R., Shao, S., Wang, J., Zhang, X., Guo, S., Zou, L., Zhong, R., Lou, J., Zhou, J., Zhang, J., et al. (2016). Genetic variant in DIP2A gene is associated with developmental dyslexia in Chinese population. A American Journal of Medical Genetics Part B: Neuropsychiatric Genetics *171*, 203–208.

Kong, X.-F., Bousfiha, A., Rouissi, A., Itan, Y., Abhyankar, A., Bryant, V., Okada, S., Ailal, F., Bustamante, J., Casanova, J.-L., et al. (2013). A Novel Homozygous p.R1105X Mutation of the AP4E1 Gene in Twins with Hereditary Spastic Paraplegia and Mycobacterial Disease. PLoS ONE *8*, e58286.

Kooistra, C.A., and Heilman, K.M. (1988). Motor dominance and lateral asymmetry of the globus pallidus. Neurology *38*, 388–388.

Kos, M., van den Brink, D., Snijders, T.M., Rijpkema, M., Franke, B., Fernandez, G., and Hagoort, P. (2012). CNTNAP2 and Language Processing in Healthy Individuals as Measured with ERPs. PLoS ONE *7*, e46995.

Kotz, S.A., Schwartze, M., and Schmidt-Kassow, M. (2009). Non-motor basal ganglia functions: A review and proposal for a model of sensory predictability in auditory language perception. Cortex *45*, 982–990.

Kovács, K.J. (1998). Invited review c-Fos as a transcription factor: a stressful (re)view from a functional map. Neurochemistry International *33*, 287–297.

Kraft, S.J. (2010). Genome-wide association study of persistent developmental stuttering.

Kriegstein, A.R., and Noctor, S.C. (2004). Patterns of neuronal migration in the embryonic cortex. Trends in Neurosciences *27*, 392–399.

Krug, A., Nieratschker, V., Markov, V., Krach, S., Jansen, A., Zerres, K., Eggermann, T., Stöcker, T., Shah, N.J., Treutlein, J., et al. (2010). Effect of CACNA1C rs1006737 on neural correlates of verbal fluency in healthy individuals. NeuroImage *49*, 1831–1836.

Kwasnicka-Crawford, D.A., Carson, A.R., Roberts, W., Summers, A.M., Rehnström, K., Järvelä, I., and Scherer, S.W. (2005). Characterization of a novel cation transporter ATPase gene (ATP13A4) interrupted by 3q25-q29 inversion in an individual with language delay. Genomics *86*, 182–194.

Lai, C.S.L. (2003). FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. Brain *126*, 2455–2462.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519–523.

Lambert, N., Lambot, M.-A., Bilheu, A., Albert, V., Englert, Y., Libert, F., Noel, J.-C., Sotiriou, C., Holloway, A.K., Pollard, K.S., et al. (2011). Genes Expressed in Specific Areas of the Human Fetal Cerebral Cortex Display Distinct Patterns of Evolution. PLoS ONE *6*, e17753.

Lan, J., Song, M., Pan, C., Zhuang, G., Wang, Y., Ma, W., Chu, Q., Lai, Q., Xu, F., Li, Y., et al. (2009). Association between dopaminergic genes (SLC6A3 and DRD2) and stuttering among Han Chinese. Journal of Human Genetics *54*, 457–460.

Landsberg, R.L., Awatramani, R.B., Hunter, N.L., Farago, A.F., DiPietrantonio, H.J., Rodriguez, C.I., and Dymecki, S.M. (2005). Hindbrain Rhombic Lip Is Comprised of Discrete Progenitor Cell Populations Allocated by Pax6. Neuron *48*, 933–947.

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

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics *24*, 719–720.

Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? PLoS Computational Biology *7*, e1001057.

Larsen, C.W., Zeltser, L.M., and Lumsden, A. (2001). Boundary formation and compartition in the avian diencephalon. Journal of Neuroscience *21*, 4699–4711.

Le, W., Sayana, P., and Jankovic, J. (2014). Animal Models of Parkinson's Disease: A Gateway to Therapeutics? Neurotherapeutics *11*, 92–110.

Lee, H.K. (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Research *14*, 1085–1094.

Lee, C.K., Weindruch, R., and Prolla, T.A. (2000). Gene-expression profile of the ageing brain in mice. Nature Genetics *25*, 294–297.

Lee, C.-K., Sunkin, S.M., Kuan, C., Thompson, C.L., Pathak, S., Ng, L., Lau, C., Fischer, S., Mortrud, M., Slaughterbeck, C., et al. (2008). Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. Genome Biology *9*, R23.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lennon, G.G. (2000). High-throughput gene expression analysis for drug discovery. Drug Discovery Today *5*, 59–66.

Lennon, P.A., Cooper, M.L., Peiffer, D.A., Gunderson, K.L., Patel, A., Peters, S., Cheung, S.W., and Bacino, C.A. (2007). Deletion of 7q31.1 supports involvement of FOXP2 in language impairment: clinical report and review. American Journal of Medical Genetics Part A *143A*, 791–798.

Lim, Y., and Golden, J.A. (2007). Patterning the developing diencephalon. Brain Research Reviews *53*, 17–26.

Lim, C.K.P., Ho, C.S.H., Chou, C.H.N., and Waye, M.M.Y. (2011). Association of the rs3743205 variant of DYX1C1 with dyslexia in Chinese children. Behavioral and Brain Functions *7*, 16.

Lind, P.A., Luciano, M., Wright, M.J., Montgomery, G.W., Martin, N.G., and Bates, T.C. (2010). Dyslexia and DCDC2: normal variation in reading and spelling is associated with DCDC2 polymorphisms in an Australian population sample. European Journal of Human Genetics *18*, 668–673.

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

Lovick, T.A. (1985). Ventrolateral medullary lesions block the antinociceptive and cardiovascular responses elicited by stimulating the dorsal periaqueductal grey matter in rats: Pain *21*, 241–252.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B.A. (2004). Gene regulation and DNA damage in the ageing human brain. Nature *429*, 883–891.

Luciano, M., Lind, P.A., Duffy, D.L., Castles, A., Wright, M.J., Montgomery, G.W., Martin, N.G., and Bates, T.C. (2007). A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading and spelling ability. Biological Psychiatry *62*, 811–817.

Ludlow, C.L., Rosenberg, J., Salazar, A., Grafman, J., and Smutok, M. (1987). Site of penetrating brain lesions causing chronic acquired stuttering. Annals of Neurology *22*, 60–66.

MacDermot, K.D., Bonora, E., Sykes, N., Coupe, A.-M., Lai, C.S.L., Vernes, S.C., Vargha-Khadem, F., McKenzie, F., Smith, R.L., Monaco, A.P., et al. (2005). Identification of FOXP2 Truncation as a Novel Cause of Developmental Speech and Language Deficits. The American Journal of Human Genetics *76*, 1074–1080.

Maestrini, E., Pagnamenta, A.T., Lamb, J.A., Bacchelli, E., Sykes, N.H., Sousa, I., Toma, C., Barnby, G., Butler, H., Winchester, L., et al. (2010). High-density SNP association study and copy number variation analysis of the AUTS1 and AUTS5 loci implicate the IMMP2L–DOCK4 gene region in autism susceptibility. Molecular Psychiatry *15*, 954–968.

Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M.J., and Lelieveldt, B.P.F. (2015). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. Methods *73*, 79–89.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research *27*, 209–220.

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C.H.S., Beaton, A., Desmond, J., De Witte, E., Fawcett, A.J., Hertrich, I., Küper, M., et al. (2013). Consensus Paper: Language and the Cerebellum: an Ongoing Enigma. The Cerebellum.

Marino, C., Giorda, R., Luisa Lorusso, M., Vanzin, L., Salandi, N., Nobile, M., Citterio, A., Beri, S., Crespi, V., Battaglia, M., et al. (2005). A family-based association study does not support DYX1C1 on 15q21.3 as a candidate gene in developmental dyslexia. European Journal of Human Genetics *13*, 491–499.

Marino, C., Citterio, A., Giorda, R., Facoetti, A., Menozzi, G., Vanzin, L., Lorusso, M.L., Nobile, M., and Molteni, M. (2007). Association of short-term memory with a variant within DYX1C1 in developmental dyslexia. Genes, Brain and Behavior *6*, 640–646.

Marseglia, G., Scordo, M.R., Pescucci, C., Nannetti, G., Biagini, E., Scandurra, V., Gerundino, F., Magi, A., Benelli, M., and Torricelli, F. (2012). 372 kb microdeletion in

18q12.3 causing SETBP1 haploinsufficiency associated with mild mental retardation and expressive speech impairment. European Journal of Medical Genetics *55*, 216–221.

Marshall, C.R., Denmark, T., and Morgan, G. (2006). Investigating the underlying causes of SLI: A non-sign repetition test in British Sign Language. Advances in Speech Language Pathology *8*, 347–355.

Martin, T.A., Keating, J.G., Goodkin, H.P., Bastian, A.J., and Thach, W.T. (1996). Throwing while looking through prisms: I. Focal olivocerebellar lesions impair adaptation. Brain *119*, 1183–1198.

Martínez-Cerdeño, V., Noctor, S.C., and Kriegstein, A.R. (2006). Estradiol stimulates progenitor cell division in the ventricular and subventricular zones of the embryonic neocortex. European Journal of Neuroscience *24*, 3475–3488.

Mascheretti, S., Bureau, A., Battaglia, M., Simone, D., Quadrelli, E., Croteau, J., Cellino, M.R., Giorda, R., Beri, S., Maziade, M., et al. (2013). An assessment of gene-by-environment interactions in developmental dyslexia-related phenotypes. Genes, Brain and Behavior *12*, 47–55.

Mason, K., Rowley, K., Marshall, C.R., Atkinson, J.R., Herman, R., Woll, B., and Morgan, G. (2010). Identifying specific language impairment in deaf children acquiring British Sign Language: Implications for theory and practice. British Journal of Developmental Psychology *28*, 33–49.

Matsson, H., Tammimies, K., Zucchelli, M., Anthoni, H., Onkamo, P., Nopola-Hemmi, J., Lyytinen, H., Leppanen, P.H.T., Neuhoff, N., Warnke, A., et al. (2011). SNP Variations in the 7q33 Region Containing DGKI are Associated with Dyslexia in the Finnish and German Populations. Behavior Genetics *41*, 134–140.

Matsuda, S., Miura, E., Matsuda, K., Kakegawa, W., Kohda, K., Watanabe, M., and Yuzaki, M. (2008). Accumulation of AMPA Receptors in Autophagosomes in Neuronal Axons Lacking Adaptor Protein AP-4. Neuron *57*, 730–745.

McAllister, A.K., Katz, L.C., and Lo, D.C. (1999). NEUROTROPHINS AND SYNAPTIC PLASTICITY. Annual Review of Neuroscience *22*, 295–318.

McClung, C.A., and Nestler, E.J. (2003). Regulation of gene expression and cocaine reward by CREB and DeltaFosB. Nature Neuroscience *6*, 1208–1215.

Meng, H., Smith, S.D., Hager, K., Held, M., Liu, J., Olson, R.K., Pennington, B.F., DeFries, J.C., Gelernter, J., O'Reilly-Pol, T., et al. (2005a). DCDC2 is associated with reading disability and modulates neuronal development in the brain. Proceedings of the National Academy of Sciences *102*, 17053–17058.

Meng, H., Hager, K., Held, M., Page, G.P., Olson, R.K., Pennington, B.F., DeFries, J.C., Smith, S.D., and Gruen, J.R. (2005b). TDT-association analysis of EKN1 and dyslexia in a Colorado twin cohort. Human Genetics *118*, 87–90.

Menon, V., Kwon, H., Eliez, S., Taylor, A.K., and Reiss, A.L. (2000). Functional brain activation during cognition is related to FMR1 gene expression. Brain Research *877*, 367–370.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., and Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel) *4*, 1180–1211.

Miller, J.A., Horvath, S., and Geschwind, D.H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proceedings of the National Academy of Sciences *107*, 12698–12703.

Mody, M., Cao, Y., Cui, Z., Tay, K.Y., Shyong, A., Shimizu, E., Pham, K., Schultz, P., Welsh, D., and Tsien, J.Z. (2001). Genome-wide gene expression profiles of the developing mouse hippocampus. Proceedings of the National Academy of Sciences *98*, 8862–8867.

Montandon, G., Qin, W., Liu, H., Ren, J., Greer, J.J., and Horner, R.L. (2011). PreBotzinger Complex Neurokinin-1 Receptor-Expressing Neurons Mediate Opioid-Induced Respiratory Depression. Journal of Neuroscience *31*, 1292–1301.

Moreno-De-Luca, A., Helmers, S.L., Mao, H., Burns, T.G., Melton, A.M.A., Schmidt, K.R., Fernhoff, P.M., Ledbetter, D.H., and Martin, C.L. (2011). Adaptor protein complex-4 (AP-4) deficiency causes a novel autosomal recessive cerebral palsy syndrome with microcephaly and intellectual disability. Journal of Medical Genetics *48*, 141–144.

Morris, J.A., Jordan, C.L., and Breedlove, S.M. (2004). Sexual differentiation of the vertebrate nervous system. Nature Neuroscience *7*, 1034–1039.

Myers, E.M., Bartlett, C.W., Machiraju, R., and Bohland, J.W. (2015). An integrative analysis of regional gene expression profiles in the human brain. Methods *73*, 54–70.

NCBI Resource Coordinators (2015). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research *43*, D6–D17.

Newbury, D.F., Bishop, D.V.M., and Monaco, A.P. (2005). Genetic influences on language impairment and phonological short-term memory. Trends in Cognitive Sciences *9*, 528–534.

Newbury, D.F., Winchester, L., Addis, L., Paracchini, S., Buckingham, L.-L., Clark, A., Cohen, W., Cowie, H., Dworzynski, K., Everitt, A., et al. (2009). CMIP and ATP2C2 Modulate Phonological Short-Term Memory in Language Impairment. The American Journal of Human Genetics *85*, 264–272.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F., Talcott, J.B., and Monaco, A.P. (2011). Investigation of Dyslexia and SLI Risk Variants in Reading- and Language-Impaired Subjects. Behavior Genetics *41*, 90–104.

Ng, L., Pathak, S.D., Chihchau Kuan, Lau, C., Hongwei Dong, Sodt, A., Chinh Dang, Avants, B., Yushkevich, P., Gee, J.C., et al. (2007). Neuroinformatics for Genome-Wide 3-D Gene Expression Mapping in the Mouse Brain. IEEE/ACM Transactions on Computational Biology and Bioinformatics *4*, 382–393.

Ng, L., Bernard, A., Lau, C., Overly, C.C., Dong, H.-W., Kuan, C., Pathak, S., Sunkin, S.M., Dang, C., Bohland, J.W., et al. (2009). An anatomic gene expression atlas of the adult mouse brain. Nature Neuroscience *12*, 356–362.

Nicolson, R.I., and Fawcett, A.J. (2007). Procedural learning difficulties: reuniting the developmental disorders? Trends in Neurosciences *30*, 135–141.

Nielsen, C.K., Simms, J.A., Li, R., Mill, D., Yi, H., Feduccia, A.A., Santos, N., and Bartlett, S.E. (2012). -Opioid Receptor Function in the Dorsal Striatum Plays a Role in High Levels of Ethanol Consumption in Rats. Journal of Neuroscience *32*, 4540–4552.

Oberheim, N.A., Wang, X., Goldman, S., and Nedergaard, M. (2006). Astrocytic complexity distinguishes the human brain. Trends in Neurosciences *29*, 547–553.

Oberheim, N.A., Takano, T., Han, X., He, W., Lin, J.H.C., Wang, F., Xu, Q., Wyatt, J.D., Pilcher, W., Ojemann, J.G., et al. (2009). Uniquely Hominid Features of Adult Human Astrocytes. Journal of Neuroscience *29*, 3276–3287.

O'Brien, E.K., Zhang, X., Nishimura, C., Tomblin, J.B., and Murray, J.C. (2003). Association of Specific Language Impairment (SLI) to the Region of 7q31. The American Journal of Human Genetics *72*, 1536–1543.

Oldham, M.C., Horvath, S., and Geschwind, D.H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences *103*, 17973–17978.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D.H. (2008). Functional organization of the transcriptome in human brain. Nature Neuroscience *11*, 1271–1282.

O'Leary, D.D.M., Chou, S.-J., and Sahara, S. (2007). Area patterning of the mammalian cortex. Neuron *56*, 252–269.

Olsson, M., Björklund, A., and Campbell, K. (1998). Early specification of striatal projection neurons and interneuronal subtypes in the lateral and medial ganglionic eminence. Neuroscience *84*, 867–876.

Pagnamenta, A.T., Bacchelli, E., de Jonge, M.V., Mirza, G., Scerri, T.S., Minopoli, F., Chiocchetti, A., Ludwig, K.U., Hoffmann, P., Paracchini, S., et al. (2010). Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia. Biological Psychiatry *68*, 320–328.

Paracchini, S. (2006). The chromosome 6p22 haplotype associated with dyslexia reduces the expression of KIAA0319, a novel gene involved in neuronal migration. Human Molecular Genetics *15*, 1659–1666.

Paracchini, S., Steer, C.D., Buckingham, L.-L., Morris, A.P., Ring, S., Scerri, T., Stein, J., Pembrey, M.E., Ragoussis, J., Golding, J., et al. (2008). Association of the KIAA0319 dyslexia susceptibility gene with reading skills in the general population. American Journal of Psychiatry *165*, 1576–1584.

Paracchini, S., Ang, Q.W., Stanley, F.J., Monaco, A.P., Pennell, C.E., and Whitehouse, A.J.O. (2011). Analysis of dyslexia candidate genes in the Raine cohort representing the general Australian population. Genes, Brain and Behav *10*, 158–165.

Pavlidis, P., and Noble, W.S. (2001). Analysis of strain and regional variation in gene expression in mouse brain. Genome Biology *2*, RESEARCH0042.

Peeva, M.G., Guenther, F.H., Tourville, J.A., Nieto-Castanon, A., Anton, J.-L., Nazarian, B., and Alario, F.-X. (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. NeuroImage *50*, 626–638.

Peñagarikano, O., and Geschwind, D.H. (2012). What does CNTNAP2 reveal about autism spectrum disorder? Trends in Molecular Medicine *18*, 156–163.

Peter, B., Raskind, W.H., Matsushita, M., Lisowski, M., Vu, T., Berninger, V.W., Wijsman, E.M., and Brkanac, Z. (2011). Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. Journal of Neurodevelopmental Disorders *3*, 39–49.

Peter, B., Matsushita, M., Oda, K., and Raskind, W. (2014). De novo microdeletion of *BCL11A* is associated with severe speech sound disorder. American Journal of Medical Genetics Part A *164*, 2091–2096.

Pfenning, A.R., Hara, E., Whitney, O., Rivas, M.V., Wang, R., Roulhac, P.L., Howard, J.T., Wirthlin, M., Lovell, P.V., Ganapathy, G., et al. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. Science *346*, 1256846–1256846.

Pletikos, M., Sousa, A.M.M., Sedmak, G., Meyer, K.A., Zhu, Y., Cheng, F., Li, M., Kawasawa, Y.I., and Sestan, N. (2014). Temporal specification and bilaterality of human neocortical topographic gene expression. Neuron *81*, 321–332.

Plomin, R., and Kovas, Y. (2005). Generalist genes and learning disabilities. Psychological Bulletin *131*, 592–617.

Poelmans, G., Engelen, J.J.M., Van Lent-Albrechts, J., Smeets, H.J., Schoenmakers, E., Franke, B., Buitelaar, J.K., Wuisman-Frerker, M., Erens, W., Steyaert, J., et al. (2009). Identification of novel dyslexia candidate genes through the analysis of a chromosomal deletion. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics *150B*, 140–147.

Poelmans, G., Buitelaar, J.K., Pauls, D.L., and Franke, B. (2011). A theoretical molecular network for dyslexia: integrating available genetic findings. Molecular Psychiatry *16*, 365–382.

Pruunsild, P., Sepp, M., Orav, E., Koppel, I., and Timmusk, T. (2011). Identification of cis-Elements and Transcription Factors Regulating Neuronal Activity-Dependent Transcription of Human BDNF Gene. Journal of Neuroscience *31*, 3295–3308.

Puelles, L., and Rubenstein, J.L.R. (2003). Forebrain gene expression domains and the evolving prosomeric model. Trends in Neuroscience *26*, 469–476.

Ramachandra, N., Saviour, P., Kumar, S., Kiran, U., Ravuri, R., and Rao, V. (2008). Allelic variants of <i> DYX1C1</i> are not associated with dyslexia in India. Indian Journal of Human Genetics *14*, 99.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell *140*, 744–752.

Ray, R.S., and Dymecki, S.M. (2009). Rautenlippe Redux—toward a unified view of the precerebellar rhombic lip. Current Opinion in Cell Biology *21*, 741–747.

Raza, M.H., Mattera, R., Morell, R., Sainz, E., Rahn, R., Gutierrez, J., Paris, E., Root, J., Solomon, B., Brewer, C., et al. (2015). Association between Rare Variants in AP4E1, a Component of Intracellular Trafficking, and Persistent Stuttering. American Journal of Human Genetics *97*, 715–725.

Redgrave, P., Prescott, T.J., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? Neuroscience *89*, 1009–1023.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Research *44*, W83–W89.

Rice, M.L., Smith, S.D., and Gayán, J. (2009). Convergent genetic linkage and associations to language, speech and reading measures in families of probands with Specific Language Impairment. Journal of Neurodevelopmental Disorders *1*, 264–282.

Richiardi, J., Altmann, A., Milazzo, A.-C., Chang, C., Chakravarty, M.M., Banaschewski, T., Barker, G.J., Bokde, A.L.W., Bromberg, U., Büchel, C., et al. (2015). BRAIN NETWORKS. Correlated gene expression supports synchronous activity in brain networks. Science *348*, 1241–1244.

Riecker, A., Mathiak, K., Wildgruber, D., Erb, M., Hertrich, I., Grodd, W., and Ackermann, H. (2005). fMRI reveals two distinct cerebral networks subserving speech motor control. Neurology *64*, 700–706.

Ritchie, M.E., Dunning, M.J., Smith, M.L., Shi, W., and Lynch, A.G. (2011). BeadArray Expression Analysis Using Bioconductor. PLoS Computational Biology *7*, e1002276.

Rizzi, T.S., van der Sluis, S., Derom, C., Thiery, E., van Kesteren, R.E., Jacobs, N., Van Gestel, S., Vlietinck, R., Verhage, M., Heutink, P., et al. (2013). FADS2 Genetic Variance in Combination with Fatty Acid Intake Might Alter Composition of the Fatty Acids in Brain. PLoS ONE *8*, e68000.

Roll, P. (2006). SRPX2 mutations in disorders of language cortex and cognition. Human Molecular Genetics *15*, 1195–1207.

Roll, P., Vernes, S.C., Bruneau, N., Cillario, J., Ponsole-Lenfant, M., Massacrier, A., Rudolf, G., Khalife, M., Hirsch, E., Fisher, S.E., et al. (2010). Molecular networks implicated in speech-related disorders: FOXP2 regulates the SRPX2/uPAR complex. Human Molecular Genetics *19*, 4848–4860.

Roth, R.B., Hevezi, P., Lee, J., Willhite, D., Lechner, S.M., Foster, A.C., and Zlotnik, A. (2006). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics *7*, 67–80.

Rumsey, J.M., Donohue, B.C., Brady, D.R., Nace, K., Giedd, J.N., and Andreason, P. (1997). A magnetic resonance imaging study of planum temporale asymmetry in men with developmental dyslexia. Archives of Neurology *54*, 1481–1489.

Sacktor, T. (2012). Memory maintenance by PKMζ — an evolutionary perspective. Molecular Brain *5*, 31.

Sanes, D.H., Reh, T.A., and Harris, W.A. (2012). Development of the nervous system (Amsterdam ; Boston : Burlington, MA: Elsevier ; Academic Press).

Sansom, S.N., and Livesey, F.J. (2009). Gradients in the Brain: The Control of the Development of Form and Function in the Cerebral Cortex. Cold Spring Harbor Perspectives in Biology *1*, a002519–a002519.

Sarkisyan, D., Hussain, M.Z., Watanabe, H., Kononenko, O., Bazov, I., Zhou, X., Yamskova, O., Krishtal, O., Karpyak, V.M., Yakovleva, T., et al. (2015). Downregulation of the endogenous opioid peptides in the dorsal striatum of human alcoholics. Frontiers in Cellular Neuroscience *9*.

Scerri, T.S. (2004). Putative functional alleles of DYX1C1 are not associated with dyslexia susceptibility in a large sample of sibling pairs from the UK. Journal of Medical Genetics *41*, 853–857.

Scerri, T.S., Morris, A.P., Buckingham, L.-L., Newbury, D.F., Miller, L.L., Monaco, A.P., Bishop, D.V.M., and Paracchini, S. (2011). DCDC2, KIAA0319 and CMIP Are Associated with Reading-Related Traits. Biological Psychiatry *70*, 237–245.

Scholz, V.H., Flaherty, A.W., Kraft, E., Keltner, J.R., Kwong, K.K., Chen, Y.I., Rosen, B.R., and Jenkins, B.G. (2000). Laterality, somatotopy and reproducibility of the basal ganglia and motor cortex during motor tasks11Published on the World Wide Web on 28 August 2000. Brain Research *879*, 204–215.

Schulte, E.C., Stahl, I., Czamara, D., Ellwanger, D.C., Eck, S., Graf, E., Mollenhauer, B., Zimprich, A., Lichtner, P., Haubenberger, D., et al. (2013). Rare Variants in PLXNA4 and Parkinson's Disease. PLoS ONE *8*, e79145.

Schumacher, J., Anthoni, H., Dahdouh, F., König, I.R., Hillmer, A.M., Kluck, N., Manthey, M., Plume, E., Warnke, A., Remschmidt, H., et al. (2006). Strong Genetic Evidence of DCDC2 as a Susceptibility Gene for Dyslexia. The American Journal of Human Genetics *78*, 52–62.

Seeger, M., Tear, G., Ferres-Marco, D., and Goodman, C.S. (1993). Mutations affecting growth cone guidance in drosophila: Genes necessary for guidance toward or away from the midline. Neuron *10*, 409–426.

Settembre, C., Fraldi, A., Medina, D.L., and Ballabio, A. (2013). Signals from the lysosome: a control centre for cellular clearance and energy metabolism. Nature Reviews Molecular Cell Biology *14*, 283–296.

Shaywitz, S.E., and Shaywitz, B.A. (2005). Dyslexia (Specific Reading Disability). Biological Psychiatry *57*, 1301–1309.

Shi, W., Oshlack, A., and Smyth, G.K. (2010). Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Research *38*, e204–e204.

Sia, G.M., Clem, R.L., and Huganir, R.L. (2013). The Human Language-Associated Gene SRPX2 Regulates Synapse Formation and Vocalization in Mice. Science *342*, 987–991.

Sigmund, C.D. (2000). Viewpoint: are studies in genetically altered mice out of control? Arteriosclerosis Thrombosis and Vascular Biology *20*, 1425–1429.

Simmons, T.R., Flax, J.F., Azaro, M.A., Hayter, J.E., Justice, L.M., Petrill, S.A., Bassett, A.S., Tallal, P., Brzustowicz, L.M., and Bartlett, C.W. (2010). Increasing Genotype-Phenotype Model Determinism: Application to Bivariate Reading/Language Traits and Epistatic Interactions in Language-Impaired Families. Human Heredity *70*, 232–244.

Sir, J.-H., Barr, A.R., Nicholas, A.K., Carvalho, O.P., Khurshid, M., Sossick, A., Reichelt, S., D'Santos, C., Woods, C.G., and Gergely, F. (2011). A primary microcephaly protein complex forms a ring around parental centrioles. Nature Genetics *43*, 1147–1153.

SLI Consortium (2002). A genomewide scan identifies two novel loci involved in specific language impairment. American Journal of Human Genetics *70*, 384–398.

Smith, J.D., Meehan, M.H., Crean, J., and McCann, A. (2011). Alpha T-catenin (CTNNA3): a gene in the hand is worth two in the nest. Cellular and Molecular Life Sciences *68*, 2493–2498.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (New York, NY: Springer New York), pp. 397–420.

Stager, S.V., Calis, K., Grothe, D., Bloch, M., Berensen, N.M., Smith, P.J., and Braun, A. (2005). Treatment with medications affecting dopaminergic and serotonergic mechanisms: effects on fluency and anxiety in persons who stutter. Journal of Fluency Disorders *30*, 319–335.

Stern, C.D. (2005). Neural induction: old problem, new findings, yet more questions. Development *132*, 2007–2021.

Stoodley, C.J. (2012). The Cerebellum and Cognition: Evidence from Functional Imaging Studies. The Cerebellum *11*, 352–365.

Strand, A.D., Aragaki, A.K., Baquet, Z.C., Hodges, A., Cunningham, P., Holmans, P., Jones, K.R., Jones, L., Kooperberg, C., and Olson, J.M. (2007). Conservation of Regional Gene Expression in Mouse and Human Brain. PLoS Genetics *3*, e59.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.

Sun, T. (2005). Early Asymmetry of Gene Transcription in Embryonic Human Left and Right Cerebral Cortex. Science *308*, 1794–1798.

Suto, F., Ito, K., Uemura, M., Shimizu, M., Shinkawa, Y., Sanbo, M., Shinoda, T., Tsuboi, M., Takashima, S., Yagi, T., et al. (2005). Plexin-a4 mediates axon-repulsive activities of both secreted and transmembrane semaphorins and plays roles in nerve fiber guidance. Journal of Neuroscience *25*, 3628–3637.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., Haltia, T., Myllyluoma, B., Lyytinen, H., Muller, K., Kaaranen, M., Lindsberg, P.J., Hannula-Jouppi, K., et al. (2003). A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. Proceedings of the National Academy of Sciences *100*, 11553–11558.

Tani, T., and Sakai, Y. (2011). Analysis of five cases with neurogenic stuttering following brain injury in the basal ganglia. Journal of Fluency Disorders *36*, 1–16.

Thevenon, J., Callier, P., Andrieux, J., Delobel, B., David, A., Sukno, S., Minot, D., Mosca Anne, L., Marle, N., Sanlaville, D., et al. (2013). 12p13.33 microdeletion including ELKS/ERC1, a new locus associated with childhood apraxia of speech. European Journal of Human Genetics *21*, 82–88.

Theys, C., De Nil, L., Thijs, V., van Wieringen, A., and Sunaert, S. (2013). A crucial role for the cortico-striato-cortical loop in the pathogenesis of stroke-related neurogenic stuttering: Neural Network of Neurogenic Stuttering. Human Brain Mapping *34*, 2103–2112.

Thompson, C.L., Ng, L., Menon, V., Martinez, S., Lee, C.-K., Glattfelder, K., Sunkin, S.M., Henry, A., Lau, C., Dang, C., et al. (2014). A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. Neuron *83*, 309–323.

Tikhonov, D.B., and Magazanik, L.G. (2009). Origin and Molecular Evolution of Ionotropic Glutamate Receptors. Neuroscience and Behavioral Physiology *39*, 763–773.

Tjen-A-Looi, S.C., Li, P., and Longhurst, J.C. (2007). Role of medullary GABA, opioids, and nociceptin in prolonged inhibition of cardiovascular sympathoexcitatory reflexes during electroacupuncture in cats. American Journal of Physiology: Heart and Circulatory Physiology *293*, H3627–H3635.

Tolosa, A., Sanjuán, J., Dagnall, A.M., Moltó, M.D., Herrero, N., and de Frutos, R. (2010). FOXP2 gene and language impairment in schizophrenia: association and epigenetic studies. BMC Medical Genetics *11*.

Tukey, J.W. (1977). Exploratory data analysis (Reading: Addison-Wesley).

Ueda, S., Fujimoto, S., Hiramoto, K., Negishi, M., and Katoh, H. (2008). Dock4 regulates dendritic development in hippocampal neurons. Journal of Neuroscience Research *86*, 3052–3061.

Ullman, M.T., and Pierpont, E.I. (2005). Specific Language Impairment is not Specific to Language: the Procedural Deficit Hypothesis. Cortex *41*, 399–433.

Vallipuram, J., Grenville, J., and Crawford, D.A. (2010). The E646D-ATP13A4 Mutation Associated with Autism Reveals a Defect in Calcium Regulation. Cellular and Molecular Neurobiology *30*, 233–246.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research *9*, 85.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics *10*, 252–263.

Vargha-Khadem, F., Watkins, K.E., Price, C.J., Ashburner, J., Alcock, K.J., Connelly, A., Frackowiak, R.S.J., Friston, K.J., Pembrey, M.E., Mishkin, M., et al. (1998). Neural basis of an inherited speech and language disorder. Proceedings of the National Academy of Sciences *95*, 12695–12700.

Venkatesh, S.K., Siddaiah, A., Padakannaya, P., and Ramachandra, N.B. (2011). An examination of candidate gene SNPs for dyslexia in an Indian sample. Behavioral Genetics *41*, 105–109.

Venkatesh, S.K., Siddaiah, A., Padakannaya, P., and Ramachandra, N.B. (2013a). Analysis of genetic variants of dyslexia candidate genes KIAA0319 and DCDC2 in Indian population. Journal of Human Genetics *58*, 531–538.

Venkatesh, S.K., Siddaiah, A., Padakannaya, P., and Ramachandra, N.B. (2013b). Lack of association between genetic polymorphisms in ROBO1, MRPL19/C2ORF3 and THEM2 with developmental dyslexia. Gene *529*, 215–219.

Vernes, S.C., Spiteri, E., Nicod, J., Groszer, M., Taylor, J.M., Davies, K.E., Geschwind, D.H., and Fisher, S.E. (2007). High-Throughput Analysis of Promoter Occupancy Reveals Direct Neural Targets of FOXP2, a Gene Mutated in Speech and Language Disorders. The American Journal of Human Genetics *81*, 1232–1250.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcón, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., et al. (2008). A Functional Genetic Link between Distinct Developmental Language Disorders. New England Journal of Medicine *359*, 2337–2345.

Villanueva, P., Nudel, R., Hoischen, A., Fernández, M.A., Simpson, N.H., Gilissen, C., Reader, R.H., Jara, L., Echeverry, M.M., Francks, C., et al. (2015). Exome Sequencing in an Admixed Isolated Population Indicates NFXL1 Variants Confer a Risk for Specific Language Impairment. PLOS Genetics *11*, e1004925.

Wang, V.Y., Rose, M.F., and Zoghbi, H.Y. (2005). Math1 Expression Redefines the Rhombic Lip Derivatives and Reveals Novel Lineages within the Brainstem and Cerebellum. Neuron *48*, 31–43.

Wang, Y., Paramasivam, M., Thomas, A., Bai, J., Kaminen-Ahola, N., Kere, J., Voskuil, J., Rosen, G.D., Galaburda, A.M., and Loturco, J.J. (2006). DYX1C1 functions in neuronal migration in developing neocortex. Neuroscience *143*, 515–522.

Watkins, K.E., Vargha-Khadem, F., Ashburner, J., Passingham, R.E., Connelly, A., Friston, K.J., Frackowiak, R.S.J., Mishkin, M., and Gadian, D.G. (2002). MRI analysis of an inherited speech and language disorder: structural brain abnormalities. Brain *125*, 465–478.

Weckerly, J., Wulfeck, B., and Reilly, J. (2001). Verbal Fluency Deficits in Children With Specific Language Impairment: Slow Rapid Naming or Slow to Name? Child Neuropsychology (Neuropsychology, Development and Cognition: Section C) *7*, 142–152.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., Somerville, C., and Loraine, A. (2006). Transcriptional Coordination of the Metabolic Network in Arabidopsis. Plant Physiology *142*, 762–774.

White, J.M., and Irvine, R.J. (1999). Mechanisms of fatal opioid overdose. Addiction *94*, 961–972.

Whitehouse, A.J.O., Bishop, D.V.M., Ang, Q.W., Pennell, C.E., and Fisher, S.E. (2011). *CNTNAP2* variants affect early language development in the general population. Genes, Brain and Behavior *10*, 451–456.

Wiegreffe, C., Simon, R., Peschkes, K., Kling, C., Strehle, M., Cheng, J., Srivatsa, S., Liu, P., Jenkins, N.A., Copeland, N.G., et al. (2015). Bcl11a (Ctip1) Controls Migration of Cortical Projection Neurons through Regulation of Sema3c. Neuron *87*, 311–325.

Wigg, K.G., Couto, J.M., Feng, Y., Anderson, B., Cate-Carter, T.D., Macciardi, F., Tannock, R., Lovett, M.W., Humphries, T.W., and Barr, C.L. (2004). Support for EKN1 as the susceptibility locus for dyslexia on 15q21. Molecular Psychiatry *9*, 1111–1121.

Wilcke, A., Weissfuss, J., Kirsten, H., Wolfram, G., Boltze, J., and Ahnert, P. (2009). The role of gene DCDC2 in German dyslexics. Annals of Dyslexia *59*, 1–11.

Wildgruber, D., Ackermann, H., and Grodd, W. (2001). Differential Contributions of Motor Cortex, Basal Ganglia, and Cerebellum to Speech Motor Control: Effects of Syllable Repetition Rate Evaluated by fMRI. NeuroImage *13*, 101–109.

Wingate, R.J. (2001). The rhombic lip and early cerebellar development. Current Opinion in Neurobiology *11*, 82–88.

Wolock, S.L., Yates, A., Petrill, S.A., Bohland, J.W., Blair, C., Li, N., Machiraju, R., Huang, K., and Bartlett, C.W. (2013). Gene × smoking interactions on human brain gene expression: finding common mechanisms in adolescents and adults. Journal of Child Psychology and Psychiatry *54*, 1109–1119.

Worthey, E.A., Raca, G., Laffin, J.J., Wilk, B.M., Harris, J.M., Jakielski, K.J., Dimmock, D.P., Strand, E.A., and Shriberg, L.D. (2013). Whole-exome sequencing supports genetic heterogeneity in childhood apraxia of speech. Journal of Neurodevelopmental Disorders *5*, 29.

Yang, J., Seo, J., Nair, R., Han, S., Jang, S., Kim, K., Han, K., Paik, S.K., Choi, J., Lee, S., et al. (2011). DGKι regulates presynaptic release during mGluR-dependent LTD. EMBO Journal *30*, 165–180.

Yip, A.M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics *8*, 22.

Zapala, M.A., Hovatta, I., Ellison, J.A., Wodicka, L., Del Rio, J.A., Tennant, R., Tynan, W., Broide, R.S., Helton, R., Stoveken, B.S., et al. (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. Proceedings of the National Academy of Sciences *102*, 10357–10362.

Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L., et al. (2012). Large-Scale Cellular-Resolution Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures. Cell *149*, 483–496.

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology *4*.

Zhang, Y., Li, J., Tardif, T., Burmeister, M., Villafuerte, S.M., McBride-Chang, C., Li, H., Shi, B., Liang, W., Zhang, Z., et al. (2012). Association of the DYX1C1 dyslexia susceptibility gene with orthography in the Chinese population. PLoS ONE *7*, e42969.

Zirlinger, M., Kreiman, G., and Anderson, D.J. (2001). Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. Proceedings of the National Academy of Sciences *98*, 5270–5275.

(2013). Diagnostic and statistical manual of mental disorders: DSM-5 (Washington, D.C: American Psychiatric Association).

CURRICULUM VITAE