

Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

**COMPARATIVE
NEUROTRANSCRIPTOMICS IN
MAMMALS AND BIRDS**

T. Grant Belgard

NEW COLLEGE

OXFORD

MICHAELMAS TERM

2011

COMPARATIVE NEUROTRANSCRIPTOMICS IN MAMMALS AND BIRDS

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

T. Grant Belgard, New College, Michaelmas Term 2011

Abstract

In this thesis I apply new sequencing technologies and analytical methods derived from genomics and computer science to the neuroanatomy of gene expression. The first project explores characteristics of gene expression across adult neocortical layers in a representative mammal – the mouse. Amongst the thousands of genes and transcripts differentially expressed across layers, I found common functional characteristics of genes that define certain layers, candidate cases of isoform switching, and over a thousand apparent long intergenic non-coding RNA transcripts. The second project compares patterns of gene expression in the structurally diverged adult derivatives of the pallium in mice and chickens. Overall, gene expression levels were moderately correlated between the two species. While expression patterns of ‘marker’ genes were only poorly conserved in these regions, there nevertheless was significant conservation of cross-species marker genes for homologous structures, cell types and functionally analogous regions. Many aspects of these data from both projects can now be easily browsed and searched from custom-built web interfaces. In addition to generating unprecedented genome-wide resources for the neuroscience community to explore the functional and structural dimensions of gene expression amongst different pallial regions in mammals and birds, this work also provides new insights into the widespread evolutionary shuffling of adult marker gene expression.

Acknowledgements

First and foremost I am sincerely grateful for my supervisors Prof. Chris Ponting and Dr. Elliott Margulies for their encouragement, ideas, time, and advice.

I thank my collaborators, without whom these data would not exist to analyse and the predictions would not have been validated. We also had many valuable discussions. In particular, I thank Prof. Zoltán Molnár, Dr. Ana Marques, Dr. Peter Oliver, Dr. Tamara Sirey, Hatice Ozel Abaan, Dr. Anna Hoerder-Suabedissen, Dr. Juan Montiel, Dr. Wei-Zhi Wang, and Dr. Fernando Garcia-Moreno.

I am also grateful to Prof. Ponting and Dr. Montiel for their helpful feedback on this thesis.

I further thank Dr. Andreas Heger, Dr. Leo Goodstadt and Dr. Martin Goodson for helpful discussions and programming advice and Dr. Caleb Webber, Dr. Rebecca Chodroff, Dr. Chris Nellaker and Dr. Gerton Lunter for helpful discussions throughout the project.

I thank my family and Angel Zhang for their moral support. I appreciate the many members of the Ponting and Margulies groups, past and present, for helping to make my graduate experience so enjoyable.

Finally, I acknowledge the Marshall Scholarship, New College, the Intramural Research Program of the National Human Genome Research Institute and the MRC Functional Genomics Unit for my primary financial support. I also thank the NIH-Oxford-Cambridge Scholars Program; the Oxford Department of Physiology, Anatomy and Genetics; the Guarantors of Brain; the NIH Office of Intramural Training and Education; the Foundation for Advanced Education in the Sciences; the International Biomedical Research Alliance; and the Okinawa Institute of Science and Technology for providing travel and housing support for conferences, workshops and courses.

Table of Contents

Chapter 1 - Introduction	6
1.1 What is a Transcriptome?	6
1.2 Traditional Technologies for Measuring the Transcriptome	12
1.3 RNA-Seq: Short-Read Shotgun Sequencing cDNA	14
1.4 Navigating the Vertebrate Central Nervous System	17
1.5 Anatomy of the Neocortex	21
1.6 Prior Lines of Work in Comparative Neurotranscriptomics	24
1.7 Project Aims and Thesis Structure	28
Chapter 2 - Materials and Methods	30
2.1 RNA-Seq with Illumina Sequencing by Synthesis	30
2.2 Short Read Mapping with Tophat	33
2.3 Expression Quantification with Cufflinks	38
2.4 General Computational Tools	44
2.5 Sources of Functional Annotations	45
2.6 <i>In Situ</i> Hybridisation and the Allen Mouse Brain Atlas	46
Chapter 3 - Transcriptomes of Mouse Neocortical Layers	50
Abstract	50
3.1 Introduction	51
Publication and contributions	51
Background	51
Brief introduction to the naïve Bayes classifier	54
Chapter structure	58
3.2 Materials and Methods	58

Mice and initial dissections	58
Sequencing, read mapping, transcript building and quantification for the initial dissections.....	59
Sets of curated genes	63
Comparison with microarray data from juvenile mice	63
Classifier training and validation.....	64
Other machine learning methods considered.....	67
Classifier predictions of genes and transcripts	68
Estimating the prevalence of patterned genes	69
Functional differences	69
Simulating the null distributions for assessing functional differences in individual predicted sets	71
Correcting for multiple testing in testing functional differences.....	74
Replications in dorsal and lateral cortex.....	74
3.3 Results	78
Gene expression levels across dissected samples reflect known neocortical layers	78
Thousands of genes are patterned across neocortical layers	84
Isoform switching across layers.....	98
Biological associations for genes expressed highly in specific layers	103
Gene sets and functional differences replicated in dorsal cortex and lateral cortex	110
3.4 Discussion	119
Chapter 4 - Gene Expression Evolution in the Brains of Mammals and Birds.....	122
Abstract	122
4.1 Introduction.....	123
Gene expression is a phenotype that can be studied in an evolutionary context....	123

Types of biological similarity and sameness	124
The amniote pallium and debates over homologies.....	126
Comparative neurotranscriptomics in mouse and chick	139
WGCNA: Weighted Gene Co-expression Network Analysis	140
Chapter structure.....	142
4.2 Materials and Methods.....	143
Dissections and RNA extraction.....	143
Sequencing and gene expression quantification	147
Quality control for batch effects	149
Determining correspondence of the mouse cortical layer dissections	152
Confirming accuracy of the mouse dissections in other regions	153
Selecting genes for the cross-species comparison	163
Identifying gene co-expression modules	163
Comparing gene co-expression modules between species	165
Annotating gene co-expression modules	176
4.3 Results	191
Functional annotation enrichments amongst co-expressed genes in mouse and chick	191
Limited and weak conservation of gene co-expression patterns between mouse and chick.....	197
Matched co-expression modules in chick and mouse.....	199
Modules mark chick and mouse striatum	201
Modules mark chick and mouse hippocampus	202
Modules mark functionally analogous thalamic recipients in chick and mouse ...	202
Modules mark oligodendrocytes in chick and mouse.....	207
Batch effects confound interpretation of the final module overlap	207

4.4 Discussion	209
Gene connectivity is less conserved than median expression level.....	209
Functional enrichments guide module interpretation	210
Some modules are significantly conserved.....	213
Conserved striatal genes are functionally important in striatum	214
Many conserved hippocampal genes are presently of unknown function.....	216
Shared expression in functionally analogous areas may be evidence of convergent evolution	217
Conserved amniote markers of oligodendrocytes.....	220
Batch effects confound interpretation of the final module	220
Genes specifically expressed in mouse but not in chicken that have spatially-relevant functional consequences	221
Conclusions.....	222
Chapter 5 - Conclusions and Future Perspectives.....	225
References.....	229
Appendix A – Supplemental Methods for Chapter 3.....	253
Computational LincRNA Analysis Methods.....	253
Identification of cortical lincRNAs	253
LincRNA substitution rates	253
LincRNA associations with genomic features.....	254
Correlation of lincRNA expression with neighbouring protein-coding genes	255
<i>In Situ</i> Hybridisation Methods	255
Appendix B - Supplemental Figures and Tables for Chapter 3	257
Appendix C - Supplemental Figures and Tables for Chapter 4.....	279

Chapter 1 - Introduction

1.1 What is a Transcriptome?

Here I define a transcriptome as the complete set of RNAs found in a cell or group of cells at a given time. Transcriptomes are stochastic. The production, processing, trafficking, and breakdown of RNAs depend upon inherently stochastic molecular interactions. Transcriptomes are temporally regulated. This can be periodic, as with circadian (with a roughly 24 hour period, such as the human sleep-wake cycle), infradian (with a longer period, such as a menstrual cycle) and ultradian (with a shorter period, such as human appetite) rhythms (Bustos et al. 2011; Kawasaki et al. 2009; McMaster et al. 2011); or throughout stages of the cell cycle (Lenz and Sogaard-Andersen 2011). This can also be driven by various developmental factors. These factors can be signalled exogenously or endogenously and are processed in context of the epigenetic substrate of that particular cell's lineage (such as DNA methylation and chromatin modifications). This results in reproducible transcriptome changes at equivalent anatomical positions and developmental milestones (Kang et al. 2011). Transcriptomes can respond to affect, physiological insult, and pathological conditions. Even controlling for anatomical location, rhythmic effects, developmental time point, and environmental history, and allowing for stochastic changes, transcriptomes will not be equivalent for all members of an outbred species, neither in sequence nor in transcript abundance. Transcriptomes are also spatially organised within a cell. Some transcripts are primarily nuclear, others cytoplasmic. Some are trafficked to distant reaches of the cell (Mikl et al. 2010), while others form macromolecular complexes. Finally, a transcript's abundance in the

transcriptome is not necessarily proportional to its rate of transcription from the genome since transcripts are degraded at different rates (Hayles et al. 2010).

Elements of the transcriptome have an astounding variety of functions. Some join with proteins in macromolecular complexes. For example, ribosomal RNAs (rRNAs) are critical elements of the ribosomal complex; signal recognition particle RNA (SRP RNA) is part of a ribonucleoprotein critical for protein trafficking and secretion (Leung and Brown 2010); and telomerase RNA component (TERC) is part of the telomerase ribonucleoprotein involved in telomere extension (Zhang et al. 2011). Transfer RNAs (tRNAs) are adapters that translate trinucleotide DNA 'codons' to amino acids. RNAs known as ribozymes can catalyse chemical reactions. For example, ribonuclease P (RNase P) is a ribozyme that catalyses RNA cleavage (Jarrous and Gopalan 2010). Some small nuclear RNAs (snRNAs) join small nuclear ribonucleoprotein complexes (snRNPs) involved in RNA splicing (Douglas and Wood 2011). Small nucleolar RNAs (snoRNAs) mediate nucleoside modifications in rRNA (Gerbi et al. 2001). MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are both involved in the RNA interference (RNAi) pathway, crucial for post-transcriptional gene silencing (Liu and Paroo 2010). Other RNA classes, such as the products of endogenous retroviruses (ERVs) and retrotransposons, are often neutral or deleterious to the organism built from the host genome; as such there are numerous mechanisms to prevent the activity of these elements (Maksakova et al. 2008). One of these, piwi-interacting RNAs (piRNAs), form protein complexes with piwi proteins in an attempt to suppress retrotransposons in germ line cells (Senti and Brennecke 2010). Even this is not an exhaustive list of functions or RNA classes.

This thesis is concerned with two classes: messenger RNAs (mRNAs) and long intergenic noncoding RNAs (lincRNAs). Perhaps the best-known RNA class, mRNAs are ‘translated’ into proteins. Protein sequence is typically encoded in the “open reading frame” by triplet codons, but the noncanonical amino acid selenocysteine is encoded by a structural motif in the mRNA (Copeland and Driscoll 2001). Messenger RNA can also be ‘edited’ post-transcriptionally to change the protein sequence (Hogg et al. 2011). One mRNA ‘primary transcript’ can be capable of encoding multiple protein isoforms through a process known as alternative splicing, in which different segments of the primary transcript are removed prior to translation. Some mRNAs even contain overlapping open reading frames out of phase with one another (Chung et al. 2007; Nekrutenko and He 2006). Protein-coding sequence is frequently highly evolutionarily conserved, reflecting the functional constraints of the encoded proteins. Regulating their expression is important too: as described above, many ‘noncoding’ RNAs regulate levels of mRNAs. Promoters, enhancers, silencers and insulators are genomic elements that encode regulatory information rather than protein sequence. Variation in abundance of mRNAs is moderately correlated with variation in the abundance of the proteins they encode (Maier et al. 2009). This correlation should not be overstated, as there are numerous levels of regulation of protein abundance beyond the transcriptome including transcript sequestration, rate of translation, and protein longevity.

The second class of RNA discussed in this thesis is the lincRNA. I define these RNAs to have the following characteristics: (1) not translated, (2) longer than 200 nucleotides (nt), (3) not included in another recognised class of noncoding RNA (such as rRNA), and (4) transcribed from loci that do not overlap protein-coding loci (on either strand) (Ponting

and Belgard 2010). LincRNAs are a subset of a broader class of long noncoding RNAs (lncRNAs) for which the last constraint is relaxed. As with proteins encoded by mRNAs and the other classes of noncoding RNAs described above, lncRNAs have a large variety of functions, often associated with regulating protein-coding gene expression (Figure 1.1). One recurring function is in the regulation of protein-coding gene expression. There are also genomic signatures of functionality for long noncoding RNAs. LincRNAs are evolutionarily conserved as a group, but their exonic sequence is generally less conserved than that of protein-coding genes (Guttman et al. 2009). Further evidence of functionality is provided by the conservation of their promoters and dinucleotide splicing motifs (Chodroff et al. 2010; Ponjavic et al. 2007) and their enrichment in predicted secondary structures (Marques and Ponting 2009; Ponjavic et al. 2009).

Measuring transcriptomic correlates with different variables (for example, anatomical structure) can thus help one to generate hypotheses regarding the myriad molecular mechanisms that may underlie corresponding biological differences.

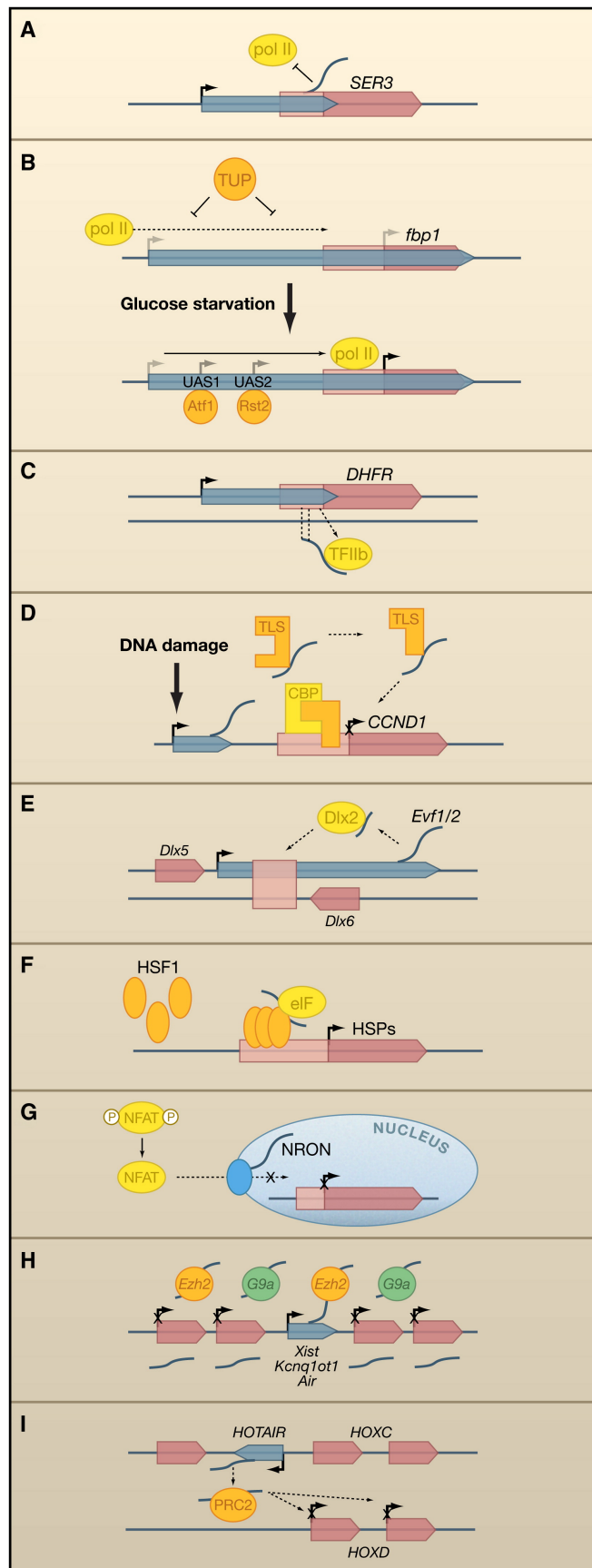


Figure 1.1: Mechanisms of lncRNA function in transcriptional regulation. LncRNAs are blue and protein-coding genes are pink, with pale pink regions indicating promoter/enhancer elements. (A) Transcriptional interference. Transcription of the lncRNA *SRG1* through the promoter of the adjacent *SER3* gene (Martens et al. 2004). (B) Initiation of chromatin remodelling. RNA pol II processivity upstream of *fbp1* is normally repressed by Tup proteins, however, rare lncRNAs are transcribed. Upon glucose starvation, the Atf1 activator binds to the UAS1 element, facilitating chromatin remodelling by RNA pol II and the subsequent binding of Rst2 to a second UAS2 element. As further lncRNAs are transcribed, the chromatin structure around the *fbp1* initiation site is then accessible to the transcriptional machinery allowing induction of the gene to occur (Hirota et al. 2008). (C) Promoter inactivation by binding to basal transcription factors. Formation of a complex between an lncRNA and both the *DHFR* promoter and TFIIB prevents normal preinitiation of transcription (Martianov et al. 2007). (D) Activation of an accessory protein. In response to stress, lincRNAs upstream of *CCND1* form a complex with an RNA-binding protein TLS (translocated in liposarcoma) in which the inactive conformation of the protein is altered, facilitating repression of *CCND1* via chromatin-binding protein (CBP) (Wang et al. 2008). (E) Activation of transcription factors. The lncRNA *Evf2* cooperates with the *Dlx2* homeodomain protein to activate the *Dlx5/6* enhancer (Feng et al. 2006). (F) Oligomerisation of an activator protein. In response to heat shock, a lincRNA assists the trimerisation of the HSF1 protein, which in turn forms a complex with the translation factor EIF to facilitate HSP expression (Shamovsky et al. 2006). (G) Transport of transcription factors. Dephosphorylated NFAT is prevented from

translocating to the nucleus and activating its targets due to interactions between the lincRNA NRON and importin proteins (Willingham et al. 2005). (H) Epigenetic silencing of gene clusters by lincRNAs. The *Xist*, *Kcnq1ot1*, and *Air* RNAs establish a nuclear domain (or “coating”) for gene silencing of genes in *cis*. The lincRNAs may directly or indirectly attract epigenetic modifiers such as histone methyltransferases (G9a or Ezh2) to bring about repressive epigenetic marks in the cluster (Nagano et al. 2008; Pandey et al. 2008; Zhao et al. 2008). (I) Epigenetic repression of genes by a lincRNA in trans. *HOTAIR* RNA, transcribed within the *HOXC* cluster, interacts with the Polycomb repressor complex 2 (PRC2) resulting in the methylation and silencing of several genes in the *HOXD* locus (Rinn et al. 2007). Figure reproduced and caption adapted from Figure 4 of Ponting et al. (2009) with permission from Elsevier.

1.2 Traditional Technologies for Measuring the Transcriptome

Here I sketch the most commonly used technologies to measure different sequence- and abundance-related aspects of the transcriptome prior to the arrival of “second-generation sequencing”. For the sake of brevity, I shall not discuss methods that are only used to measure the shorter end of the transcriptome, such as miRNAs. Nor shall I explicitly discuss low-throughput approaches, such as northern blotting, since single measurements are too small a part of the transcriptome to deserve the suffix 'ome'.

These approaches generally rely either on hybridisation of nucleic acids to a set of predetermined 'probe' sequences or on determining the sequence of nucleic acids without

the use of predetermined probes. Some of these, most notably microarray technologies, continue to be in widespread use today.

High-throughput probe-based approaches are epitomised by the cDNA microarray. With its origins arguably dating to 1982 (Augenlicht and Kobrin 1982), expression arrays are now a well-established family of technologies that rely on the high-throughput hybridisation of cDNA (reverse transcribed from RNA) to complementary probe sequences. This hybridisation can be competitive, in which two samples are conjugated to different fluorescent probes (typically one green and the other red). On other platforms, only a single sample is hybridised. Either way, array experiments are compared to one another following one of several normalisation methods (Quackenbush 2002).

Sequencing-based approaches also have a long history. Expressed Sequence Tags (ESTs) are incomplete, unedited, single-pass Sanger sequences from cDNA libraries (Adams et al. 1991). Though traditionally sequenced from the 3' or 5' end of a transcript, they can also be sequenced from its middle (Dias Neto et al. 2000). As such, they have been important in gene discovery, gene structure, and alternative splicing identification (Nagaraj et al. 2007). In Serial Analysis of Gene Expression (SAGE) (Velculescu et al. 1995) and its elaborated variants (Gowda et al. 2004; Matsumura et al. 2005; Saha et al. 2002), a restriction enzyme is used to cut a single short segment out of (nearly) every transcript. These segments are serially ligated to one another (and adapters) and sequenced to provide a digital measure of abundance for each transcript (however, not all of these short 'tags' are unique). Massive parallel signature sequencing

(MPSS), like SAGE, yields digital counts of transcripts, but uses different biochemical preparation and sequencing protocols (Brenner et al. 2000).

Many aspects of the transcriptome can be measured by variants of these approaches. For example, Cap Analysis Gene Expression (CAGE) is used to identify the 5' boundary of a capped transcript (Shiraki et al. 2003). Special microarrays with probes spanning a transcript or for specific splice junctions can be used to measure alternative splicing, but comprehensive arrays tend to be expensive. Gene discovery is not entirely limited to sequencing approaches: with a known genome sequence, 'tiling' arrays, with their dense coverage of the genome, can even be used to discover new transcribed regions (Guttman et al. 2009).

1.3 RNA-Seq: Short-Read Shotgun Sequencing cDNA

RNA-seq uses second- (or third-) generation shotgun sequencing of cDNA (or RNA) fragments to measure the transcriptome, combining advantages of both the hybridisation-based and traditional sequence-based approaches described above (Table 1.1; Figure 1.2). It is a relatively new technique, first reported in 2008 (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008). The resulting cDNA sequence fragments are aligned back to the genome as I describe in the next chapter (Figure 1.2).

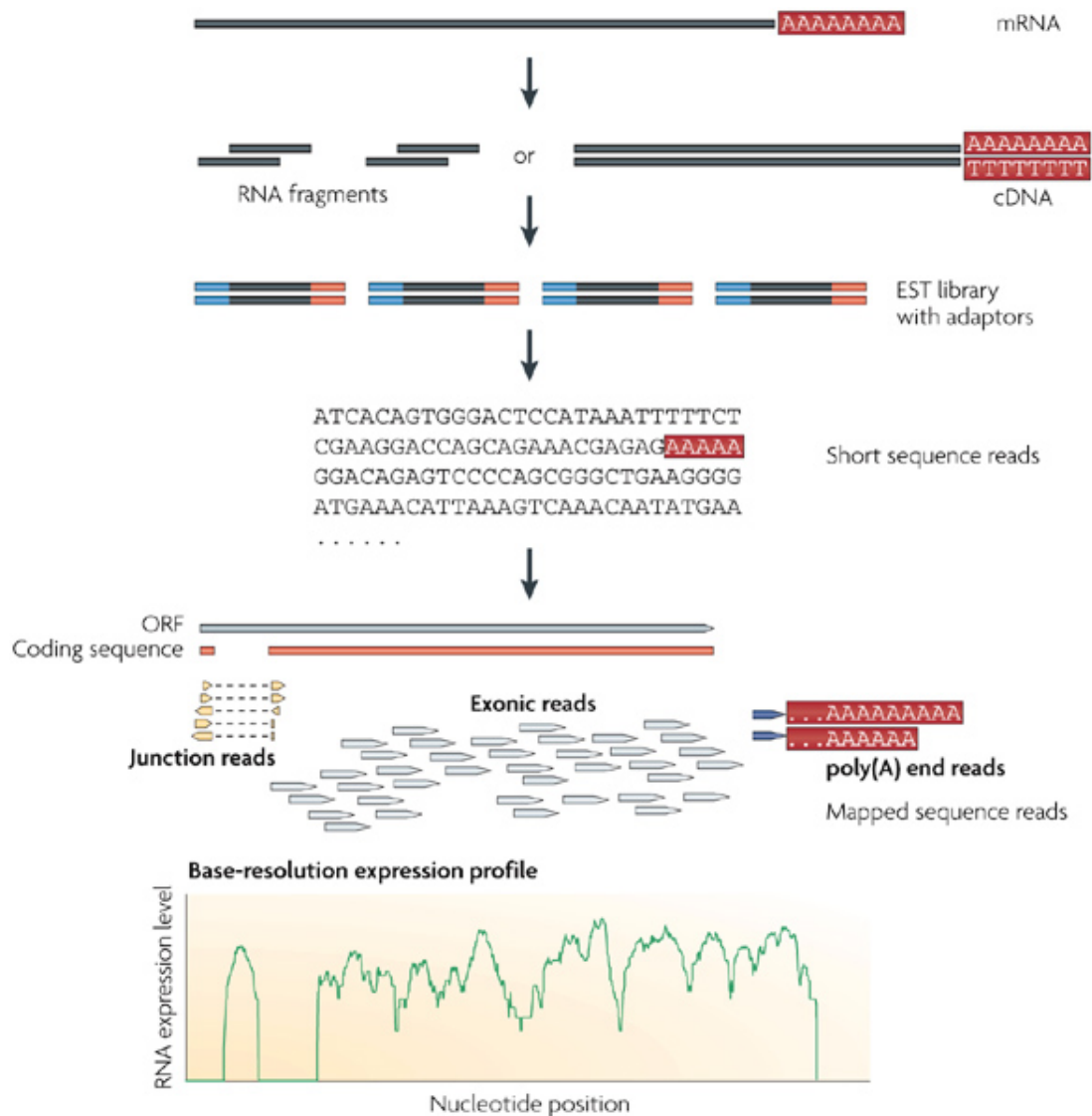


Figure 1.2: A typical RNA-seq experiment. RNAs >200 nt are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue and red) are subsequently added to each cDNA fragment and a short sequence is obtained from each fragment using a high-throughput sequencing technology. The resulting sequences are ‘mapped’ to their proper positions on the reference genome or transcriptome, and can be classified as one of three types: exonic reads, junction reads and poly(A) end-reads. These three

types are used to generate a base-resolution expression profile for each gene; a yeast open reading frame with one intron is shown. Figure and caption reproduced from Figure 1 of Wang et al. (2009) with permission from the Nature Publishing Group.

Table 1.1: Advantages of RNA-Seq compared with other transcriptomics methods.

Table reproduced from Table 1 of Wang et al. (2009) with permission from the Nature Publishing Group.

<i>Technology</i>	Tiling microarray	cDNA or EST sequencing	RNA-Seq
<i>Technology specifications</i>			
<i>Principle</i>	Hybridisation	Sanger sequencing	High-throughput sequencing
<i>Resolution</i>	From several to 100 bp	Single base	Single base
<i>Throughput</i>	High	Low	High
<i>Reliance on genomic sequence</i>	Yes	No	In some cases
<i>Background noise</i>	High	Low	Low
<i>Application</i>			
<i>Simultaneously map transcribed regions and gene expression</i>	Yes	Limited for gene expression	Yes
<i>Dynamic range to quantify gene expression level</i>	Up to a few- hundredfold	Not practical	>8,000-fold
<i>Ability to distinguish different isoforms</i>	Limited	Yes	Yes
<i>Ability to distinguish allelic expression</i>	Limited	Yes	Yes
<i>Practical issues</i>			
<i>Required amount of RNA</i>	High	High	Low
<i>Cost for mapping transcriptomes of large genomes</i>	High	High	Relatively low

1.4 Navigating the Vertebrate Central Nervous System

Early in vertebrate development the neural groove deepens and the neural folds elevate and then join to form the neural tube, the developmental progenitor of the central nervous system (CNS). Three swellings of the neural tube will develop into the three major subdivisions of the adult vertebrate brain - the hindbrain, midbrain and forebrain - while the remainder will form the spinal cord (Figure 1.3).

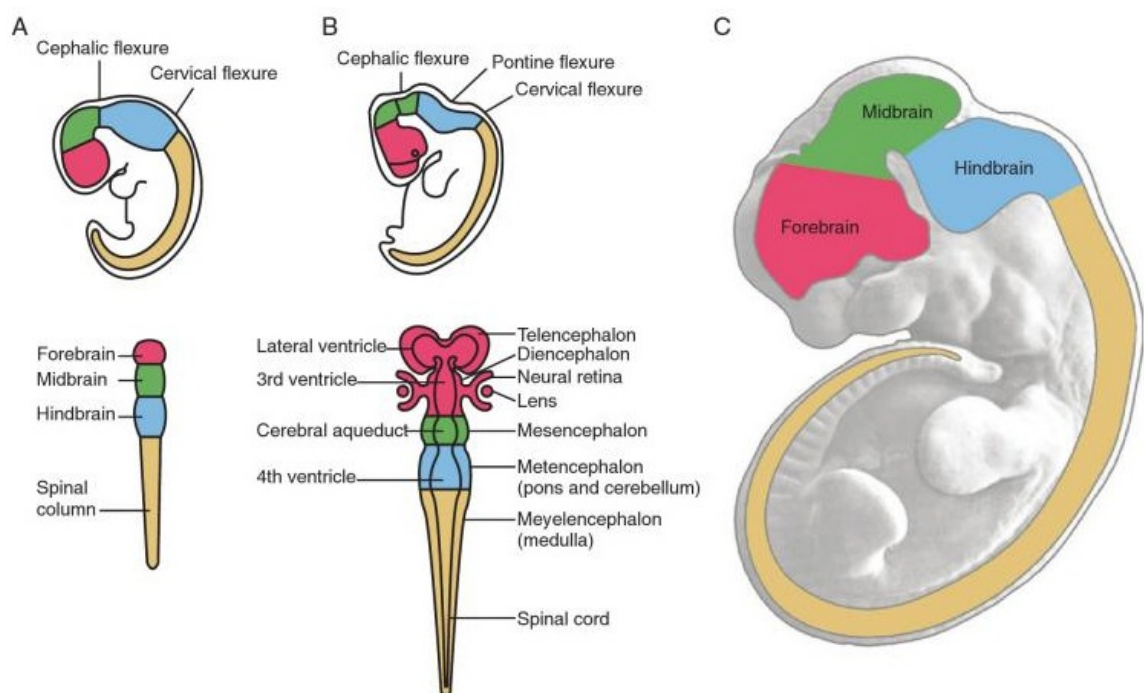


Figure 1.3: The vertebrate brain and spinal cord develop from the neural tube. Shown here as views from the side (at top) and from behind (at bottom) of human embryos at successively older stages of embryonic development (A, B, C). A. The primary three divisions of the brain occur as three brain vesicles or swellings of the neural tube, known as the forebrain (prosencephalon), midbrain (mesencephalon), and hindbrain (rhombencephalon). B. The next stage of brain development results

in further subdivisions such as the forebrain vesicle becoming subdivided into the paired telencephalic vesicles and the diencephalon. C. These basic brain divisions can be related to the overall anatomical organisation of the embryo. Figure reproduced and caption adapted from Figure 2.2 of Sanes et al. (2012) with permission from Elsevier.

The neural tube is divided not just into segments, but also into four 'plates' within each segment: roof, alar, basal, and floor (Figure 1.4). These segments and plates are reflected not just in the spinal cord, but also in the brain itself (Figure 1.5). The alar-basal boundary runs from the base of the spinal cord to the hypothalamus; its axis is bent within the brain, and it also has a kink in the diencephalon. The three axes within the CNS can be defined with respect to this boundary (note that not all neuroanatomists strictly define directionality with respect to this axis). The direction towards the roof plate is known as 'dorsal', while the opposite direction towards the floor plate is known as ventral. 'Rostral' or 'anterior' refers to the direction of the neural tube that terminates in the hypothalamus, while 'caudal' or 'posterior' refers to the direction of the neural tube towards its terminus of the spinal cord. Finally, 'medial' is nearer the midline plane, while 'lateral' is farther from the midline. I discuss images in two sectioning planes in this thesis: coronal (defined by the dorsal-ventral and medial-lateral axes) and sagittal (defined by the dorsal-ventral and rostral-caudal axes).

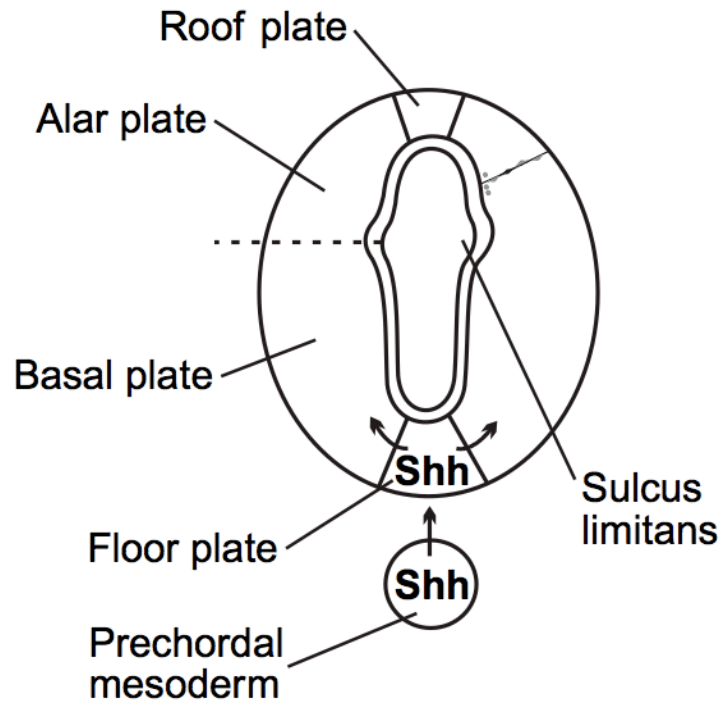


Figure 1.4: Cross-section of the developing neural tube. The sonic hedgehog (Shh) signalling pathway is one critical requirement for proper induction of the floor plate (Chiang et al. 1996). Once induced, floor plate cells express Shh (Wilson and Maden 2005), creating a concentration gradient emanating from the floor plate (Marigo and Tabin 1996). The organisation of these plates (floor, basal, alar and roof) is reflected throughout CNS development. Figure reproduced from Figure 3-7 of Butler and Hodos (2005) with permission from John Wiley and Sons.

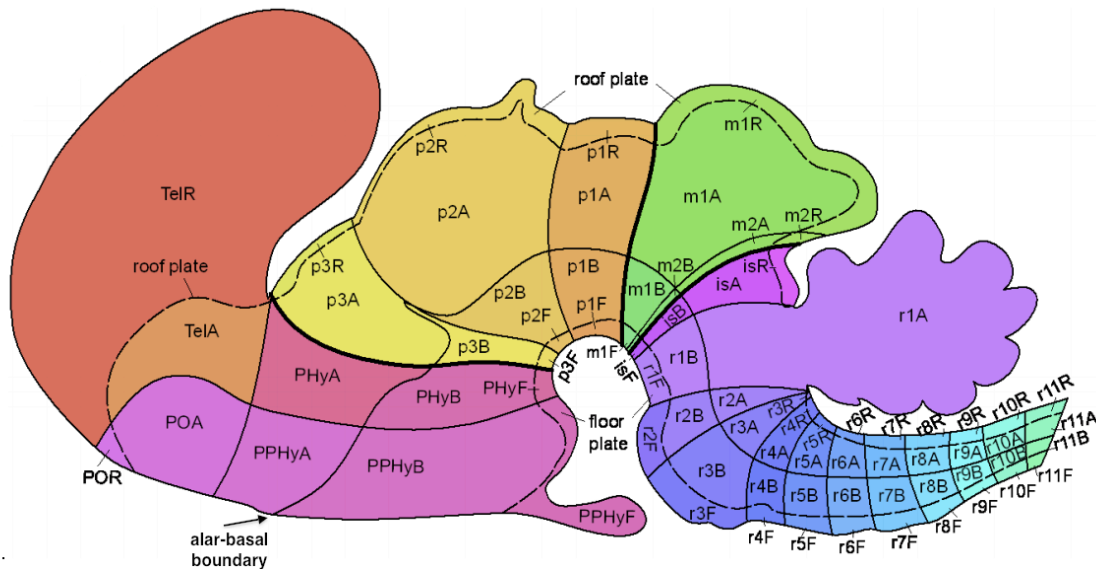


Figure 1.5: The prosomeric model of the developing mouse brain. Dotted lines and the alar-basal boundary separate, from dorsal to ventral, derivatives of roof plate, alar plate, basal plate, and floor plate. Thick black lines delineate (from rostral to caudal) secondary prosencephalon (telencephalon + hypothalamus), diencephalon, midbrain, and hindbrain. Note the segmentation within each of these regions into roof-, alar-, basal- and floor-derived components. Figure drawn by Luis Puelles and adapted from "Technical White Paper: Allen Developing Mouse Brain Reference Atlas" (<http://developingmouse.brain-map.org/docs/ReferenceAtlas.pdf>) under the Allen Institute Citation Policy.

The dorsal-most portion of the roof-derived telencephalon is known as the pallium. In mammals, this gives rise to several regions, including the neocortex (an anatomical feature specific to mammals). I shall discuss additional derivatives of the pallium (and its subdivisions) in mammals and birds in Chapter 4.

1.5 Anatomy of the Neocortex

The neocortex is, by convention, composed of six cytoarchitectonic layers enumerated from I near the edge of the brain to VI above the white matter (Hindu-Arabic numerals and Roman numerals are used interchangeably to denote the layers). Depending on the area and species, these layers can have distinct sublayers (layers IVa, IVb and IVc in primate primary visual cortex) or they can be merged (layers II/III in mice). Developmentally, neurons migrate along radial glia to sequentially populate layers VI, V, IV, III and finally II (Figure 1.6). The termination of their migration is guided by reelin, a protein secreted by Cajal-Retzius cells already present in layer I (Rice and Curran 2001). I also discuss layer VIb in this thesis, which contains adult forms of subplate neurons, and thus developmentally precedes neurons in layers II-VI (Torres-Reveron and Friedlander 2007). (Both the Cajal-Retzius and subplate neurons are found earlier in the primordial plexiform layer of the preplate, which splits to produce what become the most and least superficial layers, I and VIb.) These layers are named for their relative locations and the cytoarchitecture of their discriminating constituent cells (in reality they often are heterogeneous mixtures): the neuron-poor "molecular layer" (I), external and internal granular layers (II and IV, respectively), external and internal pyramidal layers (III and V, respectively), and the polymorphic or multiform layer (VI) (Butler and Hodos 2005; Shipp 2007).

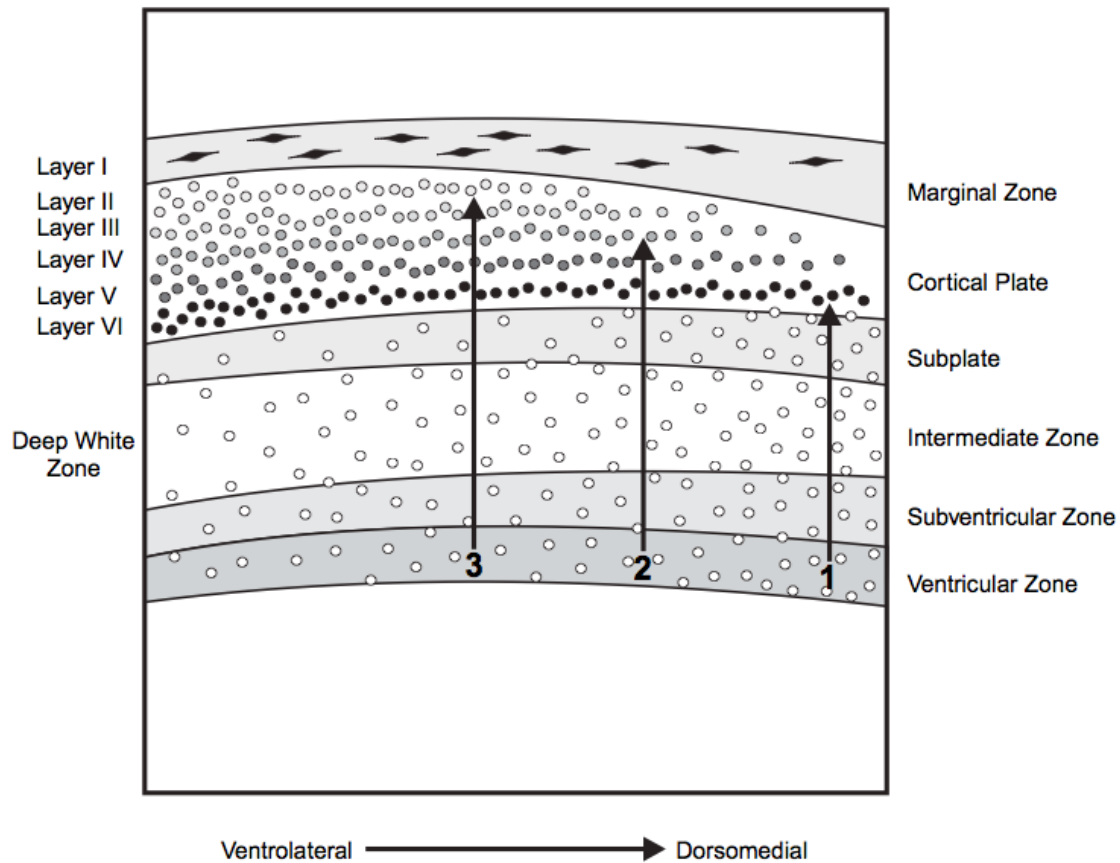


Figure 1.6: Drawing showing the progressive formation of neocortical layers VI–II within the cortical plate, with the neurons that form the deepest layer, layer VI, being generated earliest (arrow 1), those in the intermediate layers V–III generated subsequently and in deep to superficial order (arrow 2), and the most superficial of these layers, layer II, generated last (arrow 3). Cajal-Retzius cells are shown in the marginal zone, which becomes layer I. Neurons migrating radially into the cortical plate stop when they encounter the reelin signal produced by the Cajal-Retzius cells. The drawing reflects the ventrolateral (left) to dorsomedial (right) gradient of maturation of the neocortex. Figure reproduced and caption adapted from Figure 3-8 of Butler and Hodos (2005) with permission from John Wiley and Sons.

The neocortex is not only organised into layers. To an extent, it also has a repeated columnar organisation of interconnected neurons (Hirsch and Martinez 2006). These neurons are frequently connected in canonical (and less commonly, noncanonical) circuits in a manner partially determined by layer identity (Figure 1.7). Finally, the neocortex is organised into areas distinct in both cytoarchitectonic structure and function, characterised by gradients of gene expression in development that become more discrete patterns in adults (Sansom and Livesey 2009).

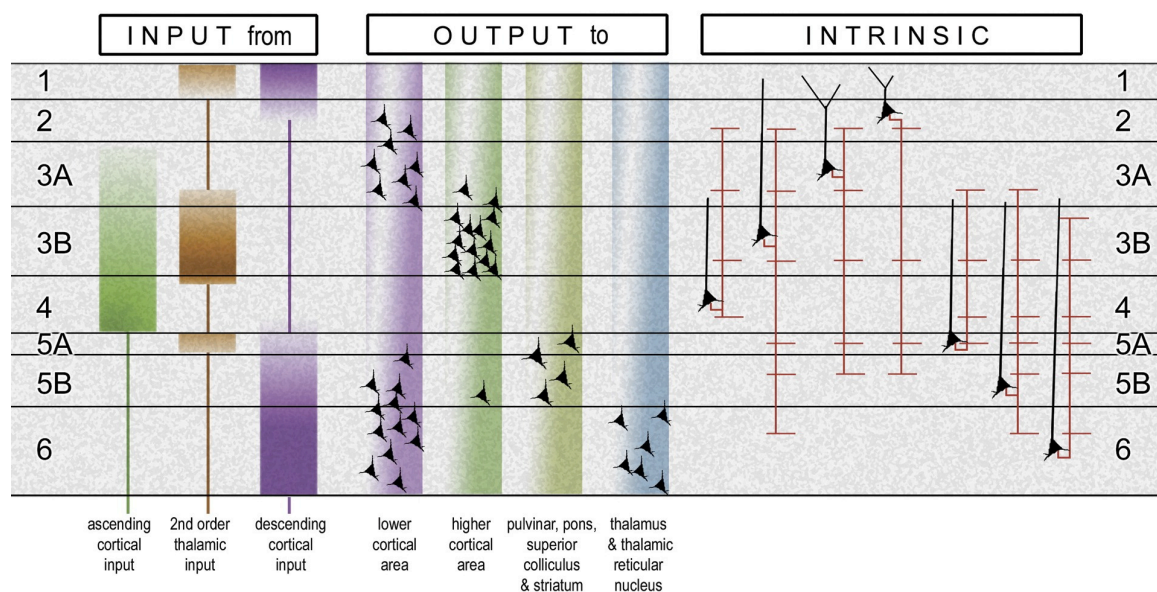


Figure 1.7: Summary of inputs, outputs and intrinsic excitatory connections of a generic, nonprimary visual area of primate cerebral cortex. Figure and caption reproduced from Shipp (2007) with permission from Elsevier.

1.6 Prior Lines of Work in Comparative Neurotranscriptomics

One could arguably define "comparative neurotranscriptomics" rather broadly as comparisons of nervous system-derived transcriptomes involving any variables (for example, developmental time point, time of day or disease state). In this thesis, I discuss only a small subset of cases that would fall under this broad definition. Specifically, I consider comparisons of brain-derived polyadenylated transcriptomes in which the only variables under investigation are anatomical structure and/or species. (I do not investigate population variability in this thesis.) Within a species, such work can reveal molecular differences that may underlie some differences in physiology. Between species, it can suggest possible molecular underpinnings of differences and similarities. (Any mechanistic hypotheses suggested by such high-throughput approaches necessarily remain hypotheses until tested.) I discuss the reasoning underlying cross-species work in more detail in Chapter 4.

Much can be learned from transcriptome comparisons of different anatomical structures within a species. Comparisons within species can identify new molecular markers that differentiate different regions. A two-tiered effort involving microarrays for discovery followed by large-scale *in situ* hybridisation in mice revealed numerous stable markers of hypothalamic nuclei (Shimogori et al. 2010). Such comparisons can also reveal interesting relationships between different regions. The transcriptomes of adult mouse brain regions reflect the developmental segments along the neural tube from which they derived (Zapala et al. 2005). In rodents, correlation of gene expression profiles between two regions is even weakly correlated with structural connectivity (French and Pavlidis 2011). Anatomical comparative neurotranscriptomic studies have also provided insights

into the expression patterns of disease genes and cell types (Sunkin and Hohmann 2007). Even without explicitly including cross-species transcriptome data, such studies can suggest directions for future evolutionary experiments. For example, comparisons of thirteen different brain regions in human have shown that putative *cis*-regulatory elements of differentially expressed genes between those regions are characterised by accelerated sequence evolution in the human lineage (Johnson et al. 2009).

Very recently, there has been an explosion of such anatomical data sets for humans. For example, one resource reports the array-based transcriptomes of 1,340 tissue samples from 16 regions of 57 human brains (Kang et al. 2011). BrainSpan (<http://www.brainspan.org>) provides RNA-seq expression profiles of 16 brain regions. The Human Brain Map from the Allen Institute (<http://human.brain-map.org>) provides expression profiles for hundreds of regions from three human brains. The potential analyses with such resources are vast, and extend well beyond the initial papers describing them.

There have also been numerous transcriptome comparisons between species. Compensatory changes in post-transcriptional regulatory processes are unlikely to consistently balance transcriptome changes since significant transcriptome differences between humans and chimps are strongly reflected in those species' proteomes (Fu et al. 2007). One would expect this to be even truer when comparing patterns of differentially expressed genes between species, since this would require non-parsimonious concerted spatiotemporal changes in gene expression and post-transcriptional regulation. Comparative studies of transcriptome evolution for different organs among vertebrates (Chan et al. 2009), mammals (Brawand et al. 2011) and apes (Khaitovich et al. 2005)

have all found that nervous system tissues have the slowest rate of transcriptome evolution. Intergenic transcription patterns in the brain show more evolutionary divergence than exonic transcription (Xu et al. 2010). This matches the pattern seen in genome evolution, in which even functional non-protein-coding elements are turned over much more rapidly than protein-coding elements (Ponting et al. 2011).

Unsurprisingly, there is also conservation in transcriptomic patterns among different brain regions. Genes specific to the cerebral cortex, caudate nucleus or cerebellum in human are generally specific to that same region both in mouse (Strand et al. 2007) and in chimp (Khaitovich et al. 2004). A large-scale meta-analysis of microarray data found that most (but not all) modules of co-expressed genes in humans and mice were conserved (Miller et al. 2010). Another study found that these gene co-expression relationships were more conserved between humans and chimps in cortical areas than in subcortical regions, although this may depend upon the particular regions chosen (Oldham et al. 2006).

Many studies have looked at the relationships between primary sequence and transcription. Human genes preferentially expressed in cortical areas have slightly lower evolutionary rates (when compared to chimp and mouse) than genes preferentially expressed in subcortical areas (Tuller et al. 2008). The frequency of mammal- and primate-specific genes is also slightly higher amongst genes expressed subcortically (Tuller et al. 2008). Genes differentially expressed between human and mouse cerebral cortex, caudate nucleus and/or cerebellum also have higher sequence constraint between the two species (Strand et al. 2007).

Microarrays and sequencing aren't the only methods available for "comparative neurotranscriptomics". High-throughput imaging methods provide a level of anatomical detail impractical to attain by dissection-based methods. The Allen Institute for Brain Science has produced high-throughput *in situ* hybridisation image series for tens of thousands of probes covering multiple species (<http://www.brain-map.org>). The Gene Expression Nervous System Atlas (GENSAT) generates high-throughput *in situ* hybridisations and images from a large set of BAC-EGFP reporter mice (Heintz 2004). A major drawback of these industrial-scale probe-based imaging methods is their considerable expense. Furthermore, the restricted availability of primate brains (human and otherwise) of acceptable quality makes a hypothesis-free effort impractical for that clade. As such, reasonably comprehensive (for protein-coding genes) vertebrate data only exists for mice, complicating genome-wide comparative efforts.

One exciting new way of doing comparative neurotranscriptomics involves imaging and overlaying the expression patterns of carefully chosen genes in a species with highly stereotypic and synchronous development across individuals (Tomer et al. 2010). Using two-channel confocal microscopy, this team imaged both an immunostain highlighting axon bundles and an *in situ* hybridisation with dozens of genes of interest in *Platynereis* worms. These images were computationally registered using the axon bundles as landmarks, and then combined to give three-dimensional maps of gene expression at different developmental stages. They demonstrated that the mammalian pallium and the insect mushroom body both derive from a similar morphogenetic position. Of course, to use this method one must first know the relevant developmental genes. Identifying

candidate marker genes is one area in which more 'hypothesis-free' approaches such as microarrays and sequencing can make an impact.

1.7 Project Aims and Thesis Structure

In this thesis, I present novel applications of comparative neurotranscriptomics using RNA-seq. It is comprised of two phases: first, I compare the transcriptomes of distinct but heterogeneous cell groups of cortical layers within a species; second, I compare the transcriptomes of pallial structures and cell groups between two species.

In **Chapter 2**, I review the materials and methods used in both phases. They shared a common sequencing platform, computational tools for the expression analysis, and sets of functional annotations. They also both used data from *in situ* hybridisations.

In **Chapter 3**, I explore the transcriptomes of mouse neocortical layers. This was the first study across neocortical layers to use an unbiased genome-wide approach. As such, I explored the functional annotations of genes enriched in various layers, found candidates for isoform switching across layers and identified long noncoding RNAs that were differentially expressed across layers.

In **Chapter 4**, I compare the transcriptomes of several cerebral regions in mice and chicks with a co-expression approach. The results reflect both known homologies and functional analogies.

In **Chapter 5**, I conclude with a general discussion of how this work expands the emerging field of comparative neurotranscriptomics, and of the doors it opens for future research.

Chapter 2 - Materials and Methods

2.1 RNA-Seq with Illumina Sequencing by Synthesis

“Second-generation sequencing” (SGS), also known as “next-generation sequencing”, refers to a family of ‘short-read’ DNA shotgun sequencing technologies, each capable of producing massive numbers of short sequence ‘reads’ from PCR-amplified DNA templates. Illumina currently manufactures the most widely used SGS platform. The other major second-generation sequencers are variants of Roche’s 454 platform, which provides longer reads but lower throughput, and Life Technology’s SOLiD, which produces higher-quality but shorter reads in ‘colour space’ by sequencing each base twice using redundant fluorophores. There are also less mature “third-generation sequencing” technologies, such as the platforms manufactured by Helicos BioSciences and Pacific Biosciences. These sequencers rely on measurements from single molecules instead of the ensembles of molecules measured by second-generation sequencers.

Here, I briefly discuss Illumina’s “sequencing by synthesis” technology as applied to produce the data discussed in this thesis (see http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf). For long RNAs, one first selects some subset of the transcriptome from total RNA (since the overwhelming majority of standing RNA is rRNA). This transcriptome can be rRNA-depleted total RNA, DSN normalised RNA, or poly(A)-selected RNA. Next, one either fragments RNA and reverse transcribes it to make cDNA, or else reverse transcribes RNA to make cDNA that is then fragmented (different orders produce

differing coverage profiles across the length of the transcript) (Wang et al. 2009). In this project, RNA was poly(A) selected, fragmented, and then made into cDNA with ‘random’ hexamer primers (these primers do not randomly prime the transcriptome with equal efficiency; see Figure 2.5). Adapters are ligated to the fragment ends (a different sequence on either end) before the fragments are run on a gel for size selection. Sequencing templates from the excised band (typically having a median fragment size of a few to several hundred bases) are PCR amplified to produce the final ‘sequencing library’.

Once the sequencing libraries are generated, their molecular constituents must be represented on a ‘flow cell’ which consists of eight ‘lanes’ sandwiched between two panes of glass. Each lane has a hole at either side to permit the entry and exit of reagents as needed. Each lane is replete with a ‘lawn’ of oligonucleotide sequences that are complementary to the Illumina adapters previously ligated during library preparation. The library templates are then added to a lane at a carefully chosen concentration, and the adapters are given time to bind to their complements. (Several libraries, uniquely identified by an additional molecular barcode, can be ‘multiplexed’ on a single lane. This was not done for the projects discussed in this thesis.) These single molecules are then PCR bridge-amplified on the flow cell using primers against the appropriate adapters, producing homogeneous ‘clusters’ of DNA attached to the flow cell. A flow cell may contain different types of libraries (e.g. RNA-seq, genomic DNA, etc.), but these libraries will typically be sequenced in the same way (in terms of number of bases and if the bases are read from both ends of the fragment or only one).

Following cluster generation in its lanes, a flow cell is ready to be ‘imaged’ on a ‘run’ of the sequencer. Depending on the number of ‘cycles’ (one base sequenced per cycle), and if the sequence is generated from both of the ends (‘paired-end’ or ‘mate-paired’) or only one, a run can take from several days to over a week. To begin the first sequencing cycle, a solution of the four canonical nucleotides (adenine, cytosine, guanine and thymine), modified to include a distinctive fluorophore for each base and conjugated to reversible terminators, is added to the flow cell along with primers and DNA polymerase. After washing, most molecules in each cluster are tagged by the same fluorophore (of course not all of the reactions go to completion, which produces some ‘dephasing’ within the cluster and is a limiting factor of read lengths). Two lasers at different frequencies are tuned to excite the fluorophores (two per laser). Two filters are used over a camera or cameras to facilitate algorithmic deconvolution of the emission spectra from a cluster under each of these conditions. In older sequencing systems with one camera (such as the Illumina GAIIx used in this thesis) this is done sequentially. The newer HiSeq 2000 is more parallelised with four cameras, two flow cells, and the ability to read clusters from the top and bottom of each flow cell (Minoche et al. 2011). After terminators are cleaved, this process is repeated as desired. To read from the second end of a DNA fragment, the alternative primer is used.

Proprietary image recognition software is used to identify distinct clusters (overlapping clusters that cannot be computationally separated with confidence are discarded). The four grey scale images produced in each cycle are used to assign the probability that a particular base call is wrong. When there is evidence in a single cycle that the read has been strongly dephased, the read is said to fail the ‘chastity filter’.

2.2 Short Read Mapping with Tophat

Once generated, RNA-seq short reads can be ‘assembled’ to create a transcriptome, or they can be ‘mapped’ back to a reference genome (or a fully annotated transcriptome). Since both of the species used in this project have annotated genomes available (and neither has a ‘fully’ annotated transcriptome), I used the latter, which permits quantification of genes at moderate to low coverage. There are several approaches to and dozens of tools for mapping short reads to a genome. In this thesis I use tophat, which was one of the first tools available for splice junction discovery in short reads (Trapnell et al. 2009). Tophat uses bowtie (Langmead et al. 2009), an efficient short read mapper that indexes the genome using the Burrows-Wheeler transform (Figure 2.1) and then traverses this index to map reads (Figure 2.2). (I have also experimented with other short read mappers and traditional aligners not discussed here due to space constraints.)

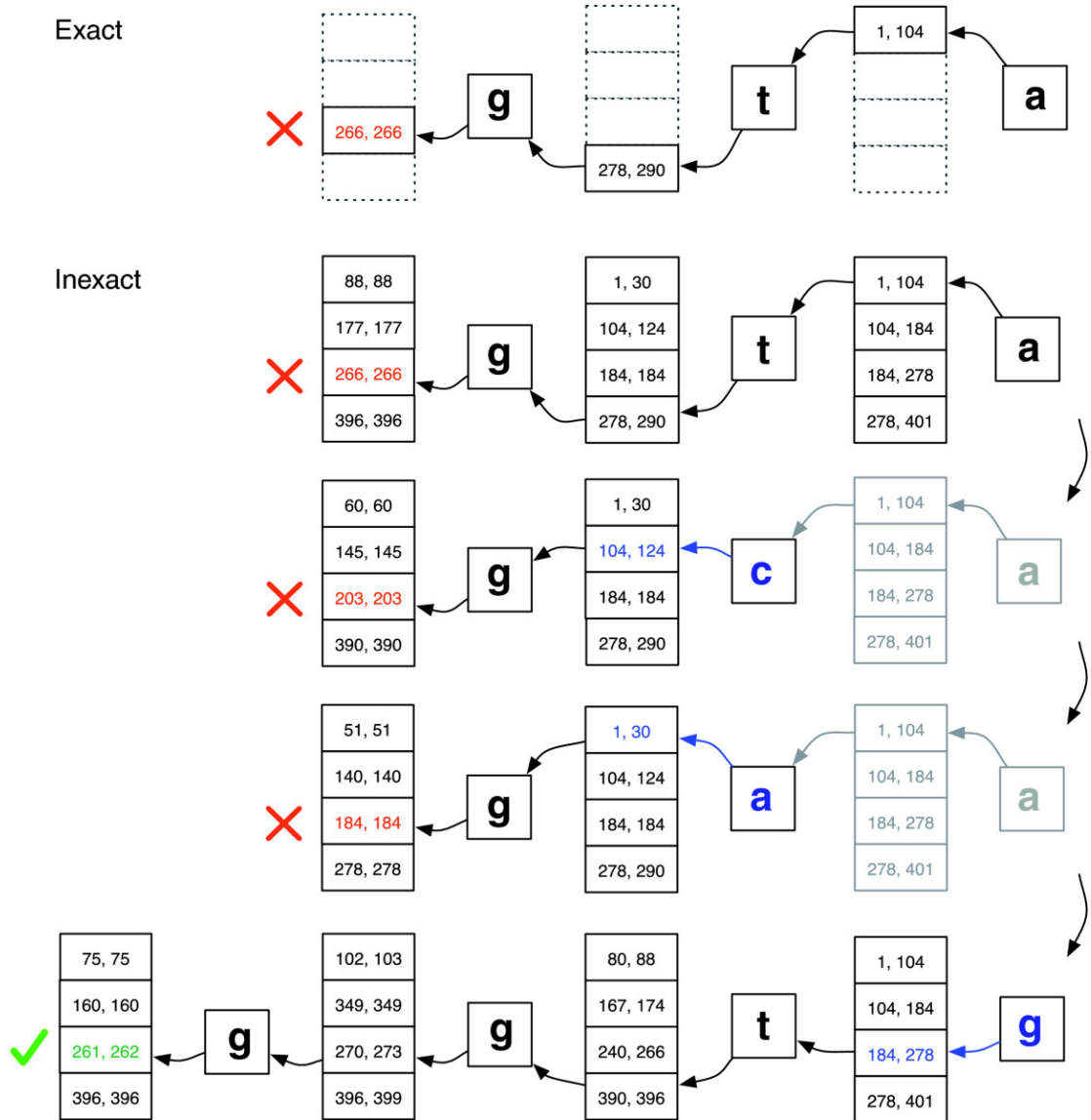


Figure 2.2: Exact matching versus inexact alignment. Illustration of how an exact matching algorithm (top) and the mapper used by bowtie and tophat (bottom) proceed when there is no exact match for query 'ggta' but there is a one-mismatch alignment when 'a' is replaced by 'g'. Boxed pairs of numbers denote ranges of matrix rows beginning with the suffix observed up to that point. A red X marks where the algorithm encounters an empty range and either aborts (as in an exact matching algorithm) or backtracks (as in the inexact algorithm implemented in

bowtie). Note that the a→g backtrack yields an empty range as well, and would be attempted before the successful backtrack. The substituted base is determined by a pseudo-random number generator so the alignments can be replicated when the same seed is used. A green check marks where the algorithm finds a nonempty range delimiting one or more occurrences of a reportable alignment for the query. Figure reproduced and caption adapted from Langmead et al. (2009) under the Creative Commons Attribution License 2.0.

Of course, bowtie is not itself a spliced mapper; the spliced mappings require the pipeline implemented in tophat. There are many options and variants of this pipeline, but I will only sketch those used in this thesis, described in detail in Trapnell et al. (2009) and the supplementary information of Trapnell et al. (2010). First, tophat removes low-complexity reads (for example, a string of As). Tophat then attempts to map reads to the genome using bowtie and subsequently identifies genomic ‘islands’ where reads were able to map. Using these islands, tophat then generates a list of potential splice sites that could connect islands within 500 kb of one another (20 kb in older versions) using canonical splice donor and acceptor signals (Trapnell et al. 2009) (<http://tophat.cbc.umd.edu>). Tophat then attempts to map reads that could not map to the genome to this set of potential splice sites (which the user can extend using a customised list of possible junctions). To report a new splice site, tophat requires a read to span the junctions on both sides by at least some number of nucleotides (8 by default). Tophat currently reports up to 20 equally valid mappings per read in SAM format (Li et al. 2009). These mappings are more generally termed ‘alignments’, since the presence of

splice junctions allows part of a read to map to one genomic location and the other part of the read to map to another genomic location, allowing an intron in between.

If both reads from a mate pair align uniquely, those fragment alignments are reported. If one or both reads of a mate pair does not align uniquely, knowledge of the fragment distribution can sometimes resolve the ambiguity. Fragment alignments are ranked, and only the highest-ranking alignments of each read are reported, as described in Trapnell et al. (2010):

Let x and y be fragment alignments. Then $x < y$ if any of the following (applied in order) are true:

- (1) x is a singleton, and y has both ends mapped,*
- (2) x crosses more splice junctions than y ,*
- (3) the reads from x map significantly farther apart in the genome than expected according to the library's fragment length distribution (≥ 3 standard deviations), and the reads from y do not,*
- (4) the reads from x are significantly closer together than expected according to the library's fragment length distribution, and the reads from y are not,*
- (5) the reads from x map more than 100 bp farther apart than the reads from y ,*
- (6) x and y both span an intron, and x spans a longer one,*
- (7) x has more mismatches than y to the genome.*

Fragments that have multiple equally good alignments according to the above rules are all reported. If there are n alignments for a fragment, each is assigned a probability of only $1/n$ of being correct. The SAM format encodes this probability in the mapping quality field, which is later used by Cufflinks to reduce the contribution of multiply mapping fragments (to $1/n$ of a uniquely mappable read) in FPKM calculations. (Trapnell et al. 2010)

2.3 Expression Quantification with Cufflinks

Cufflinks is a program used to identify *de novo* transcripts and quantify expression abundance based on tophat-aligned read pairs, as described in Figure 2.3 (Trapnell et al. 2010). This normalised abundance is estimated in terms of Fragments Per Kilobase of exon model per Million reads mapped (FPKM).

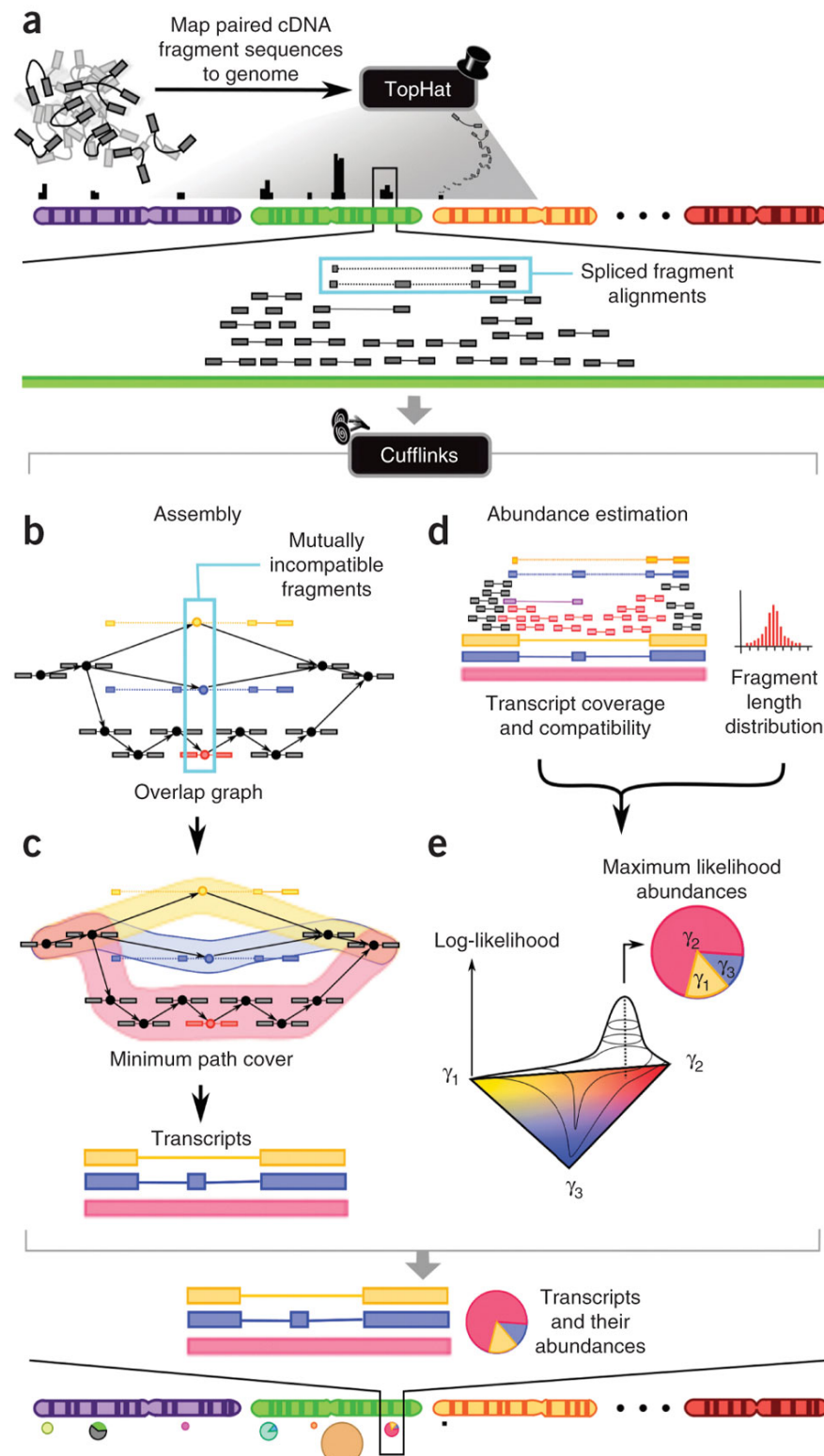


Figure 2.3: Overview of cufflinks. (a) The algorithm takes as input cDNA fragment sequences that have been aligned to the genome by tophat. **(b–e)** With paired-end

RNA-Seq, cufflinks treats each pair of fragment reads as a single alignment. The algorithm assembles overlapping 'bundles' of fragment alignments (b, c) separately, which reduces running time and memory use, because each bundle typically contains the fragments from no more than a few genes. Cufflinks then estimates the abundances of the assembled transcripts (d, e). The first step in fragment assembly is to identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms (b). Fragments are connected in an 'overlap graph' when they are compatible and their alignments overlap in the genome. Each fragment has one node in the graph, and an edge, directed from left to right along the genome, is placed between each pair of compatible fragments. In this example, the yellow, blue and red fragments must have originated from separate isoforms, but any other fragment could have come from the same transcript as one of these three. Isoforms are then assembled from the overlap graph (c). Paths through the graph correspond to sets of mutually compatible fragments that could be merged into complete isoforms. The overlap graph here can be minimally 'covered' by three paths, each representing a different isoform. Dilworth's Theorem states that the number of mutually incompatible reads is the same as the minimum number of transcripts needed to 'explain' all the fragments. Cufflinks implements a proof of Dilworth's Theorem (given in the appendix of Trapnell et al. (2010)) that produces a minimal set of paths that cover all the fragments in the overlap graph by finding the largest set of reads with the property that no two could have originated from the same isoform. Next, transcript abundance is estimated (d). Fragments are matched (denoted here using colour) to the transcripts from which they could have originated. The violet fragment could have originated from the blue or red isoform.

Grey fragments could have come from any of the three shown. Cufflinks estimates transcript abundances using a statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated. Because only the ends of each fragment are sequenced, the length of each may be unknown. Assigning a fragment to different isoforms often implies a different length for it. Cufflinks can incorporate the distribution of fragment lengths to help assign fragments to isoforms. For example, the violet fragment would be much longer, and very improbable according to the cufflinks model, if it were to come from the red isoform instead of the blue isoform. Last, the program numerically maximizes a function that assigns a likelihood to all possible sets of relative abundances of the yellow, red and blue isoforms ($\gamma_1, \gamma_2, \gamma_3$) (e), producing the abundances that best explain the observed fragments, shown as a pie chart. Figure reproduced and caption adapted from Trapnell et al. (2010) with permission from the Nature Publishing Group.

Unsurprisingly, the accuracy of quantifications produced by cufflinks increases for more highly expressed genes since these quantifications tend to be based on a greater number of reads (Figure 2.4).

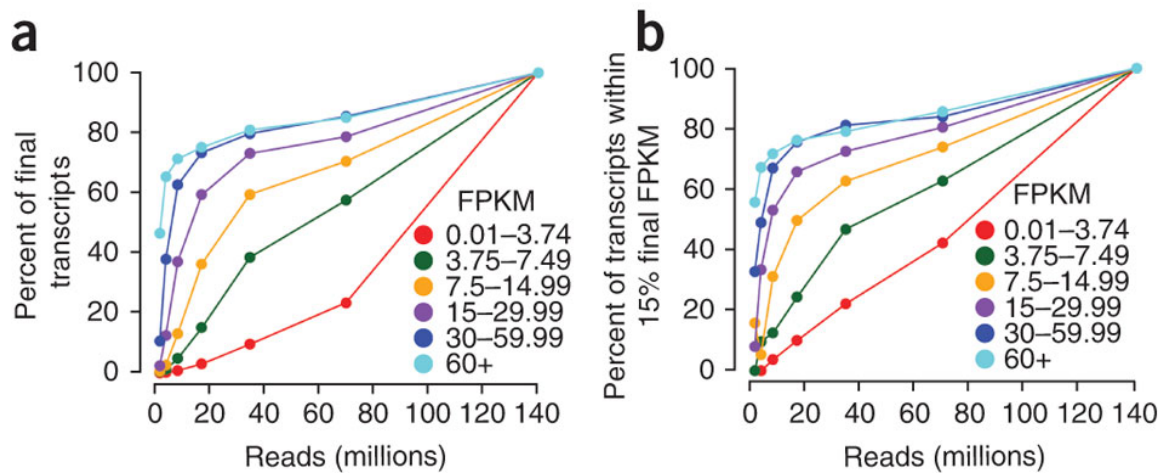


Figure 2.4: Robustness of assembly and abundance estimation as a function of expression level and depth of sequencing. Subsets of a large read set were mapped with tophat and assembled with cufflinks, and the resulting assemblies were compared for structural and abundance agreement with the assembly from the full data set. Coloured lines show the results obtained at different depths of sequencing in the full assembly; for example, the light blue line tracks the performance for transcripts with FPKM >60. (a) The fraction of transcript fragments fully recovered increases with additional sequencing data, although nearly 75% of moderately expressed transcripts (≥ 15 FPKM) are recovered with fewer than 40 million 75-bp paired-end reads (20 million fragments), a fraction of the data generated by a single run of the sequencer used in this experiment. (b) Abundance estimates are similarly robust. At 40 million reads, transcripts determined to be moderately expressed using all reads were estimated at within 15% of their final FPKM values. Figure reproduced and caption adapted from Trapnell et al. (2010) with permission from the Nature Publishing Group.

I also used a module incorporated into cufflinks to correct for sequence-specific biases observed at the 5' and 3' ends of fragments (Figure 2.5) (Roberts et al. 2011). Such bias, at least partially due to primers, affects sequence coverage across a transcript and confounds estimates of gene expression levels (Roberts et al. 2011). The bias-correction module uses the empirically determined positional bias weights to reduce or increase the number of reads found at a particular position for purposes of calculating FPKMs of genes and transcripts.

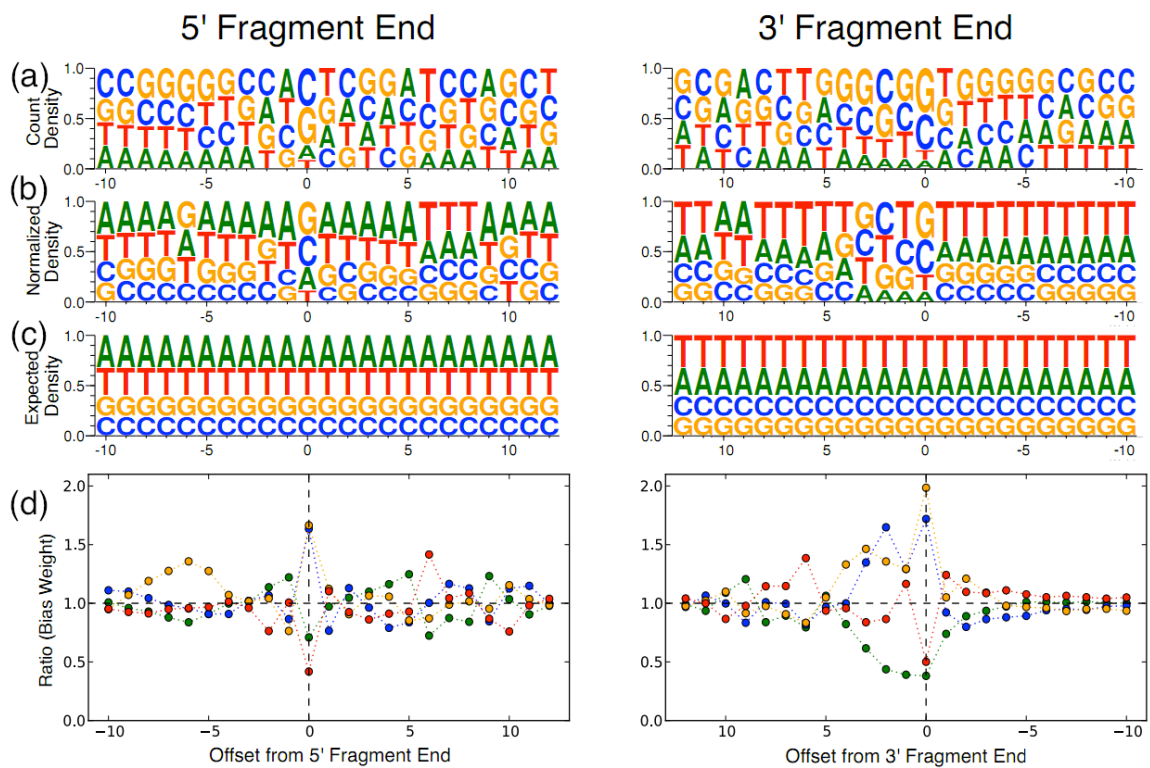


Figure 2.5: Nucleotide distribution surrounding fragment ends and calculation of bias weights. (a) Sequence logos showing the distribution of nucleotides in a 23 bp window surrounding the ends of fragments from an RNA-seq experiment in yeast primed with Illumina's 'random hexamers' (Levin et al. 2010). The 3' end sequences are complemented (but not reversed) to show the sequence of the primer

during first-strand synthesis. The offset is calculated so that zero is the ‘first’ base of the end sequence and only non-negative values are internal to the fragment. Counts are taken only from transcripts mapping to single-isoform genes. (b) Sequence logo showing normalized nucleotide frequencies after reweighting by initial (not bias corrected) FPKM in order to account for differences in abundance. (c) The background distribution for the yeast transcriptome, assuming uniform expression of all single-isoform genes. The difference in 5′ and 3′ distributions are due to the ends being primed from opposite strands. Comparing (c) to (a) and (b) shows that while the bias is confounded with expression in (a), the abundance normalization reveals true bias to extend upstream and downstream of the fragment end. Taking the ratio of the normalized nucleotide frequencies (b) to the background (c) for the random hexamer dataset gives bias weights (d). Although the weights here are based on independent frequencies, the authors found correlations among sites in the window and took these into account in the full model to produce more informative weights. Figure reproduced and caption adapted from Roberts et al. (2011) under the Creative Commons Attribution License 2.0.

2.4 General Computational Tools

I explicitly name special-purpose modules and programs used in this thesis as appropriate, but in this section I list the computational tools I used on a day-to-day basis that are not acknowledged elsewhere. All the servers I used for computational analyses at Oxford and the NIH ran distributions of GNU/Linux. (My personal workstations

throughout the DPhil ran either Mac OS X or GNU/Linux.) For version control, I used Mercurial; for a web server, Zope; for databases, MySQL, PostgreSQL and Oracle; for dynamically creating interactive web images, the Google Chart API; for statistics and plotting, the python module SciPy or R; for computational pipelines, Ruffus (Goodstadt 2010); and for basic formatting and text parsing, custom python scripts or the tools native to GNU/Linux. From time to time I read, modified or wrote code in C, C++, perl, R or Java, but python (and its wide range of modules) was my language of choice for daily activities. I used samtools (Li et al. 2009) for viewing or manipulating SAM/BAM files (the standard file format for short read alignments). For working with short read alignments within my scripts I used pysam, an open-source python-based wrapper of samtools I developed during my DPhil with Drs. Andreas Heger, Martin Goodson, Gerton Lunter, and Leo Goodstadt (<http://code.google.com/p/pysam/>).

2.5 Sources of Functional Annotations

I used a wide range of sources for the functional annotations in this thesis, and linked all of them through Ensembl gene IDs. I downloaded all Gene Ontology (Ashburner et al. 2000) annotations from Ensembl BioMart (Smedley et al. 2009) (release 56) and treated each type of annotation (biological process, molecular function and cellular component) independently. For the mouse knockout phenotypes (Blake et al. 2011), I translated Ensembl IDs to Mouse Genome Informatics (MGI) IDs using data from ftp://ftp.informatics.jax.org/pub/reports/MRK_ENSEMBL.rpt, translated MGI IDs into low-level phenotypes using data in

ftp://ftp.informatics.jax.org/pub/reports/MGI_PhenoGenoMP.rpt, and finally associated these phenotypes to all relevant phenotype descriptors in the ontology using data in ftp://ftp.informatics.jax.org/pub/reports/MPheno_OBO.ontology. I downloaded mouse protein complex associations from Ensembl BioMart (Reactome database, release 56) (Croft et al. 2011). For pathway information from KEGG (Kanehisa et al. 2004), I translated Ensembl IDs into KEGG gene IDs using data in ftp://ftp.genome.jp/pub/kegg/genes/organisms/mmu/mmu_ensembl-mmu.list (downloaded June 17, 2010); KEGG gene IDs were associated to pathways with data in ftp://ftp.genome.jp/pub/kegg/genes/organisms/mmu/mmu_pathway.list (downloaded June 17, 2010) and pathways were assigned names with the data in ftp://ftp.genome.jp/pub/kegg/pathway/map_title.tab (downloaded June 17, 2010). I downloaded additional human disease associations from Ensembl Variation (release 57; downloaded March 9, 2010) (Chen et al. 2010) via BioMart. I removed phenotypes with fewer than ten associated genes. Then I translated these human gene IDs to mouse gene IDs using Ensembl BioMart (keeping only the genes having a one-to-one mapping). In all gene-gene translations described above, no more than one mapping was taken for any gene to avoid multiple counting.

2.6 *In Situ* Hybridisation and the Allen Mouse Brain Atlas

I personally ran no *in situ* hybridisation (ISH) experiments for the projects in this thesis, but I did need to interpret data from them. Collaborators performed some *in situ* hybridisations but most of the ISH data I use in this thesis came from the Allen Mouse

Brain Atlas (AMBA) (Lein et al. 2007). Hence, I will describe here the methods used to generate the AMBA (the specific conditions used for figures in this thesis are described in their respective chapters, but conceptually they were generated in the same way as images in the AMBA).

ISH provides spatial information on a sequence of interest by hybridising to it a complementary radio-, fluorescent- or antigen-labelled oligonucleotide. All the ISH in this thesis is concerned with localising RNA, and uses antigen-labelled bases (specifically, the plant steroid digoxigenin) such that immunohistochemistry with anti-digoxigenin antibodies shows the qualitative distribution of the RNA of interest. Briefly, the ISH images in the Allen Mouse Brain Atlas are created as follows (Lein et al. 2007).

First, riboprobes are designed to interrogate the gene of interest (genes are drawn from RefSeq and other sequence databases). Riboprobes are PCR amplified from one of three types of templates: cDNA clones from the NIH's Mammalian Gene Collection (MGC) or RIKEN's FANTOM3; pooled cDNA made from whole mouse brain total RNA; or genomic DNA derived from the tips of mouse tails. Primers for the PCR amplification are designed to avoid repetitive or homologous sequence and to have a moderate GC content. The probes flanked by these primers are required to be within 3kb of a poly(A) tail if from a pooled cDNA template, and wholly within single exons if from a genomic DNA template. Where the source is a cDNA clone, the riboprobe is PCR-amplified using these two primers. Where the source is pooled cDNA or genomic DNA, a larger version of the riboprobe is amplified using these primers, and the riboprobe itself is amplified in a subsequent PCR reaction using a 'nested' primer. Importantly the reverse (or nested, as appropriate) primer includes the binding sequence for the bacteriophage-

derived, DNA-dependent SP6 RNA polymerase as the antisense primer (ensuring that the transcribed probes will be reverse complements of the sequence being probed). The PCR products are confirmed by Sanger sequencing. To produce corresponding single-stranded RNA probes labelled with digoxigenin (on average at every 20-25th nucleotide), the probe sequence is then *in vitro* transcribed from these PCR templates using SP6 RNA polymerase with a mix of ATP, GTP, CTP, UTP and digoxigenin-conjugated UTP. The labelled riboprobes are purified and their quality is assessed with electrophoretograms from a Bioanalyzer.

When the labelled riboprobe is prepared, the ISH can begin. Brains are removed from recently killed mice, frozen in OCT, and cryosectioned in either coronal or sagittal planes. All riboprobes are hybridised on sagittal sections, and a subset of genes with interesting expression patterns or of known or suspected neurological importance are also hybridised on coronal sections. Some sections are stained for Nissl (for extranuclear RNA granules) and others are used for ISH with different genes. Sections are fixed in paraformaldehyde, acetylated to reduce non-specific probe binding effects that can happen with amine and carboxylate functional groups, and dehydrated on glass slides. Tissues are treated to slow endogenous peroxidase activity and to increase tissue permeability to the labelled oligonucleotide probe. Before the colorimetric action is begun, the tissue is treated with several agents to block endogenous protein activity that could interfere with the colorimetric reaction. Probes are then imaged in a four-step procedure: (1) add a solution of anti-digoxigenin antibodies conjugated to horseradish peroxidase (HRP); (2) incubate the tissue with biotin-conjugated tyramide, which HRP converts to a reactive oxidised intermediate that covalently bonds to nearby proteins; (3)

add neutravidin-conjugated alkaline phosphatase, which binds to biotin via neutravidin;
(4) add 5-bromo-4-chloro-3-indolyl phosphate (BCIP) and nitroblue tetrazolium (NBT) so alkaline phosphatase can cleave a phosphate group from BCIP producing indoles that react with NBT to form a blue precipitate. The tissue sections are then washed and fixed with paraformaldehyde for imaging on a background of white light.

When reviewing the results of these reactions one must be mindful of artefacts that can occur in the tissue preparation or at any stage in the colorimetric reaction. With experience, and by comparing with adjacent images in the same series or with results of different probes for the same gene, one can begin to separate fact from artefact. All images were reviewed by expert staff at the Allen Institute and are marked as 'failed' if they appear to have artefacts that could easily be mistaken by a non-expert for real transcription (images having more obvious artefacts, such as bubbles, are not removed).

Chapter 3 - Transcriptomes of Mouse Neocortical Layers

Abstract

In the mammalian neocortex, neurons and glia form a patterned structure across six traditional layers whose complex cytoarchitectonic arrangement, connectivity and physiological properties contribute to cognition. In collaboration with the Molnár laboratory and sequencing facilities at the NIH, I analysed transcriptomes sequenced from layers 1-6b of the adult (postnatal day 56) mouse somatosensory cortex, and of dorsal and lateral areas of neocortex, to understand the transcriptional levels and functional repertoires of coding and noncoding loci for cells constituting these layers. I predict 5,835 protein-coding genes and 66 noncoding RNA loci to be differentially expressed ('patterned') across the layers, based on a machine-learning model (naïve Bayes) approach. Some layers are associated with specific functional and disease annotations that may provide insights into their biological functions. This new resource (<http://genserv.anat.ox.ac.uk/layers>) extends previously available resources, such as the Allen Mouse Brain Atlas and microarray data sets, by providing quantitative expression levels, by being genome-wide, by including novel loci, and by identifying candidate alternatively spliced transcripts that are differentially expressed across layers.

3.1 Introduction

Publication and contributions

This chapter is based on the work described in my first-author paper published in *Neuron* (Belgard et al. 2011). As such, the following text is a considerably modified and expanded version of the published manuscript (with permission from Elsevier, the copyright holder). Contributions from other co-authors are explicitly recognised as such in the text that follows. Briefly, I performed all the computational analysis (and most of the interpretation) for all sections except the lincRNA computational analysis. I also wrote the great majority of the published manuscript. After the stage of transcript building and filtering, the lincRNA computational analyses were performed by Dr. Ana Marques with considerable advisory input from myself and Prof. Chris Ponting. Outside the lincRNA section, the ideas for nearly all analyses came from either Prof. Chris Ponting or myself. Prof. Zoltán Molnár, Dr. Anna Hoerder-Suabedissen, Dr. Fernando Garcia-Moreno, Dr. Tamara Sirey, Dr. Elliott Margulies and Ms. Hatice Ozel Abaan were responsible for sample collection and generating raw sequence. Dr. Peter Oliver was responsible for all the *in situ* hybridisation validations. Dr. Zoltán Molnár, Prof. Chris Ponting and I designed the sample collection and processing. The *in situ* hybridisation and lincRNA analysis methods are detailed in Appendix A.

Background

Comparative and pathological studies suggest that the mammalian cerebral cortex is a critical anatomical substrate in higher cognitive functions including language, episodic

memory and voluntary movement (Jones and Rakic 2010; Kaas 2008; Rakic 2009). The cerebral cortex has a roughly uniform laminar structure that historically has been divided into six layers (Brodmann 1909). The upper layers (1 through 4) form localized intra-cortical connections (Gilbert and Wiesel 1979; Toyama et al. 1974) and are thought to process information locally. The deep layers of the cortex, 5 and 6, form longer distance projections to subcortical targets and to the opposite hemisphere. Some layer 5 neurons are among the largest cells of the brain and exhibit the longest-range connections. Layer 6b in mouse neocortex is a distinct sublayer with characteristic connections, gene expression patterns and physiological properties (Hoerder-Suabedissen et al. 2009; Kanold and Luhmann 2010).

Understanding how neurons and glia are organized into layers to assemble into functional microcircuits (Douglas and Martin 2004) is important to relate cellular characteristics to physiological properties. Sub-classes of pyramidal neurons and interneurons populate specific layers, each characterized by a different depth in the cortex with a specific pattern of dendritic and axonal connectivity (Jones 2000; Lorente de No 1949; Peters and Yilmaz 1993). However, analysing these laminar differences is difficult and often suffers from subjectivity (Zilles and Amunts 2010).

The currently available repertoire of markers that allow the distinction of cortical layers and of many neuronal and glial subtypes is rapidly improving due to developments in cell sorting and gene expression analysis (Doyle et al. 2008; Heintz 2004; Miller et al. 2010; Molyneaux et al. 2007; Monyer and Markram 2004; Nelson et al. 2006; Thomson and Bannister 2003; Winden et al. 2009). These molecular tags can allow highly specific classes of neurons and glia to be monitored, modulated or eliminated, thereby providing

greater insights into cortical neurogenesis and the classification of lamina specific sub-classes of cells.

Laminar molecular markers were first identified by studying single protein-coding genes (Hevner et al. 2006; Molyneaux et al. 2007; Yoneshima et al. 2006), but more recently high-throughput *in situ* hybridisation (Hawrylycz et al. 2010; Lein et al. 2007; Ng et al. 2010) and microarrays (Arlotta et al. 2005; Hoerder-Suabedissen et al. 2009) have enhanced our understanding of mammalian cortical expression patterns for thousands of genes. However, these probe-based approaches are necessarily tied to existing gene models, and exclude thousands of noncoding loci and most alternative splice variants. Despite these limitations and the quality-control challenges posed by industrialized automated histology platforms, the Allen Mouse Brain Atlas is a superb resource for qualitative information on mouse brain gene expression. While microarrays are more quantitative than *in situ* hybridisations, they exhibit a narrow dynamic range compared to the six orders of magnitude that can be spanned with RNA-seq data (Bradford et al. 2010; Trapnell et al. 2009). Instead, we sought to provide accurate and comprehensive genome-wide profiles of transcript and gene expression across cortical cell layers by deep sequencing both coding and noncoding polyadenylated transcripts across adult mouse neocortical layer samples. Owing to cells traversing layer boundaries and dissection limitations, I built naïve Bayes classifiers to infer patterns of layer-specific expression based on the sequencing data.

Brief introduction to the naïve Bayes classifier

Machine learning is a branch of artificial intelligence concerned with forming predictions of a characteristic C based on properties learned from ‘training’ data associated with known C . There are many algorithms used for such ‘supervised learning’, among them Support Vector Machines, Random Forests, and naïve Bayes. For the sake of brevity, I will only discuss the last of these, since it was the only model I used to produce the results in this chapter.

A naïve Bayes (NB) classifier uses a simplistic model to estimate the probability of a binary characteristic C given properties $P_1, P_2, \dots$. These properties can be continuous or discrete (and there can be a mix of the two), but all properties used here are continuous. Importantly, NB classifiers *naïvely* assume that each property *independently* contributes to the likelihood of characteristic C . I will explain the application through an example.

Consider predicting the likelihood, $p(C_R)$, of a gene’s apparent enrichment of expression in a region R based on relative gene expression specificities S_1, S_2 and S_3 (equivalent to the properties P_1, P_2 , etc. mentioned above) from three unknown and possibly overlapping dissections (where specificity, S_i , is the proportion of total expression observed across the three dissections, $S_i = D_i / (D_1 + D_2 + D_3)$, where D_1 is the FPKM in dissection 1). After training this classifier, we wish to ask the following question: “Given that gene X has relative expression pattern S_1, S_2 and S_3 across the dissected samples, what is the likelihood that gene X is a marker of region R ?” A trained classifier will accept specificities S_1, S_2 and S_3 , and use a ‘density estimator’ to calculate from this

the probability $p(C_R)$ that gene X is a marker of region R (see Figure B.1A for an example of this in practice).

This ‘density estimator’ is trained in some way from the data, and explicitly incorporates both a prior probability (“What is the probability that a random gene is a marker of region R ?”) and several conditional probability distributions (e.g. “Given that gene X is a marker of region R , what is the likelihood that it has relative expression specificity S_l ?”). These probability distributions can be modelled in many ways. For example, one can fit a function (for example, normal, lognormal, gamma, or Poisson) to the training data globally. Alternatively, one can approximately model the bivariate relationship without explicitly using a parametric model.

We must then estimate the prior probability and the conditional probability distributions, based on a training set. In this case, a training set would consist of (randomly sampled) genes with known status as markers or non-markers of region R . I estimate the prior probability using the relative frequency in a training set (“What proportion of a random set of genes mark region R ?”). In keeping with the application in this chapter, I will avoid an explicit parameterized model in this toy example and instead estimate the three conditional probability distributions using LOESS (locally weighted scatterplot smoothing) (Cleveland and Devlin 1988).

A function of the conditional probability distribution for S_l (see below) could be modelled as follows: construct a plot of S_l versus probability; on this plot, add points (S_l observed for that gene, $marker_status$) for all the training genes where $marker_status=1$ if the gene is a marker and $marker_status=0$ if not; and fit a highly smoothed Loess

Curve to these points. The conditional probability distribution for S_2 and S_3 would be modelled similarly.

As the name implies, LOESS is a method for fitting a locally weighted smoothed curve to scatterplot data. To do this, Orange (Demšar et al. 2004) calculates a ‘smoothed’ y-value for each x-value corresponding to a training point (when there are enough training examples). This y-value is calculated based on a local, weighted linear regression. To calculate a y-value for a point with x-value x , any points whose x-value, x' , is more distant than half a given window size, $h/2$ (in this case, $h=0.5$; note that this window denotes a proportion of the x-range), are given zero weight in the regression. Any other point, at x' , is weighted using the bi-cubic weight function $(1-|(x-x')/h|)^3$. Y-values for x-values between these points are linearly interpolated.

These are all the data we need to arrive at a final estimated probability. (Note that this probability may need to be recalibrated, but this recalibration does not affect the generalisability metrics of the binary predictions.) According to Bayes’ Theorem,

$$p(R|S_1, S_2, S_3) = \frac{p(R)p(S_1, S_2, S_3 | R)}{p(S_1, S_2, S_3)} \text{ where } p(R|S_1, S_2, S_3) \text{ is the } a \text{ posteriori probability}$$

that a gene is a marker of region R given observed specificities S_1 , S_2 , and S_3 ; $p(R)$ is the *a priori* probability that a gene is a marker of region R ; $p(S_1, S_2, S_3 | R)$ is the probability that a gene has specificities S_1 , S_2 and S_3 given it is a marker of region R ; and $p(S_1, S_2, S_3)$ is the *a priori* probability of S_1 , S_2 , and S_3 . Invoking the ‘naïve’ assumption that properties S_1 , S_2 and S_3 contribute independently of one another, this can be rewritten as

$$p(R|S_1, S_2, S_3) = \frac{1}{p(S_1, S_2, S_3)} p(R) \prod_{i=1}^3 p(S_i | R) \text{ where } p(S_i | R) \text{ is the probability that a gene}$$

has specificity S_i given it is a marker of region R . Likewise, the probability, $p(nR)$ that gene X is *not* a marker of region R is given by

$$p(nR | S_1, S_2, S_3) = \frac{1}{p(S_1, S_2, S_3)} p(nR) \prod_{i=1}^3 p(S_i | nR). \text{ By definition, the two probabilities}$$

must sum to one: $p(nR | S_1, S_2, S_3) + p(R | S_1, S_2, S_3) = 1$ and $p(nR) + p(R) = 1$. Following some algebraic manipulation, one can rewrite the above equations as

$$p(R | S_1, S_2, S_3) = \frac{1}{1 + \frac{1 - p(R)}{p(R) \prod_{i=1}^3 \frac{p(S_i | R)}{p(S_i | nR)}}}. \text{ This is now simple to compute, as } p(R) \text{ is}$$

simply the proportion of marker genes for region R in the training data and

$\frac{p(S_i | R)p(R)}{p(S_i | nR)(1 - p(R))}$ is the quantity estimated above using the LOESS curves for S_1 , S_2 and S_3 .

Finally, one can assess the generalisability of such a model using cross-validation. There are different approaches to cross-validation, but for brevity I will only explain the one used here. First, one randomly divides the training data into ten equally sized bins. Then, for each bin in turn, one trains a naïve Bayes classifier using data in the remaining nine bins, and tests the accuracy (and other metrics of model generalisability) using data in the remaining bin. The median of the resulting ten estimates of each generalisability metric suggests how the model might be expected to perform when subsequently classifying data with an unknown solution.

Chapter structure

The remainder of this chapter is organised into the following sections: Materials and Methods, Results and Discussion. The following topics are described in the Materials and Methods and Results sections:

- layer dissections & cDNA sequencing;
- gene quantification and reference-based transcript assembly;
- assessing accuracy of the layer dissections based on known layer markers;
- comparison to microarray data;
- constructing, assessing and applying a naïve Bayes classifier to predict layer specificity based on observed expression patterns of genes and transcripts;
- estimating the overall prevalence of genes differentially expressed across layers;
- assessing common functional characteristics of genes enriched in particular layers;
- replications of the dissections and sequencing in dorsal and lateral cortex;
- and identification and computational characterisation of candidate lincRNA loci (methods in Appendix A).

3.2 Materials and Methods

Mice and initial dissections

The initial batch of brains came from eight adult male mice (56 days old; C57BL/6J strain) that were killed by cervical dislocation according to approved schedule one UK Home Office guidelines (Scientific Procedures Act, 1986). The eight comprised two

groups of four littermates each. The mice were decapitated, the skull opened down the midline and the brain removed. Newly dissected brains were rinsed in RNase free PBS, submerged in ice-cold RNAlater (Ambion) for 24 hours and stored at -20°C in RNAlater (Ambion). Dr. Anna Hoerder-Suabedissen embedded whole brains in 5% agarose and sectioned them using a vibrating microtome (Leica, VT1000S) into 200 micrometer coronal sections using a 1:1 mixture of RNAlater and PBS. Prof. Zoltán Molnár dissected out six sections corresponding approximately to cortical layers 1-3, 4, upper 5, lower 5, 6, and 6b (henceforth referred to as samples A-F, respectively) under visual guidance, using transillumination on a dissecting microscope (MZFLIII, Leica) and stored these separately in RNAlater at -80°C until all microdissection was complete. For RNA extraction samples from individual zones from the eight mice were combined and all tissue samples were processed concurrently. To provide a biological replicate sample, the RNA for sample B was extracted and pooled by litter, providing two samples each representing 4 mice (heretofore known as samples B1 and B2). Dr. Tamara Sirey extracted total RNA >200 nt using the RNeasy Lipid Tissue Mini kit (QIAGEN), following the manufacturer's instructions and using the on-column DNase digest. She assessed RNA quantity using a NanoDrop 1000 spectrophotometer (ThermoScientific) and RNA quality and integrity using a Bioanalyzer (Agilent Laboratories) (Table B.1).

Sequencing, read mapping, transcript building and quantification for the initial dissections

I shipped these samples on dry ice to Hatice Ozel Abaan, a technician at the National Human Genome Research Institute, who prepared total RNA for paired-end deep sequencing on Illumina's Genome Analyzer Iix following the manufacturer's protocol.

Briefly, poly(A) RNA was enriched using Dynal oligo(dT) beads (Invitrogen). This was fragmented using the RNA fragmentation kit (Ambion). First- and second-strand cDNA was synthesized with random hexamer primers (Invitrogen) and SuperScript II (Invitrogen). Ends were repaired using T4 DNA polymerase and Klenow DNA polymerase. A single adenosine and Illumina adapters were ligated using Klenow 3'-to-5' exo-nuclease. Following gel purification of cDNA templates, the library was enriched with 15 rounds of PCR before being added to the flow cell for paired-end sequencing. 51 nt were sequenced from either end, of which 50 nt were used for mapping for all lanes except two of the five lanes of sample F, in which 76 nt were sequenced and 75 nt were used in mapping.

These mixed read lengths had only negligible effects on the quantifications used in downstream analyses. Two of the five lanes sequenced for sample F (090723_NUTELLA_429NWAAXX.7 & 090723_NUTELLA_429NWAAXX.8) were mapped using 75 nucleotides instead of 50 nucleotides (mapping and quantification details are given below). Metrics of sample F (% genome covered, correlation to other samples, % genes covered, etc.), which included the 75 nt reads, were similar to those of other samples. Indeed, the Spearman's rank correlation (for genes covered in both sets, since the reduced set has fewer reads) between genic FPKM values from sample F including the 75 nt reads and genic FPKM values from sample F excluding the 75 nt reads ('reduced' sample F) is 0.996. This correlation was also 0.996 when only considering the reduced set of genes that were used in the naïve Bayes classifiers (see below). When I recalculated the sample E/F pairwise correlation, among the highest in Table 3.1, by removing genes not expressed in both sets, it came to 0.982, significantly

lower ($p < 10^{-300}$, Fisher r-to-z transformation) than the 0.996 correlation of ‘reduced’ sample F and the full sample F. These expression quantifications are thus far more similar to one another than the differences observed between samples. As these data were mapped to identical splice sites, we need to consider only the small difference in mappability of 2x50 nt reads versus 2x75 nt reads in the transcriptome. For transcript-level expression values, the correlation of sample F with the ‘reduced’ sample F was 0.992. Moreover, for transcripts used in the alternative splicing analysis, the correlation of sample F with the ‘reduced’ sample F was 0.995, suggesting that the transcript-level alternative splicing analysis would be affected as little as the gene-level analysis. Finally, all patterning analyses below were based on the results of the classifiers. One would expect global (but not specific) effects on gene quantification to be correspondingly reflected in the classifier generalisation metrics.

Table 3.1: Inter-sample gene expression Spearman rank correlations. These correlations use the FPKM values of Ensembl (release 57) genes as quantified by cufflinks. By definition, the table is symmetrical about the diagonal.

	A	B1	B2	C	D	E	F
A	1.000						
B1	0.921	1.000					
B2	0.942	0.965	1.000				
C	0.922	0.968	0.972	1.000			
D	0.919	0.965	0.967	0.979	1.000		
E	0.915	0.957	0.962	0.973	0.978	1.000	
F	0.907	0.950	0.950	0.961	0.966	0.972	1.000

Reads were processed with versions 1.4 and 1.6 of the Illumina pipeline. I estimated the internal insert size and standard deviation (modelling it as a Gaussian distribution) for each library by empirically calculating the full width at half max (FWHM) for an internal insert size distribution constructed using uniquely mapping reads on the same

chromosome from Illumina's Gerald pipeline. I estimated the insert size to be the midpoint of the minimum and maximum insert sizes at half (or greater) the frequency of the mode insert size, and calculated the standard deviation under a Gaussian assumption $StdDev = FWHM / (2\sqrt{2\ln 2})$. Using this insert size and standard deviation, I separately mapped lanes with tophat v1.0.13 (Trapnell et al. 2009) to the mouse reference genome (mm9, downloaded from UCSC) plus all splice junctions from the UCSC Table Browser (Rhead et al. 2010) tables all_mrna (mouse mRNAs), intronEst (mouse spliced ESTs), and xenoMrna (mRNAs mapped from other species to mouse) with the following options: butterfly search, closure search, fill gaps, microexon search, min anchor length = 5, min isoform fraction = 0.0, max intron length = 500 kb. In a second round of tophat mapping (default options except min isoform fraction = 0.0), all novel junctions identified from any sample in the previous step were given to tophat again (as raw junctions), so splice junctions could be measured across all samples in an unbiased manner (Table B.2).

I built *de novo* transcript models for each sample using cufflinks v0.8.3 (Trapnell et al. 2010), providing the average insert sizes and standard deviations calculated previously and using default options. I combined transcript models across all samples using cuffcompare with default options, and comparing against a GTF of Ensembl gene models (Flicek et al. 2011) downloaded from Ensembl (release 57) that I adapted to use the UCSC nomenclature. I then assessed expression levels of the resultant unified cuffcompare transcript models in every sample using cufflinks in quantification-only mode with default settings. Finally, I assessed expression levels of all genes and

transcripts in Ensembl (release 57) with cufflinks in quantification-only mode with default settings.

Sets of curated genes

I downloaded manually annotated layer expression patterns across neocortical layers for genes (matched for strain, sex, age, and cortical region) from the Allen Mouse Brain Atlas (<http://mouse.brain-map.org/pdf/SomatosensoryAnnotation.xls>; downloaded 3 June 2010). Gene symbols from distinctive categories containing 11 genes or more were extracted and mapped to Ensembl gene identifiers using the MGI Batch Query (<http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=batchQF>). Only those genes having Ensembl identifiers were kept. These categories were then mapped to a smaller number of categories, corresponding to genes enriched in layers 2/3, 4, 5, 6, 6b and genes showing no layer enrichment (or depletion). Corresponding categories were constructed from other genes for the inverse, e.g. “not in 4”. For example, genes enriched in layers 2/3 and 5 were marked as enriched in both of these layers. The two negative categories (“not in 4” and “not in 5”) were not included in any sets of layer-enriched genes, but only to the sets of genes that were not enriched in layers 4 and 5, respectively.

Comparison with microarray data from juvenile mice

I compared these RNA-seq results with expression levels previously obtained using microarrays for layer 6 and 6b from anterior cortex (putative S1) of postnatal day 8 mice (Hoerder-Suabedissen et al. 2009). I translated microarray summary probe IDs having a fold change greater than 1.5 and a reported *p*-value less than 0.05 to Ensembl gene IDs

using the MGI batch query. Unfortunately these p -values were not corrected for multiple testing in the original microarray analysis that was done for the previous publication, and I could not obtain the full data set that would permit a proper correction for multiple testing. Relative expression levels of these gene IDs were then compared between samples E and F. The true concordance may be even higher than the (already high) concordance reported below given the likely abundance of false positives in the microarray calls.

Classifier training and validation

To remove low-quality quantifications and to improve predictions, I only used for classification the *de novo* transcript models, *de novo* gene models, Ensembl transcript models, and Ensembl gene models if the width of their largest 95% confidence interval of expression quantification among the samples was less than or equal to 50% the average FPKM of that gene or transcript across libraries. This retained 11,410 (34%) Ensembl genes (release 57) and 10,261 (45%) Ensembl protein-coding genes.

2,200 of the ‘classifiable’ Ensembl genes had a manually annotated expression pattern across layers from the Allen Mouse Brain Atlas as described above. For each individual layer 2/3-6b and “no layer enrichment”, I used the interactive software package Orange (Demšar et al. 2004) to train a naïve Bayes classifier (using default settings) to assign, for all genes, the probability that a gene was enriched in the layer of interest (Figure B.1).

I also used Orange to produce, for each classifier, a pictorial representation of the classification model (Figure B.1), a Receiver Operating Characteristic (ROC) curve

reflecting the sensitivity-specificity tradeoff (Figure B.2), and a plot relating the classifier's estimated probability to the actual probability of enrichment based on the training set of 2,200 genes (Figure 3.4).

I subsequently recalibrated the estimated probability produced by the models to the observed probabilities by locally weighted scatterplot smoothing (LOESS) (Cleveland and Devlin 1988) using the Excel add-in available at http://peltiertech.com/images/2009-10/PTS_LOESS.xls (Figure 3.4). Points corresponded to genes of known laminar patterning (or lack thereof), with the x-coordinate equal to the classifier-produced probability and the y-coordinate either 1 or 0, corresponding to curated enrichment or not, respectively, for that layer. The smoothing parameter, alpha, was 0.33 (726 points in moving regression) for all models except layer 4 and layer 6, where alpha was raised to 0.65 (1430 points) and 0.45 (990 points) respectively, to avoid edge effects.

No 'model selection' step was necessary for the naïve Bayes classifiers, since there were no user-adjustable parameters optimized. Hence, classifier metrics based on ten-fold cross-validation are expected to generalize when applied more broadly to expression distributions across samples of genes and transcripts. This assumes that the average correspondence between apparent layer distributions derived from *in situ* hybridisations and gene expression distributions across samples derived from RNA-seq is similar for Allen Mouse Brain Atlas-curated genes and our classifiable genes and transcripts. So long as this average is consistent, specific deviations (which may arise from complex patterns of expression across diverse cell types, where quantitative expression measurement may not match qualitative distribution measures) are accounted for by the classifier's measures of accuracy.

One exception to this rule is in layer predictions of genes at FPKM<1.0, which may have elevated rates of false positives. Recall that to obtain a subset of all genes for classification, I filtered on the basis of confidence intervals as reported by cufflinks. These quantification estimates were previously validated to be accurate even below 1 FPKM using spike-ins and *in silico* simulation. However, in our experiments, the amount of total sequence gathered per library varied by as much as two fold (Table B.2). Hence, even for equivalent levels of ‘true’ normalized FPKM expression, the tightness of confidence intervals will vary amongst libraries.

Indeed, genes that never attained an FPKM>1.0 in any sample were 49% more likely to be called as patterned than the Allen Mouse Brain Atlas (AMBA) curated genes at similar levels of expression ($p=0.002$; hypergeometric test). No other expression bins showed a significant increase in patterned genes (Figure 3.6). Instead, for most bins, classifiable genes were significantly less likely than the AMBA curated genes at equivalent expression levels to be predicted as patterned. This, however, is expected: the curated AMBA genes are enriched for patterned genes (<http://mouse.brain-map.org/pdf/SomatosensoryAnnotationWhitePaper.pdf>).

Genes are 35% more likely ($p<10^{-11}$; hypergeometric test) to be called as patterned in the FPKM<1 bin than in all the other bins combined. So if one accepts the simplifying assumption that genes at all expression levels (i.e. in all bins) should have the same ratio of patterned to unpatterned genes, one could estimate that $1-1/1.35=25\%$ of these predicted patterned genes <1.0 FPKM are false positives. Since this bin accounts for 740 predicted patterned genes (11% of the total predicted patterned genes), this indicates 185 false positives, or 2.75% of the total patterned genes. Fewer than ~3% false positives

may not be accounted for in the current classifier measures of false positives. Hence, this should not substantially affect the larger set of all patterned genes.

There is a good explanation for why the cufflinks confidence interval-based filtering may not have removed these. The sequencing libraries had not inconsiderable differences in numbers of reads, which means the confidence intervals will be consistently smaller in some. For genes at very low levels of expression (which may not be overlapped in all libraries), this could exclude 'no expression' from the 95% confidence interval of a gene in one library and the same gene could not have a single overlapping read in another library (there are many examples of this), even if they were actually expressed at the same level. Accepting the well-validated cufflinks confidence intervals, this seems the most parsimonious theoretical explanation.

Hence, I suggest the 89% of total predicted patterned genes >1 FPKM may be most fruitful for experimental follow up.

Other machine learning methods considered

In addition to naïve Bayes, I considered two other machine-learning approaches: Support Vector Machines (SVMs) and random forests. I constructed SVMs with radial basis functions, where C and γ were chosen to optimize the F1 score (the harmonic mean of precision and recall) using leave-one-out cross-validation assessed using a separate test set (15% of the starting set). These had comparable classification metrics to the naïve Bayes models, so I chose the latter since they provide a per-gene probability score of enrichment as opposed to just a binary classification. Interestingly, I found random forest classifiers had comparable performance to naïve Bayes. The random

forest classifiers produced more accurate uncalibrated probabilities. In contrast to the monotonic calibration curves of the naïve Bayes classifiers, those of the random forest classifiers were sometimes not monotonic.

Classifier predictions of genes and transcripts

Once trained and validated, I applied classifiers to the laminar expression distributions of known and *de novo* genes and transcripts that met the criteria for classification as described above. A single Ensembl gene was considered to have alternatively spliced variants and to undergo isoform switching across layers if both of the following conditions were met:

1. In at least one pair of sequenced samples, the 95% confidence intervals of FPKM expression of a transcript of this gene (as calculated by cufflinks) must not overlap, indicating higher expression in one sample. Another transcript of this gene must additionally have non-overlapping 95% confidence intervals for the same two samples that indicate higher expression in the other sample. Of the 2,003 classifiable genes (17 receptors or ion channels), this retained 1,646 (82%), of which 14 encoded receptors or ion channels.
2. The Euclidean distance in calibrated layer enrichment probability space (square root of the sum of the squares of the differences of layer enrichment probabilities) was greater than 0.5 for at least one pair of its isoforms. Of the 1,646 classifiable genes showing significant differences in expression across samples, this retained 719.

Estimating the prevalence of patterned genes

Based on the unpatterned classifier (Table 3.2), and adjusting for precision and recall, I expect 5,835 of the classifiable genes to be patterned (11,410-5,575). The sum total of genes expected to be enriched in each of the layers (redundant set) is 7,972, which means we expect a gene to be enriched in, on average, 1.37 ($7,972/5,835$) layers. This is consistent with the 1.38 ($[\text{sum of genes enriched in each layer} / \text{number of unique genes enriched in any layer}] = 2,459/1,780$) average number of layers in which patterned genes we extracted from Allen Mouse Brain Atlas curations (Lein et al. 2007) were expressed. This suggests that, since one may expect these ratios to be similar, extrapolation of total number of patterned genes from the unpatterned classifier statistics is consistent with the layer-specific classifiers. This ratio further implies that most of these patterned genes are markers of single layers.

This proportion (51%) is lower than the 62% ($1,780/2,861$) of genes annotated as being patterned from *in situ* hybridisation images (Lein et al. 2007) (Online Supplementary Table 6 at http://www.fgu.anat.ox.ac.uk/~grantb/mouse_layers/tables.html). This was unsurprising, as the curated Allen Mouse Brain Atlas gene set was enriched for genes likely to be patterned in expression across layers (<http://mouse.brain-map.org/pdf/SomatosensoryAnnotationWhitePaper.pdf>).

Functional differences

I looked for two types of functional difference in these predicted gene sets: (1) functions enriched or depleted in genes predicted to be patterned across layers as compared to genes predicted to be evenly expressed, and (2) functions enriched in genes expressed in

a specific layer as compared to the set of all classifiable genes. I addressed these two questions in different ways owing to the complicating nature of the classifiers. The functional sources of these databases and more (I built an extensive functional browser available at <http://genserv.anat.ox.ac.uk/cortex>) are provided in Table B.5 and described in Section 2.5 above.

The first type of functional difference was based on a two-sided Fisher's exact test comparing the predicted set of patterned genes (all genes predicted to be patterned by at least one classifier) and the predicted set of unpatterned genes. The null hypothesis of each term-wise test is that there is no difference in the proportion of genes with that term between the patterned and unpatterned sets of genes. I only performed a test for a term if the test was sufficiently powered to detect the maximal possible difference at a p -value < 0.05 , given the frequency of that term in the union of patterned and unpatterned sets. The R package `fisher.test` was used. For 'conditional' databases (mouse knockout phenotypes (Blake et al. 2011) and Gene Ontology (GO) cellular component (Ashburner et al. 2000)), the 2x2 contingency table was only constructed with genes having at least one annotation in that database. For non-conditional databases (KEGG (Kanehisa et al. 2004) molecular pathways, mouse orthologs of human genes nearby SNPs associated with phenotypes by the Ensembl Variation database (Chen et al. 2010) and GO (Ashburner et al. 2000) biological process and molecular function, I used all genes to construct the contingency table.

I based the second type of functional difference on a simulated null distribution. The null hypothesis of each term-wise test is that there is no difference in the proportion of genes with that term between the genes with enriched expression in the layer being

considered and all cortex-expressed classifiable genes. Briefly, for each layer, I simulated genes with replacement from the set of all classifiable genes with a probability reflecting the precision of the classifier for that layer. Otherwise, genes were selected with replacement from one of the predicted sets with a likelihood that would best simulate the quantified sources of false positives for that classifier (see below). This continued until the number of simulated genes matched the number of genes in the predicted set (or the number of genes associated with a term in that functional database, for conditional databases). I empirically determined p -values for the one-sided test from 200,000 such simulations for every term included in the background distribution, including those terms having no corresponding genes in the foreground. This simulation was also done for genes predicted to have no layer enrichment.

Simulating the null distributions for assessing functional differences in individual predicted sets

I simulated a null distribution to assess functions enriched in genes expressed in a specific layer as compared to the set of all classifiable genes. The null hypothesis of each term-wise test is that there is no difference in the proportion of genes with that term between the genes with enriched expression in the layer being considered and all cortex-expressed classifiable genes. To avoid false functional enrichments in a layer due to the presence of false positives from another layer, genes were simulated with replacement as follows:

1. Have a pseudo-random number generator pick a number, x , from 0 to 1. If this number is less than the precision of the set, randomly sample with replacement a gene

from all classifiable genes for non-conditional databases, or from the set of all classifiable genes with annotations in this database for conditional databases. We seek the contribution of the true gene expression from a specific layer, although there will be differing degrees of overlap from different layers (false positives). Fortunately, we have approximate quantifications for this from the ten-fold cross-validation. So if $x > precision$, randomly sample with replacement a gene from one of the predicted sets of layer-enriched genes (or, for conditional databases, from the subset of genes having associated terms) with probabilities as calculated from an approximate solution to Equation 1.

2. Repeat step 1 until the total number of genes in the simulated set is equal to the total number of predicted genes enriched in that layer (or the number of genes with associated terms for conditional databases).
3. For each term in the database, count the number of genes (allowing repeats) with that term. Save this and go back to step 1 until there are 200,000 resamples.

Equation 1: Probabilities for simulating the false positives.

$$\begin{bmatrix} Precision_{2/3} & FP_{2/3}(4) & FP_{2/3}(5) & FP_{2/3}(6) & FP_{2/3}(6b) & FP_{2/3}(none) \\ FP_4(2/3) & Precision_4 & FP_4(5) & FP_4(6) & FP_4(6b) & FP_4(none) \\ FP_5(2/3) & FP_5(4) & Precision_5 & FP_5(6) & FP_5(6b) & FP_5(none) \\ FP_6(2/3) & FP_6(4) & FP_6(5) & Precision_6 & FP_6(6b) & FP_6(none) \\ FP_{6b}(2/3) & FP_{6b}(4) & FP_{6b}(5) & FP_{6b}(6) & Precision_{6b} & FP_{6b}(none) \\ FP_{none}(2/3) & FP_{none}(4) & FP_{none}(5) & FP_{none}(6) & FP_{none}(6b) & Precision_{none} \end{bmatrix} \begin{bmatrix} Prob_k(2/3) \\ Prob_k(4) \\ Prob_k(5) \\ Prob_k(6) \\ Prob_k(6b) \\ Prob_k(none) \end{bmatrix} = \begin{bmatrix} Prop_k(2/3) \\ Prop_k(4) \\ Prop_k(5) \\ Prop_k(6) \\ Prop_k(6b) \\ Prop_k(none) \end{bmatrix}$$

$Prob_k(i)$ is the probability used to decide from which layer i a false positive for predicted layer k should be simulated (its value is 0 when $i=k$). $FP_k(i)$ is the probability that a gene predicted to be enriched in layer i was a false positive truly enriched in layer k .

$\text{Prop}_k(i)=\text{FP}_k(i)$ for $k \neq i$ and needs to be calculated when $i=k$ (using the fact that $\text{Prob}_k(i=k)=0$). Precision_i is the precision of the classifier for layer i and is determined empirically from ten-fold cross-validation. $\text{FP}_k(i)$ is determined empirically using the falsely called known genes. We do not use ‘positives’ by themselves because we wish to capture the effects of genes which are truly enriched in one layer but also happen to be truly enriched in a second layer, but not a third. We do not use higher-order (i.e. multi-layer) categories because, due to modest recall per layer, we were unable to reassemble those genes that have a complex multilayer pattern such as 2/3/5/6b with high accuracy. Instead, we use the same six categories used for classification: 2/3, 4, 5, 6, 6b, and no layer enrichment. Hence these values and a normalization factor N are determined from Equation 2. Since the exact solution results in negative sampling probabilities, I instead numerically minimized an error function (the sum of the squares of the deviation from the proper $\text{Prop}_k(i)$ ’s) with an iterative dense random sampling of the search space that consistently converged on the same solution.

Equation 2

$$\text{Precision}_k + N \sum_j \text{FP}_k(j) = 1$$

$$\text{Hence, } \text{Precision}_k + \sum_j \text{Prob}_k(j) = 1$$

False negatives (measured by recall) are expected to adversely impact the power of this test to detect functional differences.

For all genes, the starting matrix was:

$$\begin{bmatrix} \text{Precision}_{2/3} & FP_{2/3}(4) & FP_{2/3}(5) & FP_{2/3}(6) & FP_{2/3}(6b) & FP_{2/3}(\text{none}) \\ FP_4(2/3) & \text{Precision}_4 & FP_4(5) & FP_4(6) & FP_4(6b) & FP_4(\text{none}) \\ FP_5(2/3) & FP_5(4) & \text{Precision}_5 & FP_5(6) & FP_5(6b) & FP_5(\text{none}) \\ FP_6(2/3) & FP_6(4) & FP_6(5) & \text{Precision}_6 & FP_6(6b) & FP_6(\text{none}) \\ FP_{6b}(2/3) & FP_{6b}(4) & FP_{6b}(5) & FP_{6b}(6) & \text{Precision}_{6b} & FP_{6b}(\text{none}) \\ FP_{\text{none}}(2/3) & FP_{\text{none}}(4) & FP_{\text{none}}(5) & FP_{\text{none}}(6) & FP_{\text{none}}(6b) & \text{Precision}_{\text{none}} \end{bmatrix} = \begin{bmatrix} 0.4440 & 0.2178 & 0.0381 & 0.0381 & 0.0598 & 0.1466 \\ 0.0688 & 0.3515 & 0.0196 & 0.000 & 0.0028 & 0.0114 \\ 0.1374 & 0.2275 & 0.6148 & 0.0810 & 0.1310 & 0.1949 \\ 0.0104 & 0.0339 & 0.0173 & 0.3805 & 0.1082 & 0.0465 \\ 0.0271 & 0.0097 & 0.0242 & 0.1954 & 0.3906 & 0.1115 \\ 0.3124 & 0.1597 & 0.2860 & 0.3050 & 0.3075 & 0.4890 \end{bmatrix}$$

The following probability columns were subsequently used in the simulations:

$$\begin{bmatrix} \text{Prob}_{2/3}(\text{background}) \\ \text{Prob}_{2/3}(4) \\ \text{Prob}_{2/3}(5) \\ \text{Prob}_{2/3}(6) \\ \text{Prob}_{2/3}(6b) \\ \text{Prob}_{2/3}(\text{none}) \end{bmatrix} = \begin{bmatrix} 0.4440 \\ 0.1192 \\ 0.0646 \\ 0.0013 \\ 0.0000 \\ 0.3708 \end{bmatrix}, \begin{bmatrix} \text{Prob}_4(2/3) \\ \text{Prob}_4(\text{background}) \\ \text{Prob}_4(5) \\ \text{Prob}_4(6) \\ \text{Prob}_4(6b) \\ \text{Prob}_4(\text{none}) \end{bmatrix} = \begin{bmatrix} 0.3203 \\ 0.3515 \\ 0.3130 \\ 0.0001 \\ 0.0013 \\ 0.0138 \end{bmatrix}, \begin{bmatrix} \text{Prob}_5(2/3) \\ \text{Prob}_5(4) \\ \text{Prob}_5(\text{background}) \\ \text{Prob}_5(6) \\ \text{Prob}_5(6b) \\ \text{Prob}_5(\text{none}) \end{bmatrix} = \begin{bmatrix} 0.0014 \\ 0.0438 \\ 0.6148 \\ 0.0020 \\ 0.0029 \\ 0.3350 \end{bmatrix},$$

$$\begin{bmatrix} \text{Prob}_6(2/3) \\ \text{Prob}_6(4) \\ \text{Prob}_6(5) \\ \text{Prob}_6(\text{background}) \\ \text{Prob}_6(6b) \\ \text{Prob}_6(\text{none}) \end{bmatrix} = \begin{bmatrix} 0.0008 \\ 0.0004 \\ 0.0060 \\ 0.3805 \\ 0.2593 \\ 0.3529 \end{bmatrix}, \begin{bmatrix} \text{Prob}_{6b}(2/3) \\ \text{Prob}_{6b}(4) \\ \text{Prob}_{6b}(5) \\ \text{Prob}_{6b}(6) \\ \text{Prob}_{6b}(\text{background}) \\ \text{Prob}_{6b}(\text{none}) \end{bmatrix} = \begin{bmatrix} 0.0008 \\ 0.0121 \\ 0.0048 \\ 0.1277 \\ 0.3906 \\ 0.4640 \end{bmatrix}, \begin{bmatrix} \text{Prob}_{\text{none}}(2/3) \\ \text{Prob}_{\text{none}}(4) \\ \text{Prob}_{\text{none}}(5) \\ \text{Prob}_{\text{none}}(6) \\ \text{Prob}_{\text{none}}(6b) \\ \text{Prob}_{\text{none}}(\text{background}) \end{bmatrix} = \begin{bmatrix} 0.1233 \\ 0.0008 \\ 0.2601 \\ 0.0004 \\ 0.1257 \\ 0.4890 \end{bmatrix}$$

Correcting for multiple testing in testing functional differences

To account for multiple testing, I calculated q -values (which reflect the smallest false discovery rate at which a term would be significant) from empirical p -values using QALITY v1.11 (Kall et al. 2008) on a per-database basis for both types of functional comparisons. Enrichments not meeting a q -value threshold of 0.05 were discarded, controlling false positives at or below 5%.

Replications in dorsal and lateral cortex

Dr. Fernando Garcia-Moreno dissected laminar samples from dorsal cortex (overlapping S1 and some motor cortex) and, separately, lateral cortex (partially overlapping S2 and insular cortex) from mice matched in number, strain, sex, litter distribution, and age to

the original set of mice. These additional eight adult male mice (56 days old; C57BL/6J strain) had also been killed by cervical dislocation according to approved schedule one UK Home Office guidelines (Scientific Procedures Act, 1986). These were treated exactly as described above, with the following exceptions: sample B was not split and RNA underwent two rounds of poly(A) selection. I discarded lanes with aberrant GC content and trimmed all reads to correspond to two ends of 50 base pairs each. I discarded a read fragment if any of the following conditions was met: either read failed Illumina's chastity filter, either read had eight or more bases (corresponding to tophat's default minimum anchor length) with a Phred-scaled quality score less than three, either read mapped (bowtie v0.12.7 allowing up to two mismatches) to spliced or unspliced rRNAs, tRNAs or mitochondrial rRNAs as defined by Ensembl (release 62). I estimated fragment internal insert size and standard deviation as described above. I subsequently mapped fragments on a lane-by-lane basis with tophat v1.2.0 (Trapnell et al. 2009) enabling microexon search, coverage search, butterfly search, and providing gene models from Ensembl (release 62). I then combined all splice sites found in any lane (from original or replicate dorsal cortex samples, or lateral cortex samples) for a final round of tophat mapping in which new splice site detection was disabled but all previously detected splice sites were provided. I merged all lanes of the same libraries and quantified expression levels of all Ensembl genes (release 62) with cufflinks v0.9.3 (Trapnell et al. 2010), masking rRNA, tRNA and all mitochondrial transcripts from the denominator for the FPKM calculation, and enabling the bias correction module (Roberts et al. 2011) and upper-quartile normalization.

Once quantified, I filtered genes and provided several characteristics to a naïve Bayes classifier. First, genes were required to have a minimum FPKM of at least 1 in at least one replication dorsal cortex sample. Second, at least one relative error of the FPKM estimate (as provided by cufflinks) was required to be below 50%. Finally, the gene's quantification had to be marked as 'OK' by cufflinks in all samples. The following variables were then provided to the naïve Bayes classifier for each gene: average FPKM, % GC in the gene locus, and, for every sample, the relative proportion of total expression in that sample and the relative error on that expression level. (Genes with no overlapping reads in a sample were defined to have a relative error of 10,000%, above any empirically observed relative errors for genes overlapping at least one read, in that sample to reflect to the classifier the high degree of uncertainty for such genes.) I then trained classifiers, in turn, on these variables, and reported metrics as described above.

I trained and assessed all classifiers using data curated from S1 since expression patterns are known to be largely similar between dorsal and lateral cortex (Hawrylycz et al. 2010). (An advantage of naïve Bayes classifiers is their robustness to noisy data.) Given that noisy test data will artificially depress the classifier's generalization measures, the fact that the AUCs (see Figure 3.1 caption for description of AUC) for these classifiers, given in Table 3.5, are similar to the AUCs in dorsal cortex confirms again these assumptions were good.

I confirmed functional enrichments, as reflected in Figure 3.10B/C, by performing Fisher's exact tests (one-tailed, in which the alternative hypothesis is that the gene set predicted to be enriched has more genes with that functional annotation than does the gene set predicted not to be enriched). These confirmations demonstrate that the prior

inclusion of genes below 1 FPKM cannot account for the significant enrichments. I confirmed functional differences between patterned and unpatterned genes, as reflected in Figure 3.10A, by performing two-tailed Fisher's exact tests. These were appropriately tested as 'conditional' or 'nonconditional' databases, as described above. For a term to be tested, at least two genes were required to have the annotation and there must have been sufficient power to hypothetically achieve a p -value of at most 0.05. In both cases where the numbers were too large for R's Fisher's exact test function, I instead used the Chi-squared function.

Interestingly, classifier performance was not improved when I used data from all samples from the S1 and dorsal cortex dissections simultaneously. This suggests the reported generalized classifier performance may be limited by (1) artefacts, noise or mistakes in the *in situ* hybridisation curations (for examples of these see Table S3 in Belgard et al. (2011)), (2) fundamental differences in what is measured by *in situ* hybridisation (analogue expression within individual cells) and by RNA-seq (digital expression averaged across the sample), and/or (3) systematic biases in both sets of RNA-seq data or informatic pipelines. The first of these would cause the predicted probabilities to be overly conservative, and the effects of the second and third of these are already accounted for in the calibrated probabilities.

3.3 Results

Gene expression levels across dissected samples reflect known neocortical layers

Dr. Tamara Sirey extracted polyadenylated RNA from dissections of neocortical layers from 8 male, 56-day-old, C57BL/6J mice. RNA samples A-F were derived from six adjacent laminar segments (from superficial to deep layers, respectively) of mouse primary somatosensory cortex (S1) from two sets of four littermates dissected under a binocular microscope with a microsurgical scalpel (Figure 3.1A). Samples B1 and B2 represent biological replicates of the second segment from four littermates each.

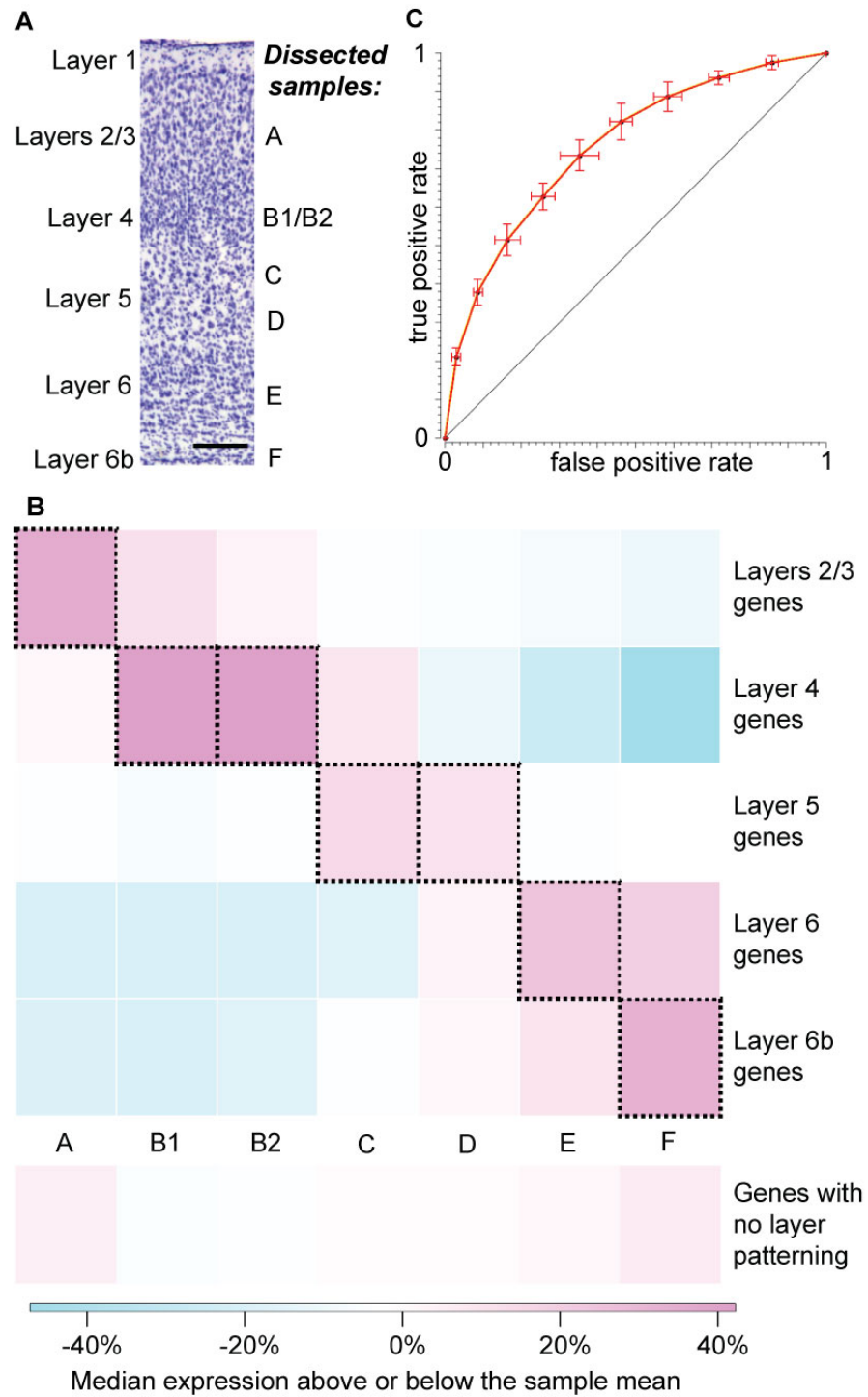


Figure 3.1: Laminar dissections (six samples A-F, including B replicates B1/B2) correspond to known cytoarchitectural layers. (A) Prof. Zoltán Molnár manually dissected samples using a microsurgical scalpel under binocular microscope

from coronal sections of primary somatosensory cortex. Nissl stain of RNA reveals layers with different cell densities across adjacent layers 1-6b and their correspondence to laminarily dissected samples (scale bar = 200 μm). (B) Genome-wide RNA-Seq gene expression patterns across samples A-F recapitulate gene expression preferences in layers 2-6b known from manual curation (Lein et al. 2007). The heatmap reflects the relative RNA-Seq expression across samples of ‘patterned’ genes known to be expressed in one (or more) neocortical layer. Demonstrated correspondences between dissected layers and samples are highlighted with outlined red panels on the descending diagonal. The lowest set of panels indicates genes that show no layer patterning. If genes g_l are annotated (Lein et al. 2007) as being preferentially expressed in layer l and f_{sgl} is the fractional expression of gene g_l in sample s relative to expression in all samples, then t_{sl} is the median of (f_{sgl}) over genes g_l . The heat map represents $[t_{sl} - \text{mean}_{\text{all layers } l}(t_{sl})] / [\text{mean}_{\text{all layers } l}(t_{sl})]$. (C) This correspondence allowed apparent layer enrichment to be predicted from expression across dissected samples using naïve Bayes classifiers. The receiver operating characteristic (ROC) curve for the layer 5 classifier is shown, as it yields a middling Area Under the ROC Curve (AUC) (Table 3.2; Figure B.2).

Subsequently, at least 50 million paired-end cDNA fragments passing Illumina’s quality filter were deep sequenced from each dissected sample on the Illumina GA IIX platform at NHGRI, thereby providing a comprehensive genome-wide view of transcription. I found that sequenced cDNA fragments spanned 25% of the mouse genome and overlapped 18,960 protein-coding genes (83%) with 16,340 protein-coding genes (72%)

expressed above a level of 0.1 Fragments per Kilobase of Exon model per Million reads mapped (FPKM) (Mortazavi et al. 2008; Trapnell et al. 2010) (Table B.3). Half of all transcripts derived from just 2% of expressed genes and the most highly expressed genes were generally of mitochondrial origin (Table B.4; Figure 3.2; Figure 3.3). 10% of mouse genome sequence located outside of known protein-coding, pseudogene, tRNA, rRNA, and short RNA gene loci, was expressed in at least one sample, including 1,055 long intergenic noncoding (lincRNA) loci (Table B.3; http://www.fgu.anat.ox.ac.uk/~grantb/mouse_layers/) (Ponting et al. 2009).

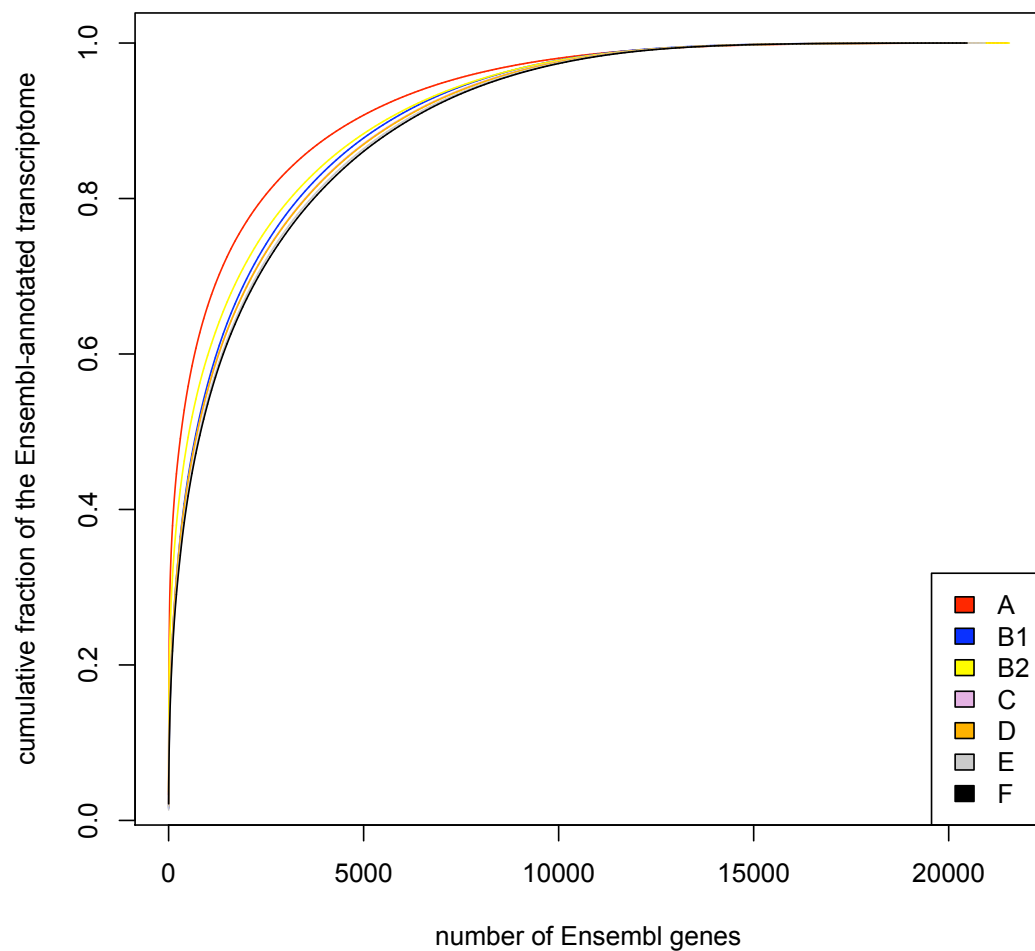


Figure 3.2: Most transcripts are expressed from only a few hundred genes. Approximately two percent of genes contribute half the transcriptome and most genes have evidence of transcription. Curve represents all expressed Ensembl genes (release 57), and cumulative fraction of the transcriptome is based on the cumulative contribution of those genes to the total FPKM from all genes.

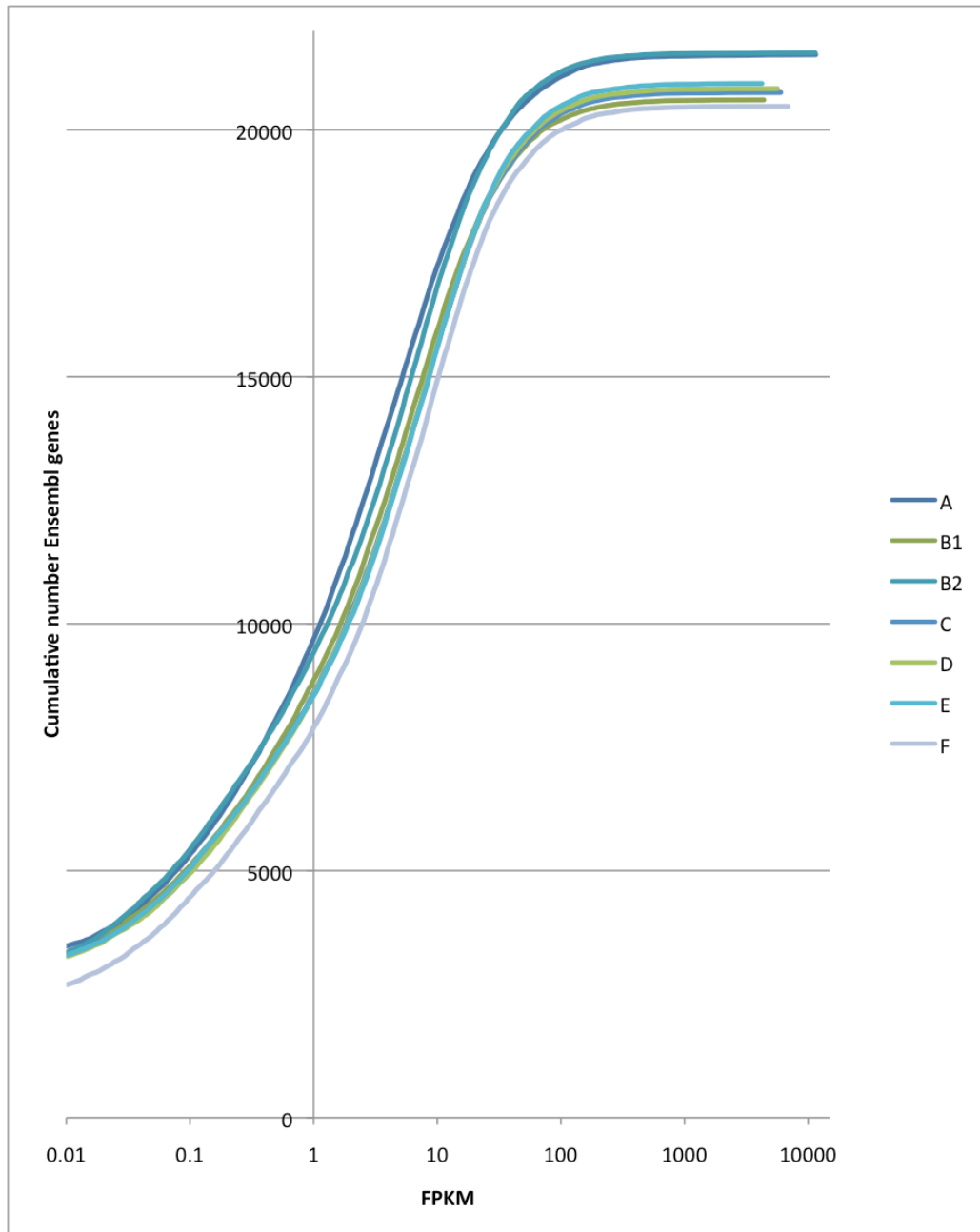


Figure 3.3: Most expressed genes have total transcript abundances within a window spanning 2-3 orders of magnitude. Cumulative graph of Ensembl genes versus expression level in FPKM. All Ensembl (release 57) genes were used.

Due to layer curvature, cells crossing multiple layers and dissection limitations it was not expected that samples should correspond precisely to individual layers. Indeed, while

Spearman's rank correlation coefficient for expression levels between B1 and B2 was high at 0.965, overlapping dissections resulted in some higher inter-pair correlations (Table 3.1). Nevertheless, known marker genes (Lein et al. 2007) for layers 2/3, 4, 5, 6, and 6b demonstrated high concordance between individual samples and specific layers (Figure 3.1B).

I compared these RNA-seq results with those previously obtained using microarrays for layer 6 and 6b from anterior cortex (putative S1) of postnatal day 8 mice (Hoerder-Suabedissen et al. 2009). RNA-seq levels for samples E and F were highly and significantly concordant with microarray expression levels for layers 6 and 6b despite methodological differences and the difference in age: 85% (147 of 173) of genes whose expression was found, using microarrays, to be lower in layer 6 versus 6b also showed lower expression in sample E versus F; significant concordance was also found for 87% (385 of 441) of genes significantly lower in layer 6b versus 6, compared with sample F versus E (each test, $p < 2 \times 10^{-16}$, two-tailed binomial test relative to a probability of 0.5).

Thousands of genes are patterned across neocortical layers

I next predicted 6,734 'patterned' genes to be preferentially expressed in one or more layers, and 5,689 'unpatterned' genes to be expressed more uniformly across all layers. For this, layer expression for 2,200 genes annotated from *in situ* hybridisation images was used to train a naïve Bayes classifier for each layer 2-6b. (Annotated marker genes were insufficient to permit training of a reliable classifier for layer 1.) These curations are generally consistent with the literature and other expression datasets (see http://mouse.brain-map.org/pdf/Top1000genesanalysisfinal_2010.pdf). A classifier was

also constructed to separate patterned from unpatterned genes. Classifier generalization accuracies were assessed using ten-fold cross-validation (Figure 3.1C; Table 3.2; Figure B.2) and smoothed calibration curves were constructed for the resulting predicted probabilities to arrive at accurate estimates of enrichment likelihood (Figure 3.4). Finally, these classifiers were applied to both known and previously unknown genes and transcripts (Table S2 in Belgard et al. (2011)). 11,410 known genes (10,261 protein-coding) were expressed at sufficiently high levels for their layer patterning to be classifiable. Predicted layer expression patterns typically recapitulated both the literature (Figure 3.5A) and the results of the high-throughput curation-based approach (Table 3.2). The most specific genes to each layer are provided in Table 3.3 and median specificities of predicted gene sets in Figure 3.7.

Table 3.2: Known genes classified as being patterned across neocortical cell layers. Three statistics from each classifier are presented: Area Under the receiver operating characteristic Curve (*AUC*), the proportion of genes with known expression patterns classified as highly expressed that were previously documented to have elevated expression in the layer (*precision*), and the proportion of the total genes documented to have elevated expression in the layer that are correctly identified by the classifier as highly expressed in that layer (*recall*). The next column contains the number of genes predicted by each classifier out of 11,410 classifiable genes. The last column represents the number of genes expected to be preferentially expressed in this layer, and is calculated as $[(\# \text{ genes predicted}) * \text{precision} / \text{recall}]$. The AUC represents the probability that the classifier will rank a randomly chosen gene that is preferentially expressed in that layer above a

randomly chosen gene that is not preferentially expressed in that layer: a random classifier has an expected AUC of 0.5. Layer enrichment probabilities of these genes, calibrated as shown in Figure 3.4, are provided in Table S2 of Belgard et al. (2011).

	AUC	precision	recall	genes predicted	genes expected
Layers 2/3	0.72	44%	42%	2598	2722
Layer 4	0.88	35%	53%	729	481
Layer 5	0.76	61%	61%	3120	3120
Layer 6	0.77	38%	41%	618	573
Layer 6b	0.75	39%	40%	1104	1076
Unpatterned	0.66	49%	50%	5689	5575

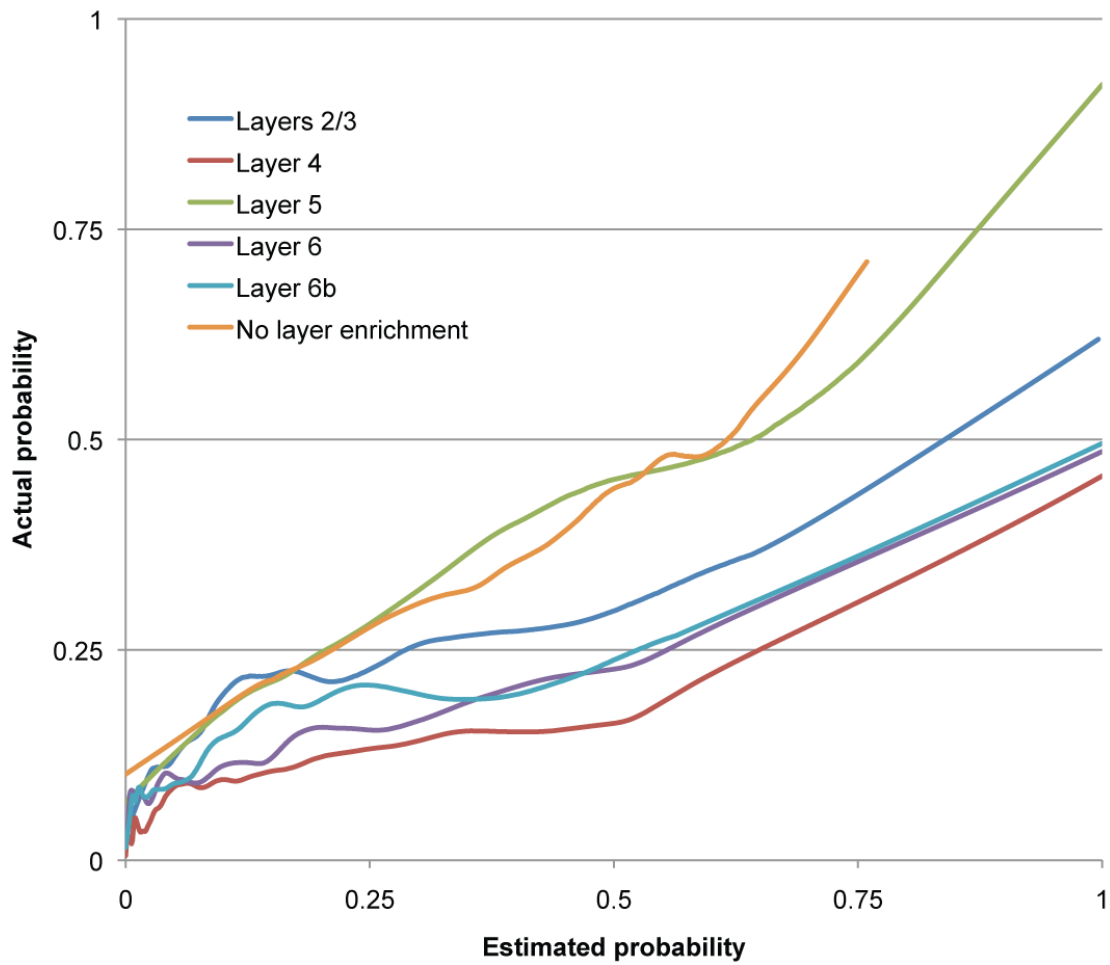


Figure 3.4: Probability calibration curves for all genes. LOESS (locally weighted scatterplot smoothing) probability calibration curves comparing the uncalibrated “estimated probability” assigned by the classifier with the cumulative “actual probability” based on the probability values of genes of known layer-enrichment for each of the layer classifiers.

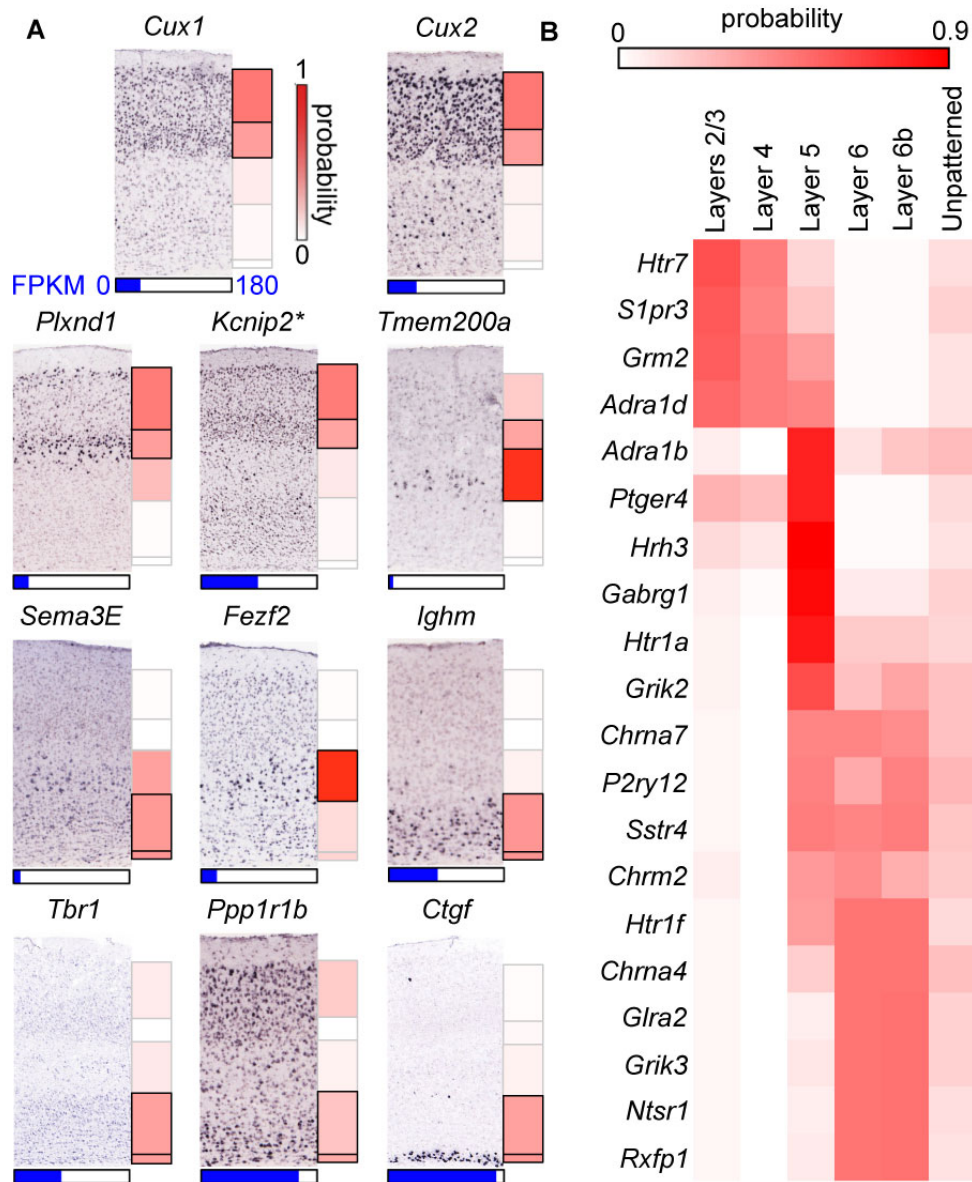


Figure 3.5: Classifiers recapitulate classical layer markers, enrich existing *in situ* hybridisations with quantitative data, and greatly expand the number of known genic layer markers. (A) Classifier predictions typically recapitulated known layer markers. Calibrated layer enrichment probabilities for eleven classical layer markers (Bulfone et al. 1995; Inoue et al. 2004; Lein et al. 2007; Molyneaux et al. 2007; Nieto et al. 2004; Ouimet et al. 1984; Watakabe et al. 2006; Xiong et al. 2004) are represented to the right of the corresponding *in situ* hybridisation image

reproduced with permission from the Allen Mouse Brain Atlas (Lein et al. 2007). Boxes outlined in black indicate that the classifier predicted the gene to be ‘enriched’ in that layer. Blue bars beneath each image represent the highest expression level of that gene in any sample, highlighting an additional dimension of information. Common synonyms for these genes include *Darpp32* (*Ppp1r1b*), *Igh6* (*Ighm*) and *C030003D03Rik* (*Tmem200a*). *I took all images from coronal sections (near position 6,000 in the AMBA) except *Kcnip2*, which was the only image from a sagittal section (near position 1,000 in the AMBA) (all other genes had a coronal image series available). (B) In addition to increasing the number of known (Lein et al. 2007) layer-patterned genes by 3-4 fold (Figure 3.6), these classifiers are also informative for genes that already have high-throughput *in situ* hybridisation images. This is illustrated by calibrated layer enrichment probabilities of the twenty particularly highly patterned genes encoding receptors, which I selected blindly to status in the Allen Mouse Brain Atlas, in which 9 were qualitatively consistent, 4 had a faint signal, 3 had a ubiquitous signal, 3 were qualitatively inconsistent, and one had a failed probe (see Table S3 of Belgard et al. (2011)). These classifiers provide an objective evaluation of cortical patterning for these genes with ambiguous images.

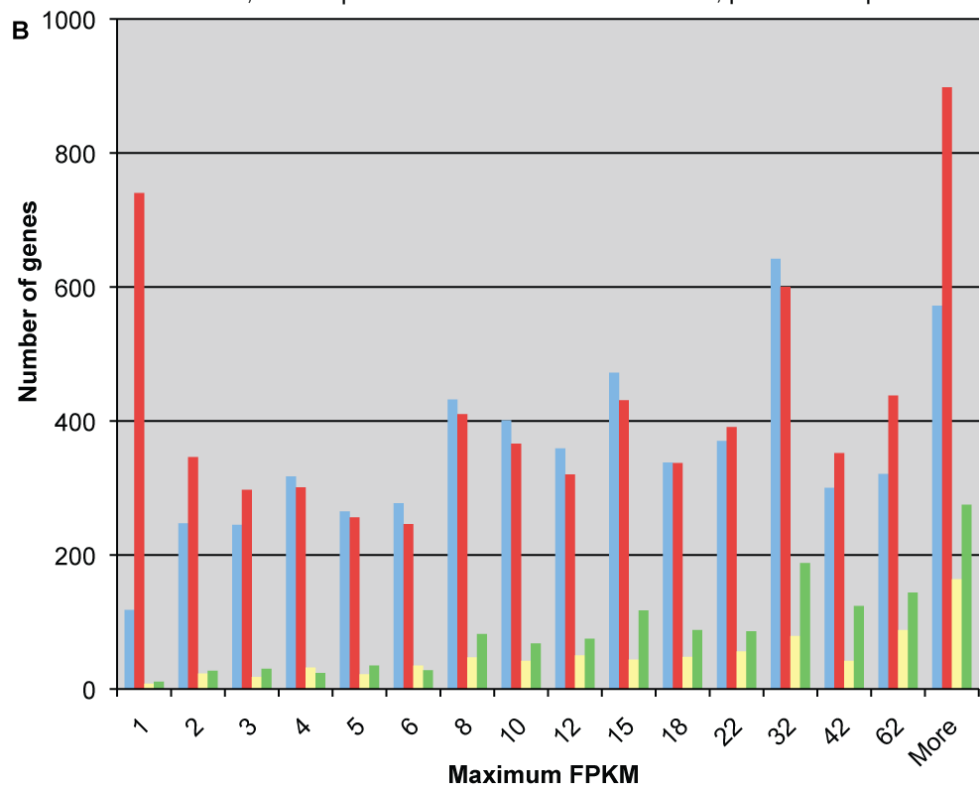
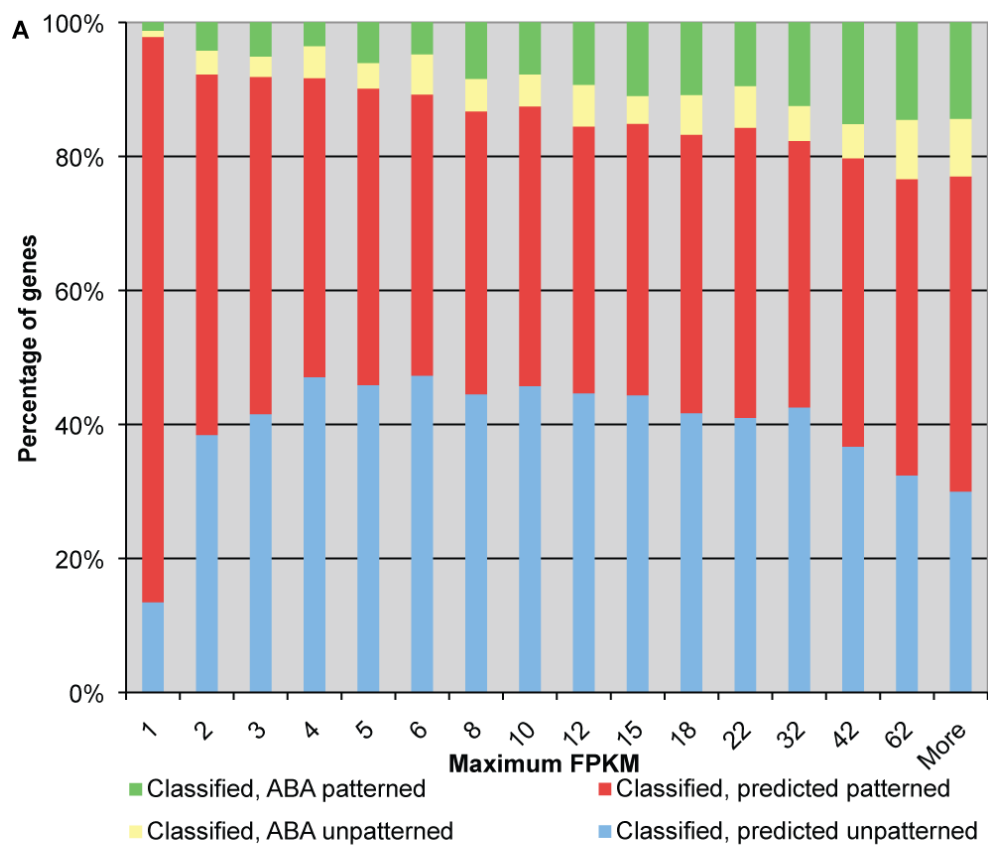


Figure 3.6: High-throughput *in situ* hybridisation curations cover relatively few genes for which we have RNA-seq-based layer classifications, especially among genes expressed at low levels. (A) Relative proportions and (B) absolute numbers of genes having different maximum FPKM values across dissected samples. Bin labels give the ceiling for that bin. For a discussion of the unusually high number of genes predicted to be patterned at <1 FPKM, see above. Out of 1,780 Allen Mouse Brain Atlas manually curated patterned genes I translated, 1,402 (79%) were classifiable. At the time of writing, a full 2,072 genes were manually curated as being patterned by the Allen Mouse Brain Atlas. With 6,734 predicted patterned genes, this increases the number of known layer-patterned genes by 3.25 [6,734/2,072] to 4.25 fold [(6,734+2,072)/2,072].

Table 3.3: The top ten genes of genes predicted to be in layers 2/3-6b. Genes predicted to be enriched in the layer by the classifiers were then sorted by expression specificity in the most relevant samples (A for layers 2/3, B1 & B2 for layer 4, C & D for layer 5, E for layer 6, F for layer 6b). An FPKM of 5 was required in each relevant sample.

Layers 2/3		Mean FPKM	fold difference with respect to mean						
Gene symbol	Gene ID		A	B1	B2	C	D	E	F
<i>Sox4</i>	ENSMUSG00000078309	5.6	5.35	0.00	1.65	0.00	0.00	0.00	0.00
<i>Cux2</i>	ENSMUSG00000072641	20.2	4.96	0.00	2.04	0.00	0.00	0.00	0.00
<i>mt-Ts1</i>	ENSMUSG00000064352	127.1	4.70	0.32	0.80	0.39	0.39	0.31	0.09
<i>Myh7</i>	ENSMUSG00000053093	3.1	4.69	0.93	0.76	0.16	0.11	0.15	0.20
<i>Tagln</i>	ENSMUSG00000032085	3.1	4.68	1.30	1.02	0.00	0.00	0.00	0.00
<i>Vps13d</i>	ENSMUSG00000073719	2.7	4.40	0.00	2.60	0.00	0.00	0.00	0.00
<i>Gm10627</i>	ENSMUSG00000074106	2.5	4.33	2.67	0.00	0.00	0.00	0.00	0.00
<i>Fau</i>	ENSMUSG00000038274	28.2	4.17	0.30	0.57	0.52	0.55	0.51	0.38
<i>mt-Te</i>	ENSMUSG00000064369	94.6	4.13	0.31	0.99	0.45	0.56	0.42	0.14
<i>Gm12254</i>	ENSMUSG00000085342	19.1	4.02	0.17	0.84	0.46	0.45	0.57	0.48

Layer 4		Mean FPKM	fold difference with respect to mean						
Gene symbol	Gene ID		A	B1	B2	C	D	E	F
<i>Rspo1</i>	ENSMUSG00000028871	9.1	0.49	3.32	2.04	0.83	0.15	0.06	0.10
<i>Gm12371</i>	ENSMUSG00000084898	2.6	1.43	2.03	2.59	0.51	0.30	0.07	0.06
<i>Tmem215</i>	ENSMUSG00000046593	3.1	1.13	2.44	2.13	0.72	0.35	0.18	0.05
<i>Gm10493</i>	ENSMUSG00000073367	23.5	0.75	3.05	1.43	1.27	0.49	0.00	0.00
<i>Ddit4l</i>	ENSMUSG00000046818	9.8	1.70	1.98	2.32	0.52	0.18	0.11	0.19
<i>Prr16</i>	ENSMUSG00000073565	3.9	0.41	2.23	2.00	1.13	0.81	0.29	0.13
<i>Fam19a2</i>	ENSMUSG00000044071	9.1	0.43	2.38	1.83	1.45	0.57	0.21	0.13
<i>Pcdh20</i>	ENSMUSG00000050505	4.9	0.50	2.47	1.72	0.78	0.48	0.57	0.47
<i>Rorb</i>	ENSMUSG00000036192	9.6	0.62	1.85	2.30	1.31	0.45	0.23	0.24
<i>Kcnh5</i>	ENSMUSG00000034402	8.0	0.26	2.61	1.53	1.35	0.52	0.42	0.32

Layer 5		Mean FPKM	fold difference with respect to mean						
Gene symbol	Gene ID		A	B1	B2	C	D	E	F
<i>Prkag1</i>	ENSMUSG00000067713	6.3	0.00	0.00	0.00	3.40	3.60	0.00	0.00
<i>Tox</i>	ENSMUSG00000041272	6.0	0.15	0.60	0.62	2.13	1.90	0.91	0.68
<i>Gm5454</i>	ENSMUSG00000047643	10.7	0.17	0.73	0.68	2.33	1.61	0.60	0.88
<i>Fezf2</i>	ENSMUSG00000021743	12.3	0.08	0.49	0.31	2.11	1.82	1.07	1.12
<i>Opn3</i>	ENSMUSG00000026525	2.8	0.14	0.76	0.58	1.99	1.92	1.03	0.59
<i>Etv1</i>	ENSMUSG00000004151	3.3	0.37	0.55	0.81	2.27	1.60	0.60	0.80
<i>Nefh</i>	ENSMUSG00000020396	25.4	0.45	0.89	0.88	2.03	1.83	0.63	0.29
<i>Slc5a5</i>	ENSMUSG00000000792	4.7	0.15	0.52	0.46	2.06	1.79	0.96	1.07
<i>Bcl6</i>	ENSMUSG00000022508	5.0	0.40	0.78	0.59	2.16	1.69	0.81	0.59
<i>Ankrd34b</i>	ENSMUSG00000045034	4.6	0.19	0.91	0.62	1.84	1.95	0.93	0.58

Layer 6		Mean FPKM	fold difference with respect to mean						
Gene symbol	Gene ID		A	B1	B2	C	D	E	F
<i>Mir1896</i>	ENSMUSG000000084590	3.9	0.00	0.00	0.68	1.49	0.45	2.87	1.51
<i>Rprm</i>	ENSMUSG000000075334	47.2	0.06	0.10	0.11	0.38	0.85	2.85	2.65
<i>Cdh9</i>	ENSMUSG000000025370	4.8	0.11	0.14	0.19	0.57	1.71	2.77	1.50
<i>Gpr21</i>	ENSMUSG000000053164	2.6	0.00	0.00	0.00	0.00	1.78	2.73	2.49
<i>Ighm</i>	ENSMUSG000000076617	29.5	0.07	0.08	0.16	0.40	1.44	2.60	2.26
<i>Grp</i>	ENSMUSG000000024517	3.8	1.00	0.19	0.23	0.54	1.15	2.56	1.33
<i>Ntsr1</i>	ENSMUSG000000027568	2.2	0.18	0.30	0.21	0.41	1.11	2.47	2.33
<i>Sla</i>	ENSMUSG000000022372	2.5	0.13	0.14	0.14	0.20	0.55	2.40	3.45
<i>Bmp3</i>	ENSMUSG000000029335	4.6	0.39	0.38	0.19	0.39	0.99	2.36	2.29
<i>Sema5a</i>	ENSMUSG000000022231	6.0	0.14	0.21	0.21	0.37	0.98	2.34	2.75

Layer 6b		Mean FPKM	fold difference with respect to mean						
Gene symbol	Gene ID		A	B1	B2	C	D	E	F
<i>Mif4gd</i>	ENSMUSG000000020743	1.7	0.00	0.00	0.00	0.00	0.00	0.00	7.00
<i>Thbs4</i>	ENSMUSG000000021702	3.5	0.11	0.03	0.01	0.01	0.02	0.36	6.46
<i>Nxph4</i>	ENSMUSG000000040258	6.9	0.06	0.01	0.01	0.02	0.05	0.89	5.96
<i>Ctgf</i>	ENSMUSG000000019997	29.5	0.04	0.02	0.03	0.05	0.12	1.06	5.68
<i>Cplx3</i>	ENSMUSG000000039714	30.8	0.30	0.10	0.10	0.07	0.09	0.99	5.33
<i>Tmem90a</i>	ENSMUSG000000071234	3.8	0.38	0.23	0.17	0.32	0.27	0.80	4.84
<i>Ttr</i>	ENSMUSG000000061808	43.1	0.59	0.02	0.03	0.03	0.18	1.52	4.62
<i>AC161876.1</i>	ENSMUSG000000055802	9.5	2.38	0.00	0.00	0.00	0.00	0.00	4.62
<i>Anln</i>	ENSMUSG000000036777	5.4	0.07	0.12	0.14	0.34	0.55	1.32	4.46
<i>Gfap</i>	ENSMUSG000000020932	29.1	1.18	0.20	0.16	0.15	0.28	0.69	4.34

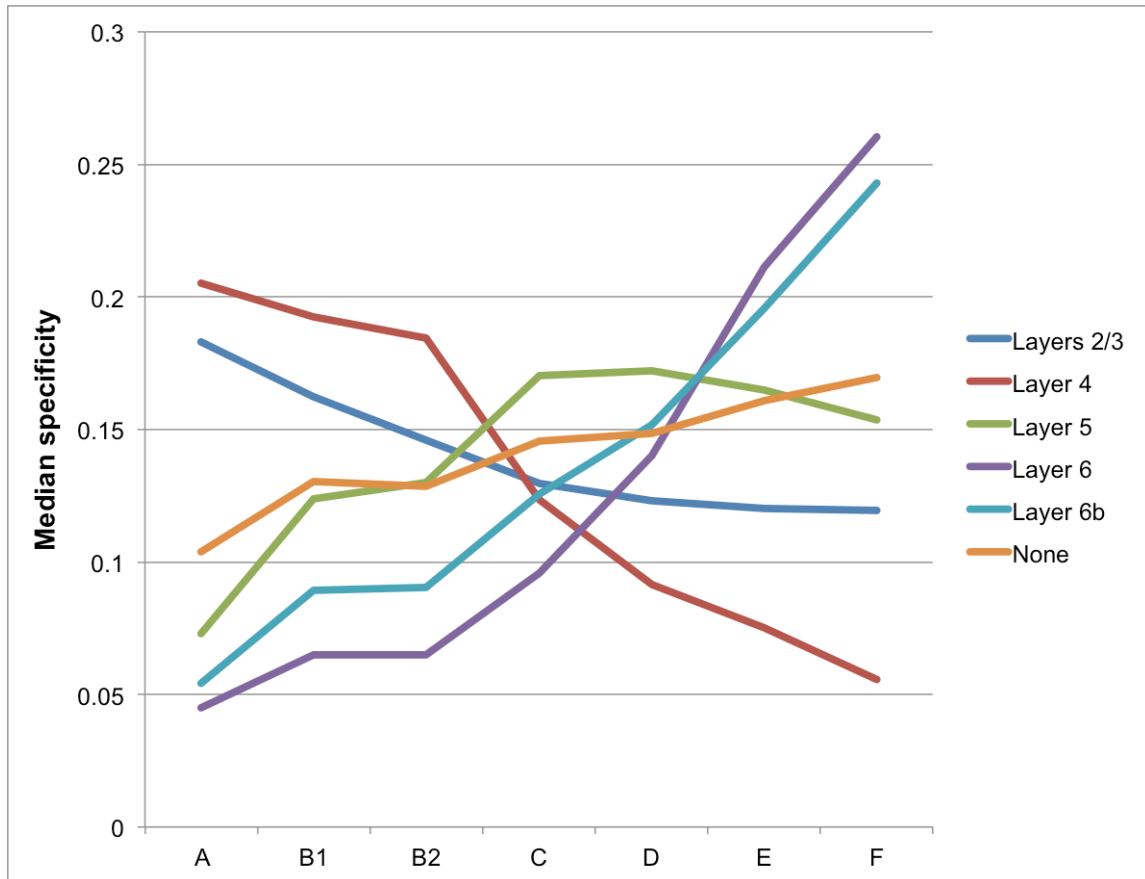


Figure 3.7: The median expression specificities in samples A-F (defined as FPKM in that sample divided by total FPKM) of predicted gene sets for different layers reflect the diffuse and overlapping relationships between layers and samples, and also the precisions of the predicted sets.

Upon correcting for false positives and false negatives, an estimated 5,835 of these 11,410 classifiable known genes (51%) were expressed preferentially in one or more layers (Table 3.2). Among protein-coding genes, 4,464 (44%) had a higher calibrated probability of being enriched in at least one layer than of being unpatterned. Transcripts from 66 lincRNA loci (see below) were classified as being patterned (see Table S7 in Belgard et al. (2011)).

Of 291 genes encoding receptors and ion channels, 108 were expressed highly enough to be classified and 82 of these were predicted to be patterned across the layers (see Table S3 in Belgard et al. (2011)). Layer enrichment probabilities of the twenty most highly patterned receptors are listed in Figure 3.5B, and are generally consistent with previous observations (see Table S3 in Belgard et al. (2011)).

Some of this patterning reflected known cell types. Neuron-enriched genes (≥ 1.5 -fold; (Cahoy et al. 2008)) were 63% more likely to be patterned than unpatterned ($p < 0.0001$; two-tailed Chi-squared test with Yates correction). There were no significant differences among the layers in their proportions of neuron-enriched genes, suggesting these differences in neuron-enriched genes may reflect laminar diversity of neuronal subtypes rather than differences in relative populations of neurons in aggregate. In contrast, astrocyte-enriched genes (≥ 1.5 -fold; (Cahoy et al. 2008)) were 17% *less* likely to be patterned than unpatterned ($p = 0.0007$; two-tailed Chi-squared test with Yates correction). Oligodendrocyte-enriched genes (Cahoy et al. 2008) were found almost exclusively in the deepest samples (Figure B.3), matching previous observations that oligodendrocytes are rare in the neocortex except in the deepest layers (Tan et al. 2009). Likewise, the specific and robust microglia marker gene F4/80 (Cucchiaroni et al. 2003; Perry et al. 1985) monotonically increased in expression with samples derived from deeper layers.

Isoform switching across layers

2,003 genes had at least two transcript isoforms that were each classifiable. 1,646 of these genes (82%) showed differential patterns of alternative splicing across sequenced

samples (see Table S5 in Belgard et al. (2011)). 719 genes, including eight encoding receptors, additionally had divergent predicted patterns of layer enrichment (Figure 3.9; see Table S4 in Belgard et al. (2011)). The differential splicing across layers of *Mtap4*, the most connected hub gene in Alzheimer's disease (Ray et al. 2008), is one example of the potential neurological relevance of this set (Figure 3.8). *Mtap4* encodes isoforms of MAP4 with differing microtubule-stabilization properties (Hasan et al. 2006) that have been proposed to regulate the dynamic behaviours of extending neurites (Hasan et al. 2006), structures lost or altered in the earliest stages of Alzheimer's disease (Knobloch and Mansuy 2008). Most Alzheimer's disease genes were enriched in either layers 2/3 or layer 5 (Figure 3.10B; see also Table S6 in Belgard et al. (2011)), which were dominated by an isoform having an additional tau domain compared to the isoform that dominated layers 6 and 6b (Figure 3.8). This is particularly striking considering that neurons in layers 3 and 5 are selectively lost in Alzheimer's disease patients (Hof and Morrison 2004). We also found that *in situ* hybridisation was sometimes unable to detect minor isoforms in the cortex that were clearly detectable by RNA-seq, which may underscore the greater dynamic range of transcriptome sequencing.

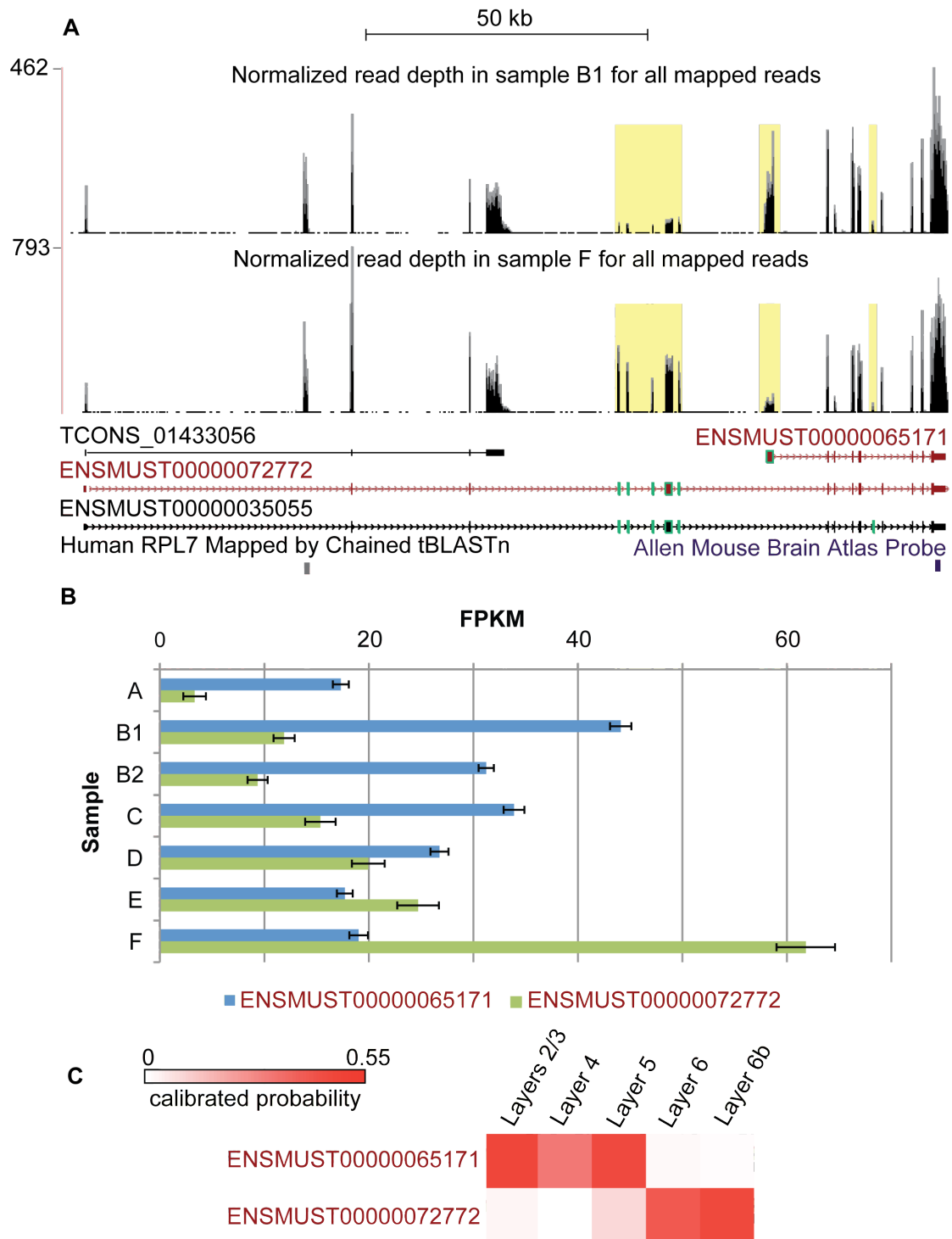


Figure 3.8: *Mtap4* shows signs of differential expression of its isoforms across layers. *Mtap4*, the most connected hub gene in Alzheimer's disease (Ray et al. 2008), encodes isoforms of MAP4 with differing microtubule-stabilization

properties (Hasan et al. 2006) that have been proposed to regulate the dynamic behaviours of extending neurites (Hasan et al. 2006). (A) Normalized read depth across samples A-F for the mouse *Mtap4* region (chr9:109,830,555-109,986,919) shows differential expression of ENSMUST00000065171 and ENSMUST00000072772. The full-length transcript, ENSMUST00000035055 (encoding UniProt protein P27546, the canonical mouse MAP4), was a minor isoform in comparison, and *de novo* transcript TCONS_01433056 lacks the MAP4 microtubule-binding domain. In contrast, the two differentially expressed transcripts marked in red, ENSMUST00000065171 (encoding P27546 isoform 4) and ENSMUST00000072772 (encoding P27546 isoform 3), encode known proteins that differ in sequence. Isoform 3 contains one fewer Tau/MAP domain than isoform 4. The number of these domains is known to affect kinesin motor activity on microtubules (Tokuraku et al. 2007). Yellow regions mark the exons, boxed in green, that discriminate among isoforms. The reads mapping to the region with similarity to human RPL7 likely derive from one of many retropseudogenes of the mouse *Rpl7* gene, which is highly expressed in all cortical samples. The probe used for *in situ* hybridisation in the Allen Mouse Brain Atlas does not discriminate between these isoforms. (B) FPKMs across samples with 95% confidence intervals suggest the two highly expressed and potentially functional isoforms are differentially patterned across samples. (C) Classifiers also predict that these two isoforms are strongly differentially patterned across layers.

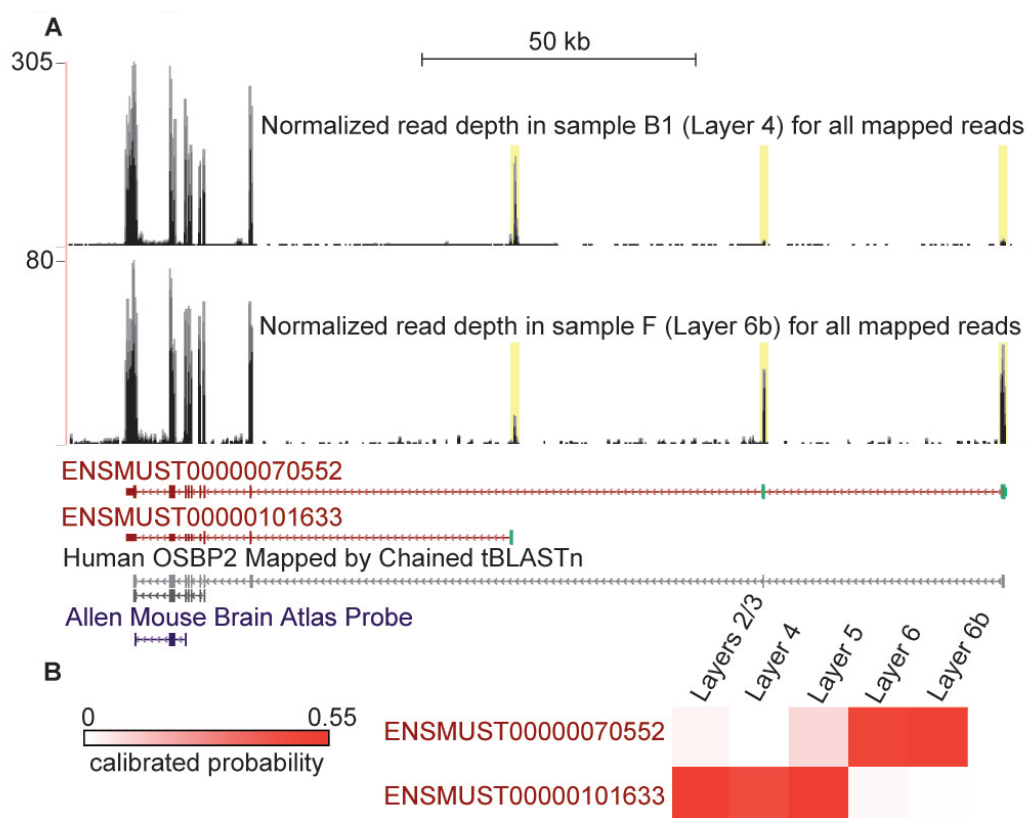


Figure 3.9: *Osbp2* shows signs of differential expression of its isoforms across layers. (A) Normalized read depth across samples A-F in the mouse chr11:3,593,238-3,771,939 region. Two isoforms account for most expression from this locus: ENSMUST00000070552 and ENSMUST00000101633. Yellow boxes highlight read coverage in mutually exclusive exons outlined in green. The FPKM of ENSMUST00000070552 was significantly higher in sample F than in sample B1, while the FPKM of ENSMUST00000101633 was significantly higher in sample B1 than in sample F, indicating that both transcripts are differentially expressed across samples. The probe used for *in situ* hybridisation in the Allen Mouse Brain Atlas (RP_050125_02_B09) does not discriminate between these isoforms. (B) Classifiers predict that ENSMUST00000070552 is enriched in deeper layers while ENSMUST00000101633 is enriched in upper layers. See Table S4 in Belgard et al.

(2011) for the layer enrichment probabilities of additional alternatively spliced genes.

The candidate set of genes with differentially expressed alternative transcripts may also reflect the action of splicing factors. For example, the splicing factor Fox-2, encoded by *Rbfox2*, is known to be sufficient for exon inclusion in *Ewsr1* in a dose-dependent manner (Underwood et al. 2005), producing a brain-enriched isoform (Melot et al. 2001). Consistent with this, I found the FPKM of *Rbfox2* across samples was strongly correlated with both the isoform fraction ($r = 0.85$; $p < 0.01$, one-tailed) and the FPKM of the brain-enriched *Ewsr1* isoform (ENSMUST00000102930; $r = 0.87$; $p < 0.01$, one-tailed). In contrast, the total FPKM of *Ewsr1* was significantly correlated neither with the isoform fraction ($r = -0.47$; not significant) nor with the FPKM of this brain-enriched isoform ($r = 0.19$; not significant).

Biological associations for genes expressed highly in specific layers

Genes that exhibited patterned expression were, in general, very different from those expressed more evenly across the cortical layers. They were more likely to encode specific receptor types, and proteins involved in synaptic transmission and ion transport (Figure 3.10A). By contrast, unpatterned genes tended to possess housekeeping cellular roles that less often contribute to cell type-specific functions. An exception to this housekeeping generalization is mitochondrial genes, which were 41% enriched in layer 5 ($q=0.003$; see also Table S6 in Belgard et al. (2011)). Note that the extremely large layer 5 pyramidal neurons, which are the only cell types that extend axons beyond the skull, are known to have high energetic demands (Guillemot et al. 2006; Molnár and Cheung

2006; Molyneaux et al. 2007). Unpatterned genes appear to be more important in early development: they were significantly more likely to result in prenatal lethality when disrupted in mice (Figure 3.10A).

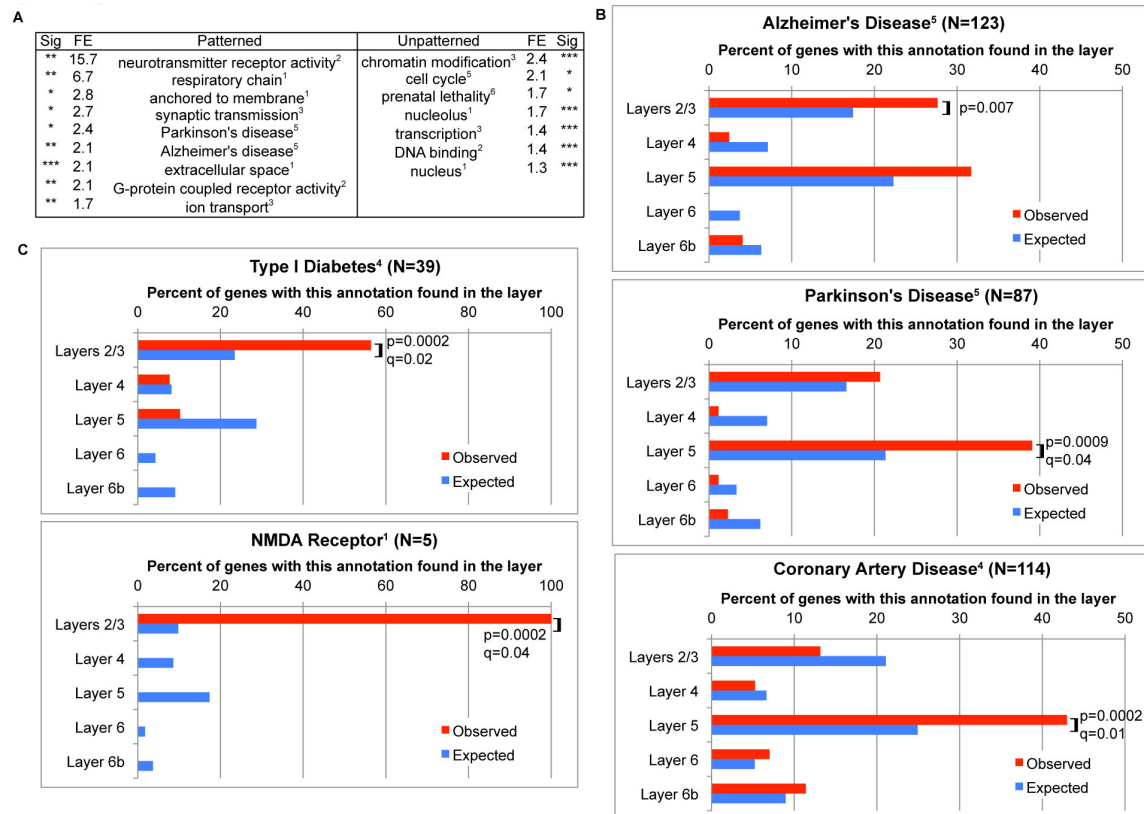


Figure 3.10: Genes enriched in expression for specific layers are significantly associated with specific diseases and functions. (A) Selected functional classes of genes whose expression across neocortical layers is significantly patterned (all genes predicted to be preferentially expressed in one or more layers) or unpatterned (predicted to have no layer enrichment). Significance was assessed using a two-tailed Fisher's exact test, followed by a q -value multiple testing correction. All of these terms were replicated with a p -value of lower than 0.05 in either dorsal or lateral cortex (see Table S6 in Belgard et al. (2011)). The full list of significant

terms is provided on the website. Sig = significance (* $q < 0.05$; ** $q < 0.01$, *** $q < 0.001$), FE = fold enrichment over expected. (B) Distributions of genes across layers of selected functions that were significantly patterned in either the original set or one of the replication sets with a Bonferroni-corrected $p < 0.05$. See Table S6 in Belgard et al. (2011) for a complete non-redundant list of enriched terms, and also for both replications. Where the observed percentage of genes in a layer differs significantly from the expected percentage of genes in a layer, I list the appropriate p - or q -values smaller than 0.05, as assessed by a bias-corrected simulation method. (C) Distributions of genes across layers of selected functions that were significantly enriched in one layer ($q < 0.05$), as assessed by a bias-corrected simulation method. Terms were drawn from the following databases: ¹Gene Ontology (Ashburner et al. 2000) Cellular Component, ²Molecular Function, and ³Biological Process; ⁴1:1 mouse orthologs of human genes previously associated with human traits and diseases (Chen et al. 2010); ⁵Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2004); and ⁶mouse knockout phenotypes (Blake et al. 2011). Genes can be found in Table 3.4.

Cells in cortical layers are proposed to have differing biological roles. I sought transcriptomic evidence for these roles by identifying, for each layer in turn, molecular annotations of genes that were more abundant than expected from a bias-corrected random sampling of all 11,410 classifiable genes. Molecular annotations were drawn from a variety of sources, including the Gene Ontology (Ashburner et al. 2000), genomic intervals associated with human diseases (Chen et al. 2010), and mouse knockouts

(Blake et al. 2011), and only results retained after application of a 5% false discovery rate threshold for each set are reported.

Table 3.4: Predicted layer enrichments of genes annotated with the functional terms in Figure 3.10B-C

Layer	Alzheimer's Disease (123 genes)
2/3	<i>Apoe, Atp2a1, Atp5d, Atp5e, Cacna1c, Calm1, Calm3, Casp7, Cdk5r1, Cox6a1, Cox6a2, Cox8a, Gapdh, Grin2a, Grin2b, Grin2c, Grin2d, Itpr3, Lpl, Lrp1, Mapt, Ndufa2, Ndufa3, Ndufa5, Ndufb7, Ndufs3, Ndufs6, Ndufs7, Ndufv1, Ndufv3, Plcb4, Ppp3ca, Psen2, Psenen</i>
4	<i>Atp2a1, Cox6a2, Lpl</i>
5	<i>1500003O03Rik, Aph1c, App, Atp2a1, Atp2a2, Atp5a1, Atp5b, Atp5c1, Atp5f1, Atp5g3, Atp5j, Bid, Capn1, Capn2, Casp3, Cox5a, Cox6a2, Cyc1, Itpr1, Nae1, Ndufa1, Ndufa6, Ndufa9, Ndufb5, Ndufb6, Ndufb8, Ndufc1, Ndufc2, Ndufs1, Ndufv2, Ppp3cb, Ppp3cc, Ppp3r1, Sdha, Sdhb, Sdhc, Uqcrc1, Uqcrc2, Uqcrcs1</i>
6	-
6b	<i>Adam10, Apaf1, Atf6, Ide, Mapk3, Ryr3</i>
Layer	Parkinson's Disease (87 genes)
2/3	<i>Atp5d, Atp5e, Cox6a1, Cox6a2, Cox8a, Lrrk2, Ndufa2, Ndufa3, Ndufa5, Ndufb7, Ndufs3, Ndufs6, Ndufs7, Ndufv1, Ndufv3, Sept5, Ube2j2, Ube2l3</i>
4	<i>Cox6a2</i>
5	<i>Atp5a1, Atp5b, Atp5c1, Atp5f1, Atp5g3, Atp5j, Casp3, Cox5a, Cox6a2, Cyc1, Lrrk2, Ndufa1, Ndufa6, Ndufa9, Ndufb5, Ndufb6, Ndufb8, Ndufc1, Ndufc2, Ndufs1, Ndufv2, Ppid, Sdha, Sdhb, Sdhc, Slc25a4, Ube2g1, Ube2g2, Ube2j1, Uqcrc1, Uqcrc2, Uqcrcs1, Vdac1, Vdac2</i>
6	<i>Gpr37</i>
6b	<i>Apaf1, Gpr37</i>
Layer	Coronary Artery Disease (114 genes)
2/3	<i>9930013L23Rik, Dab1, Gpr116, Hbegf, Hivep3, Lama4, Pamr1, Pdzd2, Prkg1, Rae1, Sf3a2, Tspan9, Tuba8, Zc3h11a, Zfp148</i>
4	<i>9930013L23Rik, Hbegf, Lama4, Nell1, Pamr1, Tuba8</i>

5	<i>A230083H22Rik, Adarb1, Ank1, Asah1, Asb3, Bmper, Clstn2, Ctnna2, Erol1b, Fam19a1, Fgf14, Fuca2, Gramd1b, Grk5, Hgf, Kcnip1, Kcns3, Lrig1, Luzp2, Mpped2, Myh10, Ncam2, Ndel1, Nedd4l, Nell1, Nhsl1, Nlgn1, Ntrk2, Pamr1, Pcdh7, Pclo, Pfk1, Pon2, Ppapdc2, Ptpm, Rit2, Scn9a, Sema3c, Slc1a1, Slc1a4, Sod2, Spdy, St3gal1, St6galnac5, Syt1, Tox2, Tspan9, Usp10, Vegfa</i>
6	<i>Ankrd28, Fam166b, Fut9, Fxyd6, Il17ra, Kcnk9, Sema5a, St3gal1</i>
6b	<i>Acs11, Ankrd28, Fam166b, Fut9, Fxyd6, Galnt10, Il17ra, Kcnk9, Plxna2, Prickle1, Ryr3, Sema5a, St3gal1</i>
Layer	Type I Diabetes (39 genes)
2/3	<i>Abcf1, Agpat1, Atf6b, Bat2, Bat3, Brd2, Cchcr1, Cdsn, Dom3z, Ehmt2, H2-DMA, Hspa11, Neu1, Nfkbil1, Rdbp, Rnf39, Stk19, Tap1, Tcf19, Vars, Vars2, Zbtb12</i>
4	<i>Cchcr1, Cdsn, Tcf19</i>
5	<i>Clec16a, Csnk2b, Nrm, Tubb5</i>
6	-
6b	-
Layer	NMDA Receptor (5 genes)
2/3	<i>Grin2a, Grin2b, Grin2c, Grin2d, Grin3b</i>
4	-
5	-
6	-
6b	-

Genes encoding subunits of the NMDA receptor were enriched in layers 2/3: expression of all five classifiable genes were identified as being significantly concentrated in these layers, a number much higher than expected simply by chance (Figure 3.10C). NMDA receptor subunits are expressed across the cortex and, despite being largely responsible for physiological responses of cells in layers 5 and 6, they are known to be preferentially expressed in layers 2/3 (Conti et al. 1997; Currie et al. 1994). In subsequent replications (see below), a majority of layer-enriched NMDA receptor subunit genes (half were layer-enriched) were enriched in layers 2/3 (see Table S6 in Belgard et al. (2011)). Genes encoding proteins localized to the extracellular space or region were expressed at significantly higher levels in layers 2/3, 4, 6 or 6b ($q < 0.05$; see Table S6 in Belgard et al. (2011)). Surface markers in brain cells are often involved in various signalling processes, from guidance to synapse formation (Maruyama et al. 2008; Uziel et al. 2006; Yamamoto et al. 2007).

I identified significant association of layer-enriched expression for genes whose human orthologs lie within genomic intervals previously associated with disease. In particular, layers 2/3-enriched genes were unusually abundant among human type 1 diabetes- and rheumatoid arthritis-associated genes (Figure 3.10C, see also Table S6 in Belgard et al. (2011)). These findings reflected nearly all of these genes' locations being within the major histocompatibility complex (MHC) region. Indeed, genes in the MHC region were 34% more likely than randomly selected genes ($p < 10^{-6}$; case resampling bootstrap) to have enriched expression in layers 2/3. Many of these examples are confirmed by *in situ* hybridisations in the Allen Mouse Brain Atlas (Figure B.4). In

subsequent replications, a majority of layer-specific MHC genes and a large minority of all MHC genes were again enriched in layers 2/3 (see Table S6 in Belgard et al. (2011)).

Another apparent disease association links mouse genes preferentially expressed in layer 5 with human genes in the Parkinson's disease pathway (Figure 3.10B). This, however, is likely to reflect the involvement of mitochondrial dysfunction in Parkinson's disease (Abou-Sleiman et al. 2006; Burbulla et al. 2010) and the prominent expression of mitochondrial and metabolic genes in this layer (Table B.6; see Table S6 in Belgard et al. (2011)). For example, *LRRK2*, which is mutated in familial Parkinson's disease (Abou-Sleiman et al. 2006), is expressed prominently in rodent neocortical layers 2/3 and 5 (Belgard et al. 2011; Lein et al. 2007), specifically pyramidal neurons, and is associated with mitochondrial markers (Biskup et al. 2006). In subsequent replications, nearly half of Parkinson's disease-related genes were enriched in layers 2/3, and nearly half in layer 5 (see Table S6 in Belgard et al. (2011)). 29 of the 36 genes with known enrichments from *in situ* hybridisation were manually curated as enriched in layer 5 (see Table S6 in Belgard et al. (2011)).

Gene sets and functional differences replicated in dorsal cortex and lateral cortex

We subsequently replicated these gene sets and functional differences in both dorsal cortex (including, but not limited to, S1) and lateral cortex (not overlapping S1; including, but not limited to, S2) in experiments involving different mice, a different dissecting anatomist (Dr. Fernando Garcia-Moreno), different technicians making the cDNA and sequencing libraries using a different protocol (the NIH Intramural Sequencing Center), updated sequencing chemistry and software, different filtering

criteria, and with additional variables used to train the classifiers as described in the methods. The overall concordance rates for genes of the two replication predictions with the original predictions ranged from 64%-96%, and a majority of genes predicted to be enriched in one experiment were also predicted to be enriched in the other (Table 3.5). All but one of the 39 significant functional differences in Table S6 (Belgard et al. 2011) were changed in the same direction in both replication sets, and 87% of these, including all disease associations, also had a replicated *p*-value smaller than 0.05 in either dorsal or lateral cortex (Table S6 in Belgard et al. (2011)). Most, but not all, genes were similar in patterning between dorsal and lateral cortex (Figure 3.11), confirming previous observations (Hawrylycz et al. 2010).

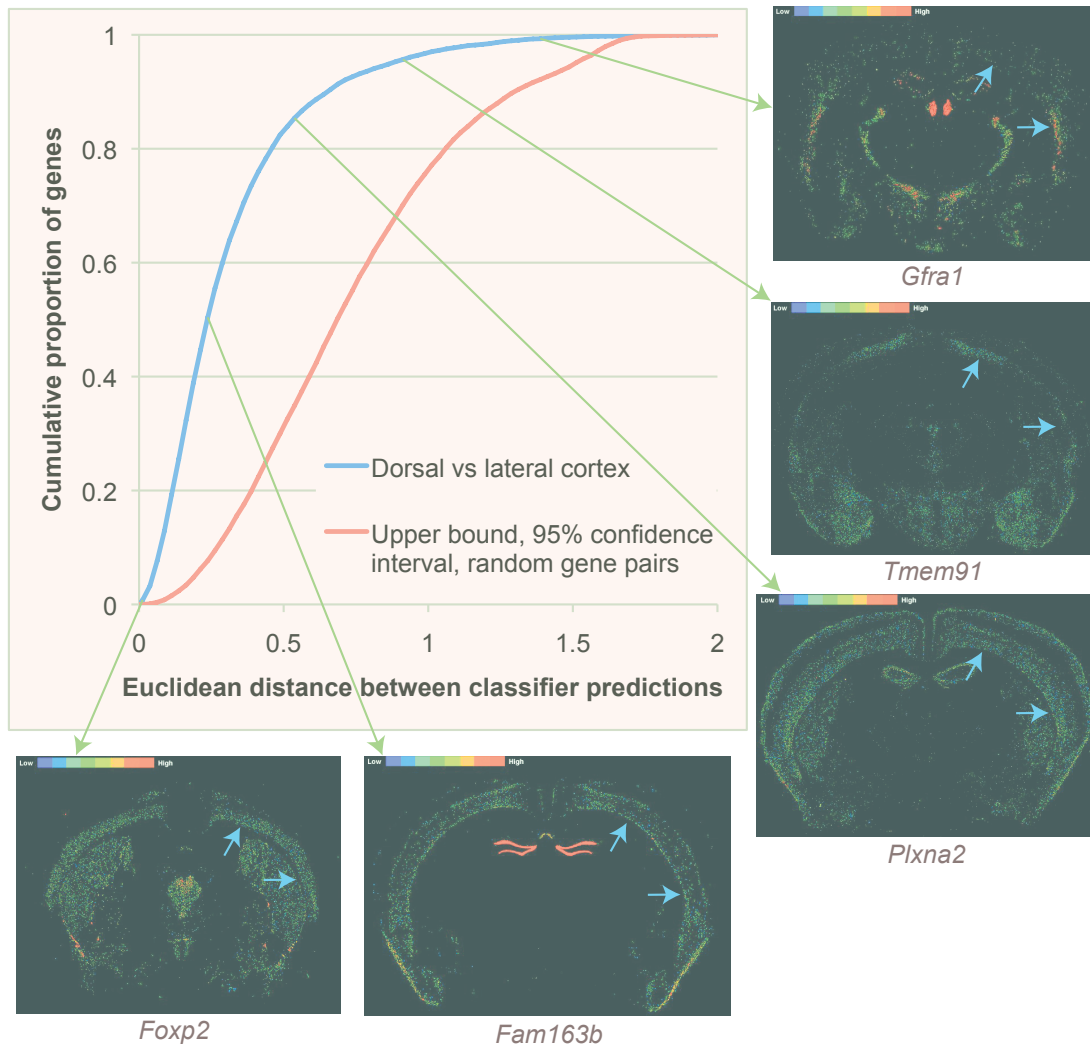


Figure 3.11: Most, but not all, genes have a similar predicted laminar expression pattern in both dorsal cortex and lateral cortex. In blue, the cumulative frequency distribution of the Euclidean distances between laminar enrichment probabilities of genes in lateral cortex and the dorsal cortex replication demonstrates that most genes have similar predictions in both regions. Most genes have a Euclidean distance smaller than 0.25; the corresponding *in situ* hybridisations suggest their predicted laminar expression pattern is similar between dorsal and lateral cortex. The red distribution represents the upper bound of the one-tailed 95% confidence interval obtained by randomly pairing genes. I calculated this as follows: 500

distributions were bootstrapped by randomly pairing, with replacement, layer predictions in lateral cortex and in the dorsal cortex replication from 12,455 genes; these distributions were sorted by numerically integrating over Euclidean distances 0 to 2, inclusive, and the resulting 95th percentile distribution is displayed. The relative positioning of these two distributions quantitatively supports the subjective observation that a difference in laminar expression patterns between dorsal and laminar cortices is the exception to the rule (which has been reported elsewhere, see (Hawrylycz et al. 2010)). If, instead, there were no relationship between expression patterns in dorsal and laminar cortex, the laminar cortex classifiers (trained on curations based on dorsal cortex) would have poor AUC values and one would instead expect the blue curve to lie beneath the red curve. Gene expression analysis of the coronal *in situ* hybridisation images from the Allen Mouse Brain Atlas for representative genes on this continuum: *Foxp2* (0.01), *Fam163b* (0.25), *Plxna2* (0.55), *Tmem91* (0.9), and *Gfra1* (1.4). The upper teal arrow indicates dorsal cortex, while the lower teal arrow indicates lateral cortex. Comparing laminar expression patterns in dorsal and lateral cortices, *Foxp2* and *Fam163b* show no differences (both genes show enrichment in deeper layers in each area), *Plxna2* shows slight differences (enriched in deeper layers in both areas, but slightly more so in laminar cortex), and *Tmem91* (heavy expression in layer 5 of dorsal cortex, sparse and relatively unpatterned expression in lateral cortex) and *Gfra1* (heavy expression in deep layers of lateral cortex, sparse and unpatterned expression in dorsal cortex) show large differences.

Table 3.5: Predicted gene sets for various layers are similar in a replication from an overlapping cortical area and also from a non-overlapping cortical area. Overall concordance is calculated as (# genes predicted enriched in both + # genes predicted not enriched in both) / (total # genes with a prediction in both). Much of the variability in the overall concordance is due to class imbalance (that is, fewer genes are predicted to be enriched in both original and replication datasets in layer 4 than in layer 5). Enriched concordance is calculated as [$\frac{\text{\# genes predicted enriched in both}}{\text{\# genes predicted enriched in both} + \text{\# genes predicted enriched in original set only}} + \frac{\text{\# genes predicted enriched in both}}{\text{\# genes predicted enriched in both} + \text{\# genes predicted enriched in replication set only}}] / 2$. I report this average to account for unequal precision and recall between original and replication classifiers. AUC, precision, recall, genes predicted and genes expected are defined in the legend of Table 3.2. 12,880 genes were predicted in the dorsal cortex replication, overlapping 9,690 genes classified in the original dataset. Similarly, 12,889 genes were predicted in the lateral cortex set, overlapping 9,700 genes classified in the original dataset. I estimated 49% of genes were patterned in both the dorsal and lateral replications, similar to the 51% in the original dataset. Functional enrichments were generally replicated in both cortical areas using these gene sets (see Table S6 in Belgard et al. (2011)). Finally, while lateral cortex was generally concordant with dorsal cortex, I also identified numerous genes that differed in laminar patterning across areas (Figure 3.11).

	Classifier	Overall concordance	Enriched concordance	AUC	precision	recall	genes predicted	genes expected
Dorsal cortex replication	Layers 2/3	79%	57%	0.73	42%	52%	3290	2657
	Layer 4	96%	50%	0.83	44%	27%	274	447
	Layer 5	74%	62%	0.76	59%	67%	4258	3750
	Layer 6	96%	61%	0.78	37%	37%	884	884
	Layer 6b	93%	58%	0.74	40%	39%	1200	1231
	Unpatterned	67%	68%	0.68	52%	51%	6434	6560
Lateral cortex replication	Layers 2/3	78%	55%	0.68	38%	44%	3279	2832
	Layer 4	96%	51%	0.85	42%	30%	405	567
	Layer 5	71%	55%	0.70	56%	60%	4178	3899
	Layer 6	94%	48%	0.83	42%	52%	1001	809
	Layer 6b	91%	52%	0.77	41%	46%	1375	1226
	Unpatterned	64%	65%	0.63	46%	43%	6144	6573

LincRNA expression varies across neocortical lamination

Dr. Ana Marques and I next considered the 1,055 lincRNA loci, expressing 1,879 multiexonic transcripts, which we identified as being expressed in the cortical cell layers, usually at low levels. Dr. Marques found a large majority (67%) of these loci were novel, in that they had no overlap with annotated noncoding RNA loci (Rhead et al. 2010), likely due to their low expression. Sequence constraint, an indicator of functionality, for these loci was similar to that seen for other lincRNA locus sets (Marques and Ponting 2009) (Figure B.5; <http://genserv.anat.ox.ac.uk/layers>). These lincRNA loci also exhibited experimental hallmarks of transcribed loci: they significantly overlapped DNase I hypersensitivity sites and histone methylation marks that are associated with active transcription in neuronal precursor cells (Figure B.5).

Some lincRNA loci have been predicted (Ponjavic et al. 2009), and experimentally verified (Orom et al. 2010), to act in *cis* by regulating the expression of genomically proximal protein-coding genes. Some lincRNAs expressed in our cortical samples may also possess such functions, particularly in the regulation of patterns of expression across cortical layers. This is supported by patterned protein-coding gene loci being 39% more likely to be adjacent to one or more of these lincRNA loci than expected by chance ($p < 10^{-4}$), and lincRNA expression across cortical layers being more often correlated, positively or negatively, with expression of their protein-coding genomic neighbours if these genes were patterned than if they were unpatterned ($p < 0.02$). Furthermore, we found a significant enrichment of cortical lincRNA transcription across enhancers defined in *in vitro* neuronal culture (Kim et al. 2010) (2-fold increase, $p < 10^{-4}$). Finally,

because Evofold-predicted (Pedersen et al. 2006) RNA structures were substantially enriched in these transcripts (2-fold increase, $p = 2 \times 10^{-3}$) the activity of some cortical lincRNAs is likely to be mediated by their secondary structures.

At least 76 of these cortical lincRNAs, from 66 loci, exhibited differential expression across cortical layers. Importantly, patterned lincRNAs were 63% more likely to be adjacent to patterned protein-coding genes than expected by chance ($p < 0.01$) (Figure 3.12) supporting a role of patterned lincRNAs in the regulation of cortical genetic architecture. As illustrated in Figure B.5 and Figure 3.12, some patterned lincRNAs transcripts are *bona fide* layer specific markers. At least one lincRNA was highly expressed outside of cortex (subventricular zone, dentate gyrus and Purkinje cells of the cerebellum; Figure B.5D), suggesting that some of these lincRNAs may be of broader interest to neuroscientists.

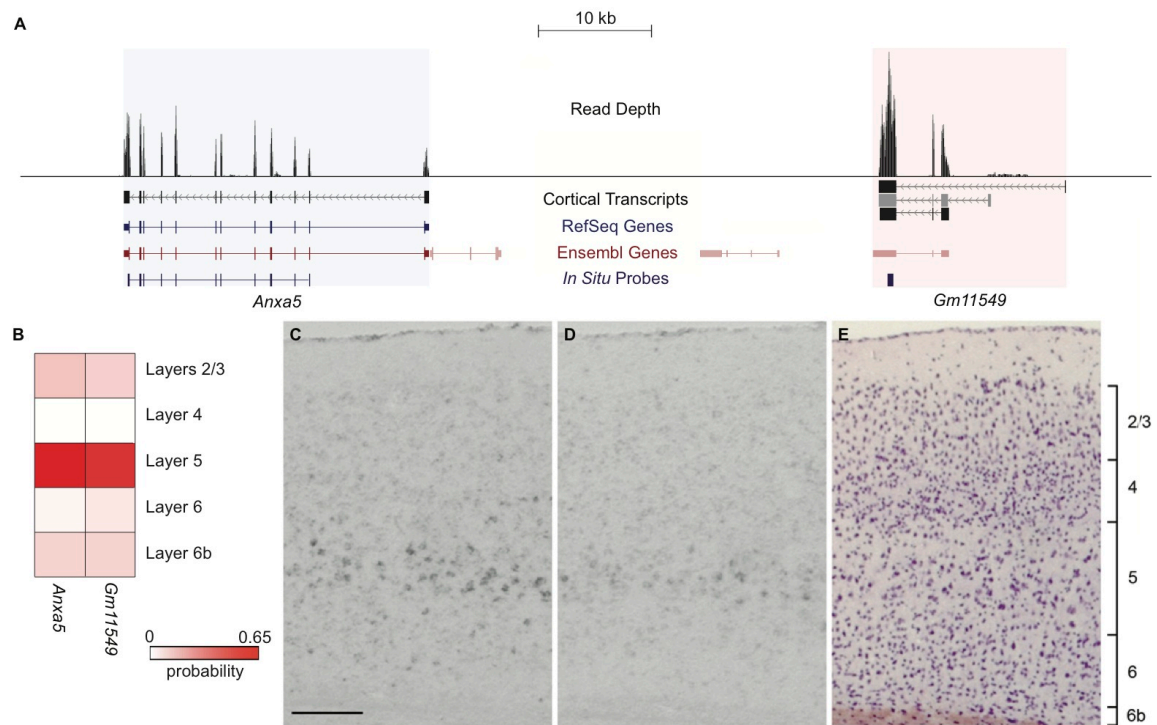


Figure 3.12: *Anxa5* cortical expression overlaps with that of its neighbouring cortical lincRNA (*Gm11549*). (A) Read depth from all cortical RNA-seq data, and multiexonic *de novo* transcripts, in the mouse chr3:36,347,839-36,431,074 region. Of all predicted Ensembl [coding (red) and noncoding (pink)] and RefSeq gene annotations in this region only *Anxa5* and *Gm11549* were expressed. All transcripts overlapping *Gm11549* were predicted with high confidence to be noncoding except for the one marked in grey, which had a short open reading frame covering less than 8% of the mostly repeat-derived transcript. Locations of the *in situ* probes are indicated in purple. (B) Calibrated cortical layer enrichment probabilities for *Anxa5* (ENSMUSG00000027712) and *Gm11549* (*de novo* model XLOC_901616). Both genes were predicted to be enriched in layer 5 (uncalibrated probability > 0.5). *In situ* hybridisations for (C) *Anxa5* and (D) *Gm11549* confirmed that their expression in the mouse cortex was as predicted. (E) Cortex stained for Nissl. The

maximum read depth at any position in this region, combining all samples, was 2,451. Scale bar = 200 μ m. Dr. Ana Marques and I designed the figure, and Dr. Peter Oliver provided panels C-D.

3.4 Discussion

We have demonstrated differential expression of protein coding or noncoding alternative transcripts and genes across expression levels spanning over six orders of magnitude for individual layers of the mouse neocortex with what is the first published deep sequencing of transcriptomes from separate mammalian neocortical layers. I built an interactive interface to explore the data, and provided links to download them, at <http://genserv.anat.ox.ac.uk/layers>. All raw data are also available in GEO and the SRA under accession number GSE27243. This resource should assist future studies that seek a detailed molecular and functional taxonomy of cortical layers and neuronal cell types. These results may increase by three to four-fold the number of known (Lein et al. 2007) layer-enriched marker genes (Figure 3.6), and furthermore introduce 66 lincRNA loci as new likely markers. These markers can assist studies of cortical cell types, neurodevelopment and comparative neuroanatomy. The data even augment known marker genes by providing a more objective grounding for their laminar classifications on the basis of quantitative expression level. They also reveal novel observations on each layer's neurological functions that lead to new lines of enquiry, for example regarding the roles of Alzheimer's disease genes or MHC genes in layers 2/3 or of mitochondrial biology in layer 5. Our findings in mouse are expected to be highly relevant to human biology owing to these species' strong similarities in brain transcriptomes (Strand et al. 2007) and to the similarities of layer markers between

mouse and ferret (Rowell et al. 2010), which is an evolutionary out group to rodents and primates. Even unexpected findings, such as the significant and replicated association between coronary artery disease and layer 5 expression, may reflect genetic underpinnings of previously-described clinical associations between vascular and neurological disease (Beerli et al. 2006; Santos et al. 2009).

My application of a machine learning classifier to carefully annotated high-throughput *in situ* hybridisations (Lein et al. 2007) yields expression levels and predictions of laminar patterning that are based on transcripts, as well as on genes, and on noncoding loci, as well as on known genes. The expression levels assessed by RNA-seq are more sensitive to smaller differences, and these can be explored on the website. Results reflect the genome as a whole (except for repetitive sequence within which mapping of reads is problematic), rather than for the limited sequence targeted by microarray probe sets. This revealed numerous lincRNA transcripts, mostly novel, which were evolutionarily constrained, sometimes imprinted (Gregg et al. 2010), and at least one that was most strongly expressed outside of cortex, opening new avenues for research into their extra-cortical functions. Additionally, I found transcripts from the same gene exhibiting expression divergence across neocortical layers, which should be investigated for potential physiological consequences. None of this would have been possible with currently available microarray based methods.

Nevertheless, this approach will be limited by imperfections in dissection, and by contributions to one layer of transcripts emanating from radial processes of cells whose soma lie in another. These limitations will degrade the classifier's performance, and hence will contribute to a large number of genes (56%) whose maximum predicted

probabilities lie below 0.5. Nevertheless, the approach still provides at least a ten-fold difference in the relative probability of enrichment in different layers for over 10,857 (95%) classifiable genes and thus is effective at inferring transcriptional levels amongst mixed populations of cells in their *milieu*, rather than for cells that have been sorted, purified or microdissected (Markram et al. 2004; Molyneaux et al. 2007; Nelson et al. 2006). Indeed, there is a recent demand for integration of neuronal, glial and vascular interactions on a molecular and cellular level within the same neuronal structures (Neuwelt et al. 2011). These findings make possible future comparisons of whole transcriptomes across both isolated cell-types and cell layers that should yield further insights into the molecular components of the neuronal circuitry underlying higher brain functions. Finally, this dataset has allowed me to begin to compare gene expression in the brains of mammals and birds, as described in Chapter 4.

Chapter 4 - Gene Expression Evolution in the Brains of Mammals and Birds

Abstract

Homologies between the mammalian and avian pallium, derivatives from the dorsal part of the developing telencephalon, are among the thorniest issues in comparative neurobiology. Pallial derivatives in both clades process and integrate sensory information and are involved in voluntary movement. However, the anatomical structures responsible for this sensory processing are derived from different developmental fields. This is curious since the stem amniote presumably had the capability to process this information. We add to this long-standing debate yet another phenotype – genome-wide patterns of gene expression, assessed through RNA-seq, in adult mouse and chick. Structures of known homology (hippocampus and a subpallial structure, striatum) share similar gene expression profiles, as do two functionally analogous structures (mammalian layer IV neocortex and avian nidopallial subdivisions). The latter may be a case of convergent evolution of gene expression since layer IV is a dorsal pallial derivative while nidopallium is a ventral pallial derivative and since both adult structures are clade-specific. Many of the candidate cross-species markers of homologous structures identified in this project have known functional importance, while there is presently little functional information about others. Among the structures of debated homology, I find no clear transcriptomic evidence of classical one-to-one homology in the pallium outside of hippocampus. To the contrary, the genes most

specific to any controversial pallial structure in one species show no consistent pattern of expression in the other species. This might be taken to support the idea that the progenitors of these structures are homologous as a developmental field, but that the structures themselves are not homologous in the classical sense.

4.1 Introduction

Gene expression is a phenotype that can be studied in an evolutionary context

Evolution can be studied at many levels. Comparative genomics, with its focus on genotype, lies at one end of the spectrum. At the other end, behavioural ecology explores the adaptive (and, perhaps, maladaptive) qualities of animal behaviour in its changing social and environmental context. Many fields fill the intervening gap, including evolutionary developmental biology, comparative anatomy and evolutionary physiology. Studying evolution at multiple levels helps one to properly appreciate how genotype is translated into adaptive or maladaptive phenotypes.

Gene expression is a phenotype sometimes thought to be more proximal to genotype. While this may be true, ‘the transcriptome’ is nevertheless a dynamic ensemble phenotype affected by anatomical position, epigenetics, social status, diet, physiological state, biological rhythms, transcript degradation rate and many other factors. Despite this complexity, there are steady-state features of the transcriptome that vary primarily by anatomical location. This can reflect, among other things, a shared developmental lineage or neighbourhood environmental cues. Thus, ‘marker genes’ can be fruitfully

used to delineate regions or cell types. Given the promiscuity of many mechanisms of transcriptional control, the mere regional specificity of a marker gene to a region need not necessarily guarantee that the gene has a functional role in those cells.

Comparative transcriptomic studies might add confidence to such inferences. *A priori*, one may expect that a marker gene whose expression specificity is conserved in a homologous region across long evolutionary time is more likely than a non-conserved gene (in expression) to be used in core, conserved features of that homologous region. While it is theoretically possible that the common ancestor did not express this shared gene in the homologous region – i.e. that the observation is a case of convergent evolution – a more parsimonious explanation is that this specificity is ancestral. Naturally, there will be genes whose expression is specific to the homologous region in one species and not in the other that will nevertheless be important in function for the former. However, without an out-group it is impossible to determine if such specificity is ancestral (and was lost in the second species) or derived along the lineage of the first species.

Types of biological similarity and sameness

‘Homology’ is broadly understood to mean ‘sameness’ (typically phyletic continuity of adult structure or of the developmental fields that give rise to those structures), though its precise definition varies. Hobart Smith (1967) described two types of phyletic homology germane to the present discussion: in discrete homology, one-to-one comparisons are made between individual structures in related species; in field homology, derivatives of a common morphogenetic developmental field are compared. Thus, statements of

homology must include precisely what is being homologised. Furthermore, similarity of function is distinct from homology – instead, it is said to establish ‘analogy’. An illustrative example from Butler and Hodos (2005) is comparing the wings of a bat to those of a bird. These two structures are analogous as wings, in that they serve a similar functional purpose. They are also homologous as derivatives of the forelimb. However, they are not homologous as wings since the most recent common ancestor of bats and birds did not have wings (violating phyletic continuity).

Since the brain, like other soft tissues, does not fossilize, comparative neuroanatomists often infer homology from study of extant species. Butler and Hodos (2005) list several examples of similarity that might be used in making a case for homology: topological, topographical, axonal connectivity, relationship to some consistent feature of the two species, embryological derivation, neuron morphology, neurochemical attributes, physiological properties, and behavioural outcomes of neuronal activity. The last two criteria are controversial, since they are based on function rather than structure, but the authors insist that these examples of functional similarity can strengthen a structural argument for homology. Inconsistencies among the criteria can be difficult to reconcile.

Even similarity of structure by itself cannot be said to constitute homology: without phyletic continuity this is instead referred to as ‘homoplasy’. This can occur for one of three reasons: convergence, in which similar adaptive pressures create comparable structures from different ancestral structures; parallelism, in which comparable structures arise independently from the same ancestral structure (this can sometimes be difficult to distinguish from homology, the more parsimonious explanation); and reversal, in which

an ancestral structure disappears and later reappears on a lineage (Butler and Hodos 2005).

Alternative approaches include biological homology and generative homology (Butler and Hodos 2005). These usually overlap with historical homology, which is based on phyletic continuity (Table 4.1). The emphasis in biological homology is on morphological identity; in practice one identifies structures that share a set of developmental processes and constraints. Generative homology requires genetic and morphogenetic identity. This approach allows a character to come and go along a lineage (as in a reversal) but allows for the characters to be homologous if the underlying genetic and morphogenetic identity is the same.

Table 4.1: Historical homology versus generative homology. Adapted with permission from John Wiley and Sons from Table 1-3 of Butler and Hodos (2005).

	Components	Biological Basis	Phenotypic Expression	Opposite
Historical Homology	Historical homology cases only	Inheritance from common ancestor of the character itself	Required to be consistent along and/or across the taxon	Homoplasy: parallelism, reversals, and convergence
Generative Homology	Most cases of historical homology plus parallelism and reversals	Inheritance from common ancestor of genetic and/or morphogenic basis for the character	Can be either consistent or inconsistent along and/or across the taxon	Allogeny (= convergence)

The amniote pallium and debates over homologies

One particularly contentious debate in comparative neuroanatomy concerns a subset of cerebral homologies between synapsids (represented today by mammals) and sauropsids (represented today by reptiles and birds), which together are called ‘amniotes’. More

specifically, this particular debate concerns characters in the ‘pallium’, which develops from the dorsal portion of the telencephalon. The ventral portion of the telencephalon develops into what is known as the ‘subpallium’.

In mammals, the pallium includes the hippocampus, neocortex and the claustramygdalar formation. In birds, the pallium includes the hippocampus, wulst (also called ‘hyperpallium’) and the dorsal ventricular ridge. Both clades have a subpallial striatum and pallidum. While both mammals and birds have a pallial medial hippocampal formation and a pallial lateral (piriform) cortex, the other pallial contents in the two clades appear strikingly different (Figure 4.1).

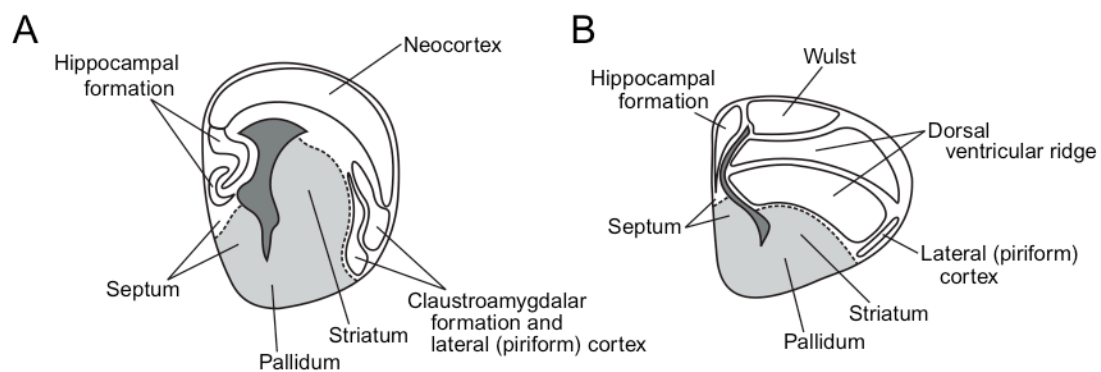


Figure 4.1: Diagram of coronal hemisections through the telencephalon to show the major pallial and subpallial regions in a mammal (A) and a bird (B). The ventricle is indicated by dark shading and the subpallium by lighter shading. In both the mammal and the bird telencephalon hemisections, all of the area in white is pallium. Figure and caption reproduced from Figure 3-10 of Butler and Hodos (2005) with permission from John Wiley and Sons.

The pallium is not monolithic. In development across mammals, birds, reptiles and amphibians, it is composed of four morphogenetically delineated components: medial, dorsal, lateral and ventral (until split by recent molecular evidence, these last two were combined in what might be called lateroventral pallium; some fields exist in fish and lampreys as well) (Bachy et al. 2002; Brox et al. 2002; Brox et al. 2003, 2004; Fernandez et al. 1998; Murakami et al. 2005; Osorio et al. 2005; Wullimann and Mueller 2004). Moreover, the relative sizes of these developmental fields differ between mammals and birds (Figure 4.2). Specifically, lateroventral pallium is expanded in birds while dorsal pallium is expanded in mammals.

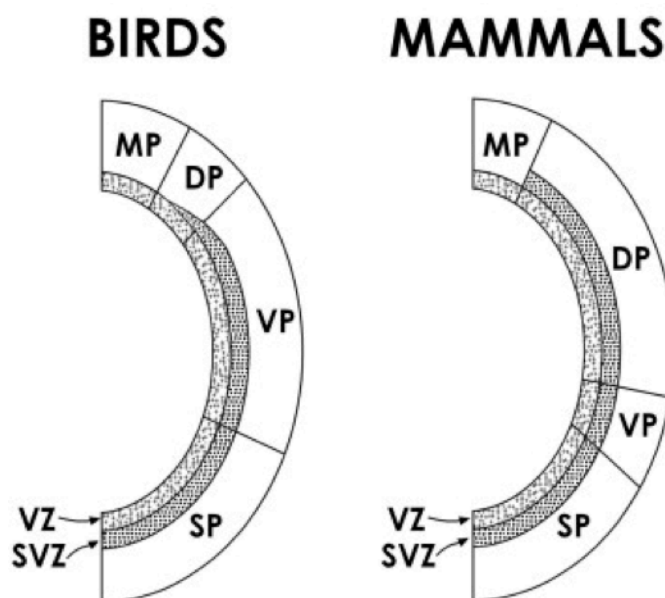


Figure 4.2: Pallial components in different amniotes. In birds, the subventricular zone (SVZ) is observed mostly in the lateroventral pallium (VP) and in the subpallium (SP). The presence of an SVZ-like structure in the dorsal pallium (DP) of birds is controversial. In mammals, the SVZ strongly expands into the DP. MP,

medial pallium; VZ, ventricular zone. Figure reproduced and caption adapted from Figure 4 of Aboitiz (2011) with permission from Elsevier.

These components are delineated by developmental marker genes as shown in Figure 4.3 (Medina et al. 2005). All pallial areas preferentially express *Tbr1* and *Pax6*. Lateral, dorsal and medial pallium express *Emx1*, while ventral pallium expresses *Dbx1*. In mouse, medial pallium and dorsal pallium express *Lmo2*, while *Lmo3* is found in dorsal, lateral and ventral pallium. *Dlx1*, *Nkx2.1*, *Emx1* have a conserved expression pattern at developmental stages across tetrapods (Fernandez et al. 1998; Smeets et al. 2000); and *Pax6* and *Shh* have a conserved expression pattern across amniotes (Frowein et al. 2002; Puellas et al. 1999). However, it is important to note that there is not yet a confirmed genetic marker that divides dorsal pallium, medial pallium and lateral pallium across long evolutionary timescales. Thus these regions are typically defined topologically and by the functions performed by their adult regions, and genes such as *Lmo2* and *Lmo3* are assigned to the appropriate pallial region on this basis. Topological inferences are drawn since the local radial glia both differentiate into and guide the outward migration of excitatory neurons (there are, however, occasional cases of tangential migration). Subpallium is marked by *Dlx-2* expression. Beyond their status as markers, most of the cross-species genes listed here have a demonstrated functional importance in development of the telencephalon, and thus cannot be explained away as non-functional but patterned expression (Bedogni et al. 2010; Bishop et al. 2003; Chou et al. 2009; Favaro et al. 2009; Flandin et al. 2010; Georgala et al. 2011; Han et al. 2011; Hong et al. 2007; Shinozaki et al. 2004; Suh et al. 2009; Tuoc et al. 2009).

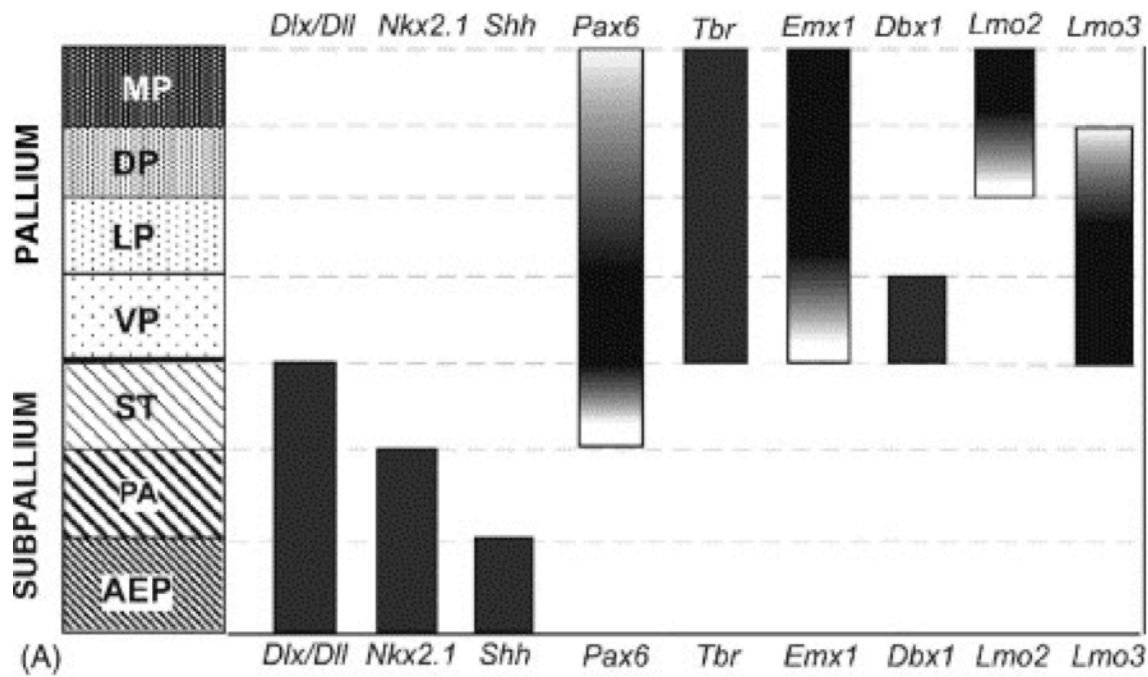


Figure 4.3: Diagrams showing the telencephalic divisions and subdivisions in tetrapods. Each telencephalic division/subdivision shows expression of a unique combination of developmental regulatory genes. Many of the genes represented have been analysed in mouse, chick and *Xenopus* (*Dlx/Dll*, *Nkx2.1*, Sonic hedgehog or *Shh*, *Pax6*, *Tbr1/2*, and *Emx1*). Some of these have also been analysed in turtle (*Dlx*, *Pax6*, *Emx1*). Their expression patterns in the different species, in combination with other anatomical and molecular data, suggest that all tetrapods show the same (field homologous) molecularly distinct divisions and subdivisions in the telencephalon. Some of the genes shown have only been analysed in mouse (*Dbx1*, *Lmo2*, *Lmo3*) and more studies are needed in non-mammalian species. Abbreviations: AEP, anterior entopeduncular area; DP, dorsal pallium; LP, lateral pallium; MP, medial pallium; PA, pallidum; ST, striatum; VP, ventral pallium. Figure and caption adapted from Figure 1 of Medina et al. (2005) with permission from Elsevier.

Some domains give rise to structurally and functionally similar adult structures in both mammals and birds. For example, the hippocampal formation (and the transient cortical hem) in both species arises from medial pallium. Likewise, lateral (piriform) cortex in both species arises from a combination of lateral pallium and (especially) ventral pallium (Puelles et al. 2007).

In contrast, some derivatives of the same pallial domains in different clades appear rather different. Here I use the anatomical and functional definitions of dorsal and lateral cortex, but it is important to keep in mind that since there is not yet a genetic marker validated across distantly related species it may be difficult to confidently discriminate between cross-species equivalent dorsal pallium and lateral pallium. For example, the hyperpallium in birds derives from the dorsal pallium, which is universally understood to give rise to the mammalian dorsal cortex. The CDL in birds comes from dorsal and/or lateral pallium. Avian mesopallium topologically corresponds to lateral pallium. Due in part to robust morphogenetic markers, structures deriving from ventral pallium are less ambiguous. These include, in chick, nidopallium, arcopallium, and most of piriform cortex. In mouse, these include most of piriform cortex and parts of both claustrum-endopiriform complex and pallial amygdala (collectively known as the claustramygdaloid complex). Piriform cortex in both species and the mammalian claustramygdaloid complex (Figure 4.4) are also thought to partially derive from lateral pallium.

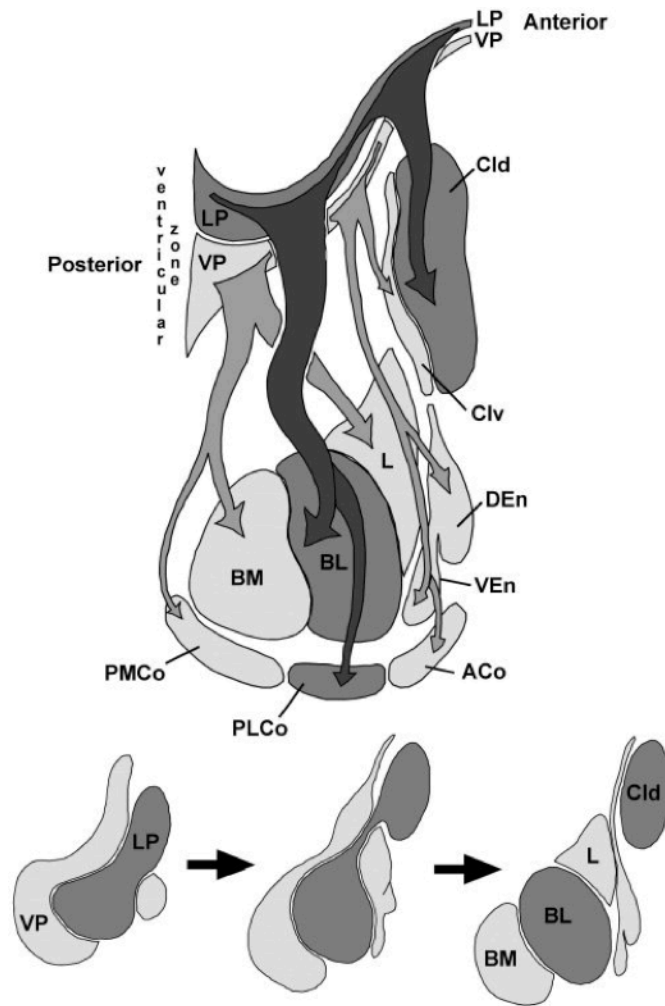


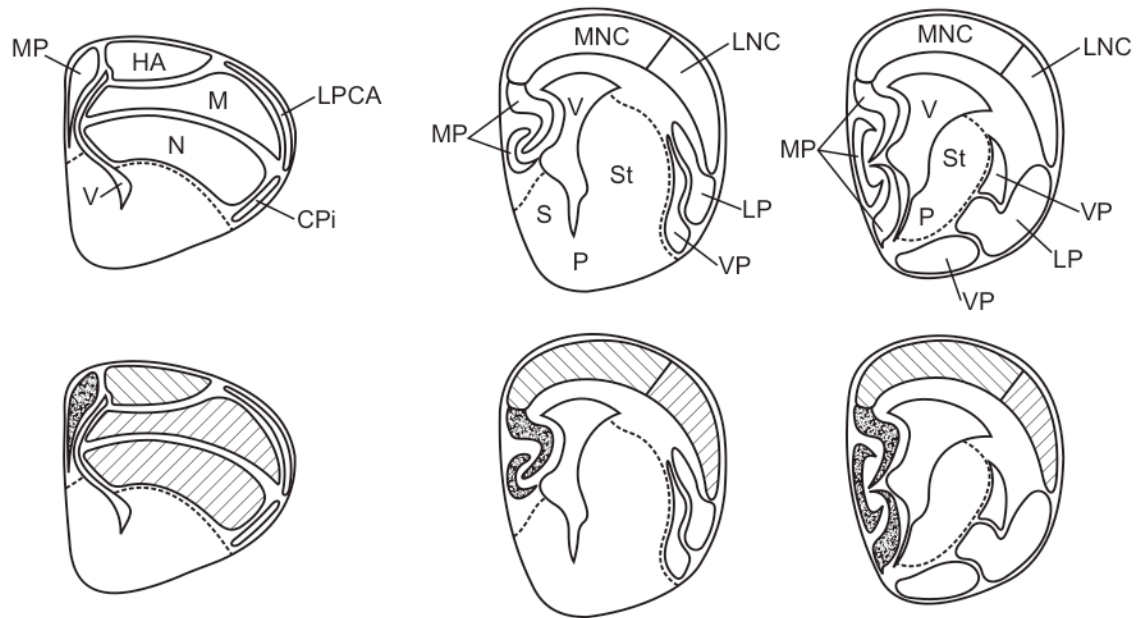
Figure 4.4: Schema showing the lateral pallial or ventral pallial origin of the nuclei/areas of the claustramygdaloid complex. Based on their selective and strong expression of Cadherin 8 plus *Emx1*, the claustrum proper (or dorsolateral claustrum; Cld), the basolateral amygdalar nucleus (BL) and the posterolateral cortical amygdalar area (PLCo) may be derivatives of rostral or caudal levels of the lateral pallium. On the other hand, based on their specific and strong expression of Neurogenin 2 and/or Semaphorin 5A, the ventromedial claustrum (Clv), the endopiriform nuclei (DEn and VEn), the lateral and basomedial amygdalar nuclei (L, BM), plus the anterior and posteromedial cortical amygdalar areas (ACo,

PMCo) may derive from different rostrocaudal levels of the ventral pallium. At the bottom, a hypothetical model of development of LP and VP parts of the claustramygdaloid complex is shown. This tries to explain the tridimensional configuration and interdigitation of the different components of the complex observed in late embryos and in the adult. Figure and caption reproduced from Figure 11 of Medina et al. (2004) with permission from John Wiley and Sons.

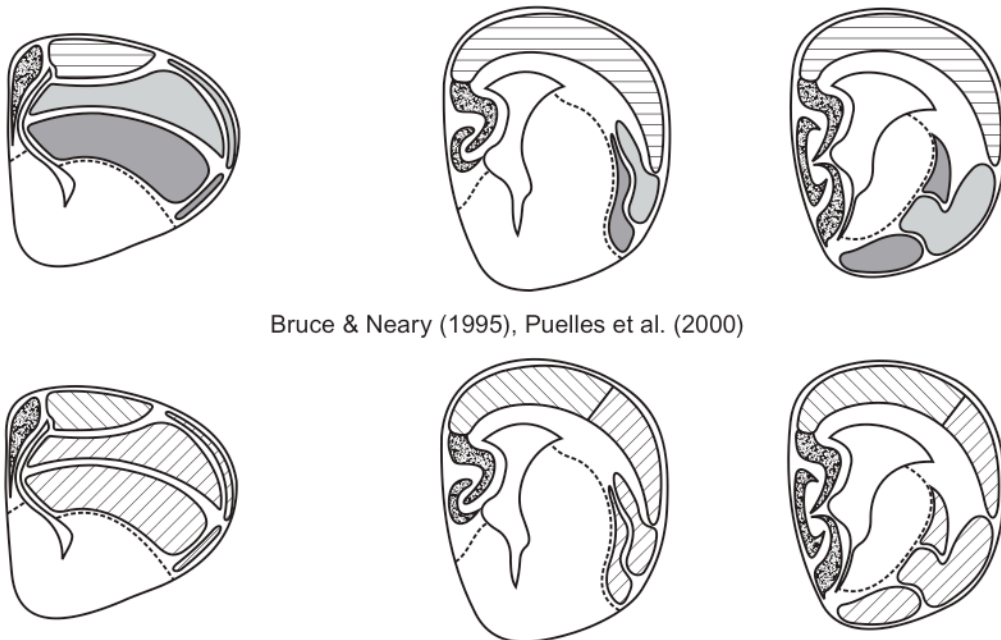
Given their clear boundary delineation and demonstrated functional importance, these markers have been used to make inferences of homology. Some of the suggested homologies are straightforward and widely accepted. The structures described above that are structurally and functionally similar, in the proper topographical organisation and derived from the matching developmental field are widely accepted as homologous. This includes the piriform cortex and parts of the hippocampal formation (as well as other medial pallial derivatives).

These markers have also been used to advance controversial arguments of homology (Figure 4.5). For example, Puelles and others homologise avian nidopallium and arcopallium with ventral pallium derivatives in mammals (Bruce and Neary 1995; Martinez-Garcia et al. 2002; Medina et al. 2005; Puelles 2001; Puelles et al. 2000; Striedter 1997). As suggested above, the presumed cross-species marker of ventral pallium (lack of *Emx1*) is supportive of these homologies. In this line of reasoning, avian mesopallium is homologised with mammalian lateral pallium derivatives. Finally, the whole of mammalian neocortex is homologised with hyperpallium. This is justified on the basis of the structural similarity across different areas of neocortex, and the fact that the genetic organisation of adult neocortex is primarily laminar rather than areal.

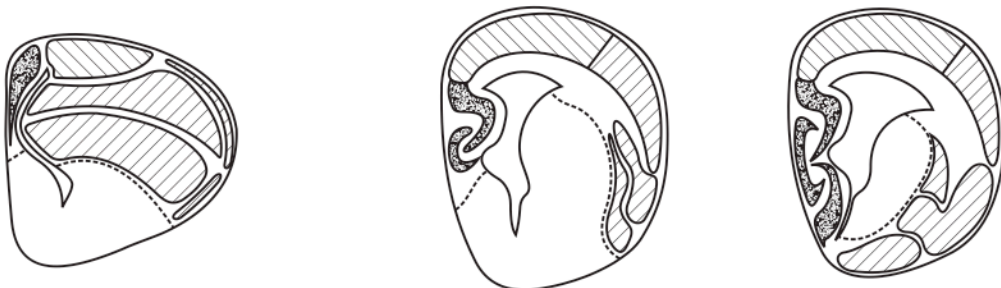
Furthermore, in the absence of a solid cross-species marker of dorsal pallium, it is difficult to say with great confidence where the avian-equivalent dorsal pallium ends and lateral pallium begins.



Karten (1969), Reiner (1993, 2000), Butler (1994)



Bruce & Neary (1995), Puelles et al. (2000)



Butler and Molnár (2002)

Figure 4.5: Overview of three current theories on the evolution of the dorsal pallium in amniotes. Diagrammatic representations based on Puelles (personal communication to Butler and Hodos). In each row, a drawing of a coronal hemisection through the telencephalon of a developing bird is shown to the left, and two drawings of coronal hemisections through the telencephalon of a developing mammal are shown to the right, with the far right drawing being the more caudal one and through the level of the amygdala. In the top row, the structures are identified. The second row illustrates the anterior dorsal ventricular ridge (ADVR)-lateral neocortex hypothesis (Butler 1994; Reiner 1993), derived from the equivalent cell hypothesis of Karten (1969). The third row illustrates the ADVR-claustroamygdalar hypothesis (Bruce and Neary 1995; Striedter 1997) (Martinez-Garcia et al. 2002; Medina et al. 2005; Puelles 2001; Puelles et al. 2000). The fourth row illustrates the ADVR-lateral neocortex plus claustroamygdalar field homology hypothesis (Butler and Molnar 2002). Shading and fill patterns are used to indicate comparative structures for each hypothesis. Since the piriform cortex is not shown as a separate entity in the mammalian figures but rather is included in the LP/VP regions, it is not shaded in most cases. All of these hypotheses basically agree on a discrete homology of most or all of piriform cortex across amniotes. Abbreviations are V, lateral ventricle; HA, hyperpallium apicale; M, mesopallium; N, nidopallium; CPi, piriform cortex; MNC, medial (lemnopallial) part of neocortex; LNC, lateral (collopallial) part of neocortex; S, septal nuclei; LP, lateral pallium; LPCA, lateral pallial cortical area; MP, medial pallium; P, pallidum; St, striatum; VP, ventral pallium. Figure reproduced and caption adapted from Figure 25-14 of Butler and Hodos (2005) with permission from John Wiley and Sons.

Other classes of proposed homologies try to account for hodological evidence, that is, data on connectivity within networks of similar function. These draw from Harvey Karten's "equivalent circuit hypothesis", based on the observation of similar connectivity of neocortical layers in the visual and audial systems in mammals and in the connectivity within pallial regions of birds (Figure 4.6; Figure 4.7). I will only discuss here a representative example from two major classes, as the other postulates are variations on a theme.

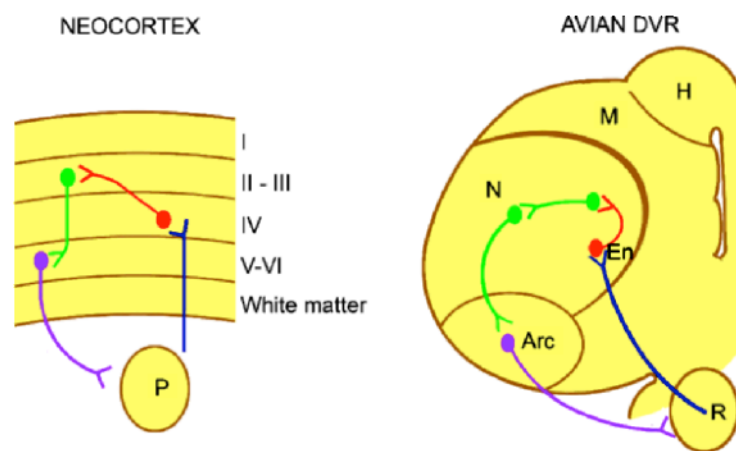


Figure 4.6: Karten's hypothesis (Karten 1969, 1997) of the equivalent circuit in the mammalian neocortex and the avian dorsal ventricular ridge (DVR). I, II, III, IV, V, VI represent distinct neocortical layers; Arc, arcopallium; En, entopallium (a component of the nidopallium); H, hyperpallium; N, nidopallium; P, pulvinar; R, nucleus rotundus. Figure and caption reproduced from Figure 12 of Aboitiz and Montiel (2007) with kind permission of Springer Science + Business Media and the authors.

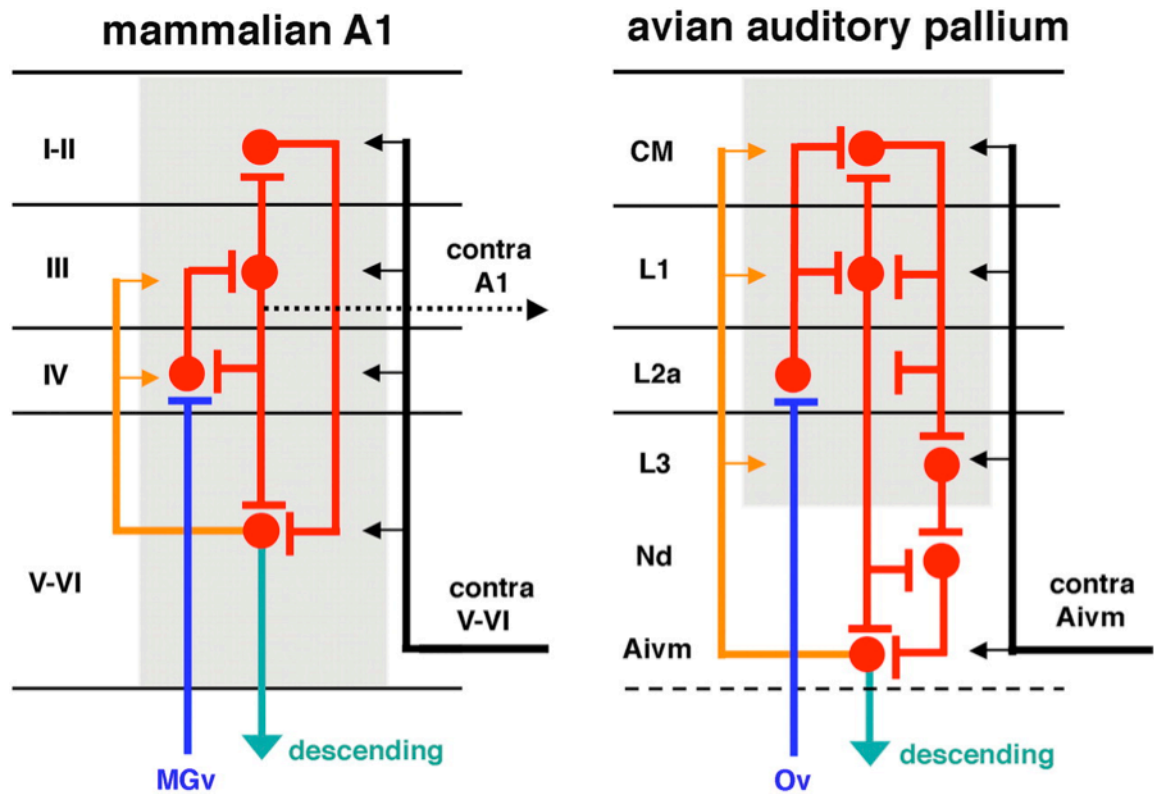


Figure 4.7: Comparable hodological organization of the mammalian A1 (left) and the avian auditory pallium (right). While there is a columnar structure in the avian pallium, the radial arrangement of its columns differs from that of columns in the neocortex. Blue, red, and green lines and arrows indicate their thalamic inputs, intrinsic connections, and descending projections, respectively. Orange lines and arrows indicate recurrent projections from the deep layers to the more superficial layers. Black lines and arrows indicate re-entrant projections from the other side of the brain. Innervation that is comparable to the projection from the mammalian layer III on the contralateral A1 (dotted black line) has not been identified in birds. Grey regions indicate columnar functional modules. Note that this schematic drawing illustrates only the major components of this intricate network. CM, caudal mesopallium; A1, primary auditory cortex; Nd, dorsal nidopallium; Aivm,

intermediate arcopallium. L1, L2a and L3 are subdivisions of field L, which is in the caudal nidopallium. Figure reproduced and caption adapted from Wang et al. (2010). Copyright 2010 by National Academy of Sciences.

The older of these (Figure 4.5) homologises both avian mesopallium and nidopallium with mammalian lateral neocortex, largely on the basis of hodological and functional similarities (Butler 1994; Karten 1969; Reiner 1993). Namely, this approach has homologous structures processing audial and visual information.

If the apparent cross-species genetic markers of ventral pallium are to be believed, this statement of homology would predict a tangential migration of neurons from the ventral pallium that would populate certain layers of the neocortex. While a subpopulation of ventral pallial neurons do indeed migrate to populate the mouse neocortex, their number is low (<5% of cortical cells at birth) and is drastically reduced by programmed apoptosis after completion of corticogenesis such that over 95% of these migratory glutamatergic neurons disappear by adulthood (Teissier et al. 2010). Hence, there is still no evidence for a migratory stream postulated over forty years ago. Nevertheless, it remains possible that such a migratory stream exists but has not yet been identified.

Butler and Molnár (2002) proposed a hypothesis that, in a rejection of statements of one-to-one homology for this thorny problem, attempts to reconcile developmental and hodological evidence through the use of field homology (Figure 4.5). Briefly, this homologises as a developmental field the debated derivatives of lateral and ventral pallium, namely avian mesopallium, nidopallium and arcopallium, and mammalian lateral neocortex and claustramygdaloid complex.

Comparative neurotranscriptomics in mouse and chick

Like homology, patterns of adult gene expression is yet another “final form” phenotype with functional consequences. New sequencing technologies make high-throughput comparative studies at the level of gene expression tractable and easier than ever before. While not impossible, such work is difficult to perform using microarrays, as they should be carefully custom designed for comparative work. Even then, uncertain isoform relative abundance in the relevant tissues in each species could complicate probe design. RNA-seq requires an annotated genome sequence, which is available for a growing number of species. (Some may argue that this is not required, but it certainly simplifies the endeavour!)

These “comparative neurotranscriptomics” approaches might be particularly informative when applied to structures of both agreed and debated homology. Thus, in close collaboration with researchers at Oxford and the NIH, I applied this to numerous pallial structures in addition to striatum, which is subpallial (but may include a small proportion of neurons derived from the pallium, see Cocas et al. (2009)), as a control.

It is important to remember that this is in adult, not through development. While developmental samples would have been ideal to address questions of homology, the methods to sequence transcriptomes from such small starting quantities had not yet been developed. Furthermore, matching developmental stages may a challenge when comparing mammals and birds.

WGCNA: Weighted Gene Co-expression Network Analysis

Weighted Gene Co-expression Network Analysis, or WGCNA, refers to a collection of published statistical network methods I use in the analysis of this project to find groups (also termed ‘modules’) of apparently co-expressed genes. I then look at the preservation of these co-expression relationships across evolutionary time. WGCNA is not the only approach used to identify co-expression relationships amongst genes (Stuart et al. 2003), but it is notable for its popularity, well-documented and flexible implementation in R, and past use in comparative studies of gene expression in the brain (Miller et al. 2010; Oldham et al. 2006).

The utility of WGCNA is grounded in the approximate ‘scale-free’ topology (Barabasi and Albert 1999) that characterises many biological phenomena including gene regulatory networks (Albert 2005). A network is characterised by a set of nodes with connections to one another (the number of connections is called the ‘degree’ of the node). Networks can be constructed randomly or they can be ordered in some way. A scale-free network is ordered in that its degree distribution follows a power law, implying that some nodes are vastly more connected than others. The biological interpretation of this for gene expression is that some genes (such as transcription factors) are highly connected central ‘nodes’ in a genetic regulatory network.

To construct a gene co-expression network, nodes must be connected in some way. Thresholds to define binary connectivity relationships lose information and can be particularly sensitive to the choice of parameters (Carter et al. 2004). Moreover, simple binary relationships are unlikely to capture the complexity of gene regulatory

relationships. A ‘soft’ thresholding approach, in which the connections are on a continuous scale, is a more robust way to describe gene co-expression (Zhang and Horvath 2005). This is the key insight implemented in WGCNA.

To implement WGCNA, one first calculates an all-genes-by-all-genes matrix of Pearson correlation coefficients based on appropriately pre-processed expression values. An adjacency matrix with the same dimensions is then calculated where the values are proportional to the gene-gene correlations on a logarithmic scale (increasing the difference between higher and lower correlations). This ‘soft power’ is chosen to maximise the mean number of connections within the domain that the network topology is approximately scale-free. The former criterion is to maximise power to detect clusters of co-expressed genes, while the latter criterion is to ensure the network properly reflects the scale-free nature of biological networks.

Next, one needs a matrix of metrics for how ‘connected’ two genes are within the wider network. One metric for this is ‘topological overlap’, which is essentially a weighted measure of the shared connections between two genes, and is defined as

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad \text{where } \omega_{ij} \text{ is the topological overlap between genes } i \text{ and } j, a_{ij} \text{ is}$$

the corresponding entry in the adjacency matrix, $l_{ij} = \sum_u a_{iu} a_{uj}$, and the node

connectivity, k_i , is given by $k_i = \sum_u a_{iu}$ (Ravasz et al. 2002; Zhang and Horvath 2005).

Topological overlap, ω_{ij} , is a similarity measure bounded by zero and one and can be thus transformed into a dissimilarity measure: $d_{ij}^{\omega} = 1 - \omega_{ij}$.

Genes are then hierarchically clustered on the basis of their topological overlap dissimilarity matrices. I divided the resulting dendrogram into clusters of genes using a flexible, automated procedure known as the ‘Dynamic Hybrid’ cut, a variant of Dynamic Tree Cut (Langfelder et al. 2008). The Dynamic Hybrid cut is two-step procedure, in which preliminary gene clusters are identified and genes previously unassigned to a cluster are subsequently assigned to a sufficiently proximal cluster. Preliminary clusters must contain a user-specified minimum number of genes and must be distinct from their surroundings on the tree. Genes nearest the tip of the branch must be tightly connected, and genes too far from a cluster are excluded, even if on the same parent branch. For brevity’s sake, the reader interested in methodological details is referred to the supplementary information of Langfelder et al. (2008). In WGCNA, these clusters of co-expressed genes are typically randomly assigned the name of a system colour (such as turquoise or darkblue) and are called ‘modules’. In this chapter, I will examine the cross-species preservation of these modules.

Chapter structure

The remainder of this chapter is organised into the following sections: Materials and Methods, Results and Discussion. Appendix C contains relevant supplemental figures and tables. The sections of this chapter are ordered as follows:

- mouse and chick dissections & cDNA sequencing;
- cDNA sequencing & gene quantification;
- quality control for batch effects;

- confirming accuracy of the mouse layer, mouse non-cortical, and chick dissections;
- selecting genes to be compared across species;
- identifying gene co-expression modules;
- comparing gene co-expression modules across species;
- annotating gene co-expression modules;
- and discussion of conserved gene co-expression modules in light of possible technical confounds and known biology.

4.2 Materials and Methods

Dissections and RNA extraction

Dr. Fernando Garcia-Moreno dissected the following regions from the brains of eight adult (postnatal day 56) male mice (C57BL/6 strain): dorsal cortex layers I-VIb and lateral cortex layers I-VIb (both are the same samples as last chapter), hippocampus, claustrum-endopiriform complex, pallial amygdala (basolateral and basomedial nuclei), and striatum (Figure 4.8). The mice were killed by cervical dislocation according to approved schedule one UK Home Office guidelines (Scientific Procedures Act, 1986). The mice were decapitated, the skull opened down the midline and the brain removed. Newly dissected brains were rinsed in RNase free PBS, submerged in ice-cold RNAlater (Ambion) for 24 hours and stored at -20°C in RNAlater (Ambion). Whole brains were embedded in 5% agarose (low melting, Bioline) and sectioned using a vibrating microtome (Leica, VT1000S) into 200 µm coronal sections using a 1:1 mixture of

RNAlater and PBS. Samples were dissected out using microsurgical scalpels (Weck, USA) under visual guidance, using transillumination on a dissecting microscope (MZFLIII, Leica) and stored separately in RNAlater at -80°C until all microdissection was complete.

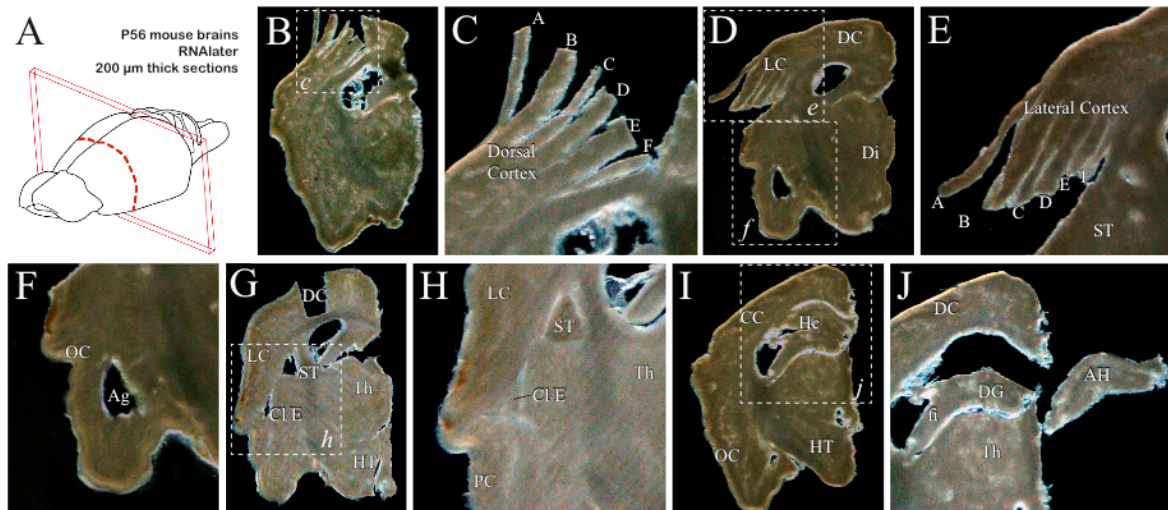


Figure 4.8: Dissection of murine tissue samples. (A) Adult (P56) male mouse brains were immersed in RNAlater, coronally sectioned (200 μ m) and dissected in a 1:1 mixture of RNAlater:PBS. (B) Dissection of dorsal cortical layers on a rostral section. (C) Magnification of the rectangle present in B, showing the cuts of the blade separating the layers. (D) Dissection of lateral cortical layers and pallial amygdala. (E) Magnification of the top rectangle present in D, showing the dissected cortical layers. (F) Magnification of the bottom rectangle depicted in D, following dissection of the pallial amygdala (Ag). (G) Dissection of the striatum (ST) and the claustrum/endopiriform complex (Cl/E). Layers of dorsal cortex (DC) were already dissected in this image. (H) Higher magnification of the rectangle in G, where both striatum and the claustrum/endopiriform complex have already been

cut, but are still present. **(I) Dissection of hippocampal (Hc) tissue samples from caudal sections. (J) Magnification from rectangle present in I, showing the removal of the Ammon's horn (AH) from the dentate gyrus (DG) and the fimbria (fi). Other abbreviations used are cerebral cortex (CC), diencephalon (Di), hypothalamus (HT), lateral cortex (LC), olfactory cortex (OC), piriform cortex (PC) and thalamus (Th). This figure is adapted from a figure made by Dr. Fernando Garcia-Moreno.**

Dr. Juan Montiel dissected the following regions from the brains of eighteen twelve-week-old chickens of mixed sex: arcopallium, dorsolateral cortex, hippocampus, mesopallium, nidopallium, hyperpallium, and striatum (not including the globus pallidus) (Figure 4.9). Zoltán Molnár, Juan Montiel, Heidy Kaune, Amanda Cheung and Wei-Zhi Wang removed the brains from freshly decapitated heads of mixed-sex, organic, free-range chickens obtained from a slaughterhouse engaged in its normal operations (24 August 2009). Brains for *in situ* hybridisation were embedded in OCT and frozen on dry ice. Brains destined for immunohistochemistry were fixed in 4% paraformaldehyde. Brains for RNA extraction were rinsed in RNase-free PBS, submerged in ice-cold RNAlater (Ambion), transported to the Molnár laboratory and stored at -20°C in RNAlater (Ambion). Whole brains were embedded in 5% agarose and sectioned using a vibrating microtome (Leica, VT1000S) into 200 micrometre coronal sections using a 1:1 mixture of RNAlater and PBS. The samples were dissected out under visual guidance, using transillumination on a dissecting microscope (MZFLIII, Leica) and stored separately in RNAlater at -80°C until all microdissection was complete.

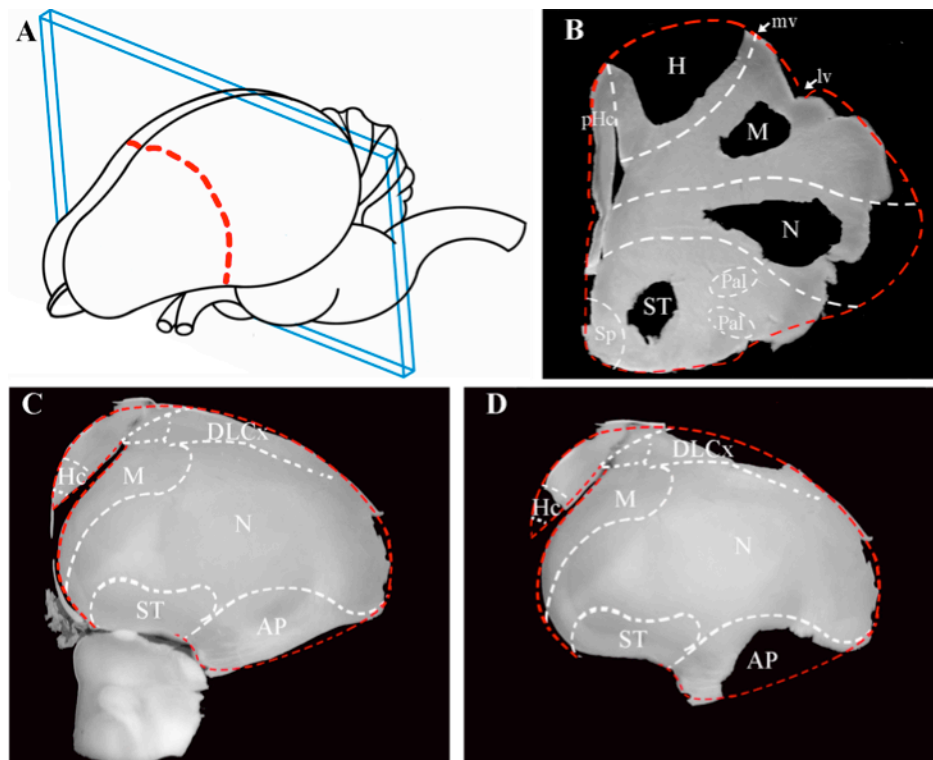


Figure 4.9: Chicken brains were coronally sectioned as depicted in (A). Representative sections from which hyperpallium (H), mesopallium (M), nidopallium (N), arcopallium (AP), striatum (ST), hippocampus (Hc) and dorsolateral cortex (DCLx) were dissected are shown in (B) and (D). (C) An undissected section with approximate boundaries marked. This figure is a reformatted version of a figure made by Dr. Juan Montiel, reproduced with permission.

For RNA extraction, samples from individual regions from the eighteen chickens and the eight mice, respectively, were combined and all tissue samples were processed concurrently. Dr. Tamara Sirey extracted total RNA using the RNeasy Lipid Tissue Mini kit (QIAGEN), following the manufacturer's instructions and using the on-column DNase digest. The silica column is reported by the manufacturer to deplete transcripts

smaller than 200 nt. RNA quantity was assessed using a NanoDrop 1000 spectrophotometer (ThermoScientific) (Table C.1; Table C.2), and, for the chicken samples, RNA quality and integrity assessed using a Bioanalyzer (Agilent Laboratories) (Table C.2).

Sequencing and gene expression quantification

Hatice Ozel Abaan, a technician in the NHGRI Genome Informatics Section, prepared the chicken samples for paired-end sequencing on Illumina's Genome Analyzer IIX by following the standard Illumina RNA-seq library preparation protocol with one round of poly(A) selection. Technicians at the NIH Intramural Sequencing Center did the same for mouse, but with two rounds of poly(A) selection. Both used the standard Illumina library preparation protocol described in Chapter 3.

Next, I filtered and mapped the reads. I computationally trimmed all reads to 50 nt for analysis and discarded three lanes with wildly aberrant GC content. The internal insert size and standard deviation were empirically estimated for each library as described in Chapter 3 (Table C.3). I required both read pairs to pass Illumina's chastity filter to be used in alignment. Using this insert size and standard deviation, each lane was separately mapped with tophat (Trapnell et al. 2009) v1.2.0 to the reference genome of mouse (NCBIM37, downloaded from Ensembl, masking the Y PAR but not simple repeats) or chicken (WASHUC2, downloaded from Ensembl, not masking simple repeats) as appropriate, using the --GTF option with corresponding GTF files from Ensembl release 61 (Table C.3) (ENSMUST00000127664 was manually removed from the mouse file as it is absurdly long and interferes with mapping and quantification).

The indel search was enabled for chicken, with 3 as the maximum insertion or deletion size, to allow more accurate quantification of genes containing indel polymorphisms. This was not necessary with the laboratory mice, since they all belonged to the C57BL/6 inbred strain that was used for the mouse genome. The minimum isoform fraction filter was disabled, as was the search for novel junctions.

For quantification purposes, and to test for potential confounds, I estimated the pre-mRNA fraction in each sample (Table C.4). To do this, I calculated the proportion of read fragments overlapping more than 3 nt of an intron for every Ensembl protein-coding gene with >200 nt exonic length having one and only one annotated transcript, and then took the median of these values for which there were at least 100 reads mapping in the locus. I used the same pre-mRNA fraction for the cufflinks quantification in each species: 10% for chicken quantifications and 5% for mouse quantifications, since those correspond to the highest values for each species rounded up to the nearest percent. The lower value in mouse may be attributable to the fact that the mouse samples underwent two rounds of poly(A) selection, while the chicken samples only underwent one.

To normalize across libraries within a species, and to equalize the variance for genes of similar expression levels, the total number of sequenced fragments was then downsampled as follows: for each protein-coding gene model (having at least one transcript with total exonic length >200 nt, as the experimental methods selected against these), the number of fragments overlapping any of its exons (by at least one base) were summed separately for each library. This included 16,734 and 22,670 protein-coding genes with a transcript >200 nt in chicken and mouse, respectively. Then, for each protein-coding gene model >200 nt total exonic length in each library, I calculated the

ratio of (fragments overlapping that model's exons in the library):(total number of fragments overlapping that model's exons across all libraries). Using only genes having at least one read in each library (14,109 in chicken, 16,329 in mouse), the median ratio across all protein-coding gene models >200 nt exonic length was found for each library. Finally, for each library, the total number of fragments overlapping an exon of a protein-coding transcript with >200 nt exonic length were summed, and I calculated the proportions of fragments within genes to the overall number of mapped fragments in each library (Table C.5). I then randomly selected read fragments without replacement to downsample each library to the following level: (total number of mapped fragments in the library)*(the minimum median of all samples)/(median ratio of that sample) (Table C.5).

Next I calculated FPKMs for all protein-coding genes (>200 nt exonic length) in each library using cufflinks (Trapnell et al. 2010) v0.9.3 with the `-GTF` option and corresponding GTF files from Ensembl release 61. Additionally, rRNAs, tRNAs and all mitochondrial transcripts (as annotated in Ensembl release 61) were explicitly masked from being used in the denominator of the normalization. The reference sequence was provided to cufflinks to improve quantification accuracy. Upper-quartile normalization was enabled to improve FPKM robustness when comparing across libraries.

Quality control for batch effects

To assess possible batch effects correlated with library, flow cell or lane (Leek et al. 2010; Marioni et al. 2008), I also calculated FPKM as described above but on a lane-by-lane basis. I then performed, separately for each species, principal component analysis

on FPKM values from all lanes in this project on all flow cells, including only genes that appeared in the top 4,000 most highly expressed in at least one library in that species. I used the R function `prcomp` on the transpose of the standard gene by sample matrix to cluster samples rather than genes, first centering the variables (samples) on zero and scaling the samples to have unit variance. For each species, I constructed biplots (using the R function `biplot`) covering every principal component that explained more than 2% of the variance in expression. For chicken, this consisted of the first six principal components, cumulatively explaining 96% of the variance across the 17 lanes (Table C.6). For mouse, this consisted of the first eight principal components, cumulatively explaining 86% of the variance across the 31 lanes (Table C.6). None of these principal components separated mainly resequenced libraries on different lanes, suggesting that flowcell and lane batch effects had no major contributions to the observed differences amongst these libraries (Figure C.1; Figure C.2).

In addition to looking at lane or flowcell effects (which implicitly takes into account the run date), I also looked explicitly in these principal components for batch effects at the level of libraries due to median pre-mRNA fraction (Table C.4), internal insert size (Table C.3), RIN (chick only; Table C.2), 260:280 nm absorbance ratios (Table C.1; Table C.2), 260:230 nm absorbance ratios (Table C.1; Table C.2), and the percentage of unique mapped fragments in protein coding genes (Table C.5). For this I performed a principal component analysis of the transpose of the gene by sample matrix considering specificities of genes in the downsampled libraries rather than the FPKMs of each individual lane using the R function `prcomp` (no scaling to unit variance; specificities were zero-centred for each sample). I only included genes that were used in the cross-

species comparison (see below). I then built biplots with the first nine components in mouse (explaining 96% of the variance; see Table C.7) and the first five components in chicken (explaining 97% of the variance; see Table C.7). I used both the library labels and the technical variable labels to identify differences in gene expression that could be driven by batch effects (Figure C.3; Figure C.4; Figure C.5; Figure C.6; Figure C.7; Figure C.8; Figure C.9; Figure C.10; Figure C.11). Some of these technical variables were correlated with the first principal component of both chicken and mouse (Figure C.12; Figure C.13; Figure C.14; Figure C.15), suggesting that RNA Integrity Number, pre-mRNA fraction, and the percentage of uniquely mapped fragments in protein-coding genes significantly affected the quantifications. In the absence of a well-validated method to control for continuous, as opposed to categorical (Johnson et al. 2007), confounds, and not having biological replicates, I took these batch effects into consideration when interpreting the results of subsequent analyses.

Since samples underwent poly(A) selection and had varying degrees of RNA quality, I also tested for any transcript 3' bias in read distribution. Note that this is affected both by the efficiency of the poly(A) selection process and by RNA degradation. I considered Ensembl-annotated protein-coding transcripts for which the isoform fraction (as identified by cufflinks) was greater than or equal to 95% and the FPKM was greater than 20. I measured read coverage at each position, starting from the 3' annotated end and working to the 5' annotated end. I normalized read coverage in each transcript such that, on average, coverage would equal one at each position of the transcript. Finally, I assessed the mean and standard deviation at each position having greater than 10 overlapping transcripts. This was done separately for each library (Figure C.16; Figure

C.17). I then looked for possible batch effects by plotting on the major principal components the maximum density within 2000 nt upstream of the 3' end, based on 9,024-9,488 transcripts in chicken and 6,373-7,900 transcripts in mouse (Figure C.18; Figure C.19). This revealed that the first principal component of the chicken samples and the first two principal components of the mouse samples were correlated with this proxy for RNA quality (Figure C.20; Figure C.21; Figure C.22). I propose that this measure of 3' bias reflects RNA quality since it is highly and significantly negatively correlated with RIN in chick ($r=-0.93$, 95% confidence interval $[-0.48, -0.99]$; two-tailed $p=0.0072$), and there is a ready theoretical explanation for why it should reflect RNA quality: the more degraded an RNA sample is, the more 3' bias should result from an equally efficient poly(A) selection. Differences in 3' bias may also be caused by differences in efficiency of the poly(A) selection process, but that is unlikely to be the sole explanation given the strong correlation with RIN observed in the chick samples.

To ensure that the final gene co-expression modules were not tainted by batch effects, I tested all resulting co-expressed gene sets for Pearson's correlations of their module eigengenes (see below) with these potential confounding technical variables, assessing significance using the online calculator at <http://faculty.vassar.edu/lowry/ch4apx.html>. I used the online calculator at <http://faculty.vassar.edu/lowry/rdiff.html> to assess the significance of the difference between two correlation coefficients.

Determining correspondence of the mouse cortical layer dissections

To determine the true correspondence of laminarily dissected samples to known cytoarchitectural layers, I created the heatmaps in Figure 4.10 as follows: if genes g_i are

annotated (Lein et al. 2007) as being preferentially *and specifically* expressed in layer l and f_{sgl} is the fractional expression of gene g_l in sample s relative to expression in all samples, then t_{sl} is the median of (f_{sgl}) over genes g_l . The heat map intensities represent $[t_{sl} - \text{mean}_{\text{all layers } l}(t_{sl})] / [\text{mean}_{\text{all layers } l}(t_{sl})]$.

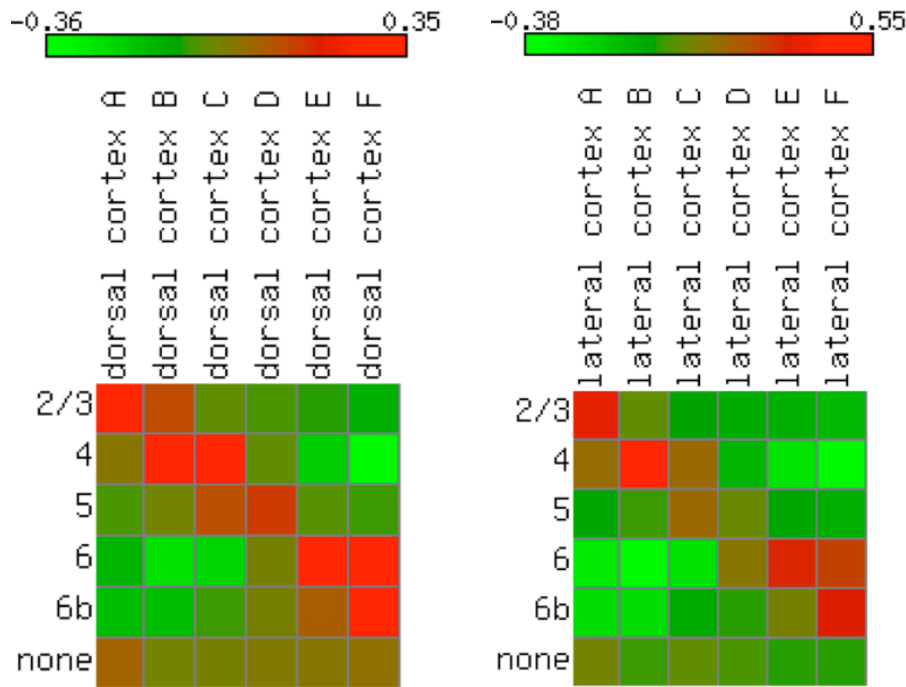


Figure 4.10: Dissected samples correspond to specific neocortical layers. Heatmaps were created using Matrix2png (<http://www.bioinformatics.ubc.ca/matrix2png/>) and reflect the calculation described above.

Confirming accuracy of the mouse dissections in other regions

To confirm the accuracy of the non-cortical dissections in the mouse, I manually reviewed images series of *in situ* hybridisations, where present, from the Allen Mouse Brain Atlas (Lein et al. 2007) of the three genes most specific to each non-cortical library amongst genes that were also used for the cross species comparisons (see below). I

defined specificity as the fraction of total expression across all libraries that occurred in the library of interest. Figure 4.11, Figure 4.12, Figure 4.13 and Figure 4.14 display an example from the top three most specific genes of each non-cortical sample, which were all consistent with the intended dissections.

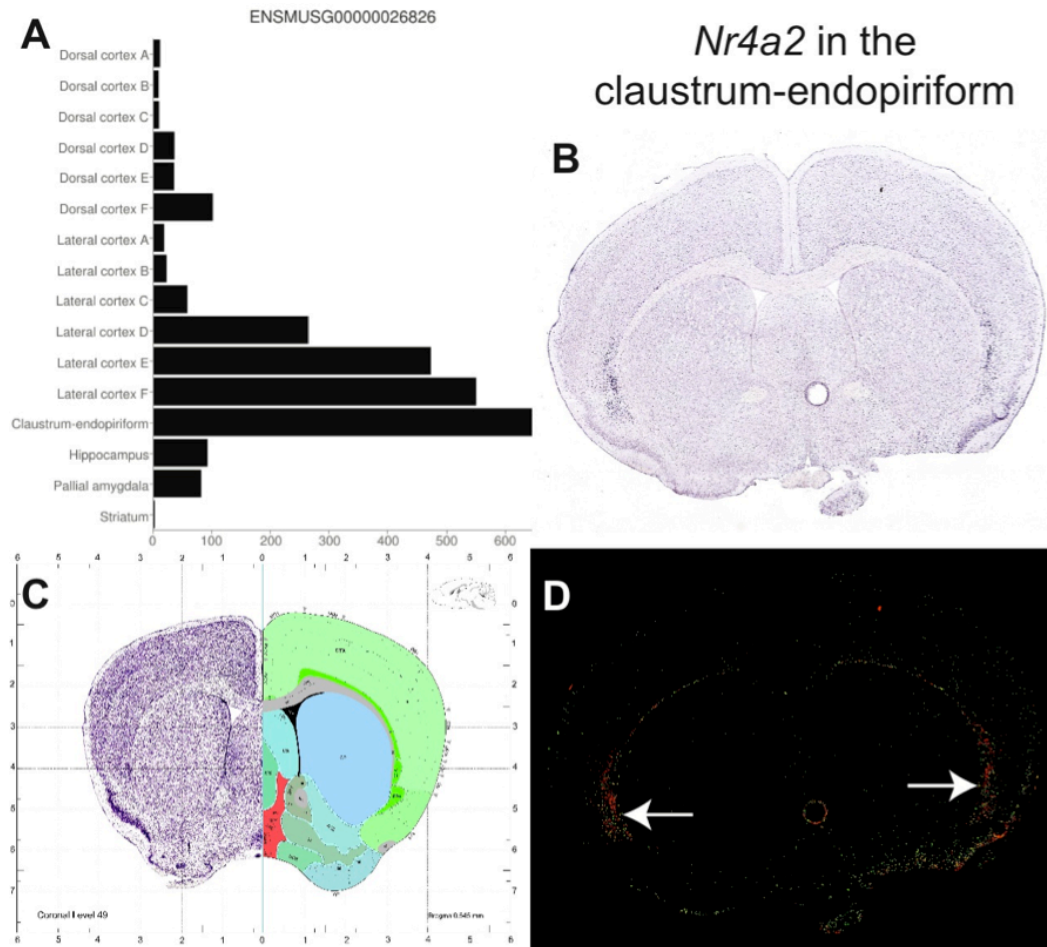


Figure 4.11: *Nr4a2* (also known as *Nurr1*), one of the three most specific marker genes in assessed regions for the claustrum-endopiriform as assessed by RNA-seq, was also specific to this region (among regions assessed) by *in situ* hybridisation. (A) gives the FPKM values of the gene in each dissected mouse region as assessed by RNA-seq, (B) is an *in situ* hybridisation image of the gene, (C) gives the corresponding coronal section of the reference atlas and (D) gives the expression analysis of the *in situ* hybridisation image with the region of interest indicated with arrows. (B)-(D) are reproduced from the Allen Mouse Brain Atlas.

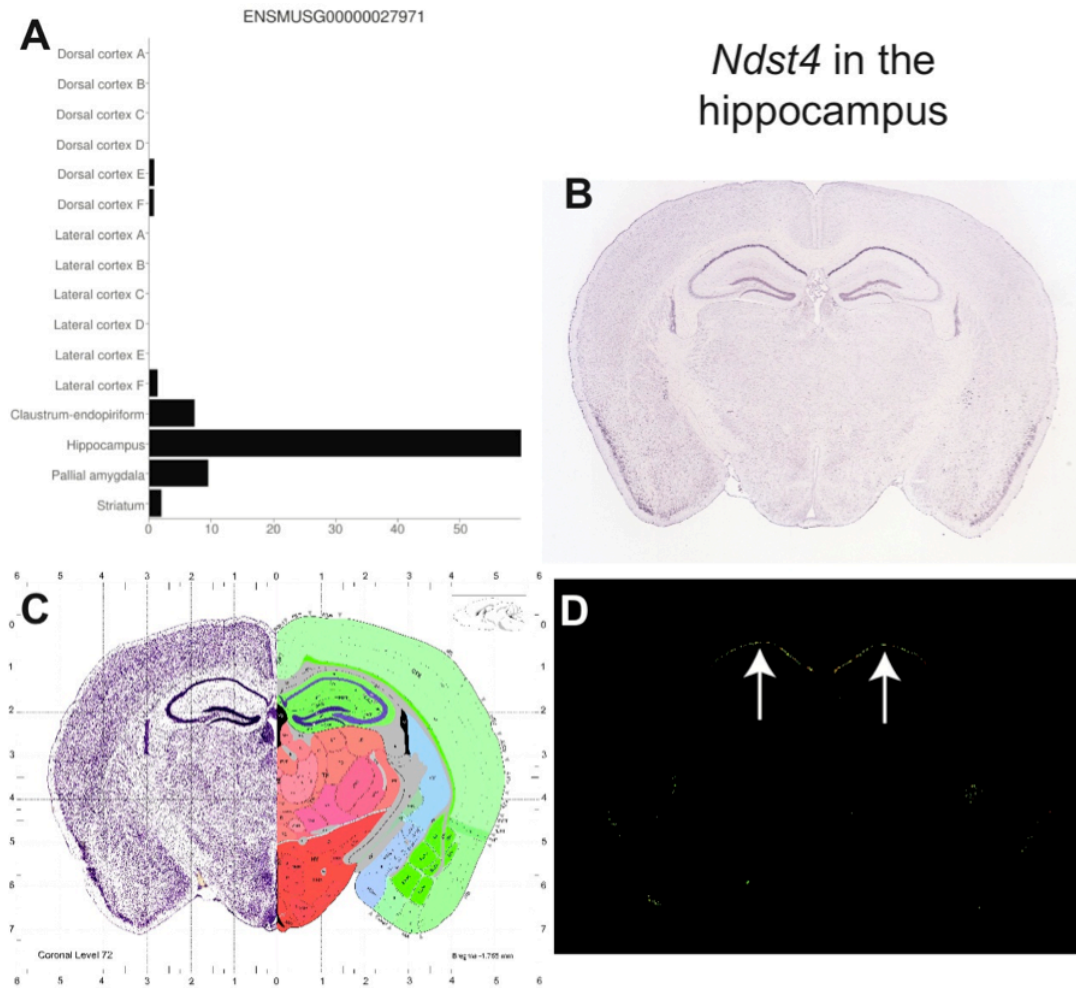


Figure 4.12: *Ndst4*, one of the three most specific marker genes in assessed regions for the hippocampus as assessed by RNA-seq, was also specific to this region (among regions assessed) by *in situ* hybridisation. (A) gives the FPKM values of the gene in each dissected mouse region as assessed by RNA-seq, (B) is an *in situ* hybridisation image of the gene, (C) gives the corresponding coronal section of the reference atlas and (D) gives the expression analysis of the *in situ* hybridisation image with the region of interest indicated with arrows. (B)-(D) are reproduced from the Allen Mouse Brain Atlas.

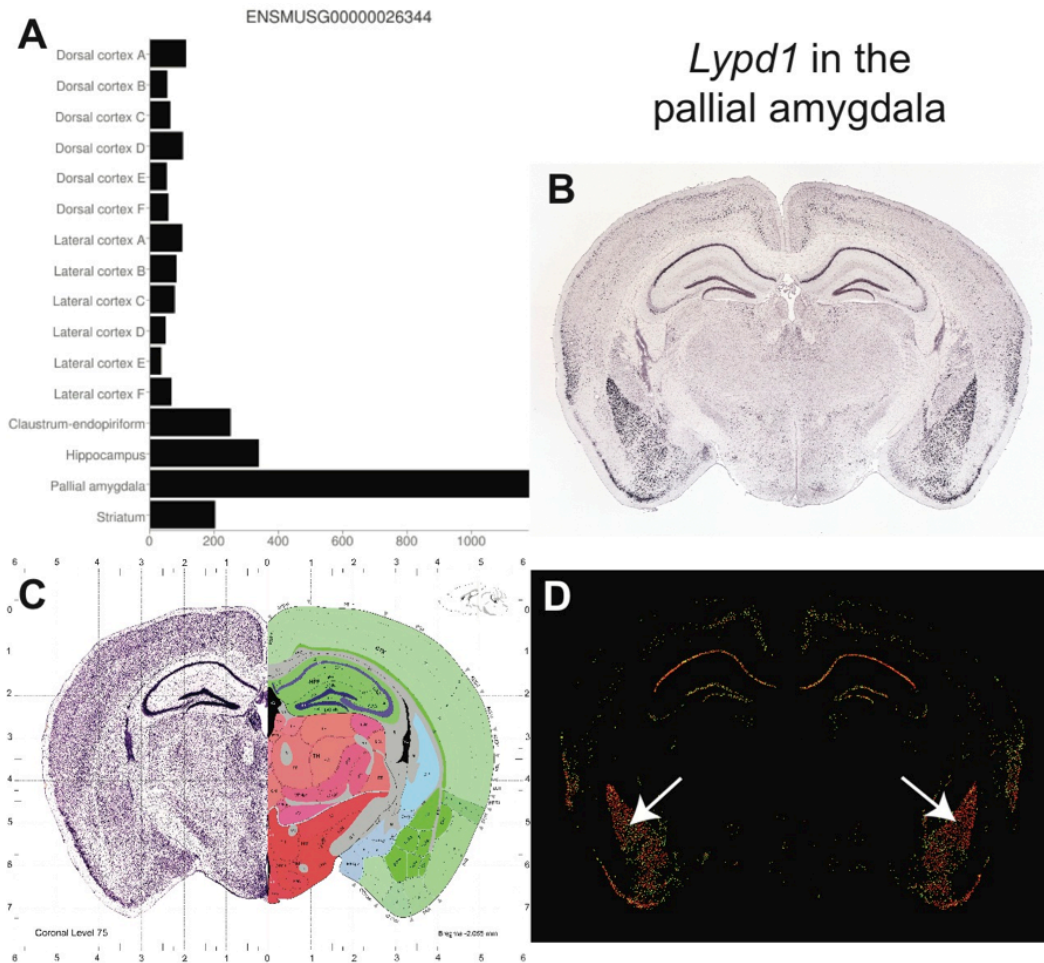


Figure 4.13: *Lypd1*, one of the three most specific marker genes in assessed regions for the pallial amygdala as assessed by RNA-seq, was also specific to this region (among regions assessed) by *in situ* hybridisation. (A) gives the FPKM values of the gene in each dissected mouse region as assessed by RNA-seq, (B) is an *in situ* hybridisation image of the gene, (C) gives the corresponding coronal section of the reference atlas and (D) gives the expression analysis of the *in situ* hybridisation image with the region of interest indicated with arrows. (B)-(D) are reproduced from the Allen Mouse Brain Atlas.

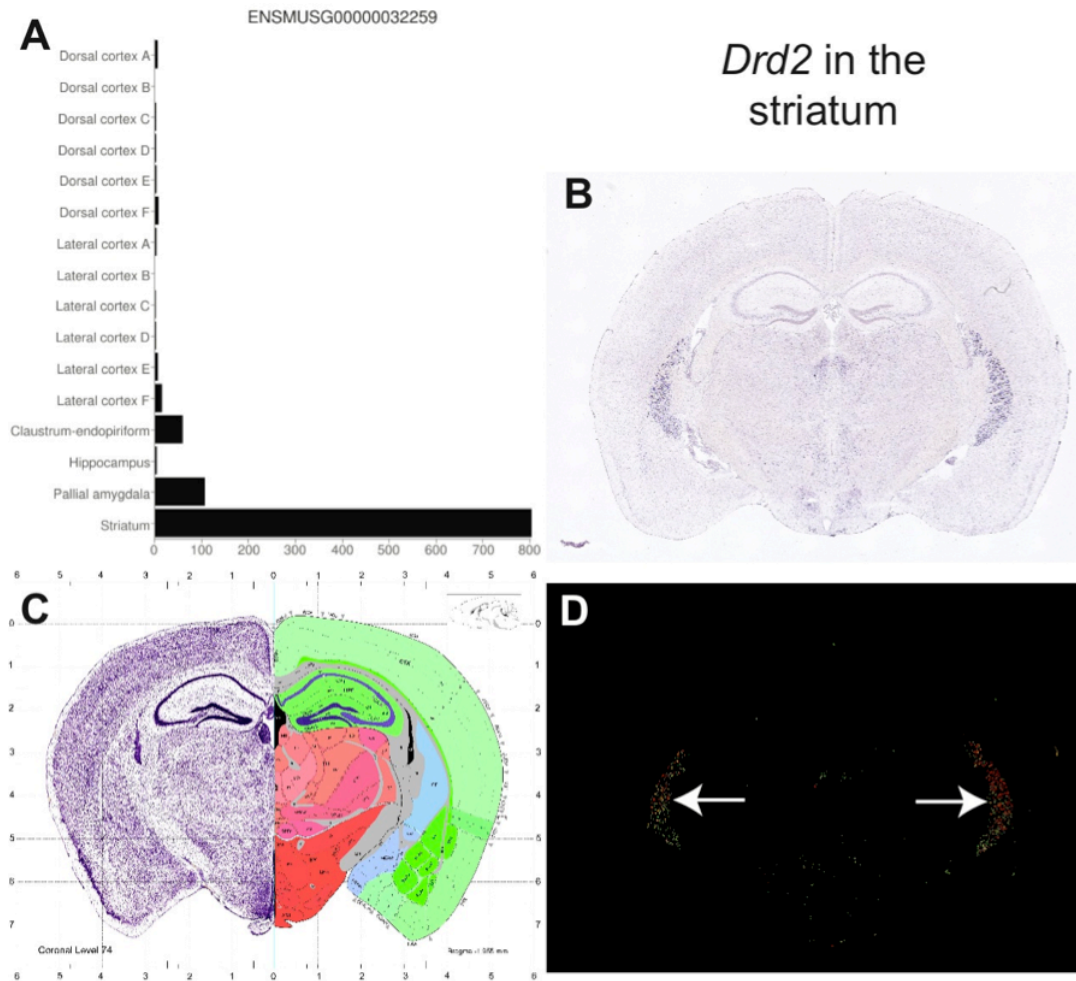
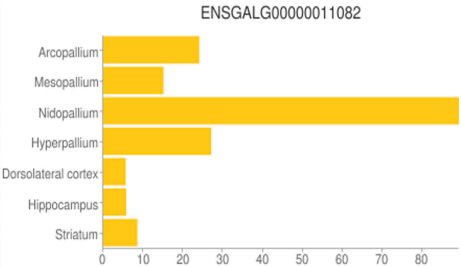
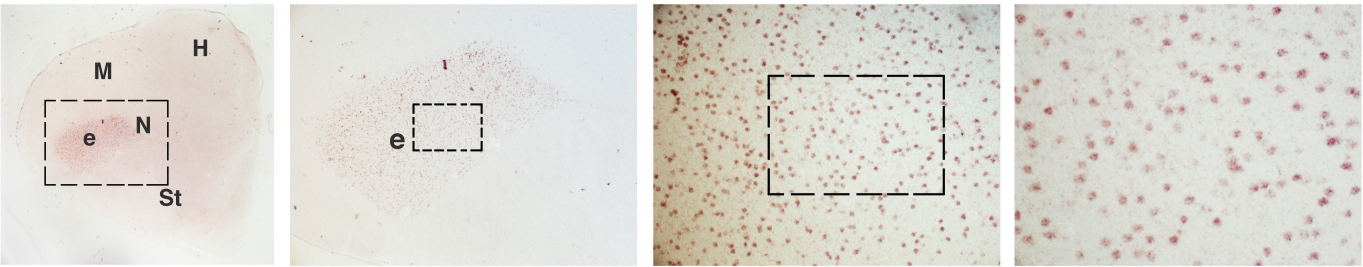


Figure 4.14: *Drd2*, one of the three most specific marker genes in assessed regions for the striatum as assessed by RNA-seq, was also specific to this region (among regions assessed) by *in situ* hybridisation. (A) gives the FPKM values of the gene in each dissected mouse region as assessed by RNA-seq, (B) is an *in situ* hybridisation image of the gene, (C) gives the corresponding coronal section of the reference atlas and (D) gives the expression analysis of the *in situ* hybridisation image with the region of interest indicated with arrows. (B)-(D) are reproduced from the Allen Mouse Brain Atlas.

Confirming accuracy of the chicken dissections

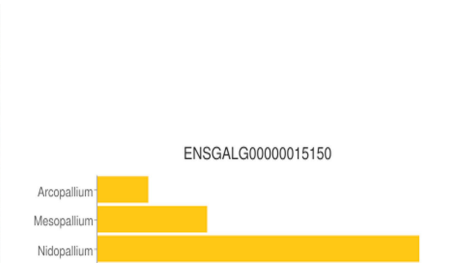
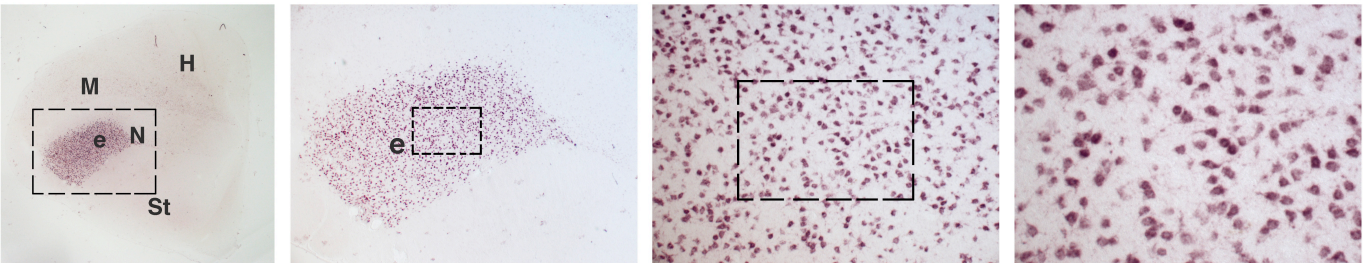
To confirm the accuracy of the chicken dissections, I asked Dr. Wei-Zhi Wang to perform *in situ* hybridisations in 20 µm coronal sections of adult chicken brains for the most specific genes to three chicken regions dissected (mesopallium, nidopallium and arcopallium), using a similar digoxigenin-based method to that described in Section 2.6 above (specifics are given in Wang et al. (2011)). The primer sequences used are provided in Table C.8. Further *in situ* hybridisations are presently ongoing for genes specific to the other regions.

Kcnh7

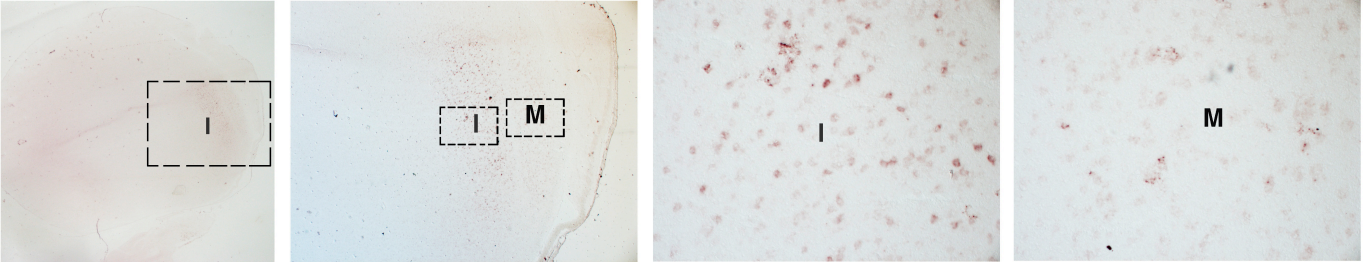


Rorβ

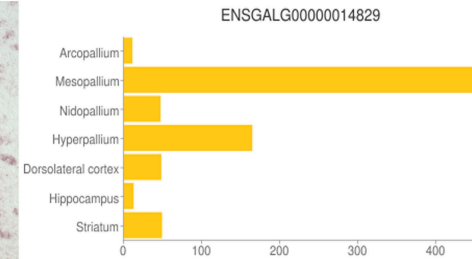
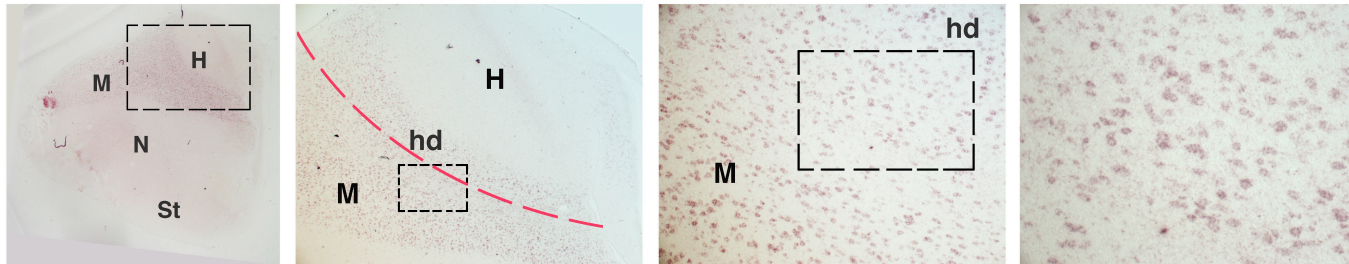
ANTERIOR



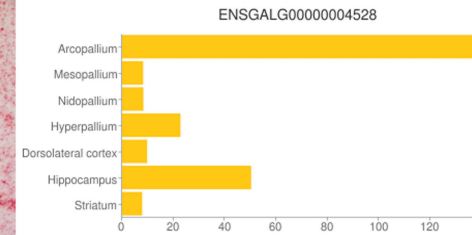
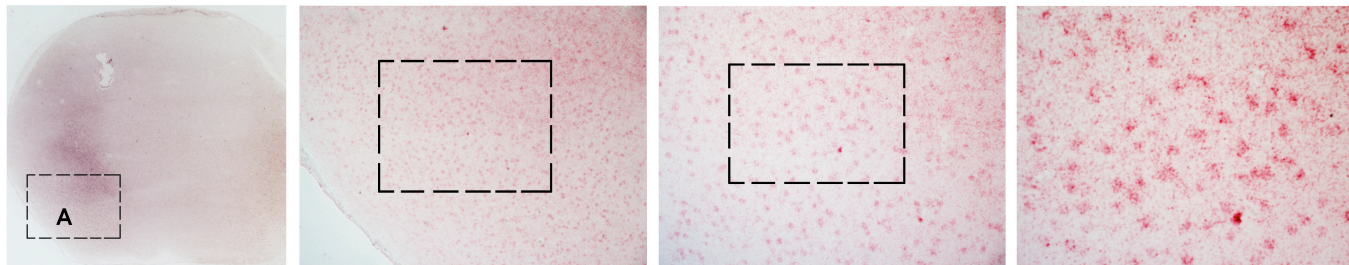
POSTERIOR



Rspo3



Tshz3



Mgat4c

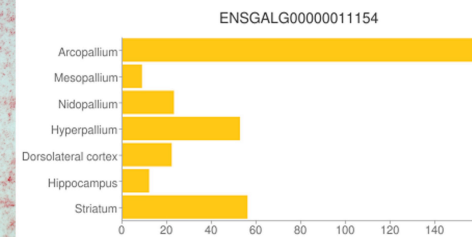
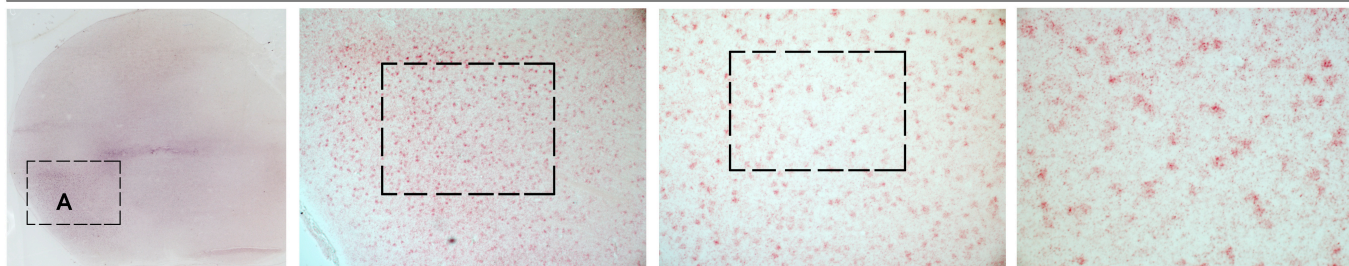


Figure 4.15: *In situ* hybridisations confirming two genes specific to chick nidopallium (*Kcnh7* and *Rorb*), one gene specific to chick mesopallium (*Rspo3*) and two genes specific to chick arcopallium (*Tshz3* and *Mgat4c*). Figures showing FPKMs in each chick sample are given at right. Figure prepared by Dr. Juan Montiel based on images provided by Dr. Wei Zhi Wang and myself. H, hyperpallium; M, mesopallium; N, nidopallium, A, arcopallium; St, striatum; e, entopallium; l, field L; hd, hyperpallium densocellulare.

Selecting genes for the cross-species comparison

For the evolutionary comparisons, I only considered one-to-one protein-coding orthologs with >200 nt exonic length in both mouse and chicken as identified by Ensembl (release 61) (Flicek et al. 2011). This comprised 11,033 genes in total. I then required a gene to be expressed in the top 5,000 genes in at least one library in one species and to have a non-zero variance in both species which left 5,170 genes ranging in expression from 115 FPKM to 106,823 FPKM or greater in at least one library.

Identifying gene co-expression modules

I next identified gene co-expression modules using weighted gene co-expression network analysis (Horvath et al. 2006; Langfelder and Horvath 2008; Zhang and Horvath 2005) with the set of highly expressed one-to-one orthologs described above. To identify the appropriate soft threshold (power) for each species in turn, I plotted the R^2 of the scale-free topology model fit against the soft threshold for powers 1-10, 12, 14, 16, 18, 20, 22, 24, 26, 28, and 30 (Table C.9; Table C.10; Figure C.23; Figure C.24). I chose 10 for mouse and 20 for chicken for the following reasons: the R^2 was high (>0.8) in both species so they both approximated a scale-free topology while simultaneously producing networks having high mean connectivity, they both had negative slopes (without which the network has no clear biological interpretation), and they produced the most preserved modules (number of modules with combined $Z_{\text{preservation}} > 3$) when comparing between the two species for several combinations (these comparisons are described below; Table C.11).

Once the soft threshold was chosen for each species, I determined the general network properties of the two sets using the softConnectivity R function (Zhang and Horvath 2005), comparing both ranked levels of expression and ranked connectivity of orthologous genes in the two species with the verboseScatterplot R function (Figure 4.16). This demonstrated that, while there was significant conservation between mouse and chicken at the level of expression ($p < 10^{-200}$), the position of a gene in a network (as a ‘hub’ or at the periphery) was not generally conserved.

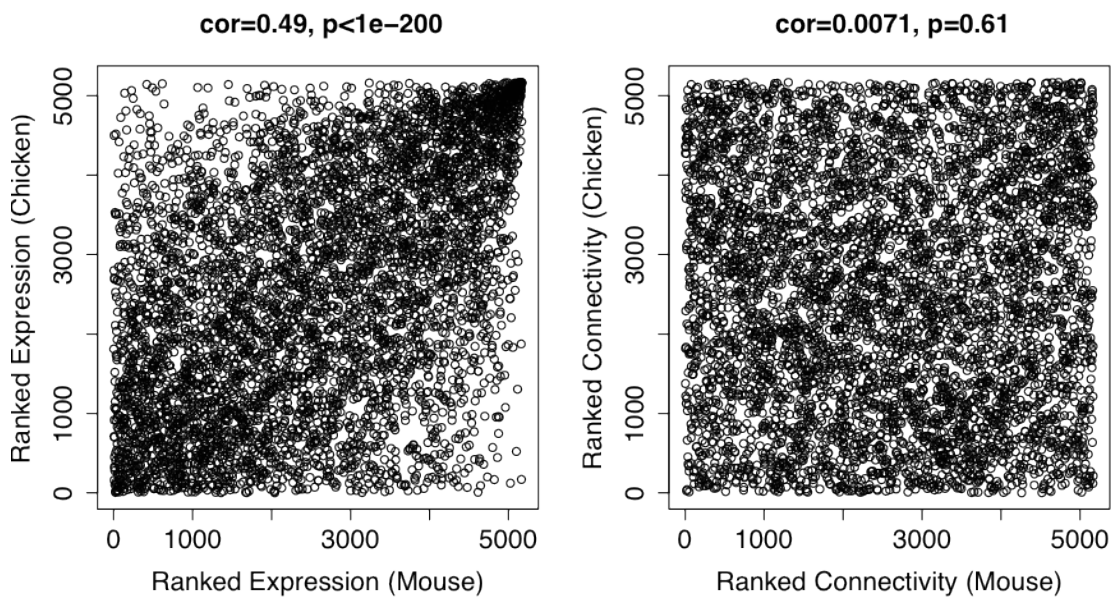


Figure 4.16: Plots of the ranked expression of orthologs in both species, and of the ranked connectivity. Correlation coefficients above represent Spearman’s rank correlation coefficient of the original variables (Pearson’s *rho* of the ranked variables) and *p*-values are two-tailed.

I built networks as described above. Briefly, I computed an adjacency matrix whose entries a_{ij} correspond to $|0.5 + \text{cor}(x_i, x_j)/2|^\beta$ where β is the soft threshold chosen above and

cor is Pearson's *rho*. The weighted adjacency between two genes i and j is thus proportional to their correlation on a logarithmic scale, and includes information on positive or negative correlation. I then set the diagonal of the adjacency matrix to zero and used it to calculate the topological overlap matrix using the signed TOMsimilarity R function. I subtracted the resulting topological overlap matrix from a matrix of ones and hierarchically clustered the result (with the flashClust R function and the average agglomeration method) to produce the dendrograms in Figure C.25 and Figure C.26.

Next, I defined modules using the dynamic tree cutting algorithm as implemented in the Dynamic Tree Cut R package (Langfelder et al. 2008) to produce five sets of modules of different sizes for both chicken and mouse (Figure C.25; Figure C.26). I used a cut height of 0.99 and a minimum cluster size of $30-3*ds$ where ds is an integer in the closed set $[0,4]$. I chose the third split in chicken and the fourth split in mouse for the same reasons the soft threshold was chosen.

I then determined the “module eigengene” for each module (Langfelder and Horvath 2007). This corresponds to the first principal component of the expression matrix of genes in the module, and is thus similar to a sort of weighted average expression profile for the module of co-expressed genes. These were hierarchically clustered to visualize similarities of different modules (Figure C.27; Figure C.28).

Comparing gene co-expression modules between species

To visualize how well individual modules were preserved between species, I projected the colours of the final mouse modules onto the chicken dendrogram (Figure 4.17) and

the colours of the final chicken modules onto the mouse dendrogram (Figure 4.18) using the R function `plotDendroAndColors`.

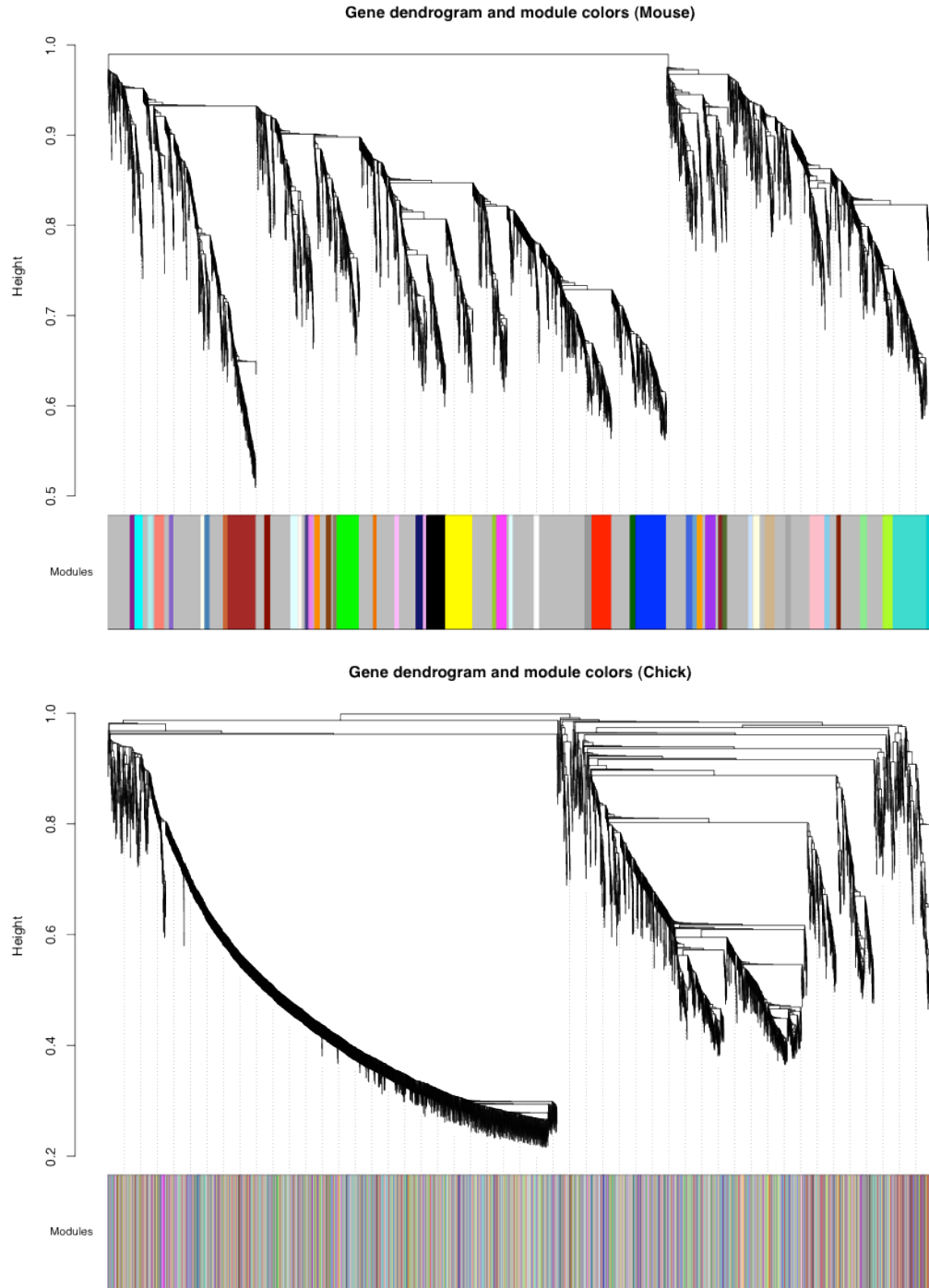


Figure 4.17: Mouse modules and dendrogram of topological dissimilarity (top) and mouse modules projected onto the corresponding chicken dendrogram (bottom).

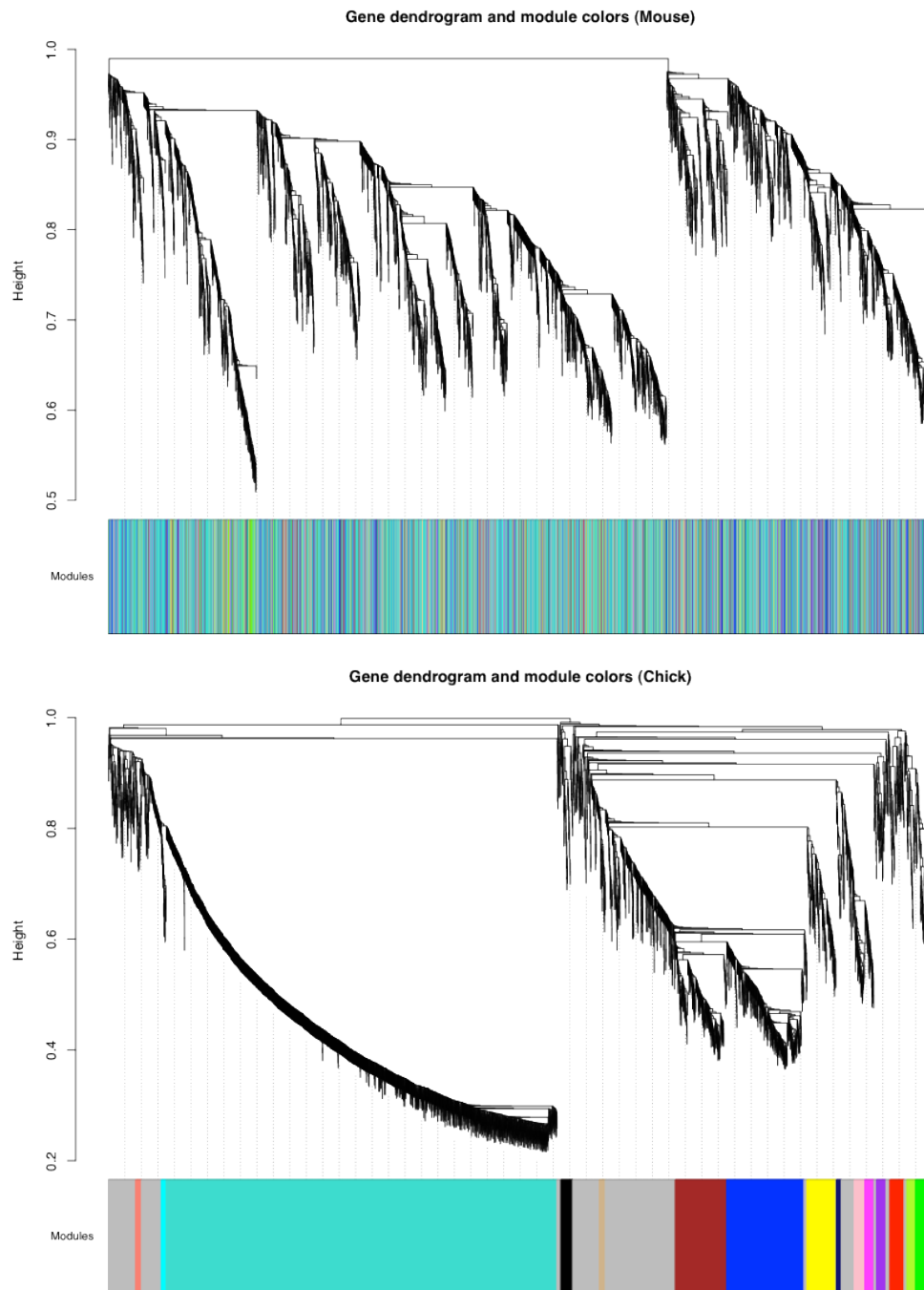


Figure 4.18: Chicken modules and dendrogram of topological dissimilarity (bottom) and chicken modules projected onto the corresponding mouse dendrogram (bottom).

To quantify how well individual modules were preserved between species, I calculated several module preservation statistics using the `modulePreservation` R function (Langfelder et al. 2011) with the following parameters: “signed” network, 30 permutations, max gold module (module of random genes) size of 30 and max module size of 400 (Table 4.2; Table 4.3). The grey module consists of genes that were not assigned to a coexpressed module. These statistics fall into two broad categories: density statistics and connectivity preservation statistics, which are averaged ($[Z_{density.pres} + Z_{connectivity.pres}]/2$) into one final Z score of cross-species module preservation $Z_{summary.pres}$. Modules with $Z_{summary.pres}$ scores between 2 and 10 are generally considered weakly to moderately preserved (Langfelder et al. 2011).

Module density statistics, summarized by $Z_{density.pres}$, reflect how densely connected genes in the module of a reference network are in the test data. This summary statistic is defined as the median of several module density metrics: $meanSignAwareCorDat$ (the mean correlation density of the module q , calculated as $mean\{vectorizeMatrix(sign(r_{ij}^{[ref](q)})r_{ij}^{[test](q)})\}$); $propVarExplained$ (proportion of the variance explained by the module eigengene in the test data, calculated as the mean squared kME); $meanSignAwareKME$ (measures the mean module membership where genes whose module memberships in the reference and test networks have opposite signs contribute negatively; is calculated as $mean_{i \in M_q}\{sign(kME_i^{[ref][q]})kME_i^{[test](q)}\}$); and $meanAdj$ (mean adjacency of genes in the module, calculated as $mean(vectorizeMatrix(A^{[test](q)}))$). Other density statistics not used to create the summary statistic, but which may nevertheless be informative, are $meanMAR$ (mean

Maximum Adjacency Ratio, $mean\left(MAR_i^{[test]} = \frac{\sum_{j \neq i} a_{ij}^2}{\sum_{j \neq i} a_{ij}}\right)$, of genes i in the module) and

$meanClusterCoeff$ (mean clustering coefficient of genes i in the module where

$$mean\left(clusterCoef_i^{[test]} = \frac{\sum_{j \neq i} \sum_{m \neq j, i} a_{ij} a_{jm} a_{mi}}{\left(\sum_{j \neq i} a_{ij}\right)^2 - \sum_{j \neq i} a_{ij}^2}\right).$$

Connectivity pattern preservation statistics, summarized by *Zconnectivity.pres*, reflect how well the co-expression connections within a module are preserved. This summary statistic is defined as the median of several connectivity pattern metrics: *cor.cor* (the correlation between pairwise correlations of expression of module genes); *cor.kME* (the correlation of correlations of genes in the module to the module eigengene in the test set); and *cor.kIM* (the correlation of intramodular connectivity for module q

$kIM_i = \sum_{\substack{j \in M_q \\ j \neq i}} a_{ij}^q$, which quantifies if hub genes i of module q in the reference network are

also hub genes of module q in the test network). Other connectivity statistics not used to create the summary statistic, but which may nevertheless be informative, are *cor.kMEall* (the correlation of correlations of all genes to the module eigengene in the test set);

cor.MAR (correlation of Maximum Adjacency Ratios, $MAR_i = \frac{\sum_{j \neq i} a_{ij}^2}{\sum_{j \neq i} a_{ij}}$, for genes i in the

module); and *cor.clusterCoeff* (correlation of the clustering coefficients,

$$clusterCoef_i = \frac{\sum_{j \neq i} \sum_{m \neq j, i} a_{ij} a_{jm} a_{mi}}{\left(\sum_{j \neq i} a_{ij}\right)^2 - \sum_{j \neq i} a_{ij}^2}, \text{ for genes } i \text{ in the module}).$$

Table 4.2: Preservation statistics of mouse modules in chicken sorted by overall preservation.

module name	size	Zsum mary. pres	Zdens ity.pr es	Zcon nectiv ity.pr es	Z.pro pVar Expla ined.p res	Z.me anSig nAwa reKM E.pre s	Z.me anSig nAwa reCor Dat.p res	Z.me anAdj .pres	Z.me anClu sterC oeff.p res	Z.me anM AR.pr es	Z.cor. kIM	Z.cor. kME	Z.cor. kMEa ll	Z.cor. cor	Z.cor. cluste rCoef f	Z.cor. MAR
salmon	61	3.42	5.78	1.05	3.26	4.80	10.13	6.77	4.17	3.84	1.05	-0.32	11.40	2.38	0.46	0.30
steelblue	31	3.23	4.18	2.28	1.04	4.15	6.75	4.21	1.61	1.19	1.54	2.28	14.05	3.39	2.08	1.46
yellow	170	1.99	3.53	0.45	1.61	3.20	4.68	3.85	2.92	2.64	0.45	-1.13	16.75	1.05	0.81	-0.19
grey	400	1.88	1.40	2.36	0.16	1.91	5.18	0.90	1.24	1.02	1.50	2.36	15.68	4.28	2.34	1.13
lightcyan1	26	1.29	4.00	-1.42	2.81	3.13	4.86	4.96	3.04	3.05	-1.42	1.70	15.56	-3.32	-1.15	-1.01
tan	62	1.27	2.65	-0.10	2.76	1.78	2.53	4.50	3.86	4.18	-1.23	-0.10	4.80	0.39	-1.20	-0.68
red	123	1.04	2.06	0.03	2.07	1.42	2.05	2.10	1.55	1.66	0.34	-0.46	15.95	0.03	0.69	0.23
grey60	45	0.93	1.54	0.32	1.48	1.56	1.52	1.80	0.93	1.01	0.32	-0.62	16.25	0.50	0.10	-0.58
white	32	0.91	3.78	-1.95	1.36	3.78	6.74	3.77	0.94	0.84	-1.95	1.16	14.30	-2.62	-0.35	-1.77
plum1	27	0.90	1.78	0.02	0.02	2.36	2.53	1.19	0.88	0.79	0.02	-0.66	16.33	0.25	0.63	-0.12
plum2	21	0.88	2.56	-0.81	1.37	2.78	4.24	2.34	0.66	0.63	-0.81	1.17	16.94	-1.80	-1.47	-0.99
turquoise	212	0.81	-0.32	1.94	1.22	-1.12	-0.85	0.21	0.81	0.37	2.46	1.33	14.31	1.94	-0.25	2.00
blue	191	0.81	2.74	-1.12	1.08	2.90	4.33	2.58	1.16	1.23	-1.12	0.42	13.48	-1.90	0.18	-0.83
darkmagenta	29	0.74	0.37	1.10	0.82	0.40	-0.07	0.35	0.09	0.21	-0.11	1.54	7.99	1.10	-1.36	-0.35
lightgreen	44	0.71	1.33	0.08	2.19	0.47	0.37	3.14	2.69	2.40	0.62	-0.64	10.17	0.08	-0.34	0.00
skyblue3	28	0.64	0.27	1.01	0.15	0.46	0.39	-0.26	-0.87	-0.69	1.01	0.84	8.52	1.03	1.88	0.27
green	140	0.61	0.54	0.69	-0.39	1.24	2.17	-0.17	-1.17	-1.38	0.15	0.69	7.50	1.37	-1.70	-0.45
cyan	50	0.58	3.41	-2.26	1.12	3.02	3.80	4.02	1.78	1.80	-2.26	1.90	10.81	-2.81	-2.24	-2.34
midnightblue	47	0.52	1.66	-0.61	0.05	1.75	2.00	1.58	0.25	-0.32	-1.02	-0.61	17.30	1.65	1.20	-1.04
darkgreen	37	0.47	3.49	-2.55	1.39	3.66	6.07	3.33	1.06	0.97	-2.55	3.28	14.93	-4.67	-2.26	-1.88
darkorange	33	0.33	1.08	-0.41	0.93	1.18	1.24	0.97	0.22	0.21	-0.41	-1.15	10.70	1.55	0.37	0.38
darkgrey	34	0.31	-0.47	1.09	0.08	-1.70	-0.92	-0.03	0.36	0.41	1.09	-1.29	6.02	1.09	0.36	0.90
lightsteelblue1	27	0.26	0.76	-0.23	0.65	0.80	0.72	0.93	1.79	1.66	0.03	-1.36	6.69	-0.23	-1.91	-1.64

module name	size	Zsummary. pres	Zdensity.pr es	Zconnectiv ity.pr es	Z.prop Var Expla ined.p res	Z.meanSig nAwareKM E.pre s	Z.meanSig nAwareCor Dat.p res	Z.me anAdj .pres	Z.me anClu sterC oeff.p res	Z.me anM AR.pr es	Z.cor. kIM	Z.cor. kME	Z.cor. kMEa ll	Z.cor. cor	Z.cor. cluste rCoef f	Z.cor. MAR
ivory	25	0.15	0.55	-0.25	1.06	0.17	-0.40	0.92	1.06	1.02	-0.25	-0.07	15.79	-0.30	-0.49	-0.49
sienna3	28	0.14	-0.65	0.92	-3.47	0.56	1.20	-1.86	-2.37	-2.22	0.90	0.92	14.00	2.33	1.53	1.32
skyblue	32	0.10	0.61	-0.41	1.25	0.01	0.53	0.69	0.67	0.83	-1.44	-0.41	16.01	0.63	-2.79	-2.40
floralwhite	24	0.09	-0.85	1.03	-1.43	-0.20	-0.27	-1.50	-1.82	-1.66	1.04	0.47	7.83	1.03	-0.17	0.21
greenyellow	62	0.04	-0.18	0.25	-1.31	0.46	1.06	-0.82	-0.58	-0.46	-0.18	0.25	14.84	0.88	0.43	-0.98
magenta	64	0.00	0.83	-0.84	-1.12	1.54	3.88	0.11	-0.95	-0.21	-0.84	-0.86	14.15	2.43	-1.94	-0.86
paleturquoise	31	-0.04	-0.44	0.36	-0.63	0.12	-0.31	-0.57	-0.83	-0.80	0.59	-1.53	10.33	0.36	0.45	1.21
lightyellow	44	-0.04	-0.48	0.41	0.39	-0.91	-0.54	-0.43	-0.35	-0.07	-1.35	1.44	3.40	0.41	-1.95	-1.43
bisque4	23	-0.05	-0.85	0.75	-0.87	-1.57	0.66	-0.83	-0.99	-0.76	1.03	-1.09	7.53	0.75	1.10	1.34
yellowgreen	28	-0.07	0.74	-0.87	-0.61	1.62	1.76	-0.14	-0.43	-0.56	-0.21	-0.87	14.27	-1.25	0.67	0.24
gold	100	-0.08	-0.45	0.30	0.28	-0.71	-0.42	-0.49	-0.61	-0.61	0.49	0.30	15.40	0.17	-0.08	0.04
darkolivegreen	30	-0.11	-1.05	0.83	-1.64	-0.85	-0.15	-1.25	-0.92	-0.93	-0.12	2.27	3.73	0.83	0.14	-0.40
mediumpurple3	27	-0.17	-0.06	-0.27	-0.24	0.13	1.10	-0.44	-0.64	-0.60	-0.28	-0.27	5.25	1.22	1.26	0.35
orangered4	27	-0.18	0.04	-0.39	-0.27	0.45	0.35	-0.64	-1.09	-0.75	-0.39	0.42	8.68	-0.40	-1.70	-1.09
saddlebrown	32	-0.20	-1.34	0.95	-1.94	-1.00	-0.79	-1.68	-1.76	-1.82	0.44	1.15	4.49	0.95	0.36	0.80
pink	94	-0.24	0.51	-0.99	0.52	0.50	0.79	0.26	-0.26	-0.16	-1.58	-0.99	14.38	2.13	-0.59	-1.83
darkturquoise	36	-0.30	0.59	-1.19	1.17	0.01	-0.29	1.30	0.75	0.94	-2.20	0.88	9.38	-1.19	-0.47	-1.68
darkred	40	-0.33	-0.07	-0.59	-0.10	-0.56	0.12	-0.03	0.36	0.59	-1.03	-0.59	9.58	1.73	0.15	-0.59
darkorange2	24	-0.38	-0.70	-0.06	-3.06	0.82	0.61	-2.02	-2.64	-2.48	-0.06	-0.97	9.03	0.79	-0.23	-0.48
royalblue	42	-0.39	-0.28	-0.50	1.12	-0.58	-0.52	-0.04	-0.69	-0.64	-0.81	-0.50	-0.15	0.46	-0.68	-1.53
brown4	24	-0.51	-0.44	-0.59	-1.30	0.21	0.45	-1.09	-1.38	-1.34	-0.59	-1.02	6.05	0.37	-0.10	-0.61
orange	33	-0.57	-1.30	0.16	-3.64	-0.47	-0.26	-2.13	-2.09	-2.06	0.16	-0.91	9.39	0.23	-1.04	-0.48
violet	30	-0.76	-1.47	-0.05	-3.04	-0.92	-0.51	-2.01	-3.04	-2.79	0.06	-0.05	8.95	-0.71	0.24	-0.19
purple	64	-0.84	-0.94	-0.75	-0.29	-1.24	-0.80	-1.08	-1.10	-1.44	-1.53	-0.75	13.78	1.03	-0.97	-1.31
lightcyan	47	-0.91	-1.33	-0.48	-5.28	0.17	0.73	-2.84	-2.38	-1.82	-0.48	-1.52	1.74	1.56	-1.52	-2.21
darkslateblue	22	-0.97	-1.13	-0.82	-1.88	-0.73	-0.28	-1.53	-1.53	-1.56	-1.16	-0.82	6.12	0.28	-0.31	-1.14

module name	size	Zsummary. pres	Zdensity.pr es	Zconnectiv ity.pr es	Z.propVar Explained.p res	Z.meanSig nAwareKM E.pres	Z.meanSig nAwareCor Dat.p res	Z.meanAdj .pres	Z.meanClu sterC oeff.p res	Z.meanM AR.p res	Z.cor. kIM	Z.cor. kME	Z.cor. kMEa ll	Z.cor. cor	Z.cor. cluste rCoef f	Z.cor. MAR
black	118	-1.19	-1.78	-0.61	-3.31	-1.31	-1.06	-2.24	-2.14	-1.60	-2.47	0.29	17.35	-0.61	-1.18	-2.24
brown	177	-1.26	-4.38	1.85	-9.80	-2.86	3.37	-5.89	-7.25	-5.48	1.85	-0.36	9.99	4.99	-0.29	0.77

Table 4.3: Preservation statistics of chicken modules in mouse sorted by overall preservation.

module name	size	Zsummary.pres	Zdensity.pr es	Zconnectiv ity.pr es	Z.pro pVar Expla ined.p res	Z.me anSig nAwa reKM E.pre s	Z.me anSig nAwa reCor Dat.p res	Z.me anAdj .pres	Z.me anClu sterC oeff.p res	Z.me anM AR.pr es	Z.cor. kIM	Z.cor. kME	Z.cor. kMEa ll	Z.cor. cor	Z.cor. cluste rCoef f	Z.cor. MAR
green	107	3.33	4.56	2.10	3.45	3.61	5.51	14.08	12.26	9.75	2.10	1.46	8.29	2.64	1.37	2.62
greenyellow	54	3.33	3.86	2.80	2.84	3.42	4.30	8.74	8.34	7.71	2.00	2.80	16.57	3.83	-0.02	2.25
cyan	33	3.01	5.02	1.01	3.72	3.70	6.31	14.79	7.92	6.67	1.01	-0.04	7.89	1.03	0.45	0.95
tan	38	2.06	3.32	0.80	1.47	2.70	3.93	4.30	3.20	2.58	0.80	0.55	5.82	1.23	1.23	0.48
red	89	1.22	2.58	-0.13	1.74	2.36	2.80	5.22	3.95	4.35	-0.13	-0.20	2.86	0.73	0.64	1.24
turquoise	400	1.13	0.93	1.33	0.34	1.38	1.63	0.49	-1.14	-1.85	1.33	-0.72	7.62	2.66	-1.60	-0.50
salmon	37	0.88	2.27	-0.51	1.43	2.21	2.41	2.33	0.34	0.46	0.83	-1.25	8.88	-0.51	1.54	0.66
purple	57	0.60	0.19	1.01	-0.44	0.29	1.44	0.09	-0.50	-0.30	1.01	0.70	3.24	1.48	0.62	0.62
black	74	0.41	0.27	0.55	0.97	-0.95	-0.01	0.56	0.91	1.14	2.44	-0.99	0.82	0.55	2.24	1.58
yellow	181	0.32	0.30	0.34	1.17	-1.42	-0.57	2.15	2.90	2.83	1.04	-0.39	4.78	0.34	-0.42	0.60
pink	65	0.06	0.83	-0.72	-2.30	1.43	1.43	0.24	0.91	0.82	-0.72	-1.09	5.51	0.20	-0.70	-0.93
midnightblue	31	-0.09	0.19	-0.38	-0.03	0.33	0.53	0.05	0.26	0.23	-0.38	-1.21	0.53	0.58	0.76	-0.51
gold	100	-0.12	-0.72	0.49	-2.21	0.34	0.54	-1.78	-1.97	-1.81	0.49	0.99	7.93	0.47	-0.33	0.09
blue	400	-0.38	-2.39	1.63	-1.39	-6.80	-2.52	-2.26	-0.98	-0.88	-0.63	1.63	10.15	1.68	0.25	-1.38
magenta	57	-0.44	-0.15	-0.72	-0.04	-0.27	-0.76	0.45	1.43	1.48	-1.37	-0.72	8.66	0.21	-0.17	-1.20
brown	322	-0.71	-1.33	-0.10	2.55	-3.48	-3.23	0.57	0.78	0.22	-0.10	-0.99	8.82	0.84	-1.09	-0.73
grey	400	-1.12	-2.56	0.31	-2.51	-3.32	0.15	-2.61	-1.81	-1.27	-1.74	2.18	10.05	0.31	-1.70	-1.78

I also compared module preservation by considering the module eigengene, as previously described (Horvath and Dong 2008). Briefly, this involved taking the gene module as defined in the first (reference) species and finding the first principal component of the expression matrix of their orthologs in the second (test) species. Once this “module eigengene” was defined in the new species, I calculated the correlation of the expression of every module gene with the expression of the module eigengene (kME) in samples from the first (reference) species and then in the second (test) species. I then plotted these correlations in each species against one another using the `verboseScatterplot` R function to find Pearson’s correlation coefficient of the two correlations plotted against one another (along with its statistical significance). Next, I did the same, but instead of only correlating module genes to the module eigengene under consideration, I correlated all genes to the module eigengene in both species.

To compare the modules in the chicken and mouse networks I used the `userListEnrichment` R function to identify the modules in each species that were significantly overlapping (Table 4.4). For every chicken module, this tested for significant enrichment in every mouse module using a hypergeometric distribution, and subsequently applied a Bonferroni correction for every pairwise comparison made (taking into account both lists).

Table 4.4: Chicken modules significantly overlapping mouse modules

chicken module	mouse module	Bonferroni-corrected <i>p</i> -value
green	brown	8.7×10^{-21}
cyan	magenta	3.0×10^{-18}
greenyellow	brown	1.4×10^{-4}
turquoise	salmon	1.6×10^{-3}
magenta	black	3.6×10^{-3}
black	orange	4.6×10^{-3}

For those modules having significant cross-species overlap, I identified genes likely to be hubs in both networks by finding the genes (in or not in the module) having the highest average kME rank (correlation with the respective module eigengene) using the modules defined in the original networks and not projecting one onto the other (Table 4.9).

Annotating gene co-expression modules

Using the same databases as described above in Section 2.5 I performed conditional one-sided Fisher's exact tests of enrichment for both the mouse modules and for the chick modules (using the mouse orthologs). The background consisted of all genes used to build modules for both chick and mouse that had any kind of annotation in the given database. To reduce the number of underpowered tests or results having limited biological interpretability, I only tested terms that (1) included at least 2 genes in the background and (2) included a sufficient number of genes in the background such that a *p*-value of 0.0005 would be theoretically possible for the given module size.

For all modules discussed in this chapter, I provide the top 10 genes most correlated with the module eigengene, their median quantile-normalized FPKM expression levels, and relative fold expression with respect to this median value in Table 4.5 and Table 4.6.

Table 4.5: Top ten correlated gene with each chicken module eigengene discussed, the median quantile-normalized FPKM for that gene, and the fold difference from the median expression in each chicken sample. DLC, dorsolateral cortex; HP, hyperpallium; MP, mesopallium; AP, arcopallium; NP, nidopallium; HIPP, hippocampus; STR, striatum.

green	107 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Kcna4</i>	ENSMUSG00000042604	93	0.86	1.20	0.99	1.01	1.00	0.81	2.29
<i>Smpd3</i>	ENSMUSG00000031906	227	1.00	0.81	0.96	0.82	1.13	1.08	2.06
<i>Foxo1</i>	ENSMUSG00000044167	44	0.94	1.12	1.00	0.92	1.14	0.57	3.57
<i>Pde10a</i>	ENSMUSG00000023868	48	0.74	1.75	0.86	2.24	1.00	0.52	10.16
<i>Drd1a</i>	ENSMUSG00000021478	147	1.62	0.85	0.92	1.00	1.22	0.03	7.09
<i>Tmem90a</i>	ENSMUSG00000071234	89	1.51	0.56	2.20	0.79	1.00	0.41	4.67
<i>Tac1</i>	ENSMUSG00000061762	248	1.47	1.42	0.00	1.00	0.45	0.00	9.03
<i>Bcr</i>	ENSMUSG00000009681	106	0.91	0.81	1.03	0.82	1.00	1.14	2.09
<i>Bcl11b</i>	ENSMUSG00000048251	69	2.23	0.67	1.62	0.46	0.75	1.00	4.80
<i>Adcy5</i>	ENSMUSG00000022840	93	1.15	0.69	0.97	1.62	0.59	1.00	6.47

greenyellow	54 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Kdm3a</i>	ENSMUSG00000053470	76	0.69	1.54	0.85	1.37	1.00	0.63	1.28
<i>Mctpl</i>	ENSMUSG00000021596	200	0.79	1.16	1.00	1.12	0.49	0.39	2.01
<i>Serinc5</i>	ENSMUSG00000021703	44	0.90	1.97	0.66	1.61	1.00	0.73	2.07
<i>Mapk1</i>	ENSMUSG00000063358	1408	1.00	0.59	1.12	0.72	1.29	0.86	2.17
<i>Pppde1</i>	ENSMUSG00000026502	37	0.85	1.99	0.89	1.74	1.00	0.77	1.93
<i>Rnf103</i>	ENSMUSG00000052656	176	0.99	0.95	1.00	1.05	1.13	0.87	1.41
<i>BC031353</i>	ENSMUSG00000034858	143	0.63	1.21	0.71	1.16	1.00	0.74	1.36
<i>Ptpn11</i>	ENSMUSG00000043733	192	0.75	1.12	0.68	1.00	1.05	0.66	1.26
<i>Fam105a</i>	ENSMUSG00000056069	151	0.80	1.30	0.65	2.84	0.93	1.00	4.52
<i>Ano3</i>	ENSMUSG00000074968	91	0.26	1.79	0.77	2.29	1.00	0.21	2.95

magenta	57 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Slco3a1</i>	ENSMUSG000000025790	63	1.00	0.91	0.73	0.97	1.08	2.38	1.13
<i>Slc16a1</i>	ENSMUSG000000032902	41	1.03	0.80	1.43	0.82	0.81	2.03	1.00
<i>Gnai3</i>	ENSMUSG000000000001	835	1.04	0.40	1.04	0.48	0.89	1.38	1.00
<i>Mtmr9</i>	ENSMUSG000000035078	351	1.03	0.49	1.00	0.67	0.64	1.37	1.07
<i>Agfg1</i>	ENSMUSG000000026159	245	1.00	0.51	1.03	0.58	0.70	1.31	1.00
<i>Uso1</i>	ENSMUSG000000029407	191	1.19	0.68	1.18	0.73	0.90	1.63	1.00
<i>Wasl</i>	ENSMUSG000000029684	82	1.34	0.64	1.00	0.53	0.55	3.22	1.75
<i>Kras</i>	ENSMUSG000000030265	168	0.96	0.67	1.34	0.59	1.00	2.04	1.45
<i>Ndst1</i>	ENSMUSG000000054008	182	1.27	0.30	1.10	0.34	0.55	1.74	1.00
<i>Rac1</i>	ENSMUSG000000001847	1751	1.30	0.55	1.34	0.62	0.84	1.94	1.00

black	74 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Dgke</i>	ENSMUSG00000000276	104	1.00	0.75	1.30	0.67	1.34	0.96	1.33
<i>Crk</i>	ENSMUSG00000017776	187	1.00	0.85	0.98	0.89	1.22	1.00	1.00
<i>Klhl24</i>	ENSMUSG000000062901	67	1.00	1.06	0.95	0.91	1.13	1.07	0.87
<i>Dnajb14</i>	ENSMUSG000000074212	653	0.98	0.61	1.06	0.73	1.19	1.12	1.00
<i>1810013L24Rik</i>	ENSMUSG000000022507	341	1.02	0.68	1.03	0.77	1.29	1.00	0.97
<i>Klhl7</i>	ENSMUSG000000028986	297	1.02	1.11	1.00	0.95	1.62	0.74	0.86
<i>Rnf11</i>	ENSMUSG000000028557	394	1.06	0.71	1.02	1.00	1.14	0.85	0.95
<i>Chmp7</i>	ENSMUSG000000034190	480	1.00	0.58	1.12	0.68	1.14	0.98	1.15
<i>Fam172a</i>	ENSMUSG000000064138	86	1.21	0.90	0.95	1.00	1.17	0.99	1.05
<i>Klhl11</i>	ENSMUSG000000048732	146	0.90	0.55	1.19	0.80	1.22	1.00	1.10

cyan	33 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Bcas1</i>	ENSMUSG00000013523	229	0.59	1.32	0.67	2.70	1.42	1.00	0.61
<i>Spg20</i>	ENSMUSG000000036580	113	0.61	1.68	0.79	2.31	1.61	1.00	0.94
<i>Pllp</i>	ENSMUSG000000031775	180	0.70	0.51	0.85	2.34	1.70	1.18	1.00
<i>Gab1</i>	ENSMUSG000000031714	108	0.67	1.09	1.00	2.60	1.68	0.94	0.98
<i>Anln</i>	ENSMUSG000000036777	34	0.02	2.64	0.57	5.30	2.75	1.00	0.81
<i>Elovl7</i>	ENSMUSG000000021696	27	0.03	3.61	1.00	4.72	2.05	0.85	0.84
<i>Slain1</i>	ENSMUSG000000055717	113	0.90	1.03	0.82	1.13	1.01	1.00	0.90
<i>Tspan2</i>	ENSMUSG000000027858	153	0.70	1.02	0.73	2.46	1.54	0.75	1.00
<i>Reep3</i>	ENSMUSG000000019873	117	0.64	1.59	0.82	2.43	1.48	0.72	1.00
<i>Mbp</i>	ENSMUSG000000041607	8687	0.50	1.00	0.71	3.19	1.94	0.77	1.00

turquoise	2452 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Gosr1</i>	ENSMUSG00000010392	119	0.96	1.96	0.92	1.73	1.16	0.90	1.00
<i>Ubqln1</i>	ENSMUSG00000005312	292	0.89	1.55	1.00	1.40	1.25	0.91	0.92
<i>March7</i>	ENSMUSG00000026977	107	1.00	2.22	0.94	1.91	1.33	0.87	1.00
<i>Necap1</i>	ENSMUSG00000030327	949	0.77	1.19	0.88	1.11	1.02	0.78	1.00
<i>Ap1g1</i>	ENSMUSG00000031731	344	0.91	1.43	0.95	1.27	1.09	0.81	1.00
<i>Tmx3</i>	ENSMUSG00000024614	184	0.84	2.61	0.92	2.26	1.40	0.78	1.00
<i>AI597479</i>	ENSMUSG00000010290	60	0.99	3.26	0.94	2.63	1.72	1.00	0.87
<i>Eif4e</i>	ENSMUSG00000028156	420	1.00	1.86	0.95	1.68	1.26	0.91	0.93
<i>Dnajc3</i>	ENSMUSG00000022136	233	0.94	2.15	1.00	1.93	1.53	0.94	1.00
<i>Heatr5b</i>	ENSMUSG00000039414	169	0.90	2.03	0.89	1.77	1.39	1.00	0.92

brown	322 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Rab28</i>	ENSMUSG00000029128	285	1.14	0.54	1.10	0.60	0.81	1.00	1.17
<i>Myo9a</i>	ENSMUSG00000039585	167	1.00	0.22	1.08	0.22	0.45	1.13	1.46
<i>Ccny</i>	ENSMUSG00000024286	206	1.11	0.48	1.06	0.43	0.72	1.00	1.11
<i>Kpna1</i>	ENSMUSG00000022905	174	1.17	0.46	1.00	0.55	0.63	1.06	1.06
<i>Rprd2</i>	ENSMUSG00000028106	173	1.16	0.36	1.07	0.33	0.57	1.21	1.00
<i>Cpsf6</i>	ENSMUSG00000055531	111	1.21	0.54	1.00	0.48	0.66	1.13	1.00
<i>Pogz</i>	ENSMUSG00000038902	74	1.03	0.05	1.03	0.39	0.37	1.00	1.43
<i>Ythdf2</i>	ENSMUSG00000040025	101	1.18	0.21	1.02	0.26	0.37	1.23	1.00
<i>Sgip1</i>	ENSMUSG00000028524	235	1.19	0.19	1.37	0.16	0.42	1.41	1.00
<i>Kpnb1</i>	ENSMUSG00000001440	228	1.13	0.38	1.09	0.40	0.63	1.01	1.00

red	89 genes		fold difference from median						
Gene symbol	Gene ID	median normalized FPKM	DLC	HP	MP	AP	NP	HIPP	STR
<i>Cul4b</i>	ENSMUSG000000031095	117	0.87	1.00	1.07	0.95	1.14	0.84	1.25
<i>Samd8</i>	ENSMUSG000000021770	163	0.72	1.00	1.25	0.91	1.52	0.84	1.37
<i>Vcpi1</i>	ENSMUSG000000045210	83	0.68	1.04	1.00	1.00	1.02	0.86	1.00
<i>Atp13a3</i>	ENSMUSG000000022533	72	0.73	1.08	0.85	1.00	1.02	0.61	1.16
<i>Rprd1a</i>	ENSMUSG000000040446	145	1.00	0.83	1.39	0.81	1.07	0.93	1.43
<i>Lmbrd2</i>	ENSMUSG000000039704	97	0.82	1.00	1.04	0.85	1.02	0.70	1.21
<i>B230219D22Rik</i>	ENSMUSG000000045767	455	1.00	0.60	1.19	0.61	1.27	0.82	1.50
<i>Gopc</i>	ENSMUSG000000019861	79	0.66	1.00	1.24	0.80	1.13	0.67	1.00
<i>Ulk2</i>	ENSMUSG000000004798	318	1.00	0.91	1.24	0.83	1.20	0.50	1.16
<i>Cul5</i>	ENSMUSG000000032030	125	0.66	1.00	1.04	0.92	1.09	0.79	1.05

Table 4.6: Top ten correlated gene with each mouse module eigengene discussed, the median quantile-normalized FPKM for that gene, and the fold difference from the median expression in each mouse sample. DC, dorsal cortex; LC, lateral cortex; HIPP, hippocampus; C/E, claustrum/endopiriform complex; PA, pallial amygdala; STR, striatum.

brown	177 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>Me2</i>	ENSMUSG00000024556	16	0.83	0.86	0.91	1.14	0.85	1.04	0.82	0.75	0.67	0.96	1.30	1.24	1.24	1.30	1.30	3.64
<i>4922501L14Rik</i>	ENSMUSG00000042943	12	0.92	0.72	1.01	0.72	1.08	0.99	1.52	0.85	0.59	1.22	1.32	0.94	0.86	1.89	2.85	3.79
<i>Pik3c2b</i>	ENSMUSG00000026447	9	1.00	0.76	0.72	0.79	1.12	1.62	1.09	1.00	0.83	1.00	1.08	1.51	1.25	0.13	0.81	3.66
<i>Adcy3</i>	ENSMUSG00000020654	43	0.98	0.76	0.87	0.96	0.85	1.01	1.31	0.88	1.12	0.99	1.06	1.16	0.93	1.23	1.37	1.62
<i>Mgea5</i>	ENSMUSG00000025220	180	0.87	1.09	1.09	0.99	0.99	0.97	0.70	0.89	0.92	1.01	1.03	1.06	0.96	1.10	1.25	1.55
<i>Wscd2</i>	ENSMUSG00000063430	49	1.70	1.02	0.98	0.83	0.86	0.83	1.67	1.30	1.12	0.89	0.97	0.90	1.98	0.96	1.15	3.06
<i>Cacna2d2</i>	ENSMUSG00000010066	39	0.51	0.69	1.42	1.45	1.11	1.02	0.55	0.98	1.24	1.31	0.74	0.64	0.42	0.81	1.50	2.69
<i>Tnrc18</i>	ENSMUSG00000039477	21	0.81	0.99	0.97	0.92	0.95	1.00	1.46	1.52	0.91	1.20	1.09	1.00	1.21	0.89	1.13	1.84
<i>Sik3</i>	ENSMUSG00000034135	113	1.01	1.06	0.98	0.99	0.95	0.98	1.13	1.07	0.99	0.92	0.85	0.91	1.49	1.01	1.12	1.99
<i>Phf15</i>	ENSMUSG00000020387	50	0.80	0.83	1.12	1.09	1.05	1.04	0.81	0.92	1.03	0.97	0.90	0.85	1.12	0.94	1.19	1.69

magenta	64 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>Lamp1</i>	ENSMUSG00000031447	292	0.76	0.87	1.01	0.94	0.97	1.47	0.89	1.18	1.13	1.04	0.99	1.18	1.09	0.92	0.90	1.16
<i>Rdx</i>	ENSMUSG00000032050	59	0.91	0.82	0.80	0.96	0.94	1.42	0.80	0.85	1.00	1.00	1.09	1.31	1.29	1.04	1.15	1.38
<i>Npc1</i>	ENSMUSG00000024413	39	1.01	0.89	0.86	1.07	1.19	1.94	0.83	0.93	1.02	0.84	0.99	1.37	1.26	0.94	0.61	1.41
<i>Dram2</i>	ENSMUSG00000027900	23	0.69	0.81	0.73	0.80	1.22	2.75	0.96	0.75	0.71	1.17	1.07	1.04	1.46	1.21	0.84	1.84
<i>Ndrgl</i>	ENSMUSG00000005125	155	0.42	0.36	0.84	1.22	1.26	2.48	0.37	0.47	0.68	1.02	0.98	1.51	1.40	0.72	1.04	1.47
<i>Rnfl3</i>	ENSMUSG00000036503	101	0.52	0.75	0.98	1.02	1.04	1.78	0.44	0.76	0.94	1.07	1.02	1.26	1.30	0.82	0.78	1.47
<i>Cln8</i>	ENSMUSG00000026317	23	0.79	0.81	1.00	1.36	1.14	1.64	0.79	0.93	1.00	1.00	1.02	1.04	1.08	0.72	0.77	1.48
<i>Elovl7</i>	ENSMUSG00000021696	11	0.74	0.73	0.94	1.20	1.82	2.98	0.69	0.71	0.74	1.06	1.30	2.07	1.61	0.57	0.53	2.51
<i>Bmpr1a</i>	ENSMUSG00000021796	39	0.60	0.65	0.86	1.01	1.01	1.37	0.44	0.60	0.66	1.04	0.97	1.11	1.37	0.99	1.13	1.31
<i>Itgbl</i>	ENSMUSG00000025809	43	0.85	0.82	1.02	1.06	1.05	1.41	0.67	0.84	0.98	0.97	0.97	1.13	1.08	0.96	1.02	1.26

tan	62 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>Pdcd6</i>	ENSMUSG00000021576	54	1.08	1.27	0.99	0.85	0.69	0.65	1.45	1.56	1.18	1.07	0.95	0.65	1.00	0.75	1.00	1.76
<i>Atox1</i>	ENSMUSG00000018585	543	1.30	1.02	0.98	1.13	0.83	0.32	1.61	2.22	0.77	1.44	0.77	0.46	1.32	0.96	0.89	1.17
<i>Srp19</i>	ENSMUSG00000014504	105	1.16	1.09	0.70	0.74	0.65	1.00	1.59	1.19	1.07	1.00	0.83	0.72	1.09	0.84	1.14	0.79
<i>H47</i>	ENSMUSG00000075701	74	1.01	1.25	1.02	0.69	0.61	0.60	1.42	1.52	1.16	0.89	0.78	0.75	1.04	0.62	0.99	1.31
<i>Acsbg1</i>	ENSMUSG00000032281	158	1.24	1.06	1.02	0.87	0.66	0.64	1.36	1.48	1.18	0.98	0.68	0.58	1.03	0.58	0.78	1.03
<i>Rpl22l1</i>	ENSMUSG00000039221	87	0.57	1.17	0.49	0.65	0.50	0.47	2.28	2.71	0.82	2.06	1.26	0.59	1.40	0.83	1.40	2.76
<i>Ndufb10</i>	ENSMUSG00000040048	409	1.20	0.93	0.79	0.76	0.70	0.63	1.50	1.36	1.15	1.07	1.10	0.89	1.12	0.82	0.88	1.35
<i>Serfl</i>	ENSMUSG00000021643	21	1.43	1.17	0.73	0.60	0.64	0.42	1.57	1.85	0.95	1.05	1.23	0.71	1.14	0.94	1.19	0.83
<i>4632415K11Rik</i>	ENSMUSG00000034105	16	1.46	1.02	0.96	0.97	0.87	1.01	2.03	1.09	1.23	0.99	0.96	0.98	1.04	0.61	0.81	1.48
<i>Park7</i>	ENSMUSG00000028964	277	1.30	1.18	1.00	0.84	0.89	0.61	1.70	1.42	1.01	0.99	0.99	0.71	1.06	0.90	1.00	1.05

black	118 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>Tars</i>	ENSMUSG00000022241	22	0.95	1.00	0.99	1.00	0.96	0.95	1.15	1.13	1.14	1.08	1.06	0.95	1.60	0.98	0.91	1.13
<i>Rnf2</i>	ENSMUSG00000026484	29	0.93	1.20	0.88	0.85	1.00	1.06	0.84	0.98	0.91	1.00	0.98	1.05	2.23	1.16	1.17	1.09
<i>Rwdd1</i>	ENSMUSG00000019782	23	0.99	0.72	0.81	1.15	1.17	1.04	1.01	0.58	0.98	0.80	0.94	1.17	1.94	1.05	1.54	0.70
<i>Ankrd46</i>	ENSMUSG00000048307	363	0.92	1.11	1.03	0.90	0.93	1.00	0.75	0.90	0.92	0.94	1.00	1.05	1.21	1.14	1.10	1.03
<i>Copb2</i>	ENSMUSG00000032458	67	0.87	1.01	0.89	1.11	1.06	0.89	0.84	0.99	0.87	1.01	1.03	0.95	1.62	1.14	1.16	0.85
<i>Acadl</i>	ENSMUSG00000026003	23	0.89	1.01	0.87	1.10	0.94	0.88	0.90	1.17	0.95	1.09	0.79	1.00	1.91	1.00	1.50	1.48
<i>Etfdh</i>	ENSMUSG00000027809	19	0.99	1.11	0.85	1.19	1.01	0.87	0.78	0.99	1.11	1.16	0.96	1.03	1.62	0.96	0.92	1.31
<i>Smu1</i>	ENSMUSG00000028409	39	0.89	0.87	0.77	0.94	0.91	0.88	1.02	0.96	0.98	1.11	1.25	1.12	1.60	1.07	1.22	1.39
<i>Slc1a3</i>	ENSMUSG00000005360	368	1.30	1.10	1.01	0.95	0.93	0.98	0.90	0.99	1.04	0.95	0.79	0.85	1.78	1.03	1.47	1.33
<i>Plrg1</i>	ENSMUSG00000027998	25	0.84	0.98	0.83	1.18	0.89	1.18	1.01	1.09	0.82	0.97	0.99	0.99	1.72	1.15	1.36	1.22

orange	33 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>ORF61</i>	ENSMUSG00000013858	208	1.03	1.44	1.15	1.03	1.07	0.88	1.44	1.21	1.24	0.89	0.94	0.97	0.81	0.95	0.77	0.86
<i>Ggt7</i>	ENSMUSG00000027603	148	1.29	1.17	1.15	0.87	0.82	0.87	1.37	1.37	1.20	0.99	1.05	1.01	0.73	0.72	0.80	0.61
<i>Atg13</i>	ENSMUSG00000027244	98	1.04	1.08	1.16	0.97	1.00	1.00	1.05	1.15	1.14	1.01	0.97	0.96	0.77	0.79	0.74	0.87
<i>Fbxo34</i>	ENSMUSG00000037536	103	1.15	1.48	1.33	1.03	0.80	0.67	1.25	1.53	1.37	0.97	0.80	0.68	1.45	0.73	0.90	0.53
<i>Zdhhc16</i>	ENSMUSG00000025157	30	1.12	1.55	1.01	1.04	0.79	0.83	1.59	1.40	1.08	1.04	0.99	0.95	0.85	0.79	0.98	0.96
<i>Spsb3</i>	ENSMUSG00000024160	234	1.33	1.55	1.27	1.04	0.98	1.02	1.45	1.36	1.42	0.98	0.90	0.89	0.74	0.82	0.61	0.95
<i>Rapgef1</i>	ENSMUSG00000039844	90	1.48	1.27	1.19	0.99	0.99	0.88	1.44	1.39	1.25	1.01	1.03	0.92	0.81	0.99	0.87	0.94
<i>Clip2</i>	ENSMUSG00000063146	134	1.31	1.10	1.21	1.00	0.91	1.00	1.44	1.60	1.36	1.08	0.90	0.98	0.78	0.88	0.69	0.52
<i>Klhl18</i>	ENSMUSG00000054792	33	1.06	1.02	1.18	1.04	0.93	0.87	1.16	1.30	1.18	1.01	0.90	0.82	0.76	0.99	0.86	0.98
<i>Dnajb5</i>	ENSMUSG00000036052	308	1.04	1.04	1.20	0.98	1.01	0.91	1.28	1.32	1.27	0.99	0.96	0.93	1.03	0.81	0.65	0.64

salmon	61 genes		fold difference from median															
Gene symbol	Gene ID	median normalized FPKM	DC A	DC B	DC C	DC D	DC E	DC F	LC A	LC B	LC C	LC D	LC E	LC F	HIPP	C/E	PA	STR
<i>Mrps33</i>	ENSMUSG00000029918	35	0.86	1.45	0.95	0.87	0.72	0.67	1.05	1.23	1.28	0.86	1.23	0.95	1.44	0.85	1.28	1.57
<i>Pdcd6</i>	ENSMUSG00000021576	54	1.08	1.27	0.99	0.85	0.69	0.65	1.45	1.56	1.18	1.07	0.95	0.65	1.00	0.75	1.00	1.76
<i>Cisd1</i>	ENSMUSG00000037710	177	0.97	1.00	0.91	0.98	0.80	0.85	1.08	1.26	1.22	1.22	1.00	0.95	1.10	0.91	1.12	1.25
<i>2310003C23Rik</i>	ENSMUSG00000027573	71	0.84	0.97	0.77	0.78	0.73	0.69	1.05	1.21	1.05	1.12	1.01	0.91	0.99	1.03	1.03	1.17
<i>Atp5f1</i>	ENSMUSG00000000563	206	0.55	1.03	0.90	1.03	0.74	0.66	0.87	1.37	1.10	1.37	1.01	0.78	1.21	0.79	0.99	1.55
<i>Rpl22l1</i>	ENSMUSG00000039221	87	0.57	1.17	0.49	0.65	0.50	0.47	2.28	2.71	0.82	2.06	1.26	0.59	1.40	0.83	1.40	2.76
<i>Cops4</i>	ENSMUSG00000035297	79	0.91	1.07	0.85	0.90	0.76	0.74	1.10	1.24	1.08	1.00	1.00	0.85	1.32	0.98	1.14	1.01
<i>Uqcrb</i>	ENSMUSG00000021520	69	0.74	1.13	0.80	0.89	0.68	0.53	0.79	1.21	1.11	1.06	0.94	0.69	1.47	1.12	1.16	1.35
<i>Atp5j</i>	ENSMUSG00000022890	320	0.68	1.04	0.80	0.91	0.50	0.65	1.00	0.87	1.43	1.11	1.09	0.64	1.82	1.00	1.23	1.78
<i>Ndufa5</i>	ENSMUSG00000023089	391	0.68	0.87	0.70	0.64	0.58	0.35	1.01	1.53	0.99	2.03	1.42	0.76	1.68	1.02	1.54	2.33

4.3 Results

I quantified gene expression in 16 regions pooled from eight adult mice and 7 regions pooled from eighteen adult chickens, and noted specific expression trends that could be tainted by batch effects. Laminar marker genes confirmed dissections of cortical layers in mouse. Non-cortical dissections in mouse were confirmed by pre-existing *in situ* hybridisations of genes I found to be among the most specific to each structure. All the completed *in situ* hybridisations in chicken are concordant with the sequencing-based predictions. Ranked expression levels between the two species were correlated with one another (Spearman's ρ , $r_s=0.49$; $p<10^{-200}$), but connectivity rank was not ($r_s=0.0071$; $p=0.61$). (However, the latter may be artificially depressed by variable RNA quality.) I then constructed modules of co-expressed genes in each species.

Functional annotation enrichments amongst co-expressed genes in mouse and chick

There were 15 distinct modules in chick and 49 in mouse (I do not include in this count the grey modules, which consists of genes that were not assigned to a module). There are at least two likely contributors to the greater number of mouse modules: first, the minimum size allowed for a module in mouse was smaller than for chick; second, there were many more samples sequenced from mouse which allowed for a greater diversity of clearly delineated gene expression patterns.

I tested for functional enrichments from several different annotation sources (see Section 2.5 above). When I applied a Bonferroni correction to account for all tests performed across all modules for a single annotation source in a single species, no modules were

significantly enriched in either species or among any of the annotation sources with a $\text{FWER} < 0.05$. (The FWER, or familywise error rate, is the probability of making at least one type I error in a set of hypotheses being tested.) When I instead applied the Bonferroni correction to account only for all tests performed from a single annotation source in a single module, there were some ‘significant’ enrichments as shown in Table 4.7 and Table 4.8.

Table 4.7: Functional enrichments in mouse modules having a Bonferroni-corrected $p < 0.05$. Modules correlated with a technical variable (two-tailed $p < 0.05$) are indicated with an asterisk. Annotation sources are drawn from the databases described in Section 2.5 – GO biological process (GO BP), GO cellular component (GO CC), GO molecular function (GO MF), and Kyoto Encyclopaedia of Genes and Genomes Pathways (KEGG).

module	single test p -value (one-tailed)	annotation (source)	$\log_2$ (fold difference)	# genes with this annotation in the module	# genes with any annotation in the module
brown	2.1×10^{-6}	signal transduction (GO BP)	1.4	28	140
brown	3.9×10^{-6}	G-protein coupled receptor protein signaling pathway (GO BP)	1.7	21	140
brown	1.5×10^{-6}	G-protein coupled receptor activity (GO MF)	2.5	13	154
brown	4.3×10^{-5}	adrenoceptor activity (GO MF)	4.2	5	154
brown	6.6×10^{-5}	signal transducer activity (GO MF)	1.6	17	154
brown	1.2×10^{-5}	Neuroactive ligand-receptor interaction (KEGG)	2.1	13	64
magenta*	3.5×10^{-6}	myelination (GO BP)	4.7	5	52
magenta*	1.6×10^{-6}	myelin sheath (GO CC)	6.3	4	57
magenta*	1.6×10^{-5}	integral to membrane (GO CC)	0.94	33	57
magenta*	3.7×10^{-5}	compact myelin (GO CC)	6.3	3	57
magenta*	3.5×10^{-5}	structural constituent of myelin sheath (GO MF)	6.3	3	55
tan*	9.1×10^{-6}	ribonucleoprotein complex (GO CC)	2.8	9	54
tan*	7.6×10^{-5}	respiratory chain (GO CC)	3.6	5	54
tan*	5.9×10^{-7}	Huntington's disease (KEGG)	3.1	9	20
tan*	1.4×10^{-4}	Parkinson's disease (KEGG)	2.9	6	20
tan*	1.5×10^{-4}	Oxidative phosphorylation (KEGG)	2.9	6	20
tan*	2.2×10^{-4}	Alzheimer's disease (KEGG)	2.8	6	20
violet	9.6×10^{-5}	membrane (GO CC)	0.95	20	25

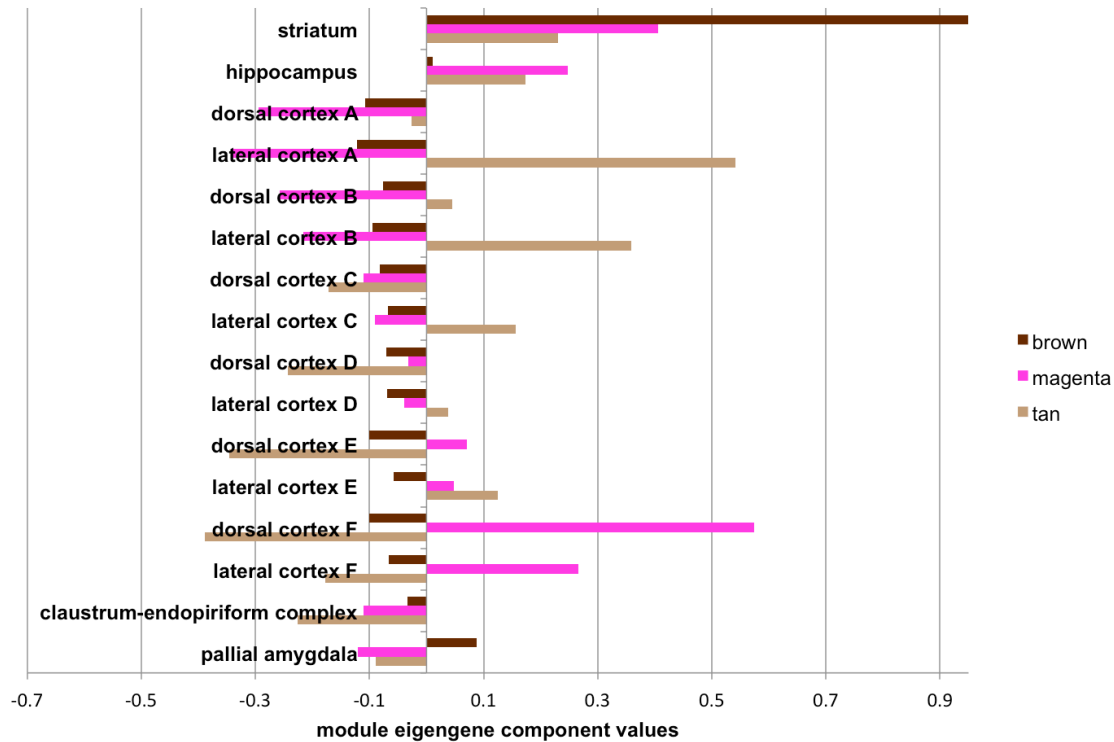


Figure 4.19: Module eigengene expression in different samples of the mouse modules with significant functional enrichments.

Three mouse modules – brown, magenta and tan – were not evenly expressed across samples and differed from one another (Figure 4.19). (I do not discuss the fourth module, violet, here since the less-than-two-fold enrichment – ‘membrane’ – is not easily biologically interpretable.) Mouse brown was specific to striatum. Mouse magenta genes tended to be expressed in striatum, hippocampus and deep cortical layers. Mouse tan was primarily found in upper layers of lateral cortex.

Two of these three mouse modules – magenta and tan – were significantly correlated with technical variables. Mouse magenta was negatively correlated with both 3' bias ($r=-0.67$; two-tailed $p=0.0048$) and pre-mRNA fraction ($r=-0.86$; two-tailed

$p=0.000016$). In contrast, mouse tan was *positively* correlated with 3' bias ($r=0.63$; two-tailed $p=0.0083$).

Table 4.8: Functional enrichments in chick modules having a Bonferroni-corrected $p<0.05$. Modules correlated with a batch variable (two-tailed $p<0.05$) are indicated with an asterisk. Annotation sources are drawn from the databases described in Section 2.5 – GO biological process (GO BP), GO cellular component (GO CC), GO molecular function (GO MF), and Kyoto Encyclopaedia of Genes and Genomes Pathways (KEGG).

module	single test p -value (one-tailed)	annotation (source)	$\log_2$ (fold difference)	# genes with this annotation in the module	# genes with any annotation in the module
brown*	7.5×10^{-5}	actin cytoskeleton (GO CC)	2.0	13	272
brown*	2.7×10^{-4}	Spliceosome (KEGG)	1.8	12	91
cyan*	4.6×10^{-5}	integral to membrane (GO CC)	1.1	20	30
green*	5.9×10^{-6}	signal transduction (GO BP)	1.6	20	87
green*	3.1×10^{-5}	calcium ion binding (GO MF)	1.6	18	96
green*	8.0×10^{-5}	potassium channel activity (GO MF)	3.2	6	96
green*	8.5×10^{-5}	adrenoceptor activity (GO MF)	4.5	4	96
red	5.0×10^{-5}	transferase activity (GO MF)	1.4	20	71

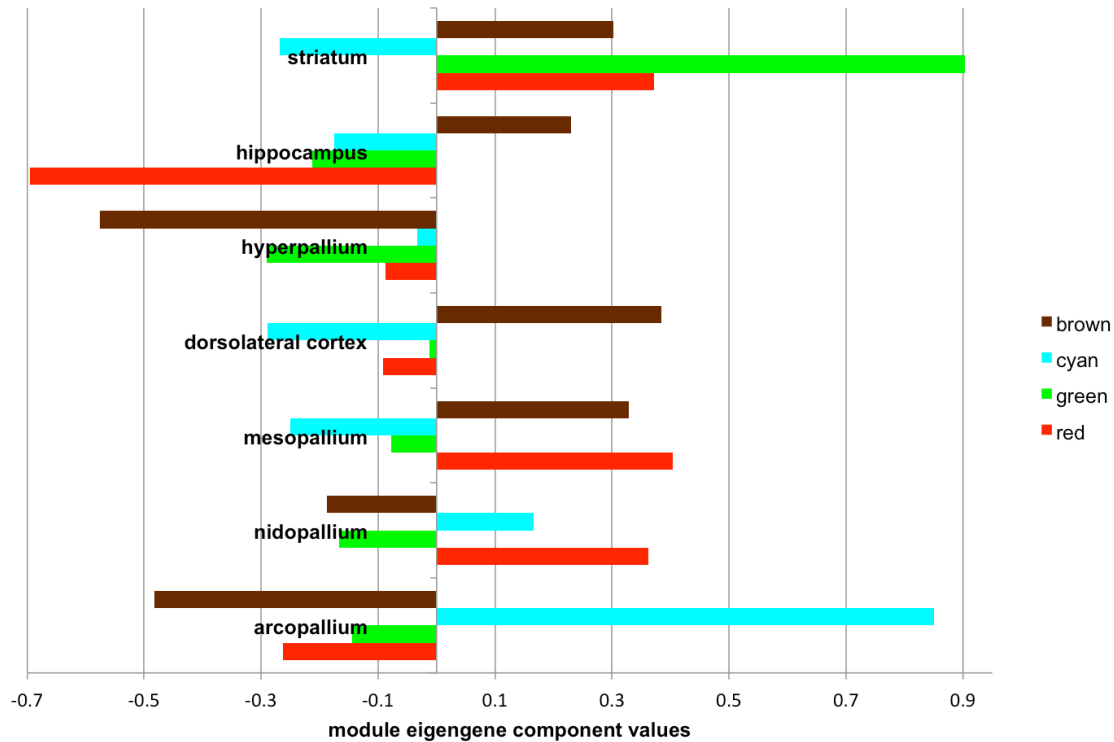


Figure 4.20: Module eigengene expression in different samples of the chick modules with significant functional enrichments.

Like the mouse modules, the four chick modules – brown, cyan, green and red – were not evenly expressed across samples and differed from one another (Figure 4.20). Chick brown was found in dorsolateral cortex, mesopallium, striatum, and hippocampus. Chick cyan was mostly in arcopallium, but in nidopallium to a lesser extent. Chick green was specific to striatum. Chick red was especially low in hippocampus.

Two of the four chick modules – brown and green – were significantly correlated with technical variables. Chick brown was positively correlated with both RNA Integrity Number ($r=0.92$; two-tailed $p=0.0097$) and pre-mRNA fraction ($r=0.82$; two-tailed $p=0.025$) and negatively correlated with the percentage of uniquely mapped reads falling

into protein-coding models ($r=-0.90$; two-tailed $p=0.0052$). Green was also positively correlated with pre-mRNA fraction ($r=0.79$; two-tailed $p=0.033$).

Two of the chick modules with functional enrichments (green and cyan), significantly overlapped two mouse modules that also exhibited functional enrichments (brown and magenta, respectively) (Table 4.4). The first of these overlaps – chick green and mouse brown – also shared two significant functional enrichments: “signal transduction” (GO BP) and “adrenoceptor activity” (GO MF). The other overlapping modules – chick cyan and mouse magenta – did not have overlapping functional enrichments after Bonferroni correction (chick cyan had only one significant enrichment), but all the mouse magenta annotations were significant in chick cyan as a single test. The most significantly enriched functional annotation for chick cyan in the GO BP database was “myelination” ($p=0.0039$), with a $\log_2(\text{fold difference})$ of 4.6. The most significantly enriched functional annotation in the GO MF database was “structural constituent of the myelin sheath” ($p=0.00028$), with a $\log_2(\text{fold difference})$ of 6.9. The second most significantly enriched annotation in the GO CC database was “compact myelination” ($p=0.0043$), with a $\log_2(\text{fold difference})$ of 6.6. The final annotation enriched in mouse magenta, “myelin sheath” (GO CC), was enriched in chick cyan with a $\log_2(\text{fold difference})$ of 5.2 ($p=0.033$).

Limited and weak conservation of gene co-expression patterns between mouse and chick

I found only limited and weak conservation of gene co-expression patterns between mouse and chick. Of the 49 modules in mouse, only two (mouse salmon and mouse

steelblue) were significantly preserved in chick at a Bonferroni-corrected $p < 0.05$ (per the summary score $Z_{\text{preservation}}$). The first of these, mouse salmon, significantly overlapped a chick module (chick turquoise) (Table 4.4). Although what appear to be housekeeping genes dominate the top five ‘consensus’ genes for the mouse-chick overlap (Table 4.9), mouse salmon was not evenly expressed across samples (Figure 4.21). The other, mouse steelblue, did not significantly overlap a chick module. It was preferentially expressed in striatum (Figure C.29).

Mouse steelblue may, however, be tainted by batch effects. Its module eigengene was negatively correlated with both a metric of 3' bias ($r = -0.63$; $p = 0.0092$) and pre-mRNA fraction ($r = -0.77$; $p = 0.00044$).

By the same criteria, only three of the fifteen modules in chick (chick green, chick greenyellow and chick cyan) were significantly preserved in mouse. All three of these modules significantly overlapped a mouse module (Table 4.4). Chick green and chick greenyellow genes were preferentially expressed in striatum, while chick cyan genes tended to be oligodendrocyte markers and expressed in arcopallium (Table 4.9; Figure 4.22).

Two of these three chick modules were correlated with technical variables. Chick green was positively correlated with pre-mRNA fraction ($r = 0.79$; two-tailed $p = 0.033$), though chick greenyellow (similar to chick green, see Figure C.27) was not significantly correlated with pre-mRNA fraction ($r = 0.26$; two-tailed $p = 0.57$). However, the difference between the correlation coefficients of chick green and chick greenyellow with pre-mRNA fraction was not significant (two-tailed $p = 0.25$). Chick cyan was the

second module correlated with a technical variable. Its module eigengene was positively correlated with the percentage of uniquely mapped fragments in protein-coding genes ($r=0.87$; two-tailed $p=0.011$).

Matched co-expression modules in chick and mouse

Most of the significantly overlapping modules included well-known marker genes for specific cell types or organelles (Table 4.9) or were specifically expressed in each species, usually in regions known to be homologous to one another (Figure 4.21; Figure 4.22). Some of these modules were also correlated with technical variables in one or both species.

Table 4.9: Top candidates for cross-species hub genes in the overlapping modules (mouse gene symbols), and descriptors of the modules; asterisk indicates possible technical confound for that module

rank	chick green*, mouse brown	chick greenyellow mouse brown	chick magenta, mouse black	chick black, mouse orange*	chick cyan*, mouse magenta*	chick turquoise*, mouse salmon
	striatum	striatum	hippocampus	layer IV/ nidopallium	oligo- dendrocytes	?
1	<i>Dlx6</i>	<i>Pde10a</i>	<i>Abi1</i>	<i>Fam19a2</i>	<i>Bcas1</i>	<i>Cisd1</i>
2	<i>Foxo1</i>	<i>Pde7b</i>	<i>Ptk2b</i>	<i>Myo16</i>	<i>Gab1</i>	<i>2310003C23Rik</i>
3	<i>Pde10a</i>	<i>Dgkb</i>	<i>2010011I20Rik</i>	<i>Dctn3</i>	<i>Anln</i>	<i>Atp5f1</i>
4	<i>Drd1a</i>	<i>Rps6ka5</i>	<i>Pcdh1</i>	<i>Rorb</i>	<i>Tspan2</i>	<i>Rpl22l1</i>
5	<i>Adcy5</i>	<i>Tac1</i>	<i>Prickle2</i>	<i>Fam73b</i>	<i>Mbp</i>	<i>Ndufa5</i>

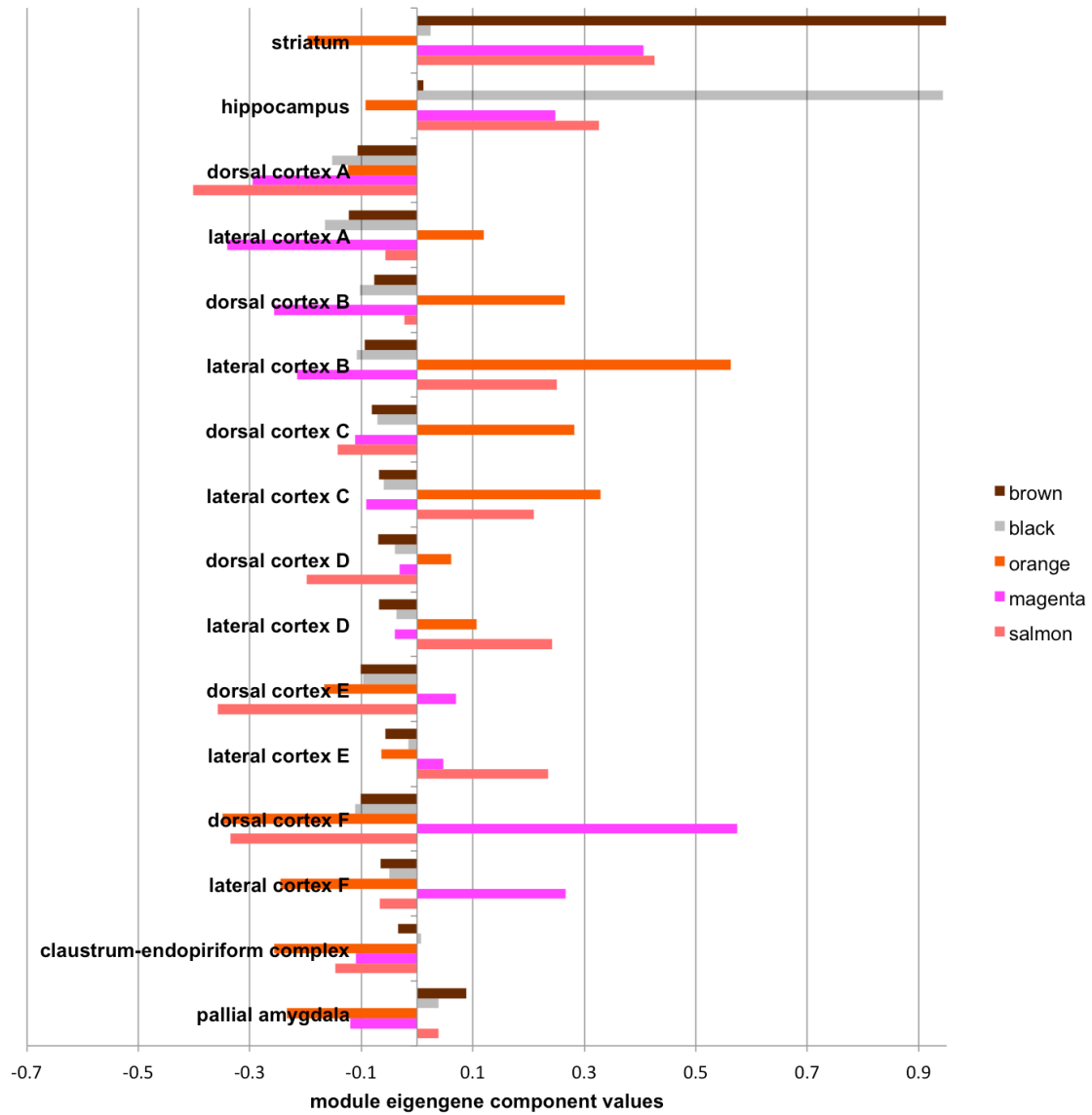


Figure 4.21: Module eigengene expression of the mouse modules that overlapped chick modules.

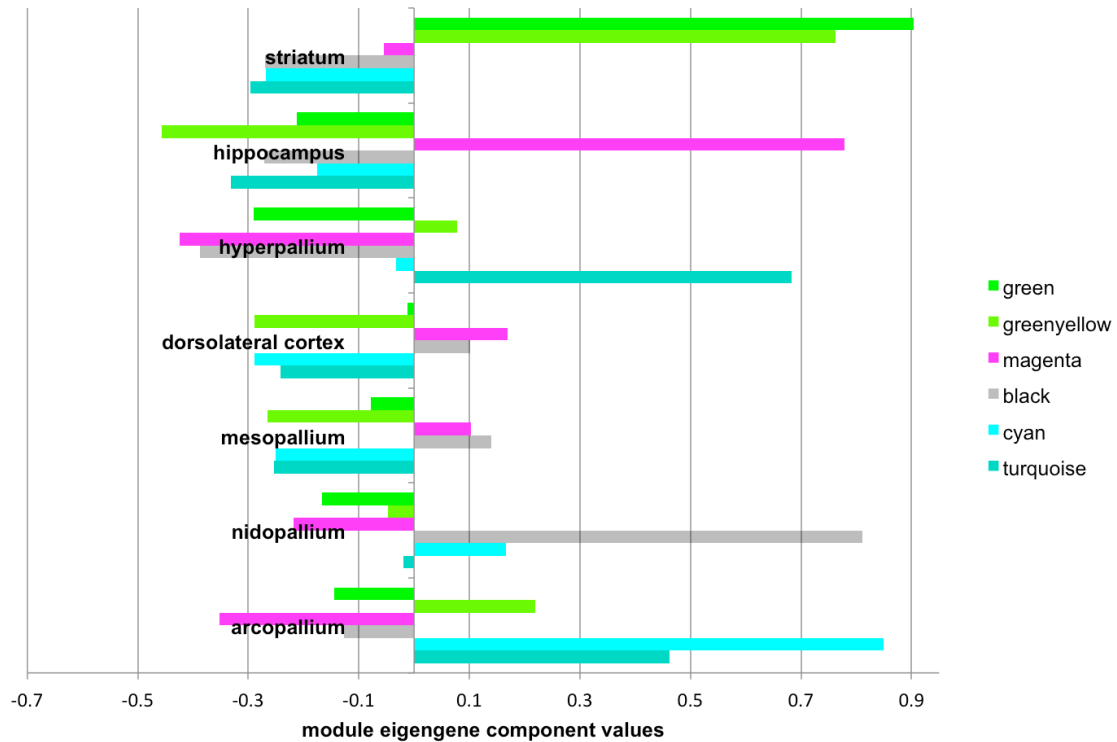


Figure 4.22: Module eigengene expression of the chicken modules that overlapped mouse modules.

Modules mark chick and mouse striatum

Genes in the chicken green/greenyellow modules and the corresponding mouse brown module were preferentially expressed in striatum (Figure 4.21; Figure 4.22). The chicken green & greenyellow modules cluster closely in the dendrogram of all chick modules, suggesting they are very similar to one another (Figure C.27). As discussed above, chick green was positively correlated with pre-mRNA fraction ($r=0.79$; two-tailed $p=0.033$), though there was no evidence that greenyellow was correlated with any technical variable.

I confirmed these predictions in mouse by examining *in situ* hybridisations (Lein et al. 2007) of the top five consensus genes from both the chick green/mouse brown and the chick greenyellow/mouse brown overlaps (Figure C.30). Seven of the nine unique genes were enriched in striatum; the probes covering the other two genes were too lightly stained to resolve expression patterns.

Modules mark chick and mouse hippocampus

Both chick magenta and mouse black were specific to hippocampus (relative, of course, to the dissected samples). I confirmed these predictions in mouse by examining *in situ* hybridisations (Lein et al. 2007) of the top five consensus genes from the chick magenta and mouse black overlap. In each image, the dissected portion of the hippocampal formation expressed the gene as highly if not higher than the rest of the section, and, of regions included in the bioinformatic analyses across all sections, none expressed the gene more highly than hippocampus (Figure C.31). However, not all regions are included in this automated analysis, and the expression intensity may be saturated in some sections. Neither chick magenta nor mouse black was significantly correlated with any technical variables.

Modules mark functionally analogous thalamic recipients in chick and mouse

According to their module eigengenes, chick black and mouse orange were primarily expressed in nidopallium and layer IV neocortex (both dorsal and lateral regions), respectively. The Allen Mouse Brain Atlas includes *in situ* hybridisations for three of the top five consensus genes. All three were highly expressed outside of neocortex;

indeed, two of the three (*Fam19a2* and *Dctn3*) were considerably more highly expressed outside neocortex (in regions that were not dissected).

In situ hybridisations from the Allen Institute demonstrate that *Rorb* is expressed in many places throughout development. At various times in development, *Rorb* is found in the eye, spinal cord (laminae 1-6 and the gray matter), hindbrain, midbrain, and forebrain (data not shown). In adult, *Rorb* is expressed in layer IV of neocortex, but is also found in several additional regions (Figure 4.23). The specificity of *Rorb* to chick nidopallium was also confirmed by *in situ* hybridisation by Dr. Wei Zhi Wang (Figure 4.15).

The extra-neocortical expression patterns of *Fam19a2* and *Dctn3* overlapped in layer 2 of piriform cortex (Figure 4.24). Beyond that, *Fam19a2* was expressed predominantly in layer 2a of entorhinal cortex, taenia tecta and the anterior olfactory nucleus; and *Dctn3* was expressed in hippocampus, motor nucleus of trigeminal, substantia nigra (compact part), pontine gray, paragigantocellular reticular nucleus, pons and spinal cord (see Allen Institute online resources) (Lein et al. 2007).

The module eigengene of mouse orange (but not of chick black) was positively correlated both with 3' bias ($r=0.56$; two-tailed $p=0.025$) and with pre-mRNA fraction ($r=0.56$; two-tailed $p=0.025$).

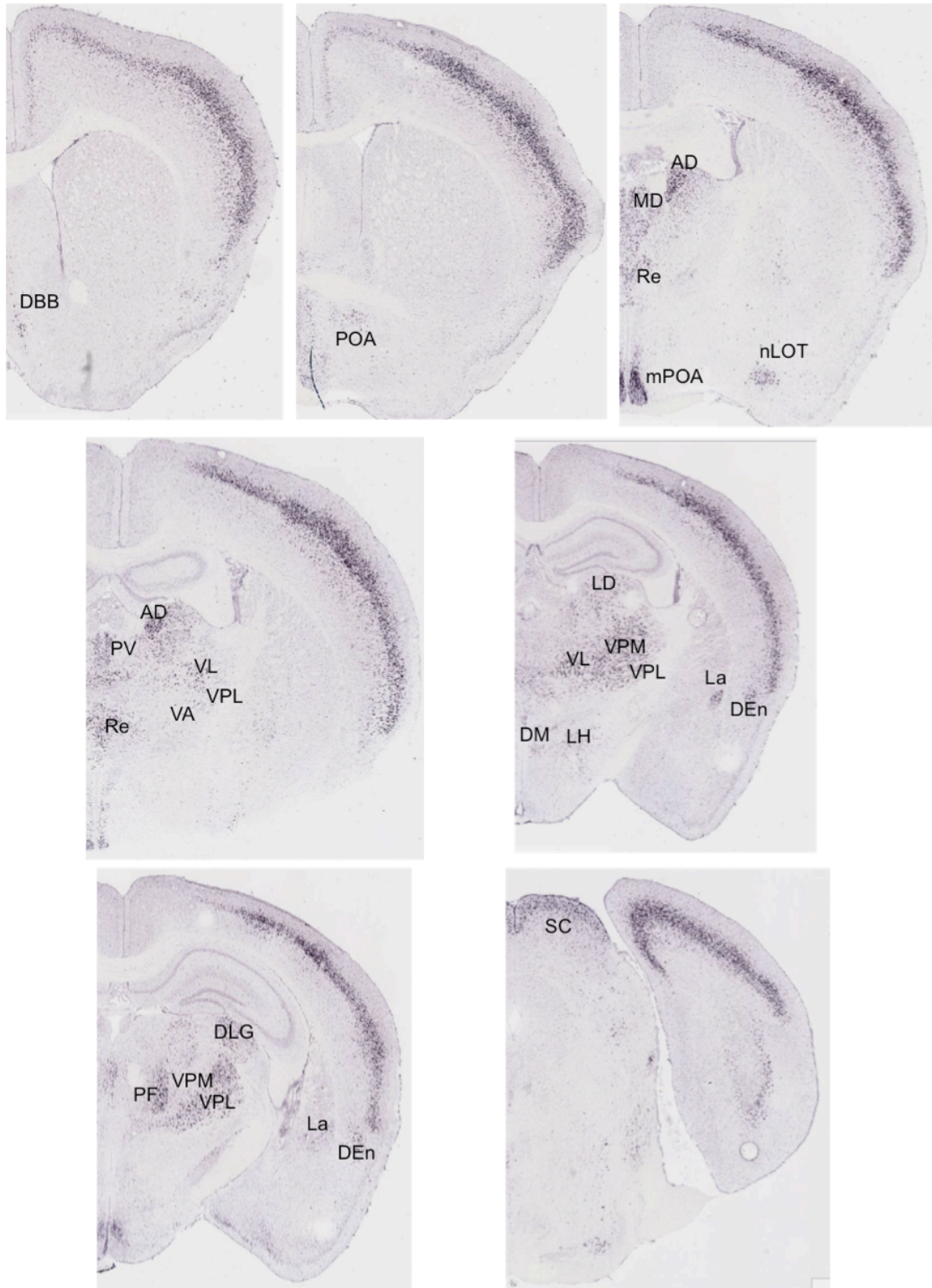


Figure 4.23: *Rorb* expression in adult mouse from the Allen Mouse Brain Atlas (Lein et al. 2007). AD: Antero dorsal Thalamic nucleus; DBB: Diagonal band of

Broca; Den: Dorsal Endopiriform nucleus; DLG: Dorsal lateral Geniculate nucleus; DM: Dorso medial Hypothalamic nucleus; La: Lateral amygdaloid nucleus; LD: Latero dorsal Thalamic nucleus; LH: Lateral Hypothalamic nucleus; MD: Medio dorsal Thalamic nucleus; mPOA: medial Preoptic area; nLOT: nucleus of the lateral olfactory tract; PF: Parafascicular nucleus of the Thalamus; POA: Preoptic area; PV: Paraventricular thalamic nucleus; Re: Reuniens nucleus; SC: Superior colliculus; VA: Ventral anterior Thalamic nucleus; VL: Ventral lateral Thalamic nucleus; VPL: Ventral postero-lateral Thalamic nucleus. Adapted from Fernando Garcia-Moreno and used with permission.

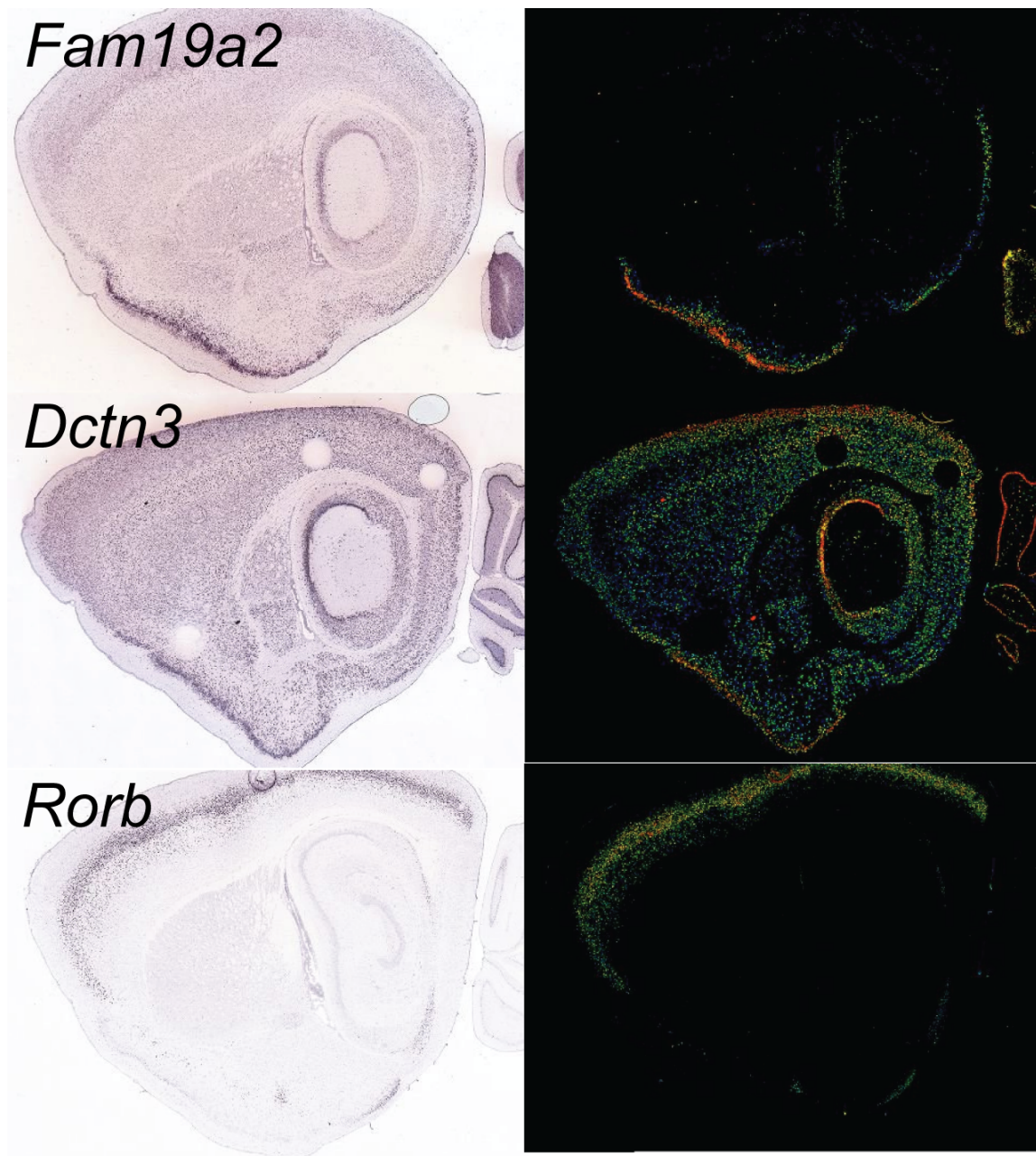


Figure 4.24: *Fam19a2* and *Dctn3* were both highly expressed in layer 2 of piriform cortex in the Allen Mouse Brain Atlas (Lein et al. 2007). *Rorb*, in contrast, was not particularly highly expressed in piriform cortex.

Modules mark oligodendrocytes in chick and mouse

All of the top five genes most correlated with the module eigengenes of chicken cyan and mouse magenta were associated with oligodendrocytes. Two are known oligodendrocyte markers: *Tspan2* (Birling et al. 1999) and *Mbp* (Boggs 2006). All five were significantly more highly enriched in oligodendrocytes than in neurons or astrocytes in a microarray-based experiment in adult mouse forebrain: *Bcas1* (34-fold), *Gab1* (4.9-fold), *Anln* (18-fold), *Tspan2* (45-fold), *Mbp* (49-fold) (Cahoy et al. 2008).

Both mouse magenta and chick cyan were correlated with technical variables. The module eigengene of mouse magenta was negatively correlated both with pre-mRNA fraction ($r=-0.86$; two-tailed $p=0.000016$) and with 3' bias ($r=-0.67$; two-tailed $p=0.0048$). Chick cyan may also be negatively correlated with pre-mRNA fraction, though the test was underpowered to be conclusive ($r=-0.74$; two-tailed $p=0.060$). Chick cyan was, however, positively correlated with the percentage of uniquely mapped fragments in protein-coding genes ($r=0.87$; two-tailed $p=0.011$).

Batch effects confound interpretation of the final module overlap

The final overlap involved chick turquoise and mouse salmon. Many of the top five genes correlated with these modules had housekeeping functions. *Cisd1* encodes mitoNEET, an outer mitochondrial membrane protein that regulates maximal mitochondrial respiratory rate (Wiley et al. 2007). *Atp5f1* (ATP synthase, H⁺ transporting, mitochondrial F0 complex, subunit B1) encodes a member of the ATP synthase complex (Higuti et al. 1991) consistently found in mouse mitochondria (Pagliarini et al. 2008). *Rpl22l1* (ribosomal protein L22 like 1) contains a ribosomal

protein L22e domain, and is thus annotated by the MGI curatorial staff as being a ribosomal subunit. The protein encoded by the human ortholog of *2310003C23Rik*, *Twa1*, is localized in the nucleus and forms a protein complex with Ran-binding protein in microtubule organising centre (RanBMP), and binder of the small GTPase Ran that is involved in nucleocytoplasmic transport of both RNA and proteins (Kobayashi et al. 2007; Umeda et al. 2003; Wennerberg et al. 2005). *Ndufa5* (NADH dehydrogenase [ubiquinone] 1 alpha subcomplex, 5) is a mitochondrial gene encoding a protein found in mitochondria of every one of fourteen mouse tissues assessed in a previous study (Pagliarini et al. 2008).

Despite this, the module eigengenes of both chick turquoise and mouse salmon were differentially expressed across samples (Figure 4.21; Figure 4.22). *In situ* hybridisations further confirmed that some of these genes were differentially expressed across samples (Figure C.32).

However, chick turquoise (by far the largest module in chick) was highly and significantly correlated with technical variables. Chick turquoise was negatively correlated with RNA Integrity Number ($r=-0.92$; two-tailed $p=0.010$) and positively correlated both with the percentage of reads uniquely mapping in protein-coding genes ($r=0.87$; two-tailed $p=0.010$) and with the 3' bias metric ($r=0.79$; two-tailed $p=0.035$). In contrast, mouse salmon was not significantly correlated with any technical variables we considered in this analysis ($r=-0.090$, two-tailed $p=0.74$ for 3' bias metric; $r=-0.29$, two-tailed $p=0.27$ for pre-mRNA fraction). In fact, its correlation with the 3' bias metric differed significantly from the correlation of chick turquoise (two-tailed $p=0.0424$).

4.4 Discussion

Gene connectivity is less conserved than median expression level

This work fits with previous studies comparing humans and chimps (Oldham et al. 2006) and mice and humans (Miller et al. 2010) that demonstrated gene connectivity is less conserved than median expression level. Unsurprisingly, the correlation of median expression level between mice and chickens ($r_s=0.49$) was significantly lower than between mice and humans ($r_s=0.60$) (two-tailed $p<2\times 10^{-14}$). The apparent lack of correlation in gene connectivity in this study, considerably lower than the $r_s=0.27$ observed between mice and humans (two-tailed $p=5\times 10^{-40}$), may be due in part to batch effects.

The study comparing transcriptional networks in mice and humans found that most, but not all, co-expression modules were shared by both species. In contrast, only a handful of modules overlapped in this study. Possible contributors to there being considerably less overlap include (1) the greater phylogenetic divergence between mice and chickens, (2) the fact that most of the anatomical structures included in this study simply do not exist in any recognisable form in the other species (while humans and mice share a much more similar basic layout), and (3) less power in this study to construct robust modules owing to fewer available samples.

Functional enrichments guide module interpretation

No modules were significantly enriched for functional annotations following a Bonferroni correction that accounted both for the number of tests performed given a fixed module and fixed annotation source and for the number of modules within a species, but not for the number of species or for the number of annotation sources. Ideally, all of these would be taken into account; however, a more stringent Bonferroni correction would necessarily produce no significant results as well. From this study, one cannot then say with confidence that genes co-expressed across these samples share some common functions more than expected by chance. However, this has been previously demonstrated in mammalian brain, where the number of dissections was larger, and the expected ‘noise’ in co-expressed modules correspondingly lower (Miller et al. 2010; Oldham et al. 2006).

It is quite conservative to control the familywise error rate in high-throughput studies since its null hypothesis is that there are no true differences. Furthermore, the assumption of independence is violated with functional annotations: all genes annotated with ‘(in) plasma membrane’ should also be annotated with ‘(in) membrane’. Considerable overlap is also to be expected between some annotation sources (for example, ‘myelination’ in GO biological process, ‘myelin sheath’ in GO cellular component and ‘structural constituent of myelin sheath’ in GO molecular function were all enriched in the mouse magenta module), while less overlap might be expected between others (for example, between MGI knockout phenotypes and GO cellular component). This dependence will make a Bonferroni-corrected familywise error rate more conservative still. By relaxing the boundaries within which the correction is

applied, one might instead identify suggestive but nevertheless inconclusive trends of functional clustering. Thus, I instead applied the Bonferroni correction to take into account the number of tests performed given a fixed module and fixed annotation source, but not for the number of modules within a species, the number of species or the number of annotation sources.

Three mouse modules and four chick modules were significantly enriched for one or more functional annotations following a Bonferroni correction using this relaxed criterion. While interpreting the results of these enrichments one must be mindful that many of the annotations were generated through the study of humans or common model organisms such as mouse, fly, worm or yeast (and carried over to orthologs or proteins sharing the same characterised domain). Few if any annotations would be derived from chick. I would expect a gene product's involvement in an ancestral biochemical process (which may only require one or two interactions) to be more conserved than its higher-order physiological effects (which may depend on highly complex signalling pathways or a similar response across multiple tissues to a physiological stressor).

I will first discuss enrichments in modules that did not significantly overlap a module in the other species. I shall discuss the other functional enrichments in context of the observed cross-species overlaps.

Mouse tan has an intriguing expression profile, being found primarily in upper layers of lateral, but not dorsal, cortex. Unfortunately, this same lateral-dorsal difference is also seen in measures of 3' bias, proposed above as a proxy for RNA quality (Figure C.19). While correlation with this metric of 3' bias does not necessarily imply that this apparent

gene co-expression module is driven in part by differences in RNA quality, the possibility cannot be excluded. It may contribute nothing to the module, or it may contribute entirely to the module. Likewise, the apparent functional enrichments could be driven by differences in genes affected by RNA quality. The largest functional enrichment (by fold difference) was “respiratory chain”. Many things could affect the apparent expression of genes for mitochondrial proteins encoded in the nuclear genome. For example, if transcripts encoding mitochondrial proteins were shorter than other transcripts, they would appear to be more highly expressed in degraded, poly(A) selected RNA. Thus, without additional confirmation, one cannot attribute this result to biology with any certainty. The Allen Mouse Brain Atlas might have provided such confirmation, but the top genes in the tan module are not represented by high-quality coronal images that would allow one to confidently compare dorsal and lateral cortex.

There were two chick modules that had functional enrichments but did not have a significant overlap with a mouse module. Chick brown is plagued by very high correlations with numerous technical variables, including RNA Integrity Number. Thus, without biological replicates, it is difficult to separate biological from technical effects. In contrast, there was no evidence that the intriguing expression profile of chick red (being particularly low in hippocampus) was tainted by correlation with technical variables. The gene symbol of the mouse ortholog of the most correlated gene with the chick red eigengene is *Cul4b*, whose human ortholog is an X-linked mental retardation gene that (in mouse) targets ubiquitylation of the H3K4 methyltransferase component WDR5 and regulates neuronal gene expression (Green and Gozani 2011). In contrast to

chick, *Cul4b* was not downregulated in mouse hippocampus, as confirmed in the AMBA (Lein et al. 2007).

Some modules are significantly conserved

Here I used two approaches to identify co-expression modules preserved in the other species. Since the specific modules produced are dependent on parameters selected, the first approach calculates various network preservation metrics in a second species of the genes in modules built in the first species (Langfelder et al. 2011). The second approach identifies statistically significant overlap of co-expression modules based on orthologous genes within those modules.

The first approach indicated two mouse modules that were preserved in chick and three chick modules that were preserved in mouse. Only two modules – mouse salmon and chick greenyellow – were not correlated with technical variables. (These two modules had significant cross-species matches in the next analysis as well.) In this type of analysis, correlation with technical variables is of particular concern because there is not a corresponding module in the other species that one could test for correlation with technical variables. Nevertheless, the significance of these two modules after a proper Bonferroni correction is strong evidence that some modules are conserved. Note that the near-zero correlation of connectivity observed might increase with more samples, which would reduce the noise in ranking connectivity in both species.

The second approach yielded five to six overlapping modules between the two species, which I will now discuss in turn.

Conserved striatal genes are functionally important in striatum

The chick greenyellow/mouse brown overlap strongly marked striatum in both species, and neither module was correlated with technical variables. A similar module to chick greenyellow, chick green, also overlapped with mouse brown. This module may be partially tainted by technical artefacts (perhaps this accounts for its split from chick green). It is not likely to be driven entirely by technical artefacts for several related reasons. First, green chick and mouse brown both shared two significant functional enrichments: “signal transduction” and “adrenoceptor activity”. Indeed, both α and β adrenoceptors are highly expressed in striatum (Hara et al. 2010; Nicholas et al. 1993; Paschalis et al. 2009; Pisani et al. 2003; Rommelfanger et al. 2009). Second, chick green and chick greenyellow have similar expression patterns. If the overlap were purely due to technical artefacts, there would be no reason chick greenyellow would have to mark striatum in chick. Third, *in situ* hybridisations in mouse generally confirmed the expression of the top consensus markers from both overlaps (Figure C.30). Fourth, homeobox transcription factor *Dlx6*, the top consensus marker for chick green/mouse brown, is a known marker of subpallium (Watson et al. 2012), and striatum was the only subpallial dissection in either species. Other chick green/mouse brown consensus markers including transcription factor *Foxo1* (Hoekman et al. 2006), phosphodiesterase *Pde10a* (Xie et al. 2006), and *Drd1a* (Matamales et al. 2009) are notably enriched in striatum. This result is as expected, since the striatum in both species is of uncontroversial homology.

One might expect the genes that are conserved in expression to be particularly important in striatal function, and indeed they are. *Dlx6* regulates molecular properties of the

striatum (Wang et al. 2011). Inhibiting *Pde10a* changes the physiology of striatal neurons (Threlfell et al. 2009). The knockout of *Drd1a*, or the pharmacological inactivation of its gene product, changes the expression of G-protein signalling transcripts in striatum (Stanwood et al. 2006). Striatal expression of *Adcy5* differed between mice selectively bred for excessive exercise or obesity (Mathes et al. 2010). *Pde7b* had altered striatal expression in a mouse model of Parkinson's disease (Kurz et al. 2010). *Rps6ka5* knockout mice show decreased locomotor sensitisation to repeated injections of cocaine; and the protein it encodes, MSK1, is a striatal kinase acting downstream of ERK that phosphorylates CREB and histone H3 and induces expression of c-Fos and dynorphin (Brami-Cherrier et al. 2005). Mouse knockouts for *Tac1*, which encodes substance P, show reduced sensitivity to intense pain (Cao et al. 1998). (Striatum is particularly important in pain modulation (Barcelo et al. 2011)). Perhaps the most direct evidence for a supraspinal-substance P behavioural link comes from a study demonstrating an altered stress-induced analgesia and reduced aggressiveness in response to a territorial challenge in homozygous mouse knockouts for the receptor of substance P (De Felipe et al. 1998). Substance P induces dopamine release in striatum, and also speeds dopamine clearance (Tang et al. 1998).

It also suggests new directions for future research. For example, mouse knockouts of *Dgkb* have impaired spatial and long-term memory, and hippocampal neurons have reduced branch- and spine-formation (Shirai et al. 2010). However, to the best of my knowledge, nobody has examined the striatum of these mice. Furthermore, the subcellular localisation (and co-localisation) of the enzyme encoded by *Dgkb* suggests a

role in the Gq-coupled receptor signalling pathway in the medium spiny neurons that dominate the striatum (Hozumi et al. 2008).

Many conserved hippocampal genes are presently of unknown function

The second and final positive control dissection, hippocampus, also gave the expected results. Chick magenta and mouse black marked hippocampus in each species, and there was no evidence that either module was tainted by technical artefacts. In mouse, this was further confirmed with the Allen Mouse Brain Atlas.

While many of these genes have been described as being in hippocampus, there has generally been little work on their potential hippocampal functions or behavioural effects. *Abil* knockout mice show embryonic lethality and malformations in their developing brains and hearts (Dubielecka et al. 2011). Uniprot annotations suggest *Ptk2b* is most abundant in the brain, especially hippocampus and amygdala, but also found in the kidney, spleen and lymphocytes. The protein encoded by *Prickle2* interacts with PSD-95 and NMDA receptors at the hippocampal postsynaptic density (Hida et al. 2011), and regulates neurite outgrowth in mouse neuroblastoma Neuro2a cells (Okuda et al. 2007). Suggestively, mutations in *Prickle2* orthologs cause seizures in flies, mice and humans (Tao et al. 2011). (Note that hippocampus has an important role in epilepsy (Kuruba et al. 2009).) The fly mushroom body, which is field homologous to the vertebrate pallium (Tomer et al. 2010), is involved in modulating sensitivity to seizures in flies (Hekmat-Safe et al. 2010). It is notable that hippocampal lesions cause a highly similar constellation of impairments in birds and mammals (Colombo and Broadbent 2000).

Shared expression in functionally analogous areas may be evidence of convergent evolution

Perhaps the most intriguing overlap associates chick nidopallium (chick black) with mouse layer IV neocortex (mouse orange). This is interesting because chick nidopallium derives from the (*Emx1*-negative) ventral pallium, while mouse neocortex is generally thought to derive from the (*Emx1*-positive) dorsal and/or lateral pallium. (It is nevertheless possible that there remains an undiscovered tangential migration in mouse from ventral pallium into neocortex, or that there is an overreliance the expression pattern of *Emx1* to best define the ventral pallium lineage in chick and that other, yet undiscovered, developmental markers may suggest nidopallium and neocortex share the same progenitors as defined by other developmental markers.)

However, I do not believe this makes a strong argument for one-to-one homology of these two regions, for several reasons. First, there were only five genes that explained this particular significant overlap (0.5 expected, yielding a ten-fold enrichment). Gene co-expression modules can sometimes mark a larger trend, since some of the genes within them should tend (by mathematical necessity) to be relatively specific to a cell type or to a process, or even to physically co-located cell types or processes (the genes needn't necessarily be regulated in the same cell type if they are expressed in cell types that co-occur). However, the significant overlap of two modules need not *necessarily* imply that there are many more genes under similar regulatory control if there is a plausible alternative explanation.

In this case, an alternative explanation could be convergence. The three genes for which the Allen Mouse Brain Atlas had *in situ* hybridisation images are frequently expressed in brain regions involved in processing or analysing streams of sensory input. For example, *Fam19a2* is found predominantly in layer 2a of entorhinal cortex, layer 2 of piriform cortex, taenia tecta and the anterior olfactory nucleus. All of these regions except the taenia tecta receive input from the lateral olfactory tract; and the taenia tecta receives input from the medial olfactory tract (Neafsey et al. 1986). These regions, which were not included in the dissections, express *Fam19a2* even more highly than does layer IV of neocortex. Some of the regions in which *Dctn3* and *Rorb* were expressed have similar sensory processing or relay functions: for example, layer 2 of piriform cortex for *Dctn3* and layer IV neocortex (especially primary somatosensory cortex), preoptic area, medial preoptic area, nucleus of the lateral olfactory tract, lateral geniculate nucleus, lateral amygdaloid nucleus, superior colliculus, laminae 1-6 of the spinal cord and the ventral postero-lateral thalamic nucleus for *Rorb*. There is also some functional evidence consistent with a sensory processing explanation of potential convergence: *Rorb* protein overexpression is sufficient to produce ‘barrel’-like clustering of layer IV neurons (which may assist in processing large streams of information), even in the absence of thalamic innervation (Jabaudon et al. 2011). *Fam19a2* is hypothesized to have a role in axonal guidance (Tom Tang et al. 2004). (Both chick nidopallium and mouse layer IV are thalamic recipients.) Finally, if these genes indicated homologous regions from an as-yet-undiscovered tangential migration of neurons into layer IV of neocortex, one might also expect these genes to be found in other ventral pallial derivatives, namely pallial amygdala. This, however, was not the case.

Since these genes have little overlap with one another in mouse outside the dissected regions, they may not be under co-ordinated regulatory control. If this were the case, genes might be recruited one at a time in a scenario of convergence. The small number of overlapping genes increases the plausibility of such a process.

A possible, though unlikely, explanation is technical confounds. Mouse orange was moderately positively correlated with 3' bias and pre-mRNA fraction (both $r=0.56$; two-tailed $p=0.025$). Note that a perfect correlation would imply highest expression in dorsal and lateral cortex sections A, while the module eigengene expression in sample A is instead middling. Chick black was not significantly correlated with any technical variable. Chick black's 3' bias correlation was -0.35 (two-tailed $p=0.44$), which may differ from mouse orange (two-tailed $p=0.080$). Note that, of all the chick module eigengenes (not just those which had a corresponding partner in the other species), chick black was the most specific to nidopallium. Likewise, mouse orange was the most specific to dorsal and lateral neocortex section B (layer IV). I consider a possible correlation with RNA quality to be a less satisfactory explanation for this overlap since it would require the chick orthologs of mouse genes whose transcripts happen to degrade faster than others to be preferentially specific to chick nidopallium. Moreover, a correlation at this level of significance may happen by chance since I conservatively did not apply a multiple-testing correction while looking for potential confounds. Thus the apparent correlation with mouse orange may not be significant and is unlikely to explain this particular result satisfactorily.

With the exception of *Rorb*, little is known about the functions of these genes. This observation is thus suggestive for future research.

Conserved amniote markers of oligodendrocytes

The fourth overlapping set (chick cyan and mouse magenta) unambiguously picks out oligodendrocyte markers. All of the top five consensus genes were oligodendrocyte markers. Mouse magenta was highly and significantly enriched for several myelin-related annotations after a Bonferroni correction, and chick cyan was also highly enriched (120-fold for “structural constituent of the myelin sheath”) for these same annotations (these were significant at a single test level). Furthermore, many of the annotations from one functional source that were enriched in mouse were the most enriched annotation for that functional source in chick. Finally, the module eigengene expression in mouse is as one would expect from oligodendrocytes (see Chapter 3 above), being found primarily in the lowest neocortical laminar sections. While both modules were correlated with technical variables this is unlikely to drive this particular overlap unless oligodendrocyte-specific transcripts globally happen to degrade *much* more slowly than other transcripts. Since less functional work has been done on oligodendrocytes than on neurons, these genes, which are specific to oligodendrocytes over long evolutionary timescales, are of particular interest for functional studies.

Batch effects confound interpretation of the final module

The final overlap, chick turquoise and mouse salmon, remains an enigma. Mouse salmon was one of only two mouse modules significantly preserved in chick by the first approach to preservation. Chick turquoise was very highly correlated with technical variables, while mouse salmon was notably not, and even differed significantly from chick turquoise. While the top five consensus genes were generally housekeeping, both

modules were differentially expressed across samples, with chick turquoise found predominantly in hyperpallium and (to a lesser degree) arcopallium and mouse salmon found in striatum, hippocampus, and *lateral* neocortex. Little is known about these genes in brain. Given the extremely high correlation of chick turquoise with technical variables and the lack of relevant functional information amongst the consensus gene set, it is difficult to interpret this result without more data. Note that chick turquoise was the extremely large set in chick. It could be that the significant overlap is a true overlap, and that the consensus genes are still driven by batch effects.

Genes specifically expressed in mouse but not in chicken that have spatially-relevant functional consequences

Several genes are specific to a structure in mouse, and functional in that structure despite being low in the corresponding homologous structure in chick. For example, *Rarb*, a retinoic acid receptor, is specific to mouse striatum but expressed at a very low level in chick striatum (Figure 4.25). Nevertheless, *Rarb* plays an important role in mouse striatum, as *Rarb* null mutant mice have defective striatal dopaminergic neurogenesis, resulting in deficits in motor behaviour (Liao et al. 2008). Like *Rarb*, *Gpr6* is specific to mouse striatum but rare in chick striatum. *Gpr6*-deficient mice produce less cAMP in their striatum and have enhanced instrumental conditioning (Lobo et al. 2007). In another example, *Npy2r*, high in mouse hippocampus but low in chick hippocampus, is thought to mediate neuropeptide Y activity in rat hippocampus, with effects on hippocampal-based spatial discrimination learning (Sorensen et al. 2008).



Figure 4.25: *Rarb* is specific to mouse striatum (below) but relatively rare in chick striatum (above). X-axis represents FPKM.

Conclusions

In summary, this dataset not only gives insight into cerebral gene expression evolution, but also suggests numerous directions for future research. First, there was no evidence for conservation of global patterns of gene co-expression connectivity between the two

species. This does not mean that there are not a handful of solid cross-species marker genes in adult (these results suggest several such markers), but it does suggest the apparent centrality of a gene to a co-expression network in these regions in mouse or chick has no predictive power to the centrality of that gene in the other species. This might not be true in brain regions that are less diverged, and it could be affected by technical variables. Importantly, the strongest marker genes of various pallial regions had fairly unpredictable expression patterns in the other species, and usually did not cluster together (these markers can be sorted and explored at <http://genserv.anat.ox.ac.uk/brainevo>).

This does not necessarily imply that there are not a limited number of adult cross-species marker genes. To the contrary, some such genes were identified above for oligodendrocytes, striatum and hippocampus. For striatum and hippocampus, which are cleanly represented by dissections in each species, this list can easily be expanded using the advanced search feature on the website (<http://genserv.anat.ox.ac.uk/brainevo>). For example, this suggests an additional two candidate cross-species hippocampal marker genes: *Nr3c2*, a mineralocorticoid receptor that regulates emotional memory in mice (Zhou et al. 2010), and *Slit1*, a signalling factor involved in cortical interneuron migration (Andrews et al. 2008) (the higher expression in hippocampus is confirmed in the Allen Mouse Brain Atlas). These conserved markers, and the others listed above, may have great functional importance, as suggested by the known functional importance of nearly all the top striatal genes.

I also gave examples of mouse genes specific to hippocampus or striatum and specific to those regions in mouse but non-specific or even relatively lowly-expressed in chick that

nevertheless has a known function in the mouse region of interest. Without an out-group, one cannot infer if this expression specificity is ancestral or derived.

Finally, I demonstrated some similarity in gene expression between two regions which are developmentally distinct but functionally analogous, namely chick nidopallium and mouse layer IV neocortex. More specifically, the *in situ* hybridisation of the *Rorb* ortholog in chick (Figure 4.15) suggests this module may be picking out the entopallium and, to a lesser extent, field L. Mouse layer IV neocortex and these subdivisions of chick nidopallium are analogous in that they both receive visual and auditory information from the thalamus. Expression of the developmental marker *Emx1* suggests chick nidopallium derives from ventral pallium while the mouse neocortex does not. By definition, similarity in a phenotypic trait (in this case, gene expression) can be attributed either to homology (by some definition, see above) or to convergence. Given these data alone, we cannot definitively answer this question. Nevertheless, the currently available developmental evidence that nidopallium is a ventral pallium derivative and neocortex is not suggests this is likely a result of convergent evolution for regions having a common function. As described above, the top consensus genes are often expressed in regions that receive or process high throughputs of sensory information. The evidence for this particular explanation is far from conclusive, but it remains a plausible scenario.

Chapter 5 - Conclusions and Future Perspectives

In this thesis I presented the first transcriptome sequencing of neocortical layers in a mammal and what may be the first comparison of distinct pallial transcriptomes between a bird and a mammal. I built web interfaces to browse and search both transcriptional and functional dimensions of both data sets. In an in-depth analysis of transcription across mouse neocortical layers, I identified significant associations between specific layers and functional annotations, candidate cases of isoform switching, and new candidate lincRNAs; suggesting new directions for experimental investigations into the molecular differences between neocortical layers, the laminar abundance and functions of proteins encoded by alternative transcript isoforms, and lincRNA functions. In a comparison of transcription in mammals and birds, I found a moderate degree of correlation in expression from brain tissues of mice and chickens. However, the apparent co-expression of pallial markers in one species was not strongly preserved in the other species. To the contrary, the most specific markers of a region in one species typically showed scattered expression in the other species. Though the degree of conservation was weak, some co-expression relationships were nevertheless significantly preserved and marked homologous tissues, cell types and even analogous regions. The conserved cross-species markers produced by this study, especially those whose function is unknown, beg for further experimental investigation. Finally, while the precise homology relationships between the mammalian and avian pallium are in dispute, the current developmental evidence

suggests we may have identified a new case of convergent evolution in a subset of the transcriptome of functionally analogous regions.

These studies not only suggest further investigation of specific genes and transcripts; they also suggest future transcriptomic studies. For example, owing to technological limitations and financial constraints, we sequenced pooled transcriptomes from many adults of each species. Recent advances have reduced the minimum quantity of starting material required for RNA-seq (Head et al. 2011; Tariq et al. 2011), which would facilitate developmental studies. Such studies in pallial development may be better placed to address questions of pallial homologies, as we have shown the genetic architecture of the adult structures to be quite diverged. Furthermore, a lack of biological replicates in the cross-species comparative study complicated the interpretation of a subset of the results that could have been attributable to batch effects. As the cost of transcriptome sequencing continues to fall, similar studies will be able to afford more biological replicates, increasing the reliability of the results.

A critical assumption in the cross-species study (and indeed in many comparative neurobiological studies) was that the avian-mammalian cross-species differences in pallial expression patterns (after the averaging effect of pooling samples) are generally larger than the inter-individual differences in pallial expression patterns. While this is likely to be true in general, a population study would allow these inferences to be strengthened for particular genes.

The inclusion of additional species would give us much greater insight into when these transcriptomic changes occurred. With only two species it is impossible to discriminate between ancestral and derived features. For example, an amphibian such as frog could serve as an out-group for such comparisons. A reptile such as turtle, representing the other major branch of the sauropsids, could allow some transcriptomic features to be parsimoniously dated within some time range. Additional representative mammals, sauropsids and amphibians would strengthen such inferences. Conducting such studies at different developmental time points could provide unprecedented insight into homology and the molecular mechanisms that underlie development of diverged anatomical characters.

Studies on human post-mortem brains could provide important insights into both human disease and normal brain function. For example, studies have demonstrated a weak but significant similarity in transcription between structurally interconnected brain regions in rodents (French and Pavlidis 2011; French et al. 2011; Wolf et al. 2011). One could overlap human transcriptome data with emerging structural (DTI) or functional (fMRI) connectivity data from the Human Connectome Project and other efforts to identify the genes most associated with specific networks. This could provide a new lens with which to look at expression patterns of candidate or known human disease genes. In a second example, traditional case-control transcriptome studies of neuropsychiatric diseases could also begin to dissect and assess the neocortex at the level of individual layers, rather than of whole neocortical areas. Such fine-scale data may be more readily interpretable.

Propelled by new technologies and declining costs, transcriptome sequencing in different brain areas of different species will continue to provide new insights into development, evolution and human disease. The projects presented in this thesis, and others I have worked on as a DPhil student, represent early efforts in this direction.

References

- Aboitiz, F. (2011) Genetic and Developmental Homology in Amniote Brains. Toward Conciliating Radical Views of Brain Evolution, *Brain Research Bulletin*, **84**, 125-136.
- Aboitiz, F. and Montiel, J. (2007) Origin and Evolution of the Vertebrate Telencephalon, with Special Reference to the Mammalian Neocortex, *Advances in Anatomy, Embryology, and Cell Biology*, **193**, 1-112.
- Abou-Sleiman, P. M., Muqit, M. M. and Wood, N. W. (2006) Expanding Insights of Mitochondrial Dysfunction in Parkinson's Disease, *Nature Reviews Neuroscience*, **7**, 207-219.
- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. and Venter, J. C. (1991) Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project, *Science*, **252**, 1651-1656.
- Albert, R. (2005) Scale-Free Networks in Cell Biology, *Journal of Cell Science*, **118**, 4947-4957.
- Andrews, W., Barber, M., Hernandez-Miranda, L. R., Xian, J., Rakic, S., Sundaresan, V., Rabbitts, T. H., Pannell, R., Rabbitts, P., Thompson, H., Erskine, L., Murakami, F. and Parnavelas, J. G. (2008) The Role of Slit-Robo Signaling in the Generation, Migration and Morphological Differentiation of Cortical Interneurons, *Developmental Biology*, **313**, 648-658.
- Arlotta, P., Molyneaux, B. J., Chen, J., Inoue, J., Kominami, R. and Macklis, J. D. (2005) Neuronal Subtype-Specific Genes That Control Corticospinal Motor Neuron Development *in Vivo*, *Neuron*, **45**, 207-221.
- 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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene Ontology: Tool for the Unification of Biology, *Nature Genetics*, **25**, 25-29.
- Augenlicht, L. H. and Kobrin, D. (1982) Cloning and Screening of Sequences Expressed in a Mouse Colon Tumor, *Cancer Research*, **42**, 1088-1093.
- Bachy, I., Berthon, J. and Retaux, S. (2002) Defining Pallial and Subpallial Divisions in the Developing *Xenopus* Forebrain, *Mechanisms of Development*, **117**, 163-172.
- Barabasi, A. L. and Albert, R. (1999) Emergence of Scaling in Random Networks, *Science*, **286**, 509-512.

Barcelo, A. C., Filippini, B. and Pazo, J. H. (2011) The Striatum and Pain Modulation, *Cellular and Molecular Neurobiology*.

Basu, S. N., Kollu, R. and Banerjee-Basu, S. (2009) AutDB: A Gene Reference Resource for Autism Research, *Nucleic Acids Research*, **37**, D832-836.

Bedogni, F., Hodge, R. D., Elsen, G. E., Nelson, B. R., Daza, R. A., Beyer, R. P., Bammler, T. K., Rubenstein, J. L. and Hevner, R. F. (2010) *Tbr1* Regulates Regional and Laminar Identity of Postmitotic Neurons in Developing Neocortex, *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 13129-13134.

Beerli, M. S., Rapp, M., Silverman, J. M., Schmeidler, J., Grossman, H. T., Fallon, J. T., Purohit, D. P., Perl, D. P., Siddiqui, A., Lesser, G., Rosendorff, C. and Haroutunian, V. (2006) Coronary Artery Disease Is Associated with Alzheimer Disease Neuropathology in *APOE4* Carriers, *Neurology*, **66**, 1399-1404.

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

Birling, M. C., Tait, S., Hardy, R. J. and Brophy, P. J. (1999) A Novel Rat Tetraspan Protein in Cells of the Oligodendrocyte Lineage, *Journal of Neurochemistry*, **73**, 2600-2608.

Bishop, K. M., Garel, S., Nakagawa, Y., Rubenstein, J. L. and O'Leary, D. D. (2003) *Emx1* and *Emx2* Cooperate to Regulate Cortical Size, Lamination, Neuronal Differentiation, Development of Cortical Efferents, and Thalamocortical Pathfinding, *The Journal of Comparative Neurology*, **457**, 345-360.

Biskup, S., Moore, D. J., Celsi, F., Higashi, S., West, A. B., Andrabi, S. A., Kurkinen, K., Yu, S. W., Savitt, J. M., Waldvogel, H. J., Faull, R. L., Emson, P. C., Torp, R., Ottersen, O. P., Dawson, T. M. and Dawson, V. L. (2006) Localization of LRRK2 to Membranous and Vesicular Structures in Mammalian Brain, *Annals of Neurology*, **60**, 557-569.

Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E. and Eppig, J. T. (2011) The Mouse Genome Database (MGD): Premier Model Organism Resource for Mammalian Genomics and Genetics, *Nucleic Acids Research*, **39**, D842-848.

Boggs, J. M. (2006) Myelin Basic Protein: A Multifunctional Protein, *Cellular and Molecular Life Sciences*, **63**, 1945-1961.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D. and Miller, C. J. (2010) A Comparison of Massively Parallel Nucleotide Sequencing with Oligonucleotide Microarrays for Global Transcription Profiling, *BMC Genomics*, **11**, 282.

Brami-Cherrier, K., Valjent, E., Herve, D., Darragh, J., Corvol, J. C., Pages, C., Arthur, S. J., Girault, J. A. and Caboche, J. (2005) Parsing Molecular and Behavioral Effects of

Cocaine in Mitogen- and Stress-Activated Protein Kinase-1-Deficient Mice, *The Journal of Neuroscience*, **25**, 11444-11454.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S. and Kaessmann, H. (2011) The Evolution of Gene Expression Levels in Mammalian Organs, *Nature*, **478**, 343-348.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays, *Nature Biotechnology*, **18**, 630-634.

Brodmann, K. (1909) *Vergleichende Lokalisationslehre Der Grosshirnrinde in Ihren Prinzipien Dargestellt Auf Grund Des Zellenbaues*. J. A. Barth, Leipzig.

Brox, A., Ferreiro, B., Puelles, L. and Medina, L. (2002) The Telencephalon of the Frog *Xenopus* Based on Calretinin Immunostaining and Gene Expression Patterns, *Brain Research Bulletin*, **57**, 381-384.

Brox, A., Puelles, L., Ferreiro, B. and Medina, L. (2003) Expression of the Genes *GAD67* and *Distal-Less-4* in the Forebrain of *Xenopus Laevis* Confirms a Common Pattern in Tetrapods, *The Journal of Comparative Neurology*, **461**, 370-393.

Brox, A., Puelles, L., Ferreiro, B. and Medina, L. (2004) Expression of the Genes *Emx1*, *Tbr1*, and *Eomes (Tbr2)* in the Telencephalon of *Xenopus Laevis* Confirms the Existence of a Ventral Pallial Division in All Tetrapods, *The Journal of Comparative Neurology*, **474**, 562-577.

Bruce, L. L. and Neary, T. J. (1995) The Limbic System of Tetrapods: A Comparative Analysis of Cortical and Amygdalar Populations, *Brain, Behavior and Evolution*, **46**, 224-234.

Bulfone, A., Smiga, S. M., Shimamura, K., Peterson, A., Puelles, L. and Rubenstein, J. L. (1995) *T-Brain-1*: A Homolog of *Brachyury* Whose Expression Defines Molecularly Distinct Domains within the Cerebral Cortex, *Neuron*, **15**, 63-78.

Burbulla, L. F., Krebiehl, G. and Kruger, R. (2010) Balance Is the Challenge--the Impact of Mitochondrial Dynamics in Parkinson's Disease, *European Journal of Clinical Investigation*, **40**, 1048-1060.

Bustos, D. M., Bailey, M. J., Sugden, D., Carter, D. A., Rath, M. F., Moller, M., Coon, S. L., Weller, J. L. and Klein, D. C. (2011) Global Daily Dynamics of the Pineal Transcriptome, *Cell and Tissue Research*, **344**, 1-11.

- Butler, A. B. (1994) The Evolution of the Dorsal Pallium in the Telencephalon of Amniotes: Cladistic Analysis and a New Hypothesis, *Brain Research Reviews*, **19**, 66-101.
- Butler, A. B. and Hodos, W. (2005) *Comparative Vertebrate Neuroanatomy: Evolution and Adaptation*. Wiley-Interscience, Hoboken, N.J.
- Butler, A. B. and Molnar, Z. (2002) Development and Evolution of the Collopallium in Amniotes: A New Hypothesis of Field Homology, *Brain Research Bulletin*, **57**, 475-479.
- 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., Thompson, W. J. and Barres, B. A. (2008) A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function, *The Journal of Neuroscience*, **28**, 264-278.
- Cao, Y. Q., Mantyh, P. W., Carlson, E. J., Gillespie, A. M., Epstein, C. J. and Basbaum, A. I. (1998) Primary Afferent Tachykinins Are Required to Experience Moderate to Intense Pain, *Nature*, **392**, 390-394.
- Carter, S. L., Brechbuhler, C. M., Griffin, M. and Bond, A. T. (2004) Gene Co-Expression Network Topology Provides a Framework for Molecular Characterization of Cellular State, *Bioinformatics*, **20**, 2242-2250.
- Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trocheset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J., Wilde, A., Brudno, M., Morris, Q. D. and Hughes, T. R. (2009) Conservation of Core Gene Expression in Vertebrate Tissues, *Journal of Biology*, **8**, 33.
- Chen, Y., Cunningham, F., Rios, D., McLaren, W. M., Smith, J., Pritchard, B., Spudich, G. M., Brent, S., Kulesha, E., Marin-Garcia, P., Smedley, D., Birney, E. and Flicek, P. (2010) Ensembl Variation Resources, *BMC Genomics*, **11**, 293.
- Chiang, C., Litington, Y., Lee, E., Young, K. E., Corden, J. L., Westphal, H. and Beachy, P. A. (1996) Cyclopia and Defective Axial Patterning in Mice Lacking Sonic Hedgehog Gene Function, *Nature*, **383**, 407-413.
- Chodroff, R. A., Goodstadt, L., Sirey, T. M., Oliver, P. L., Davies, K. E., Green, E. D., Molnar, Z. and Ponting, C. P. (2010) Long Noncoding RNA Genes: Conservation of Sequence and Brain Expression among Diverse Amniotes, *Genome Biology*, **11**, R72.
- Chou, S. J., Perez-Garcia, C. G., Kroll, T. T. and O'Leary, D. D. (2009) *Lhx2* Specifies Regional Fate in *Emx1* Lineage of Telencephalic Progenitors Generating Cerebral Cortex, *Nature Neuroscience*, **12**, 1381-1389.
- Chung, W. Y., Wadhawan, S., Szklarczyk, R., Pond, S. K. and Nekrutenko, A. (2007) A First Look at ARFome: Dual-Coding Genes in Mammalian Genomes, *PLoS Computational Biology*, **3**, e91.

- Cleveland, W. S. and Devlin, S. J. (1988) Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, **83**, 596-610.
- Cocas, L. A., Miyoshi, G., Carney, R. S., Sousa, V. H., Hirata, T., Jones, K. R., Fishell, G., Huntsman, M. M. and Corbin, J. G. (2009) *Emx1*-Lineage Progenitors Differentially Contribute to Neural Diversity in the Striatum and Amygdala, *The Journal of Neuroscience*, **29**, 15933-15946.
- Colombo, M. and Broadbent, N. (2000) Is the Avian Hippocampus a Functional Homologue of the Mammalian Hippocampus?, *Neuroscience and Biobehavioral Reviews*, **24**, 465-484.
- Conti, F., Minelli, A., DeBiasi, S. and Melone, M. (1997) Neuronal and Glial Localization of NMDA Receptors in the Cerebral Cortex, *Molecular Neurobiology*, **14**, 1-18.
- Copeland, P. R. and Driscoll, D. M. (2001) RNA Binding Proteins and Selenocysteine, *BioFactors*, **14**, 11-16.
- Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. and Stein, L. (2011) Reactome: A Database of Reactions, Pathways and Biological Processes, *Nucleic Acids Research*, **39**, D691-697.
- Cucchiaroni, M., Ren, X. L., Perides, G. and Terwilliger, E. F. (2003) Selective Gene Expression in Brain Microglia Mediated Via Adeno-Associated Virus Type 2 and Type 5 Vectors, *Gene Therapy*, **10**, 657-667.
- Currie, S. N., Wang, X. F. and Daw, N. W. (1994) NMDA Receptors in Layers II and III of Rat Cerebral Cortex, *Brain Research*, **662**, 103-108.
- De Felipe, C., Herrero, J. F., O'Brien, J. A., Palmer, J. A., Doyle, C. A., Smith, A. J., Laird, J. M., Belmonte, C., Cervero, F. and Hunt, S. P. (1998) Altered Nociception, Analgesia and Aggression in Mice Lacking the Receptor for Substance P, *Nature*, **392**, 394-397.
- Demšar, J., Zupan, B., Leban, G. and Curk, T. (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, *Knowledge Discovery in Databases: PKDD*, **3202**, 537-539.
- Dias Neto, E., Correa, R. G., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., Matsukuma, A., Baia, G. S., Simpson, D. H., Brunstein, A., de Oliveira, P. S., Bucher, P., Jongeneel, C. V., O'Hare, M. J., Soares, F., Brentani, R. R., Reis, L. F., de Souza, S. J. and Simpson, A. J. (2000) Shotgun Sequencing of the Human Transcriptome with

ORF Expressed Sequence Tags, *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 3491-3496.

Douglas, A. G. and Wood, M. J. (2011) RNA Splicing: Disease and Therapy, *Briefings in Functional Genomics*, **10**, 151-164.

Douglas, R. J. and Martin, K. A. (2004) Neuronal Circuits of the Neocortex, *Annual Review of Neuroscience*, **27**, 419-451.

Doyle, J. P., Dougherty, J. D., Heiman, M., Schmidt, E. F., Stevens, T. R., Ma, G., Bupp, S., Shrestha, P., Shah, R. D., Doughty, M. L., Gong, S., Greengard, P. and Heintz, N. (2008) Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types, *Cell*, **135**, 749-762.

Dubielecka, P. M., Ladwein, K. I., Xiong, X., Migeotte, I., Chorzalska, A., Anderson, K. V., Sawicki, J. A., Rottner, K., Stradal, T. E. and Kotula, L. (2011) Essential Role for Abi1 in Embryonic Survival and WAVE2 Complex Integrity, *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7022-7027.

Favaro, R., Valotta, M., Ferri, A. L., Latorre, E., Mariani, J., Giachino, C., Lancini, C., Tosetti, V., Ottolenghi, S., Taylor, V. and Nicolis, S. K. (2009) Hippocampal Development and Neural Stem Cell Maintenance Require Sox2-Dependent Regulation of *Shh*, *Nature Neuroscience*, **12**, 1248-1256.

Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P. and Kohtz, J. D. (2006) The *Eyf-2* Noncoding RNA Is Transcribed from the *Dlx-5/6* Ultraconserved Region and Functions as a *Dlx-2* Transcriptional Coactivator, *Genes & Development*, **20**, 1470-1484.

Fernandez, A. S., Pieau, C., Reperant, J., Boncinelli, E. and Wassef, M. (1998) Expression of the *Emx-1* and *Dlx-1* Homeobox Genes Define Three Molecularly Distinct Domains in the Telencephalon of Mouse, Chick, Turtle and Frog Embryos: Implications for the Evolution of Telencephalic Subdivisions in Amniotes, *Development*, **125**, 2099-2111.

Flandin, P., Kimura, S. and Rubenstein, J. L. (2010) The Progenitor Zone of the Ventral Medial Ganglionic Eminence Requires *Nkx2-1* to Generate Most of the Globus Pallidus but Few Neocortical Interneurons, *The Journal of Neuroscience*, **30**, 2812-2823.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Vogel, J. and Searle, S. M. (2011) Ensembl 2011, *Nucleic Acids Research*, **39**, D800-806.

- French, L. and Pavlidis, P. (2011) Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain, *PLoS Computational Biology*, **7**, e1001049.
- French, L., Tan, P. P. and Pavlidis, P. (2011) Large-Scale Analysis of Gene Expression and Connectivity in the Rodent Brain: Insights through Data Integration, *Frontiers in Neuroinformatics*, **5**, 12.
- Frowein, J., Campbell, K. and Gotz, M. (2002) Expression of *Ngn1*, *Ngn2*, *Cash1*, *Gsh2* and *Sfrp1* in the Developing Chick Telencephalon, *Mechanisms of Development*, **110**, 249-252.
- Fu, N., Drinnenberg, I., Kelso, J., Wu, J. R., Paabo, S., Zeng, R. and Khaitovich, P. (2007) Comparison of Protein and mRNA Expression Evolution in Humans and Chimpanzees, *PloS One*, **2**, e216.
- Georgala, P. A., Manuel, M. and Price, D. J. (2011) The Generation of Superficial Cortical Layers Is Regulated by Levels of the Transcription Factor *Pax6*, *Cerebral Cortex*, **21**, 81-94.
- Gerbi, S. A., Borovjagin, A. V., Ezrokhi, M. and Lange, T. S. (2001) Ribosome Biogenesis: Role of Small Nucleolar RNA in Maturation of Eukaryotic rRNA, *Cold Spring Harbor Symposia on Quantitative Biology*, **66**, 575-590.
- Gilbert, C. D. and Wiesel, T. N. (1979) Morphology and Intracortical Projections of Functionally Characterised Neurones in the Cat Visual Cortex, *Nature*, **280**, 120-125.
- Goodstadt, L. (2010) Ruffus: A Lightweight Python Library for Computational Pipelines, *Bioinformatics*, **26**, 2778-2779.
- Gowda, M., Jantasuriyarat, C., Dean, R. A. and Wang, G. L. (2004) Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis, *Plant Physiology*, **134**, 890-897.
- Green, E. M. and Gozani, O. (2011) CUL4B: Trash Talking at Chromatin, *Molecular Cell*, **43**, 321-323.
- Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D. and Dulac, C. (2010) High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain, *Science*, **329**, 643-648.
- Guillemot, F., Molnar, Z., Tarabykin, V. and Stoykova, A. (2006) Molecular Mechanisms of Cortical Differentiation, *European Journal of Neuroscience*, **23**, 857-868.
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L. and Lander, E. S. (2009) Chromatin Signature Reveals over a Thousand Highly Conserved Large Non-Coding RNAs in Mammals, *Nature*, **458**, 223-227.

- Han, W., Kwan, K. Y., Shim, S., Lam, M. M., Shin, Y., Xu, X., Zhu, Y., Li, M. and Sestan, N. (2011) TBR1 Directly Represses *Fezf2* to Control the Laminar Origin and Development of the Corticospinal Tract, *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 3041-3046.
- Hara, M., Fukui, R., Hieda, E., Kuroiwa, M., Bateup, H. S., Kano, T., Greengard, P. and Nishi, A. (2010) Role of Adrenoceptors in the Regulation of Dopamine/DARPP-32 Signaling in Neostriatal Neurons, *Journal of Neurochemistry*, **113**, 1046-1059.
- Hasan, M. R., Jin, M., Matsushima, K., Miyamoto, S., Kotani, S. and Nakagawa, H. (2006) Differences in the Regulation of Microtubule Stability by the Pro-Rich Region Variants of Microtubule-Associated Protein 4, *FEBS Letters*, **580**, 3505-3510.
- Hawrylycz, M., Bernard, A., Lau, C., Sunkin, S. M., Chakravarty, M. M., Lein, E. S., Jones, A. R. and Ng, L. (2010) Areal and Laminar Differentiation in the Mouse Neocortex Using Large Scale Gene Expression Data, *Methods*, **50**, 113-121.
- Hayles, B., Yellaboina, S. and Wang, D. (2010) Comparing Transcription Rate and mRNA Abundance as Parameters for Biochemical Pathway and Network Analysis, *PloS One*, **5**, e9908.
- Head, S. R., Komori, H. K., Hart, G. T., Shimashita, J., Schaffer, L., Salomon, D. R. and Ordoukhanian, P. T. (2011) Method for Improved Illumina Sequencing Library Preparation Using NuGEN Ovation RNA-Seq System, *BioTechniques*, **50**, 177-180.
- Heintz, N. (2004) Gene Expression Nervous System Atlas (GENSAT), *Nature Neuroscience*, **7**, 483.
- Hekmat-Scafe, D. S., Mercado, A., Fajilan, A. A., Lee, A. W., Hsu, R., Mount, D. B. and Tanouye, M. A. (2010) Seizure Sensitivity Is Ameliorated by Targeted Expression of K⁺-Cl⁻ Cotransporter Function in the Mushroom Body of the Drosophila Brain, *Genetics*, **184**, 171-183.
- Hevner, R. F., Hodge, R. D., Daza, R. A. and Englund, C. (2006) Transcription Factors in Glutamatergic Neurogenesis: Conserved Programs in Neocortex, Cerebellum, and Adult Hippocampus, *Neuroscience Research*, **55**, 223-233.
- Hida, Y., Fukaya, M., Hagiwara, A., Deguchi-Tawarada, M., Yoshioka, T., Kitajima, I., Inoue, E., Watanabe, M. and Ohtsuka, T. (2011) Prickle2 Is Localized in the Postsynaptic Density and Interacts with PSD-95 and NMDA Receptors in the Brain, *Journal of Biochemistry*, **149**, 693-700.
- Higuti, T., Tsurumi, C., Osaka, F., Kawamura, Y., Tsujita, H., Yoshihara, Y., Tani, I., Tanaka, K. and Ichihara, A. (1991) Molecular Cloning of cDNA for the Import Precursor of Human Subunit B of H⁽⁺⁾-Atp Synthase in Mitochondria, *Biochemical and Biophysical Research Communications*, **178**, 1014-1020.

- Hirota, K., Miyoshi, T., Kugou, K., Hoffman, C. S., Shibata, T. and Ohta, K. (2008) Stepwise Chromatin Remodelling by a Cascade of Transcription Initiation of Non-Coding RNAs, *Nature*, **456**, 130-134.
- Hirsch, J. A. and Martinez, L. M. (2006) Laminar Processing in the Visual Cortical Column, *Current Opinion in Neurobiology*, **16**, 377-384.
- Hoekman, M. F., Jacobs, F. M., Smidt, M. P. and Burbach, J. P. (2006) Spatial and Temporal Expression of *FoxO* Transcription Factors in the Developing and Adult Murine Brain, *Gene Expression Patterns*, **6**, 134-140.
- Hoerder-Suabedissen, A., Wang, W. Z., Lee, S., Davies, K. E., Goffinet, A. M., Rakic, S., Parnavelas, J., Reim, K., Nicolic, M., Paulsen, O. and Molnar, Z. (2009) Novel Markers Reveal Subpopulations of Subplate Neurons in the Murine Cerebral Cortex, *Cerebral Cortex*, **19**, 1738-1750.
- Hof, P. R. and Morrison, J. H. (2004) The Aging Brain: Morphomolecular Senescence of Cortical Circuits, *Trends in Neurosciences*, **27**, 607-613.
- Hogg, M., Paro, S., Keegan, L. P. and O'Connell, M. A. (2011) RNA Editing by Mammalian ADARs, *Advances in Genetics*, **73**, 87-120.
- Hong, S. M., Liu, Z., Fan, Y., Neumann, M., Won, S. J., Lac, D., Lum, X., Weinstein, P. R. and Liu, J. (2007) Reduced Hippocampal Neurogenesis and Skill Reaching Performance in Adult *Emx1* Mutant Mice, *Experimental Neurology*, **206**, 24-32.
- Horvath, S. and Dong, J. (2008) Geometric Interpretation of Gene Coexpression Network Analysis, *PLoS Computational Biology*, **4**, e1000117.
- Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liao, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F. and Mischel, P. S. (2006) Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies *ASPM* as a Molecular Target, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 17402-17407.
- Hozumi, Y., Fukaya, M., Adachi, N., Saito, N., Otani, K., Kondo, H., Watanabe, M. and Goto, K. (2008) Diacylglycerol Kinase Beta Accumulates on the Perisynaptic Site of Medium Spiny Neurons in the Striatum, *European Journal of Neuroscience*, **28**, 2409-2422.
- Inoue, K., Terashima, T., Nishikawa, T. and Takumi, T. (2004) *Fez1* Is Layer-Specifically Expressed in the Adult Mouse Neocortex, *European Journal of Neuroscience*, **20**, 2909-2916.
- Isaacs, A. M., Oliver, P. L., Jones, E. L., Jeans, A., Potter, A., Hovik, B. H., Nolan, P. M., Vitor, L., Glenister, P., Simon, A. K., Gray, I. C., Spurr, N. K., Brown, S. D., Hunter, A. J. and Davies, K. E. (2003) A Mutation in *Af4* Is Predicted to Cause

Cerebellar Ataxia and Cataracts in the Robotic Mouse, *The Journal of Neuroscience*, **23**, 1631-1637.

Jabaudon, D., S, J. S., D, J. T., M, J. G. and Macklis, J. D. (2011) *ROR β* Induces Barrel-Like Neuronal Clusters in the Developing Neocortex, *Cerebral Cortex*.

Jarrous, N. and Gopalan, V. (2010) Archaeal/Eukaryal RNase P: Subunits, Functions and RNA Diversification, *Nucleic Acids Research*, **38**, 7885-7894.

Johnson, M. B., Kawasawa, Y. I., Mason, C. E., Krsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D. H., Mane, S. M., State, M. W. and Sestan, 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.

Jones, E. G. (2000) Cortical and Subcortical Contributions to Activity-Dependent Plasticity in Primate Somatosensory Cortex, *Annual Review of Neuroscience*, **23**, 1-37.

Jones, E. G. and Rakic, P. (2010) Radial Columns in Cortical Architecture: It Is the Composition That Counts, *Cerebral Cortex*, **20**, 2261-2264.

Kaas, J. H. (2008) The Evolution of the Complex Sensory and Motor Systems of the Human Brain, *Brain Research Bulletin*, **75**, 384-390.

Kall, L., Storey, J. D. and Noble, W. S. (2008) Non-Parametric Estimation of Posterior Error Probabilities Associated with Peptides Identified by Tandem Mass Spectrometry, *Bioinformatics*, **24**, i42-48.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG Resource for Deciphering the Genome, *Nucleic Acids Research*, **32**, D277-280.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E. and Sestan, N. (2011) Spatio-Temporal Transcriptome of the Human Brain, *Nature*, **478**, 483-489.

Kanold, P. O. and Luhmann, H. J. (2010) The Subplate and Early Cortical Circuits, *Annual Review of Neuroscience*, **33**, 23-48.

Karten, H. J. (1969) The Organization of the Avian Telencephalon and Some Speculations on the Phylogeny of the Amniote Telencephalon, *Annals of the New York Academy of Sciences*, **167**, 164-179.

Karten, H. J. (1997) Evolutionary Developmental Biology Meets the Brain: The Origins of Mammalian Cortex, *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 2800-2804.

- Kawasaki, M., Sekigawa, I., Nozawa, K., Kaneko, H., Takasaki, Y., Takamori, K. and Ogawa, H. (2009) Changes in the Gene Expression of Peripheral Blood Mononuclear Cells During the Menstrual Cycle of Females Is Associated with a Gender Bias in the Incidence of Systemic Lupus Erythematosus, *Clinical and Experimental Rheumatology*, **27**, 260-266.
- Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M. and Paabo, S. (2005) Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees, *Science*, **309**, 1850-1854.
- Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H. H., Weiss, G., Enard, W., Heissig, F., Arendt, T., Nieselt-Struwe, K., Eichler, E. E. and Paabo, S. (2004) Regional Patterns of Gene Expression in Human and Chimpanzee Brains, *Genome Research*, **14**, 1462-1473.
- Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. and Greenberg, M. E. (2010) Widespread Transcription at Neuronal Activity-Regulated Enhancers, *Nature*, **465**, 182-187.
- Knobloch, M. and Mansuy, I. M. (2008) Dendritic Spine Loss and Synaptic Alterations in Alzheimer's Disease, *Molecular Neurobiology*, **37**, 73-82.
- Kobayashi, N., Yang, J., Ueda, A., Suzuki, T., Tomaru, K., Takeno, M., Okuda, K. and Ishigatsubo, Y. (2007) RanBPM, Muskelin, p48EMLP, P44ctlh, and the Armadillo-Repeat Proteins ARMC8 α and ARMC8 β Are Components of the CTLH Complex, *Gene*, **396**, 236-247.
- Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. and Gao, G. (2007) CPC: Assess the Protein-Coding Potential of Transcripts Using Sequence Features and Support Vector Machine, *Nucleic Acids Research*, **35**, W345-349.
- Kuruba, R., Hattiangady, B. and Shetty, A. K. (2009) Hippocampal Neurogenesis and Neural Stem Cells in Temporal Lobe Epilepsy, *Epilepsy & Behavior*, **14 Suppl 1**, 65-73.
- Kurz, A., Double, K. L., Lastres-Becker, I., Tozzi, A., Tantucci, M., Bockhart, V., Bonin, M., Garcia-Arencibia, M., Nuber, S., Schlaudraff, F., Liss, B., Fernandez-Ruiz, J., Gerlach, M., Wullner, U., Luddens, H., Calabresi, P., Auburger, G. and Gispert, S. (2010) A53T-Alpha-Synuclein Overexpression Impairs Dopamine Signaling and Striatal Synaptic Plasticity in Old Mice, *PloS One*, **5**, e11464.
- Langfelder, P. and Horvath, S. (2007) Eigengene Networks for Studying the Relationships between Co-Expression Modules, *BMC systems biology*, **1**, 54.
- Langfelder, P. and Horvath, S. (2008) WGCNA: An R Package for Weighted Correlation Network Analysis, *BMC Bioinformatics*, **9**, 559.

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

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.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome, *Genome Biology*, **10**, R25.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010) Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data, *Nature Reviews Genetics*, **11**, 733-739.

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., Chen, L., Chen, T. M., Chin, M. C., Chong, J., Crook, B. E., Czaplinska, A., Dang, C. N., Datta, S., Dee, N. R., Desaki, A. L., Desta, T., Diep, E., Dolbeare, T. A., Donelan, M. J., Dong, H. W., Dougherty, J. G., Duncan, B. J., Ebbert, A. J., Eichele, G., Estin, L. K., Faber, C., Facer, B. A., Fields, R., Fischer, S. R., Fliss, T. P., Frensley, C., Gates, S. N., Glattfelder, K. J., Halverson, K. R., Hart, M. R., Hohmann, J. G., Howell, M. P., Jeung, D. P., Johnson, R. A., Karr, P. T., Kawal, R., Kidney, J. M., Knapik, R. H., Kuan, C. L., Lake, J. H., Laramee, A. R., Larsen, K. D., Lau, C., Lemon, T. A., Liang, A. J., Liu, Y., Luong, L. T., Michaels, J., Morgan, J. J., Morgan, R. J., Mortrud, M. T., Mosqueda, N. F., Ng, L. L., Ng, R., Orta, G. J., Overly, C. C., Pak, T. H., Parry, S. E., Pathak, S. D., Pearson, O. C., Puchalski, R. B., Riley, Z. L., Rockett, H. R., Rowland, S. A., Royall, J. J., Ruiz, M. J., Sarno, N. R., Schaffnit, K., Shapovalova, N. V., Sivisay, T., Slaughterbeck, C. R., Smith, S. C., Smith, K. A., Smith, B. I., Sodt, A. J., Stewart, N. N., Stumpf, K. R., Sunkin, S. M., Sutram, M., Tam, A., Teemer, C. D., Thaller, C., Thompson, C. L., Varnam, L. R., Visel, A., Whitlock, R. M., Wohnoutka, P. E., Wolkey, C. K., Wong, V. Y., Wood, M., Yaylaoglu, M. B., Young, R. C., Youngstrom, B. L., Yuan, X. F., Zhang, B., Zwingman, T. A. and Jones, A. R. (2007) Genome-Wide Atlas of Gene Expression in the Adult Mouse Brain, *Nature*, **445**, 168-176.

Lenz, P. and Sogaard-Andersen, L. (2011) Temporal and Spatial Oscillations in Bacteria, *Nature Reviews Microbiology*, **9**, 565-577.

Leung, E. and Brown, J. D. (2010) Biogenesis of the Signal Recognition Particle, *Biochemical Society Transactions*, **38**, 1093-1098.

Levin, J., Adiconis, X., Yassour, M., Thompson, D., Guttman, M., Berger, M., Fan, L., Friedman, N., Nusbaum, C., Gnirke, A. and Regev, A. (2010) Development and Evaluation of RNA-Seq Methods, *Genome Biology*, **11**.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map Format and SAMtools, *Bioinformatics*, **25**, 2078-2079.

- Liao, W. L., Tsai, H. C., Wang, H. F., Chang, J., Lu, K. M., Wu, H. L., Lee, Y. C., Tsai, T. F., Takahashi, H., Wagner, M., Ghyselinck, N. B., Chambon, P. and Liu, F. C. (2008) Modular Patterning of Structure and Function of the Striatum by Retinoid Receptor Signaling, *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6765-6770.
- Liu, Q. and Paroo, Z. (2010) Biochemical Principles of Small RNA Pathways, *Annual Review of Biochemistry*, **79**, 295-319.
- Lobo, M. K., Cui, Y., Ostlund, S. B., Balleine, B. W. and Yang, X. W. (2007) Genetic Control of Instrumental Conditioning by Striatopallidal Neuron-Specific S1P Receptor Gpr6, *Nature Neuroscience*, **10**, 1395-1397.
- Lorente de No, R. (1949) Cerebral Cortex: Architecture, Intracortical Connections, Motor Projections. In Fulton, J. (ed), *Physiology of the Nervous System*. Oxford University Press, Oxford, pp. 288-330.
- Maier, T., Guell, M. and Serrano, L. (2009) Correlation of mRNA and Protein in Complex Biological Samples, *FEBS Letters*, **583**, 3966-3973.
- Maksakova, I. A., Mager, D. L. and Reiss, D. (2008) Keeping Active Endogenous Retroviral-Like Elements in Check: The Epigenetic Perspective, *Cellular and Molecular Life Sciences*, **65**, 3329-3347.
- Marigo, V. and Tabin, C. J. (1996) Regulation of Patched by Sonic Hedgehog in the Developing Neural Tube, *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 9346-9351.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays, *Genome Research*, **18**, 1509-1517.
- Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G. and Wu, C. (2004) Interneurons of the Neocortical Inhibitory System, *Nature Reviews Neuroscience*, **5**, 793-807.
- Marques, A. C. and Ponting, C. P. (2009) Catalogues of Mammalian Long Noncoding RNAs: Modest Conservation and Incompleteness, *Genome Biology*, **10**, R124.
- Martens, J. A., Laprade, L. and Winston, F. (2004) Intergenic Transcription Is Required to Repress the *Saccharomyces Cerevisiae* SER3 Gene, *Nature*, **429**, 571-574.
- Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. and Akoulitchiev, A. (2007) Repression of the Human Dihydrofolate Reductase Gene by a Non-Coding Interfering Transcript, *Nature*, **445**, 666-670.
- Martinez-Garcia, F., Martinez-Marcos, A. and Lanuza, E. (2002) The Pallial Amygdala of Amniote Vertebrates: Evolution of the Concept, Evolution of the Structure, *Brain Research Bulletin*, **57**, 463-469.

- Maruyama, T., Matsuura, M., Suzuki, K. and Yamamoto, N. (2008) Cooperative Activity of Multiple Upper Layer Proteins for Thalamocortical Axon Growth, *Developmental Neurobiology*, **68**, 317-331.
- Matamales, M., Bertran-Gonzalez, J., Salomon, L., Degos, B., Deniau, J. M., Valjent, E., Herve, D. and Girault, J. A. (2009) Striatal Medium-Sized Spiny Neurons: Identification by Nuclear Staining and Study of Neuronal Subpopulations in BAC Transgenic Mice, *PloS One*, **4**, e4770.
- Mathes, W. F., Nehrenberg, D. L., Gordon, R., Hua, K., Garland, T., Jr. and Pomp, D. (2010) Dopaminergic Dysregulation in Mice Selectively Bred for Excessive Exercise or Obesity, *Behavioural Brain Research*, **210**, 155-163.
- Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Kruger, D. H. and Terauchi, R. (2005) SuperSAGE, *Cellular Microbiology*, **7**, 11-18.
- McMaster, A., Jangani, M., Sommer, P., Han, N., Brass, A., Beesley, S., Lu, W., Berry, A., Loudon, A., Donn, R. and Ray, D. W. (2011) Ultradian Cortisol Pulsatility Encodes a Distinct, Biologically Important Signal, *PloS One*, **6**, e15766.
- Medina, L., Brox, A., Legaz, I., Garcia-Lopez, M. and Puelles, L. (2005) Expression Patterns of Developmental Regulatory Genes Show Comparable Divisions in the Telencephalon of *Xenopus* and Mouse: Insights into the Evolution of the Forebrain, *Brain Research Bulletin*, **66**, 297-302.
- Medina, L., Legaz, I., Gonzalez, G., De Castro, F., Rubenstein, J. L. and Puelles, L. (2004) Expression of *Dbx1*, *Neurogenin 2*, *Semaphorin 5a*, *Cadherin 8*, and *Emx1* Distinguish Ventral and Lateral Pallial Histogenetic Divisions in the Developing Mouse Claustroramygdaloid Complex, *The Journal of Comparative Neurology*, **474**, 504-523.
- Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R. and Lander, E. S. (2008) Genome-Scale DNA Methylation Maps of Pluripotent and Differentiated Cells, *Nature*, **454**, 766-770.
- Melot, T., Dauphinot, L., Sevenet, N., Radvanyi, F. and Delattre, O. (2001) Characterization of a New Brain-Specific Isoform of the EWS Oncoprotein, *European Journal of Biochemistry*, **268**, 3483-3489.
- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S. and Bernstein, B. E. (2007) Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells, *Nature*, **448**, 553-560.
- Mikl, M., Vendra, G., Doyle, M. and Kiebler, M. A. (2010) RNA Localization in Neurite Morphogenesis and Synaptic Regulation: Current Evidence and Novel Approaches, *Journal of Comparative Physiology A*, **196**, 321-334.

- 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 of the United States of America*, **107**, 12698-12703.
- Minoche, A. E., Dohm, J. C. and Himmelbauer, H. (2011) Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems, *Genome Biology*, **12**, R112.
- Molnár, Z. and Cheung, A. F. (2006) Towards the Classification of Subpopulations of Layer V Pyramidal Projection Neurons, *Neuroscience Research*, **55**, 105-115.
- Molyneaux, B. J., Arlotta, P., Menezes, J. R. and Macklis, J. D. (2007) Neuronal Subtype Specification in the Cerebral Cortex, *Nature Reviews Neuroscience*, **8**, 427-437.
- Monyer, H. and Markram, H. (2004) Interneuron Diversity Series: Molecular and Genetic Tools to Study GABAergic Interneuron Diversity and Function, *Trends in Neurosciences*, **27**, 90-97.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq, *Nature Methods*, **5**, 621-628.
- Murakami, Y., Uchida, K., Rijli, F. M. and Kuratani, S. (2005) Evolution of the Brain Developmental Plan: Insights from Agnathans, *Developmental Biology*, **280**, 249-259.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing, *Science*, **320**, 1344-1349.
- Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R. and Fraser, P. (2008) The *Air* Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin, *Science*, **322**, 1717-1720.
- Nagaraj, S. H., Gasser, R. B. and Ranganathan, S. (2007) A Hitchhiker's Guide to Expressed Sequence Tag (EST) Analysis, *Briefings in Bioinformatics*, **8**, 6-21.
- Neafsey, E. J., Hurley-Gius, K. M. and Arvanitis, D. (1986) The Topographical Organization of Neurons in the Rat Medial Frontal, Insular and Olfactory Cortex Projecting to the Solitary Nucleus, Olfactory Bulb, Periaqueductal Gray and Superior Colliculus, *Brain Research*, **377**, 261-270.
- Nekrutenko, A. and He, J. (2006) Functionality of Unspliced *XBPI* Is Required to Explain Evolution of Overlapping Reading Frames, *Trends in Genetics*, **22**, 645-648.
- Nelson, S. B., Sugino, K. and Hempel, C. M. (2006) The Problem of Neuronal Cell Types: A Physiological Genomics Approach, *Trends in Neurosciences*, **29**, 339-345.
- Neuwelt, E. A., Bauer, B., Fahlke, C., Fricker, G., Iadecola, C., Janigro, D., Leybaert, L., Molnar, Z., O'Donnell, M. E., Povlishock, J. T., Saunders, N. R., Sharp, F., Stanimirovic, D., Watts, R. J. and Drewes, L. R. (2011) Engaging Neuroscience to Advance

Translational Research in Brain Barrier Biology, *Nature Reviews Neuroscience*, **12**, 169-182.

Ng, L., Lau, C., Sunkin, S. M., Bernard, A., Chakravarty, M. M., Lein, E. S., Jones, A. R. and Hawrylycz, M. (2010) Surface-Based Mapping of Gene Expression and Probabilistic Expression Maps in the Mouse Cortex, *Methods*, **50**, 55-62.

Nicholas, A. P., Pieribone, V. and Hokfelt, T. (1993) Distributions of mRNAs for Alpha-2 Adrenergic Receptor Subtypes in Rat Brain: An *in Situ* Hybridization Study, *The Journal of Comparative Neurology*, **328**, 575-594.

Nieto, M., Monuki, E. S., Tang, H., Imitola, J., Haubst, N., Khoury, S. J., Cunningham, J., Gotz, M. and Walsh, C. A. (2004) Expression of *Cux-1* and *Cux-2* in the Subventricular Zone and Upper Layers II-IV of the Cerebral Cortex, *The Journal of Comparative Neurology*, **479**, 168-180.

Okuda, H., Miyata, S., Mori, Y. and Tohyama, M. (2007) Mouse *Prickle1* and *Prickle2* Are Expressed in Postmitotic Neurons and Promote Neurite Outgrowth, *FEBS Letters*, **581**, 4754-4760.

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 of the United States of America*, **103**, 17973-17978.

Orom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R. and Shiekhattar, R. (2010) Long Noncoding RNAs with Enhancer-Like Function in Human Cells, *Cell*, **143**, 46-58.

Osorio, J., Mazan, S. and Retaux, S. (2005) Organisation of the Lamprey (*Lampetra fluviatilis*) Embryonic Brain: Insights from LIM-Homeodomain, Pax and Hedgehog Genes, *Developmental Biology*, **288**, 100-112.

Ouimet, C. C., Miller, P. E., Hemmings, H. C., Jr., Walaas, S. I. and Greengard, P. (1984) DARPP-32, a Dopamine- and Adenosine 3':5'-Monophosphate-Regulated Phosphoprotein Enriched in Dopamine-Innervated Brain Regions. III. Immunocytochemical Localization, *The Journal of Neuroscience*, **4**, 111-124.

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A. and Mootha, V. K. (2008) A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology, *Cell*, **134**, 112-123.

Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D. and Kanduri, C. (2008) *Kcnqlot1* Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation, *Molecular Cell*, **32**, 232-246.

- Paschalis, A., Churchill, L., Marina, N., Kasymov, V., Gourine, A. and Ackland, G. (2009) β 1-Adrenoceptor Distribution in the Rat Brain: An Immunohistochemical Study, *Neuroscience Letters*, **458**, 84-88.
- Pavlidis, P. and Noble, W. S. (2003) Matrix2png: A Utility for Visualizing Matrix Data, *Bioinformatics*, **19**, 295-296.
- Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W. and Haussler, D. (2006) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome, *PLoS Computational Biology*, **2**, e33.
- Perry, V. H., Hume, D. A. and Gordon, S. (1985) Immunohistochemical Localization of Macrophages and Microglia in the Adult and Developing Mouse Brain, *Neuroscience*, **15**, 313-326.
- Peters, A. and Yilmaz, E. (1993) Neuronal Organization in Area 17 of Cat Visual Cortex, *Cerebral Cortex*, **3**, 49-68.
- Pisani, A., Bonsi, P., Centonze, D., Martorana, A., Fusco, F., Sancesario, G., De Persis, C., Bernardi, G. and Calabresi, P. (2003) Activation of β 1-Adrenoceptors Excites Striatal Cholinergic Interneurons through a cAMP-Dependent, Protein Kinase-Independent Pathway, *The Journal of Neuroscience*, **23**, 5272-5282.
- Ponjavic, J., Oliver, P. L., Lunter, G. and Ponting, C. P. (2009) Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain, *PLoS Genetics*, **5**, e1000617.
- Ponjavic, J., Ponting, C. P. and Lunter, G. (2007) Functionality or Transcriptional Noise? Evidence for Selection within Long Noncoding RNAs, *Genome Research*, **17**, 556-565.
- Ponting, C. P. and Belgard, T. G. (2010) Transcribed Dark Matter: Meaning or Myth?, *Human Molecular Genetics*, **19**, R162-168.
- Ponting, C. P., Nellaker, C. and Meader, S. (2011) Rapid Turnover of Functional Sequence in Human and Other Genomes, *Annual Review of Genomics and Human Genetics*, **12**, 275-299.
- Ponting, C. P., Oliver, P. L. and Reik, W. (2009) Evolution and Functions of Long Noncoding RNAs, *Cell*, **136**, 629-641.
- Puelles, L. (2001) Thoughts on the Development, Structure and Evolution of the Mammalian and Avian Telencephalic Pallium, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **356**, 1583-1598.
- Puelles, L., Kuwana, E., Puelles, E., Bulfone, A., Shimamura, K., Keleher, J., Smiga, S. and Rubenstein, J. L. (2000) Pallial and Subpallial Derivatives in the Embryonic Chick and Mouse Telencephalon, Traced by the Expression of the Genes *Dlx-2*, *Emx-1*, *Nkx-2.1*, *Pax-6*, and *Tbr-1*, *The Journal of Comparative Neurology*, **424**, 409-438.

- Puelles, L., Kuwana, E., Puelles, E. and Rubenstein, J. L. (1999) Comparison of the Mammalian and Avian Telencephalon from the Perspective of Gene Expression Data, *European Journal of Morphology*, **37**, 139-150.
- Puelles, L., Martinez-de-la-Torre, M., Paxinos, G., Watson, C. and Martinez, S. (2007) *The Chick Brain in Stereotaxic Coordinates : An Atlas Featuring Neuromeric Subdivisions and Mammalian Homologies*. Academic Press, Amsterdam ; London.
- Quackenbush, J. (2002) Microarray Data Normalization and Transformation, *Nature Genetics*, **32 Suppl**, 496-501.
- Rakic, P. (2009) Evolution of the Neocortex: A Perspective from Developmental Biology, *Nature Reviews Neuroscience*, **10**, 724-735.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002) Hierarchical Organization of Modularity in Metabolic Networks, *Science*, **297**, 1551-1555.
- Ray, M., Ruan, J. and Zhang, W. (2008) Variations in the Transcriptome of Alzheimer's Disease Reveal Molecular Networks Involved in Cardiovascular Diseases, *Genome Biology*, **9**, R148.
- Reiner, A. (1993) Neurotransmitter Organization and Connections of Turtle Cortex: Implications for the Evolution of Mammalian Isocortex, *Comparative Biochemistry and Physiology A*, **104**, 735-748.
- Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D. and Kent, W. J. (2010) The UCSC Genome Browser Database: Update 2010, *Nucleic Acids Research*, **38**, D613-619.
- Rice, D. S. and Curran, T. (2001) Role of the Reelin Signaling Pathway in Central Nervous System Development, *Annual Review of Neuroscience*, **24**, 1005-1039.
- Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. and Chang, H. Y. (2007) Functional Demarcation of Active and Silent Chromatin Domains in Human *HOX* Loci by Noncoding RNAs, *Cell*, **129**, 1311-1323.
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. (2011) Improving RNA-Seq Expression Estimates by Correcting for Fragment Bias, *Genome Biology*, **12**, R22.
- Rommelfanger, K. S., Mitrano, D. A., Smith, Y. and Weinshenker, D. (2009) Light and Electron Microscopic Localization of Alpha-1 Adrenergic Receptor Immunoreactivity in the Rat Striatum and Ventral Midbrain, *Neuroscience*, **158**, 1530-1540.

- Rowell, J. J., Mallik, A. K., Dugas-Ford, J. and Ragsdale, C. W. (2010) Molecular Analysis of Neocortical Layer Structure in the Ferret, *The Journal of Comparative Neurology*, **518**, 3272-3289.
- Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. and Velculescu, V. E. (2002) Using the Transcriptome to Annotate the Genome, *Nature Biotechnology*, **20**, 508-512.
- Sanes, D. H., Reh, T. A. and Harris, W. A. (2012) *Development of the Nervous System*. Academic, London.
- 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.
- Santos, M., Kovari, E., Hof, P. R., Gold, G., Bouras, C. and Giannakopoulos, P. (2009) The Impact of Vascular Burden on Late-Life Depression, *Brain Research Reviews*, **62**, 19-32.
- Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. and Miller, W. (2003) Human-Mouse Alignments with BLASTZ, *Genome Research*, **13**, 103-107.
- Senti, K. A. and Brennecke, J. (2010) The piRNA Pathway: A Fly's Perspective on the Guardian of the Genome, *Trends in Genetics*, **26**, 499-509.
- Shamovsky, I., Ivannikov, M., Kandel, E. S., Gershon, D. and Nudler, E. (2006) RNA-Mediated Response to Heat Shock in Mammalian Cells, *Nature*, **440**, 556-560.
- Shimogori, T., Lee, D. A., Miranda-Angulo, A., Yang, Y., Wang, H., Jiang, L., Yoshida, A. C., Kataoka, A., Mashiko, H., Avetisyan, M., Qi, L., Qian, J. and Blackshaw, S. (2010) A Genomic Atlas of Mouse Hypothalamic Development, *Nature Neuroscience*, **13**, 767-775.
- Shinozaki, K., Yoshida, M., Nakamura, M., Aizawa, S. and Suda, Y. (2004) *Emx1* and *Emx2* Cooperate in Initial Phase of Archipallium Development, *Mechanisms of Development*, **121**, 475-489.
- Shipp, S. (2007) Structure and Function of the Cerebral Cortex, *Current Biology*, **17**, R443-449.
- Shirai, Y., Kouzuki, T., Kakefuda, K., Moriguchi, S., Oyagi, A., Horie, K., Morita, S. Y., Shimazawa, M., Fukunaga, K., Takeda, J., Saito, N. and Hara, H. (2010) Essential Role of Neuron-Enriched Diacylglycerol Kinase (DGK), DGK β in Neurite Spine Formation, Contributing to Cognitive Function, *PloS One*, **5**, e11602.
- Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajski, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003) Cap Analysis Gene

Expression for High-Throughput Analysis of Transcriptional Starting Point and Identification of Promoter Usage, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15776-15781.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart - Biological Queries Made Easy, *BMC Genomics*, **10**, 22.

Smeets, W. J., Marin, O. and Gonzalez, A. (2000) Evolution of the Basal Ganglia: New Perspectives through a Comparative Approach, *Journal of Anatomy*, **196 (Pt 4)**, 501-517.

Smith, H. M. (1967) Biological Similarities and Homologies, *Systematic Zoology*, **16**, 101.

Sorensen, A. T., Kanter-Schlifke, I., Carli, M., Balducci, C., Noe, F., During, M. J., Vezzani, A. and Kokaia, M. (2008) NPY Gene Transfer in the Hippocampus Attenuates Synaptic Plasticity and Learning, *Hippocampus*, **18**, 564-574.

Stanwood, G. D., Parlaman, J. P. and Levitt, P. (2006) Genetic or Pharmacological Inactivation of the Dopamine D1 Receptor Differentially Alters the Expression of Regulator of G-Protein Signalling (Rgs) Transcripts, *European Journal of Neuroscience*, **24**, 806-818.

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.

Striedter, G. F. (1997) The Telencephalon of Tetrapods in Evolution, *Brain, Behavior and Evolution*, **49**, 179-213.

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

Suh, Y., Obernier, K., Holzl-Wenig, G., Mandl, C., Herrmann, A., Worner, K., Eckstein, V. and Ciccolini, F. (2009) Interaction between DLX2 and EGFR Regulates Proliferation and Neurogenesis of SVZ Precursors, *Molecular and Cellular Neurosciences*, **42**, 308-314.

Sunkin, S. M. and Hohmann, J. G. (2007) Insights from Spatially Mapped Gene Expression in the Mouse Brain, *Human Molecular Genetics*, **16 Spec No. 2**, R209-219.

Tan, S. S., Kalloniatis, M., Truong, H. T., Binder, M. D., Cate, H. S., Kilpatrick, T. J. and Hammond, V. E. (2009) Oligodendrocyte Positioning in Cerebral Cortex Is Independent of Projection Neuron Layering, *Glia*, **57**, 1024-1030.

Tang, F. I., Chiu, T. H. and Wang, Y. (1998) Electrochemical Studies of the Effects of Substance P on Dopamine Terminals in the Rat Striatum, *Experimental Neurology*, **152**, 41-49.

- Tao, H., Manak, J. R., Sowers, L., Mei, X., Kiyonari, H., Abe, T., Dahdaleh, N. S., Yang, T., Wu, S., Chen, S., Fox, M. H., Gurnett, C., Montine, T., Bird, T., Shaffer, L. G., Rosenfeld, J. A., McConnell, J., Madan-Khetarpal, S., Berry-Kravis, E., Griesbach, H., Saneto, R. P., Scott, M. P., Antic, D., Reed, J., Boland, R., Ehaideb, S. N., El-Shanti, H., Mahajan, V. B., Ferguson, P. J., Axelrod, J. D., Lehesjoki, A. E., Fritzsche, B., Slusarski, D. C., Wemmie, J., Ueno, N. and Bassuk, A. G. (2011) Mutations in *Prickle* Orthologs Cause Seizures in Flies, Mice, and Humans, *American Journal of Human Genetics*, **88**, 138-149.
- Tariq, M. A., Kim, H. J., Jejelowo, O. and Pourmand, N. (2011) Whole-Transcriptome RNAseq Analysis from Minute Amount of Total RNA, *Nucleic Acids Research*, **39**, e120.
- Teissier, A., Griveau, A., Vigier, L., Piolot, T., Borello, U. and Pierani, A. (2010) A Novel Transient Glutamatergic Population Migrating from the Pallial-Subpallial Boundary Contributes to Neocortical Development, *The Journal of Neuroscience*, **30**, 10563-10574.
- Thomson, A. M. and Bannister, A. P. (2003) Interlaminar Connections in the Neocortex, *Cerebral Cortex*, **13**, 5-14.
- Threlfell, S., Sammut, S., Menniti, F. S., Schmidt, C. J. and West, A. R. (2009) Inhibition of Phosphodiesterase 10A Increases the Responsiveness of Striatal Projection Neurons to Cortical Stimulation, *The Journal of Pharmacology and Experimental Therapeutics*, **328**, 785-795.
- Tokuraku, K., Noguchi, T. Q., Nishie, M., Matsushima, K. and Kotani, S. (2007) An Isoform of Microtubule-Associated Protein 4 Inhibits Kinesin-Driven Microtubule Gliding, *Journal of Biochemistry*, **141**, 585-591.
- Tom Tang, Y., Emtage, P., Funk, W. D., Hu, T., Arterburn, M., Park, E. E. and Rupp, F. (2004) TAFA: A Novel Secreted Family with Conserved Cysteine Residues and Restricted Expression in the Brain, *Genomics*, **83**, 727-734.
- Tomer, R., Denes, A. S., Tessmar-Raible, K. and Arendt, D. (2010) Profiling by Image Registration Reveals Common Origin of Annelid Mushroom Bodies and Vertebrate Pallium, *Cell*, **142**, 800-809.
- Torres-Reveron, J. and Friedlander, M. J. (2007) Properties of Persistent Postnatal Cortical Subplate Neurons, *The Journal of Neuroscience*, **27**, 9962-9974.
- Toyama, M., Maeda, T. and Shimizu, N. (1974) Detailed Noradrenaline Pathways of Locus Coeruleus Neuron to the Cerebral Cortex with Use of 6-Hydroxydopa, *Brain Research*, **79**, 139-144.
- Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: Discovering Splice Junctions with RNA-Seq, *Bioinformatics*, **25**, 1105-1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation, *Nature Biotechnology*, **28**, 511-515.

Tuller, T., Kupiec, M. and Ruppin, E. (2008) Evolutionary Rate and Gene Expression across Different Brain Regions, *Genome Biology*, **9**, R142.

Tuoc, T. C., Radyushkin, K., Tonchev, A. B., Pinon, M. C., Ashery-Padan, R., Molnar, Z., Davidoff, M. S. and Stoykova, A. (2009) Selective Cortical Layering Abnormalities and Behavioral Deficits in Cortex-Specific *Pax6* Knock-out Mice, *The Journal of Neuroscience*, **29**, 8335-8349.

Umeda, M., Nishitani, H. and Nishimoto, T. (2003) A Novel Nuclear Protein, Twa1, and Muskelein Comprise a Complex with RanBPM, *Gene*, **303**, 47-54.

Underwood, J. G., Boutz, P. L., Dougherty, J. D., Stoilov, P. and Black, D. L. (2005) Homologues of the *Caenorhabditis Elegans* Fox-1 Protein Are Neuronal Splicing Regulators in Mammals, *Molecular and Cellular Biology*, **25**, 10005-10016.

Uziel, D., Garcez, P., Lent, R., Peuckert, C., Niehage, R., Weth, F. and Bolz, J. (2006) Connecting Thalamus and Cortex: The Role of Ephrins, *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, **288**, 135-142.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial Analysis of Gene Expression, *Science*, **270**, 484-487.

Wang, B., Lufkin, T. and Rubenstein, J. L. (2011) *Dlx6* Regulates Molecular Properties of the Striatum and Central Nucleus of the Amygdala, *The Journal of Comparative Neurology*, **519**, 2320-2334.

Wang, W. Z., Oeschger, F. M., Montiel, J. F., Garcia-Moreno, F., Hoerder-Suabedissen, A., Krubitzer, L., Ek, C. J., Saunders, N. R., Reim, K., Villalon, A. and Molnar, Z. (2011) Comparative Aspects of Subplate Zone Studied with Gene Expression in Sauropsids and Mammals, *Cerebral Cortex*, **21**, 2187-2203.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M. G., Glass, C. K. and Kurokawa, R. (2008) Induced ncRNAs Allosterically Modify RNA-Binding Proteins in *Cis* to Inhibit Transcription, *Nature*, **454**, 126-130.

Wang, Y., Brzozowska-Prechtl, A. and Karten, H. J. (2010) Laminar and Columnar Auditory Cortex in Avian Brain, *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 12676-12681.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: A Revolutionary Tool for Transcriptomics, *Nature Reviews Genetics*, **10**, 57-63.

Watakabe, A., Ohsawa, S., Hashikawa, T. and Yamamori, T. (2006) Binding and Complementary Expression Patterns of Semaphorin 3E and Plexin D1 in the Mature

- Neocortices of Mice and Monkeys, *The Journal of Comparative Neurology*, **499**, 258-273.
- Watson, C., Paxinos, G. and Puelles, L. (2012) *The Mouse Nervous System*. Academic Press, London.
- Wennerberg, K., Rossman, K. L. and Der, C. J. (2005) The Ras Superfamily at a Glance, *Journal of Cell Science*, **118**, 843-846.
- Wiley, S. E., Murphy, A. N., Ross, S. A., van der Geer, P. and Dixon, J. E. (2007) MitoNEET Is an Iron-Containing Outer Mitochondrial Membrane Protein That Regulates Oxidative Capacity, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5318-5323.
- Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. and Bahler, J. (2008) Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution, *Nature*, **453**, 1239-1243.
- Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B. and Schultz, P. G. (2005) A Strategy for Probing the Function of Noncoding RNAs Finds a Repressor of NFAT, *Science*, **309**, 1570-1573.
- Wilson, L. and Maden, M. (2005) The Mechanisms of Dorsoventral Patterning in the Vertebrate Neural Tube, *Developmental Biology*, **282**, 1-13.
- Winden, K. D., Oldham, M. C., Mirnics, K., Ebert, P. J., Swan, C. H., Levitt, P., Rubenstein, J. L., Horvath, S. and Geschwind, D. H. (2009) The Organization of the Transcriptional Network in Specific Neuronal Classes, *Molecular Systems Biology*, **5**, 291.
- Wolf, L., Goldberg, C., Manor, N., Sharan, R. and Rupp, E. (2011) Gene Expression in the Rodent Brain Is Associated with Its Regional Connectivity, *PLoS Computational Biology*, **7**, e1002040.
- Wullmann, M. F. and Mueller, T. (2004) Teleostean and Mammalian Forebrains Contrasted: Evidence from Genes to Behavior, *The Journal of Comparative Neurology*, **475**, 143-162.
- Xie, Z., Adamowicz, W. O., Eldred, W. D., Jakowski, A. B., Kleiman, R. J., Morton, D. G., Stephenson, D. T., Strick, C. A., Williams, R. D. and Menniti, F. S. (2006) Cellular and Subcellular Localization of PDE10A, a Striatum-Enriched Phosphodiesterase, *Neuroscience*, **139**, 597-607.
- Xiong, H., Kovacs, I. and Zhang, Z. (2004) Differential Distribution of KChIPs mRNAs in Adult Mouse Brain, *Molecular Brain Research*, **128**, 103-111.
- Xu, A. G., He, L., Li, Z., Xu, Y., Li, M., Fu, X., Yan, Z., Yuan, Y., Menzel, C., Li, N., Somel, M., Hu, H., Chen, W., Paabo, S. and Khaitovich, P. (2010) Intergenic and Repeat

Transcription in Human, Chimpanzee and Macaque Brains Measured by RNA-Seq, *PLoS Computational Biology*, **6**, e1000843.

Yamamoto, N., Maruyama, T., Uesaka, N., Hayano, Y., Takemoto, M. and Yamada, A. (2007) Molecular Mechanisms of Thalamocortical Axon Targeting, *Novartis Foundation Symposium*, **288**, 199-208; discussion 208-111, 276-181.

Yang, Z. (1997) PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood, *Computer Applications in the Biosciences*, **13**, 555-556.

Yoneshima, H., Yamasaki, S., Voelker, C. C., Molnar, Z., Christophe, E., Audinat, E., Takemoto, M., Nishiwaki, M., Tsuji, S., Fujita, I. and Yamamoto, N. (2006) Er81 Is Expressed in a Subpopulation of Layer 5 Neurons in Rodent and Primate Neocortices, *Neuroscience*, **137**, 401-412.

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., Winrow, C., Lockhart, D. J., Reilly, J. F., Young, W. G., Bloom, F. E. and Barlow, C. (2005) Adult Mouse Brain Gene Expression Patterns Bear an Embryologic Imprint, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10357-10362.

Zhang, B. and Horvath, S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*, **4**, Article 17.

Zhang, Q., Kim, N. K. and Feigon, J. (2011) Architecture of Human Telomerase RNA, *Proceedings of the National Academy of Sciences of the United States of America*.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. and Lee, J. T. (2008) Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome, *Science*, **322**, 750-756.

Zhou, M., Bakker, E. H., Velzing, E. H., Berger, S., Oitzl, M., Joels, M. and Krugers, H. J. (2010) Both Mineralocorticoid and Glucocorticoid Receptors Regulate Emotional Memory in Mice, *Neurobiology of Learning and Memory*, **94**, 530-537.

Zilles, K. and Amunts, K. (2010) Centenary of Brodmann's Map - Conception and Fate, *Nature Reviews Neuroscience*, **11**, 139-145.

Appendix A – Supplemental Methods for Chapter 3

Computational LincRNA Analysis Methods

Identification of cortical lincRNAs

Dr. Ana Marques and I identified in our cortical transcript set 4,587 multiexonic intergenic transcripts with no overlap with Ensembl protein-coding gene annotations (release 59). I used cortical transcripts with one or more exonic base overlapping an Ensembl protein coding gene exon to expand that gene for purposes of defining intergenic space (Ponting and Belgard 2010). Dr. Ana Marques calculated, in both orientations (forward and reverse), the coding potential of all intergenic transcripts using the Coding Potential Calculator (Kong et al. 2007) and identified 1,879 intergenic noncoding transcripts longer than 200 bp (lincRNAs). These lincRNAs can be clustered into 1,055 lincRNA loci, defined as the set of transcripts that share at least one intronic or exonic base on either strand.

LincRNA substitution rates

Dr. Ana Marques used mouse-human blastZ genome alignments (Schwartz et al. 2003), available from the UCSC Genome Browser Database (Rhead et al. 2010) to identify and extract the putative orthologous sequence in human (hg19) for all lincRNA loci and transposable element-derived Ancestral Repeats (AR) (Marques and Ponting 2009). She

considered only mouse-human alignments longer than 100 bp (907 lincRNA), to ensure the accuracy of the rate estimates. Dr. Marques estimated nucleotide substitution rates (d_{loci} and d_{AR}) between orthologous mouse-human aligned sequences using baseml, from the PAML package (Yang 1997), with the REV substitution model. To obtain normalized rates ($d_{\text{loci}}/d_{\text{AR}}$) she estimated the local neutral rates by concatenating all local ARs' mouse-human alignments from a matched G+C content class (Marques and Ponting 2009). She then randomly sampled single columns from these alignments to obtain putatively neutral sequence with the exact length and nucleotide content as the sequence of interest and estimated the substitution rate. Dr. Marques used the median value from 1000 iterations to normalize d_{loci} and calculate ($d_{\text{loci}}/d_{\text{AR}}$).

LincRNA associations with genomic features

To test the lincRNA-loci genome-wide association, Dr. Ana Marques used a previously reported randomization procedure that accounts for G+C content and chromosome-specific biases (Marques and Ponting 2009). She compared the observed density of cortical lincRNA transcription across 3 different types of annotations: *i*) DNase I and histone methylation marks in neuronal precursor cells (Meissner et al. 2008; Mikkelsen et al. 2007); *ii*) RNA secondary structure predictions (Pedersen et al. 2006) and *iii*) patterned protein-coding gene territories against what would be expected based on the intergenic distribution of these annotations in the mouse genome. To define patterned protein-coding gene territories she first divided the mouse genome (mm9) into protein-coding gene (Ensembl build 59) territories by determining the mid-distance, i , between each known mouse protein-coding gene and its closest upstream and downstream protein-coding neighbour $i-1$ and $i+1$ (Ponjavic et al. 2009). A gene's territory is

defined as the interval delimited by genomic coordinates $i-1$ to $i+1$. Protein-coding genes associated with one or more patterned cortical transcript were annotated as patterned.

Correlation of lincRNA expression with neighbouring protein-coding genes

Dr. Ana Marques identified the closest upstream and downstream cortical protein-coding transcript for each cortical lincRNAs and estimated the correlation (Spearman's rank correlation) of expression (FPKM) across samples between a lincRNA and its protein-coding neighbors using R. She considered significant correlations with an associated p -value < 0.025 (see also http://www.fgu.anat.ox.ac.uk/~grantb/mouse_layers/).

***In Situ* Hybridisation Methods**

Dr. Peter Oliver separately PCR amplified and cloned into the pCR4-TOPO vector (Invitrogen) a 982 bp region of *Anxa5* (ENSMUSG00000027712) matching AMBA probe RP_040324_01_D04 (Lein et al. 2007) and a 520 bp region of its associated lincRNA (*Gm11549*) matching AMBA probe RP_060220_05_F09 (Lein et al. 2007). For transcripts in Figure B.5, he selected regions to avoid repeat elements or possible sequence homology to other transcripts: nucleotides 70-490 of *EF177380* (*Malat1*), 508-914 of *AK159400* (*Neat1*) and 290-689 of *AK140632*. He froze P56 C57BL/6 male mouse brains in OCT (Merck, Darmstadt, Germany) on dry ice, cut 14 μ m coronal cryosections (10 μ m cryosections for transcripts in Figure B.5), and mounted them on positively charged slides. Digoxigenin-labeled riboprobe synthesis and hybridisation were performed as described previously (Isaacs et al. 2003). Sense strand riboprobe hybridisation generated no detectable signal in all cases (data not shown). Dr.

Oliver exposed slides for 16 hours with the exception of those using the *Malat1* riboprobe, which were hybridised for 2 hours.

Appendix B - Supplemental Figures and Tables for Chapter 3

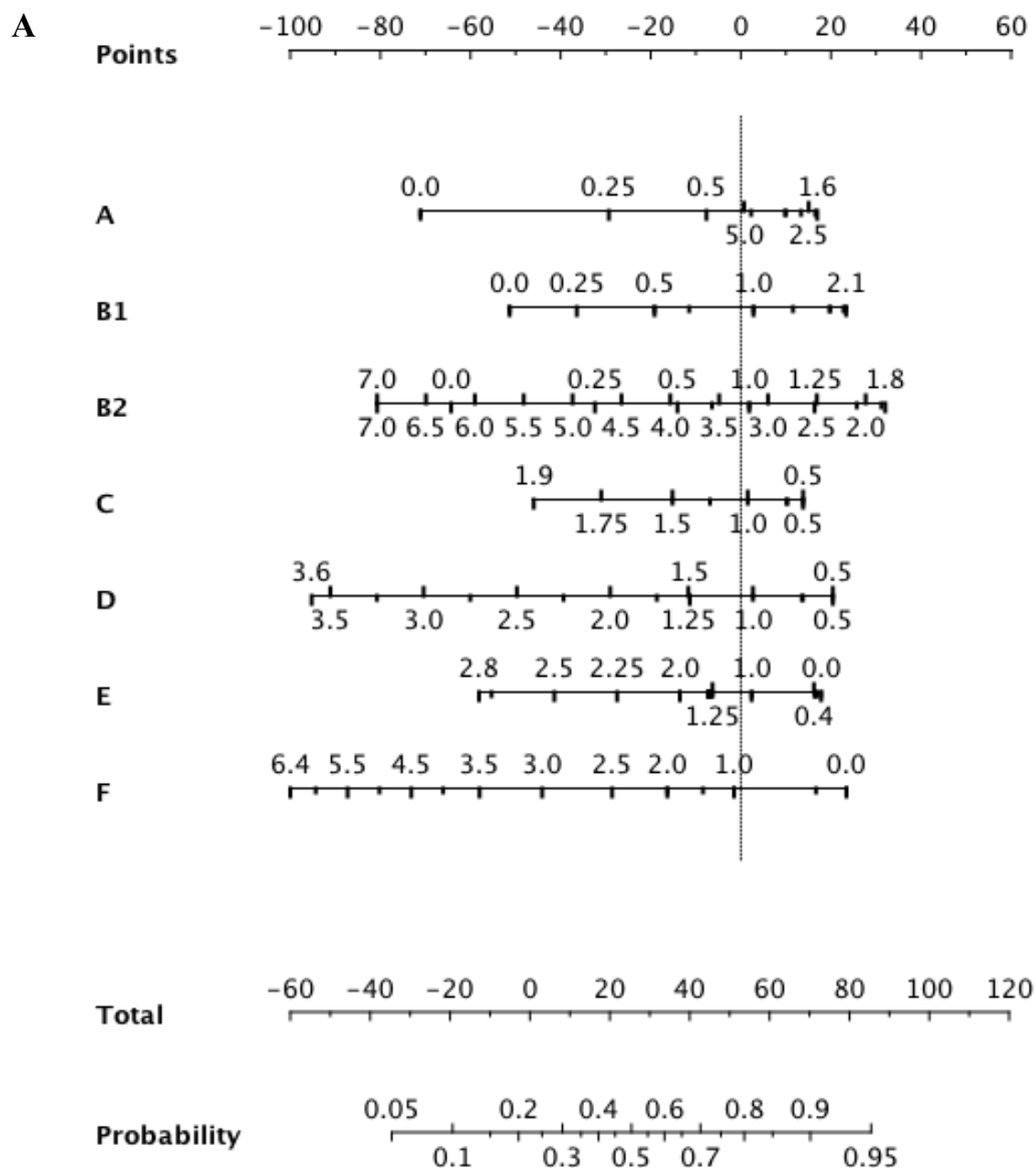
Table B.1: Integrity and quantity of total RNA from dissected samples, by litter. Dr. Tamara Sirey assessed RNA Integrity Number (RIN) using a Bioanalyzer and concentration using a NanoDrop 1000 spectrophotometer. I estimated total amount of RNA by multiplying the concentration by the 30 μL volume remaining after assessing the sample characteristics. The variation in RIN, and the low RIN in sample A, reinforce the importance of training a classifier to predict layer enrichment probabilities to obtain reasonable estimates of accuracy. *Totals are listed separately for B1 and B2, as they were made into separate libraries.

Sample	Litter 1 RIN	Litter 2 RIN	Litter 1 concentration (ng/ μL)	Litter 2 concentration (ng/ μL)	Total amount (μg)
A	3.1	2.4	88.4	132.0	6.6
B	6.6	5.7	310.9	267.0	9.3, 8.0*
C	8.2	6.7	161.0	170.9	10.0
D	8.6	7.6	155.9	152.9	9.3
E	7.8	8.1	156.3	125.2	8.4
F	6.7	8.4	41.9	51.2	2.8

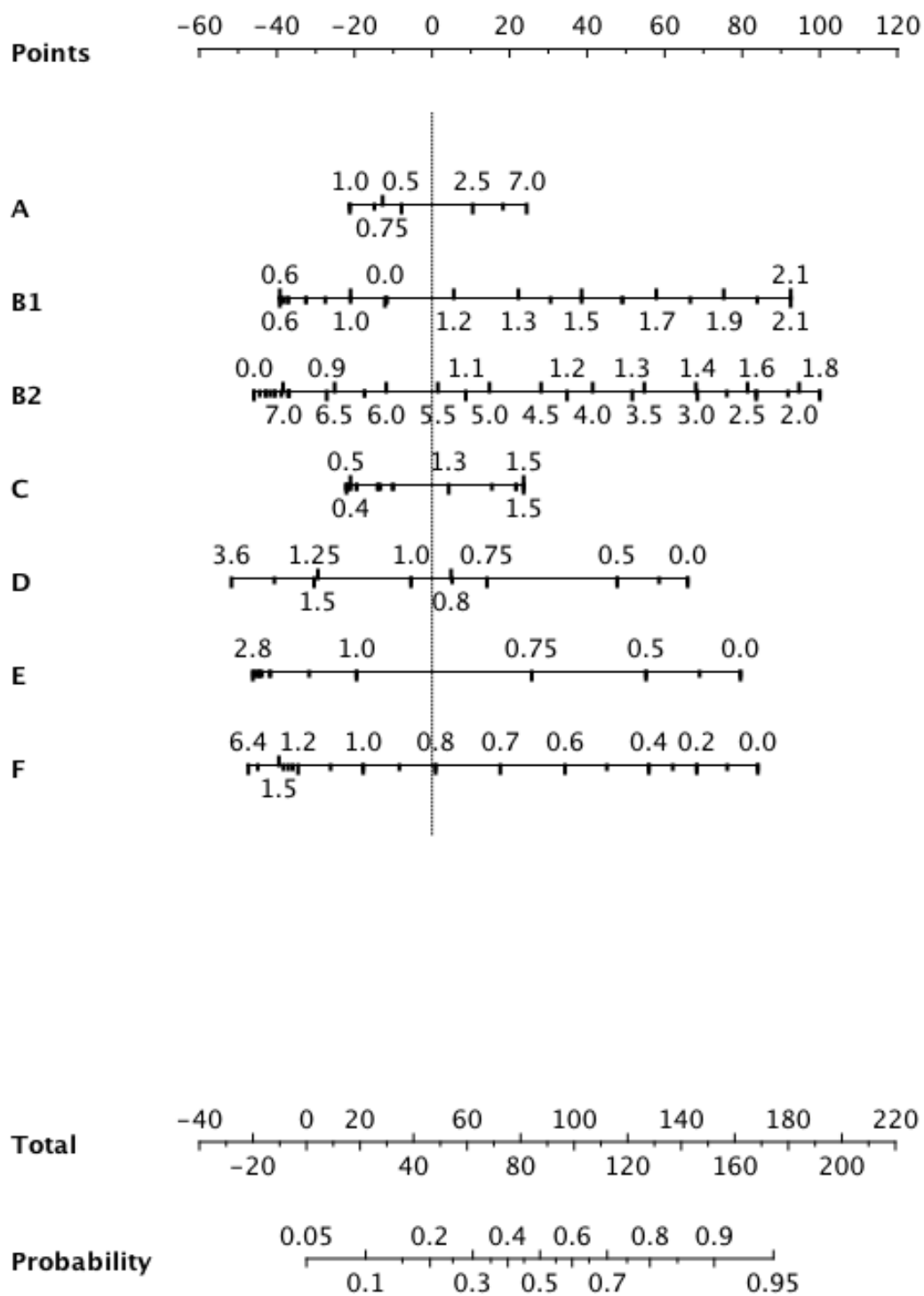
Table B.2: Fragment sizes and numbers of sequenced and aligned reads per library.

Average fragment length was the midpoint of the full width at half maximum (FWHM) of mapped paired ends based on unique ELAND mappings (from Illumina's pipeline) to the mouse genome assembly (mm9); the standard deviation was estimated using a Gaussian assumption (FWHM/2.35). Millions of clusters represents the total number of paired end fragments (2x51 (or 2x76) nucleotides, the 51st (or 76th) base was not used in subsequent analysis) imaged across all flowcells. Millions of clusters passing filter represents the number of these clusters that passed Illumina's native chastity filter. Millions of tophat alignments represents the number of alignments (in up to 40 genomic locations by default) reported by tophat, using all clusters as an input (not just those passing the chastity filter). Millions of alignable reads represents the number of reads contributing to these alignments. Millions of uniquely alignable reads represents these reads that mapped uniquely to the genome or across splice sites.

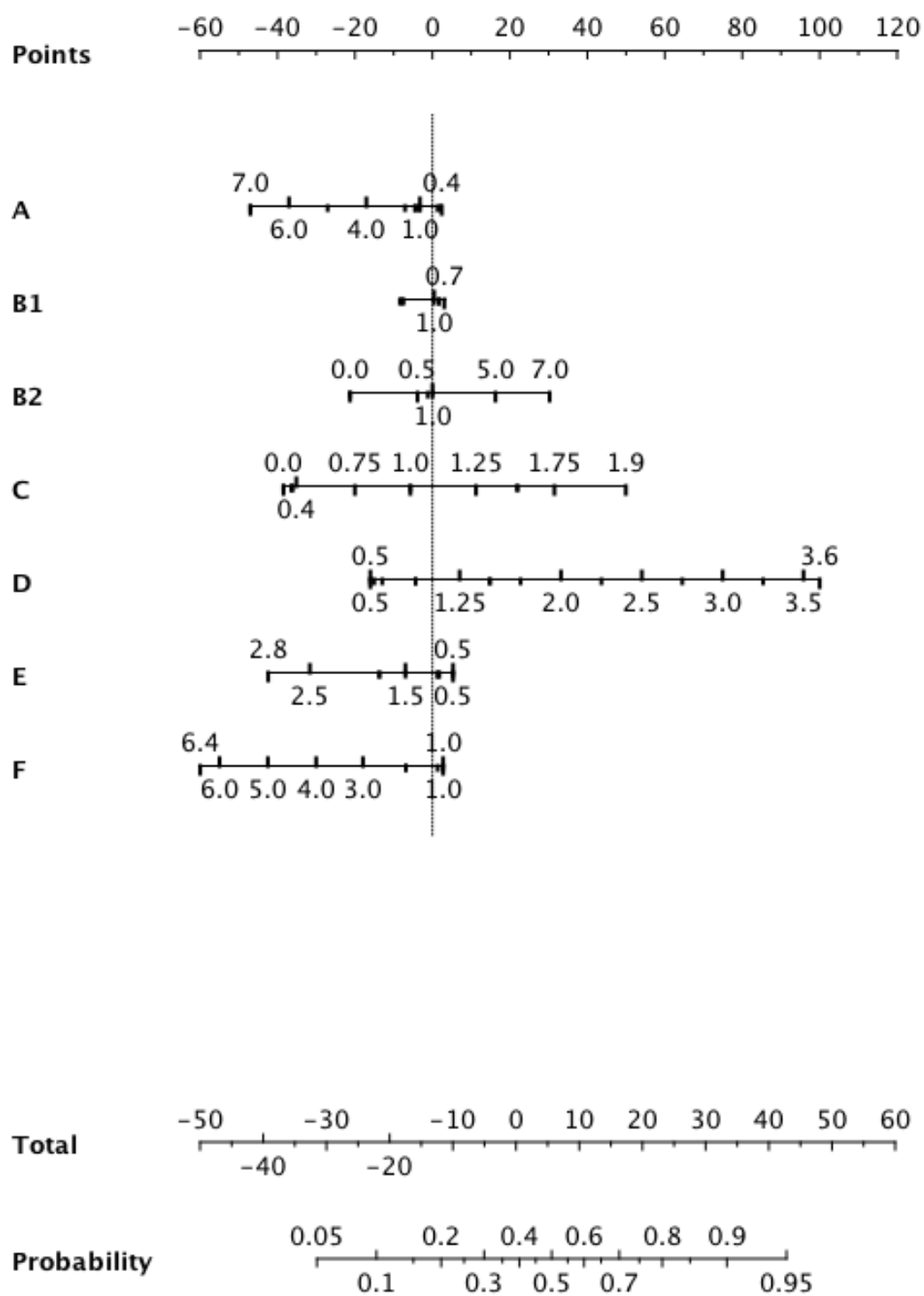
Sample	Average fragment length	Insert standard deviation	Millions of clusters	Millions of clusters passing filter	Millions of tophat alignments	Millions of alignable reads	Millions of uniquely alignable reads
A	164	27	60.5	50.8	79.7	74.1	65.7
B1	239	29	67.8	~57.7	116.7	115.0	112.7
B2	193	28	119.4	89.0	186.1	181.9	170.0
C	184	21	63.9	50.8	101.6	99.4	95.2
D	208	22	73.3	59.6	111.9	109.2	105.9
E	172	24	73.6	55.0	104.0	101.4	98.3
F	184	28	71.8	58.3	111.3	109.2	106.8
Combined	-	-	530	421	811	790	755



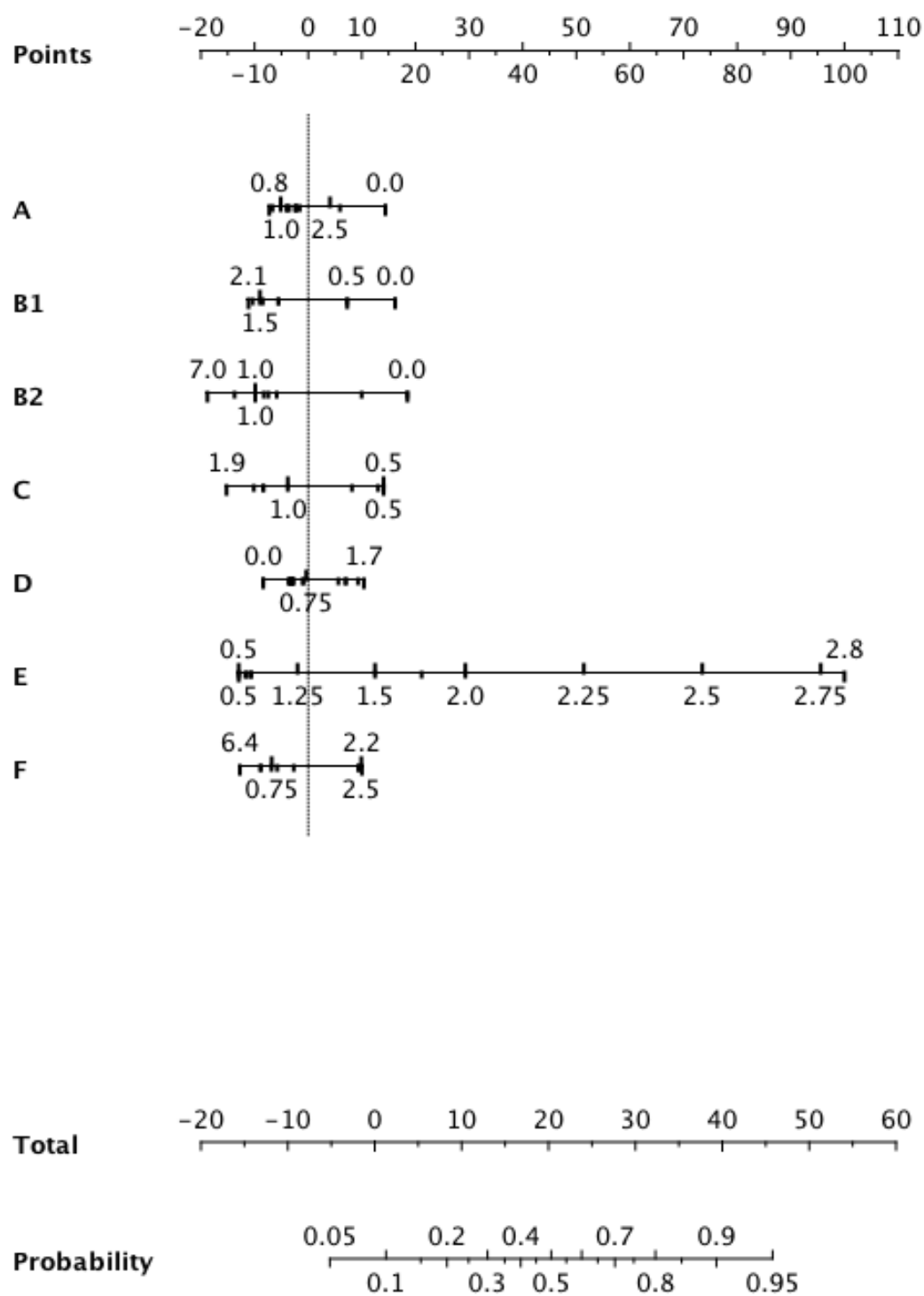
B



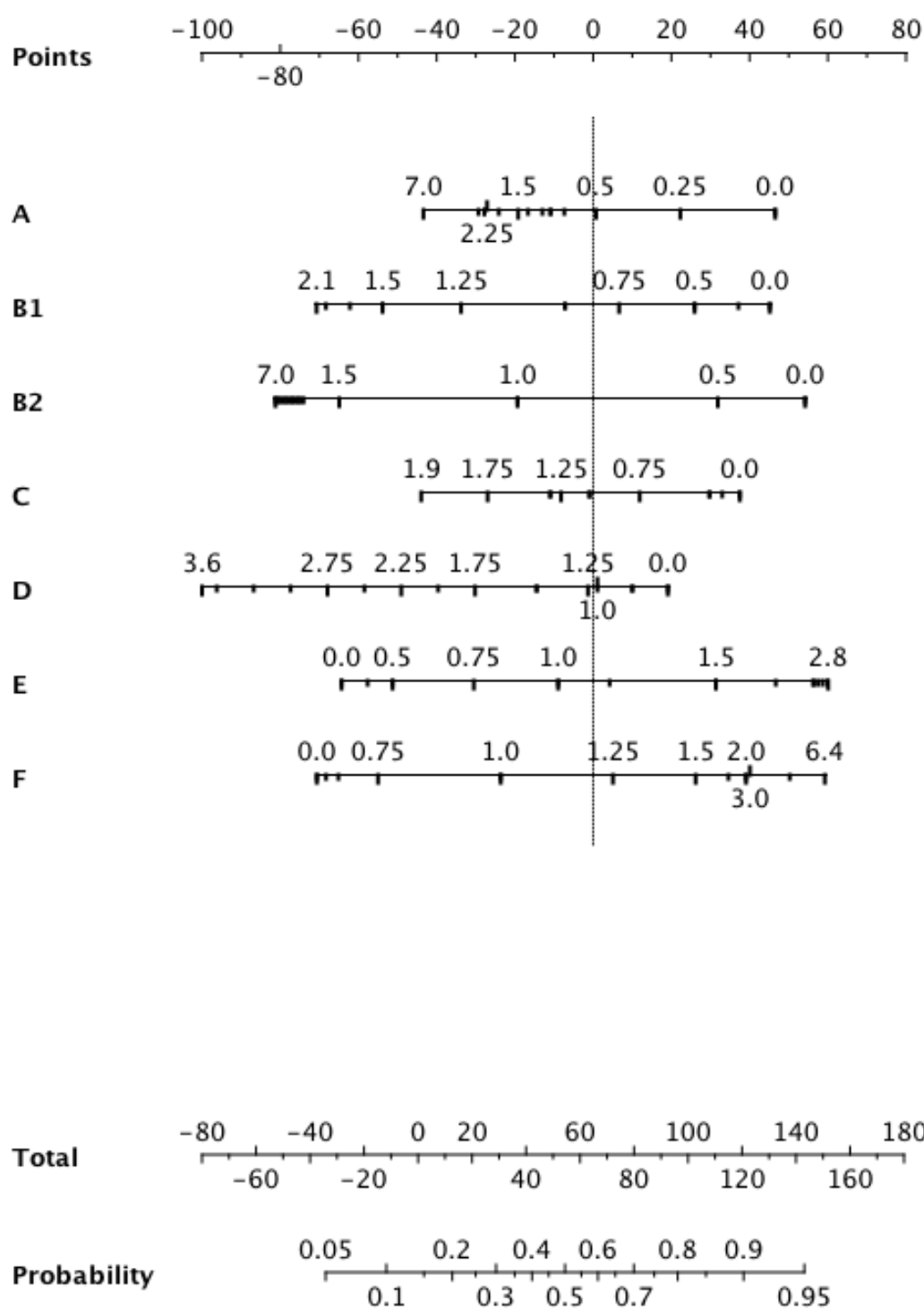
C



D



E



F

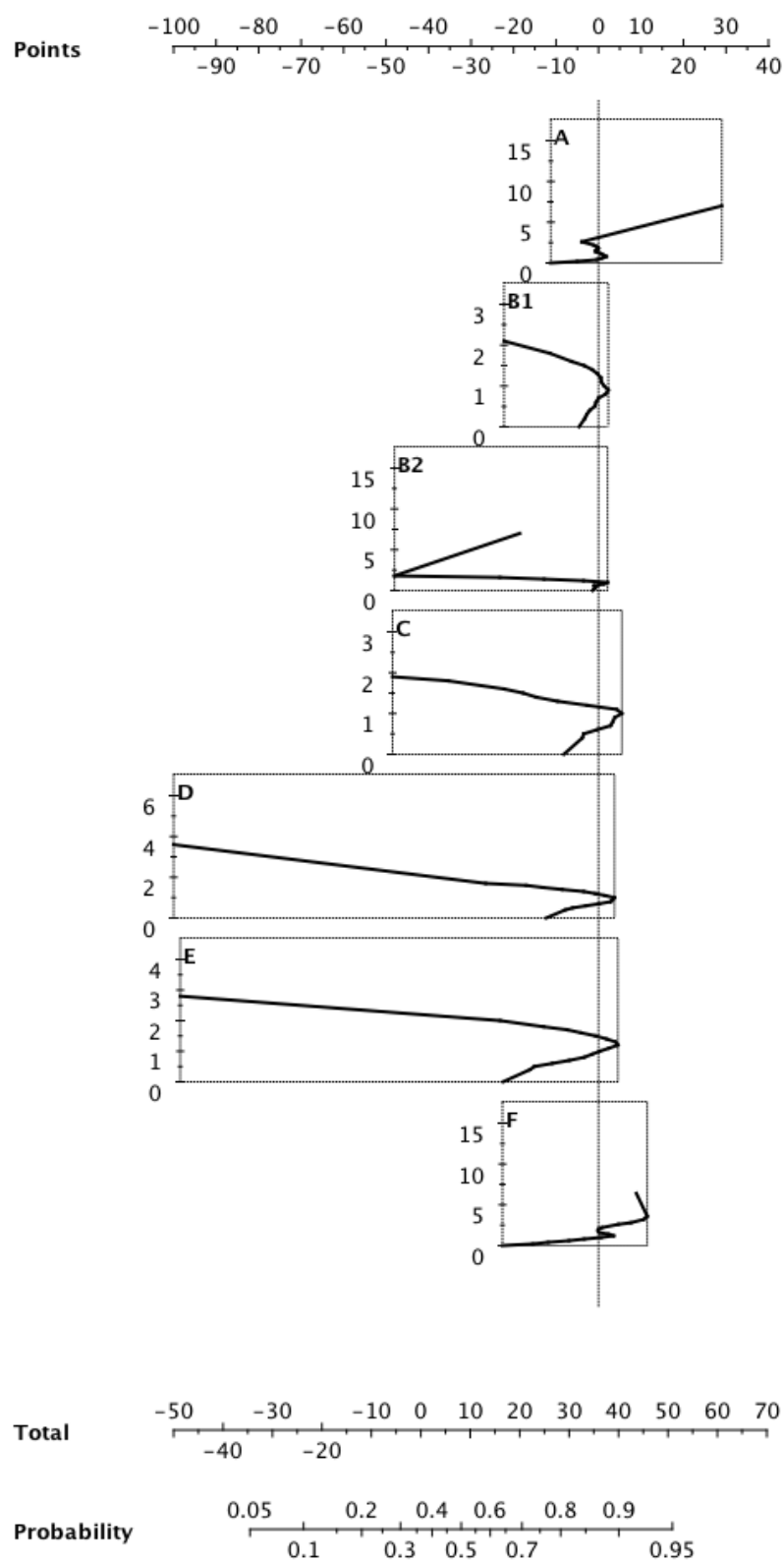


Figure B.1: Pictorial representations of the probability assignment schemes used by the classifiers for all genes. Nomograms representing classifiers for (A) layers 2/3, (B) layer 4, (C) layer 5, (D) layer 6, (E) layer 6b and (F) no layer enrichment. For relative expression in each sample (up to 7.0 = 100% expression in that sample and 0% expression in all other samples), a point value can be assigned. For any given gene (and corresponding relative expression profile across samples), the sum total of these points can be used to translate from point total into uncalibrated probability of enrichment in that layer using the scales at bottom. Consider, for example, classifier A. A gene with 7% of its expression (0.5/7.0) in A, 14% of its expression (1.0/7.0) in B1, 29% of its expression (2.0/7.0) in B2, 14% of its expression in C (1.0/7.0), 7% of its expression in D (0.5/7.0), 29% of its expression in E (2/7) and no expression in F would have a point total of $(-9)+3+28+2+20+(-12)+23=55$. Using the key at bottom this translates to an uncalibrated layer enrichment probability (in layers 2/3) of 0.8.

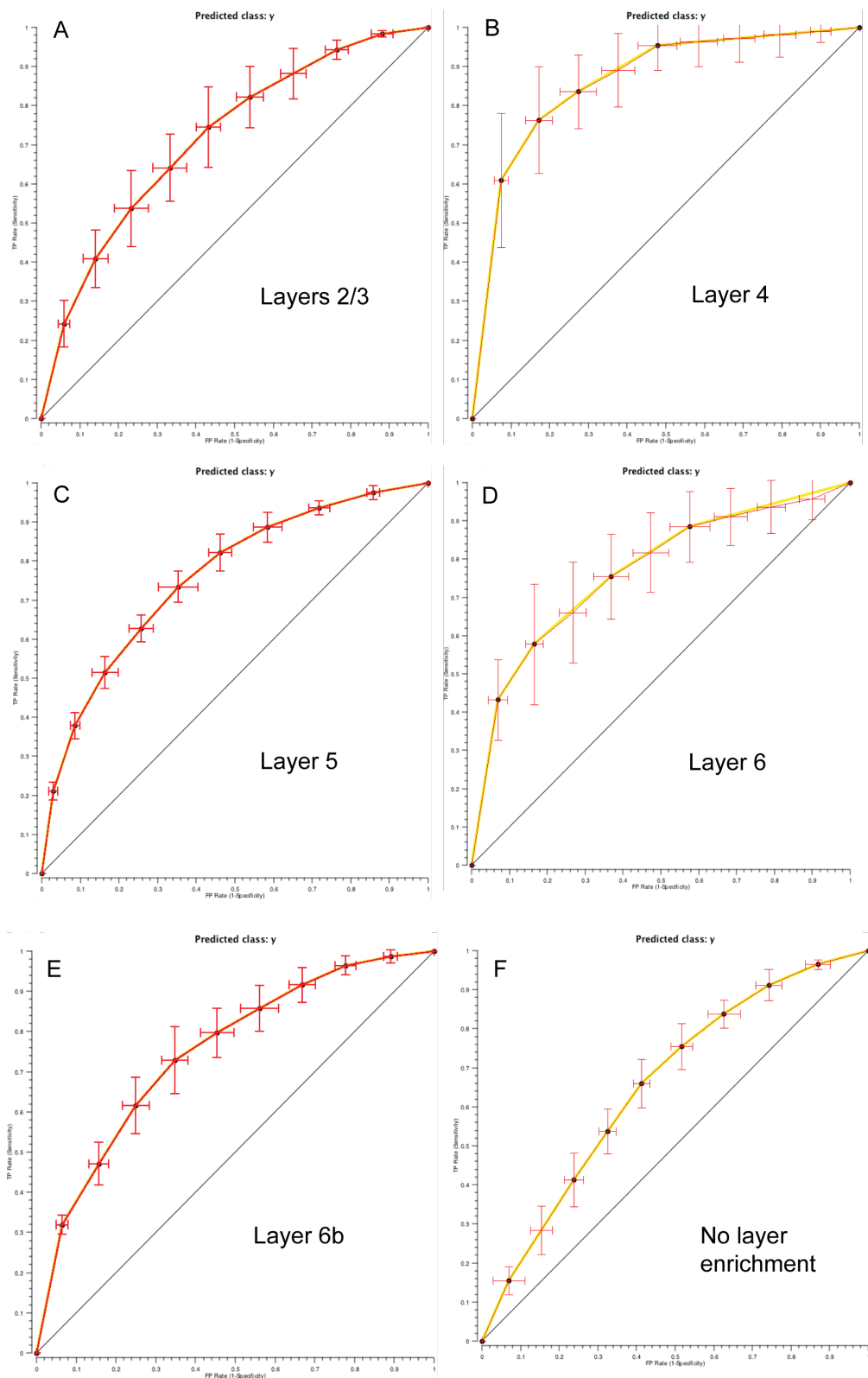


Figure B.2: Classifier ROC (receiver operating characteristic) curves for classifiers of all genes. ROC curves with error bars for each of the classifiers (A) layers 2/3, (B) layer 4, (C) layer 5, (D) layer 6, (E) layer 6b and (F) no layer enrichment. The diagonal with a slope of 1 represents the expectation value of a random classifier.

Table B.3: Coverage statistics of the genome, known genes, and intergenic regions.

Bases (% genome) covered by at least one read represents the number of bases and proportion of the 2,725,765,481 bases in mm9 used in these alignments (all chromosomes, including randoms, using the UCSC nomenclature, and chrM).

Bases (% genome) covered by at least two non-identical reads represents the number & proportion of such bases covered by two non-identical reads, where “samtools rmdup” has been run on the bam files to remove paired-end reads that have the exact same start and stop mapped locations (depleting PCR and optical duplicates).

Bases (% intergenic genome) outside non-lincRNA Ensembl loci covered by at least one read represents the number of bases, and proportion of the intergenic genome (1,704,768,106 nt remains after eliminating the 37.5% of the mouse genome inside non-lincRNA genic loci), covered by at least one read outside all loci of all classes of genes defined in Ensembl release 60, excepting lincRNAs. *% Ensembl gene loci >0.0 FPKM* represents the percentage of all Ensembl genes (release 57, which did not include lincRNA gene predictions) with any nonzero FPKM expression value called by cufflinks. Among protein-coding genes in particular, 18,960 genes (83% of the 22,806 total) had nonzero FPKMs in at least one sample. *% Ensembl genes >0.1 FPKM* represents the same as before but for genes greater than 0.1 FPKM. 16,340

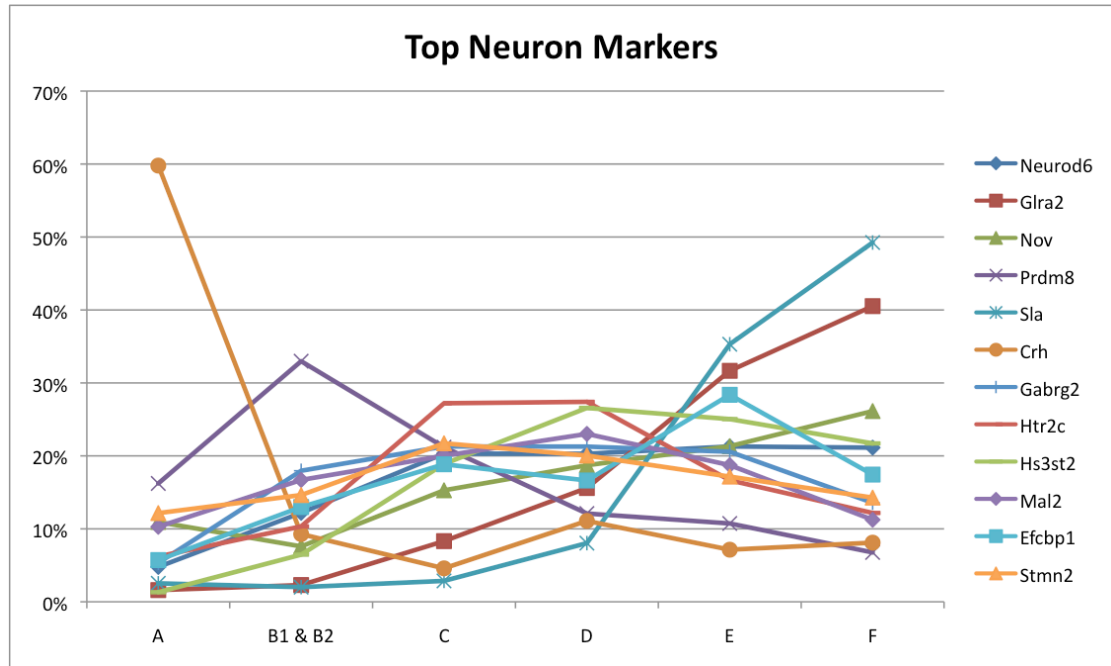
protein-coding genes (72%) were expressed at greater than 0.1 FPKM in at least one sample.

Sample	Bases (% genome) covered by at least one read	Bases (% genome) covered by at least two non-identical reads	Bases (% intergenic genome) outside non-lincRNA Ensembl loci covered by at least one read	% Ensembl genes >0.0 FPKM	% Ensembl genes >0.1 FPKM
A	268040319 (9.8%)	160004371 (5.9%)	56,436,638 (3.3%)	63%	54%
B1	290720044 (10.7%)	148490006 (5.4%)	61,498,352 (3.6%)	61%	50%
B2	372478215 (13.7%)	239652510 (8.8%)	84,657,473 (5.0%)	64%	51%
C	259025712 (9.5%)	138128102 (5.1%)	53,839,576 (3.2%)	61%	51%
D	236340872 (8.7%)	129833188 (4.8%)	50,051,652 (2.9%)	61%	51%
E	238855138 (8.8%)	126055098 (4.6%)	48,854,567 (2.9%)	62%	51%
F	211938161 (7.8%)	145046680 (5.3%)	42,521,967 (2.5%)	60%	52%
Merged	671356672 (24.6%)	365230720 (13.4%)	170,720,024 (10.0%)	71%	59%

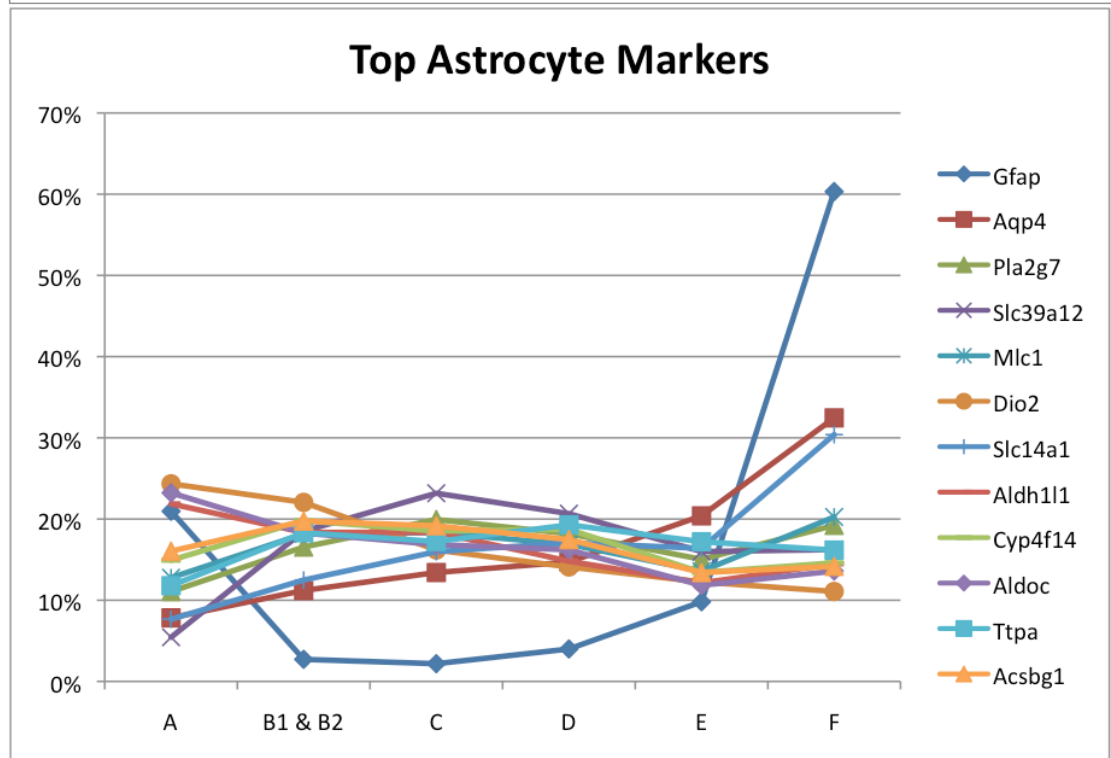
Table B.4: Ten most highly expressed known genes in each sample. FPKM values of these Ensembl genes (release 57) ranged from 11,532 to 1,291 - nearly one order of magnitude. Most genes are mitochondrially-encoded. This is unlikely to be an effect of mtDNA contamination, since the read coverage on chrM strongly stacks up in exons. MGI translations of Ensembl gene IDs are given.

A	B1	B2	C	D	E	F
<i>mt-Cytb</i>	<i>mt-Cytb</i>	<i>mt-Cytb</i>	<i>mt-Cytb</i>	<i>mt-Cytb</i>	<i>mt-Cytb</i>	<i>Mbp</i>
<i>mt-Co1</i>	<i>mt-Nd1</i>	<i>mt-Nd1</i>	<i>mt-Nd1</i>	<i>mt-Nd1</i>	<i>mt-Nd1</i>	<i>mt-Cytb</i>
<i>mt-Nd1</i>	<i>Nrgn</i>	<i>mt-Rnr2</i>	<i>mt-Rnr2</i>	<i>mt-Rnr2</i>	<i>mt-Rnr2</i>	<i>Plp1</i>
<i>mt-Rnr2</i>	<i>mt-Rnr2</i>	<i>mt-Co1</i>	<i>mt-Rnr1</i>	<i>mt-Co1</i>	<i>Mbp</i>	<i>mt-Nd1</i>
<i>mt-Rnr1</i>	<i>mt-Rnr1</i>	<i>mt-Nd5</i>	<i>mt-Nd6</i>	<i>mt-Rnr1</i>	<i>mt-Co1</i>	<i>Fth1</i>
<i>mt-Nd4</i>	<i>mt-Nd2</i>	<i>mt-Nd6</i>	<i>mt-Nd5</i>	<i>mt-Nd5</i>	<i>mt-Rnr1</i>	<i>mt-Co1</i>
<i>Nrgn</i>	<i>mt-Nd6</i>	<i>mt-Nd4</i>	<i>mt-Co1</i>	<i>mt-Nd2</i>	<i>Snap25</i>	<i>mt-Rnr2</i>
<i>mt-Nd5</i>	<i>mt-Nd4</i>	<i>mt-Nd2</i>	<i>mt-Nd2</i>	<i>mt-Nd6</i>	<i>Nrgn</i>	<i>Gm10927</i>
<i>mt-Co2</i>	<i>mt-Nd5</i>	<i>mt-Rnr1</i>	<i>mt-Nd4</i>	<i>mt-Nd4</i>	<i>mt-Nd5</i>	<i>Mobp</i>
<i>mt-Nd6</i>	<i>mt-Co1</i>	<i>Nrgn</i>	<i>Snap25</i>	<i>Snap25</i>	<i>mt-Nd2</i>	<i>mt-Nd5</i>

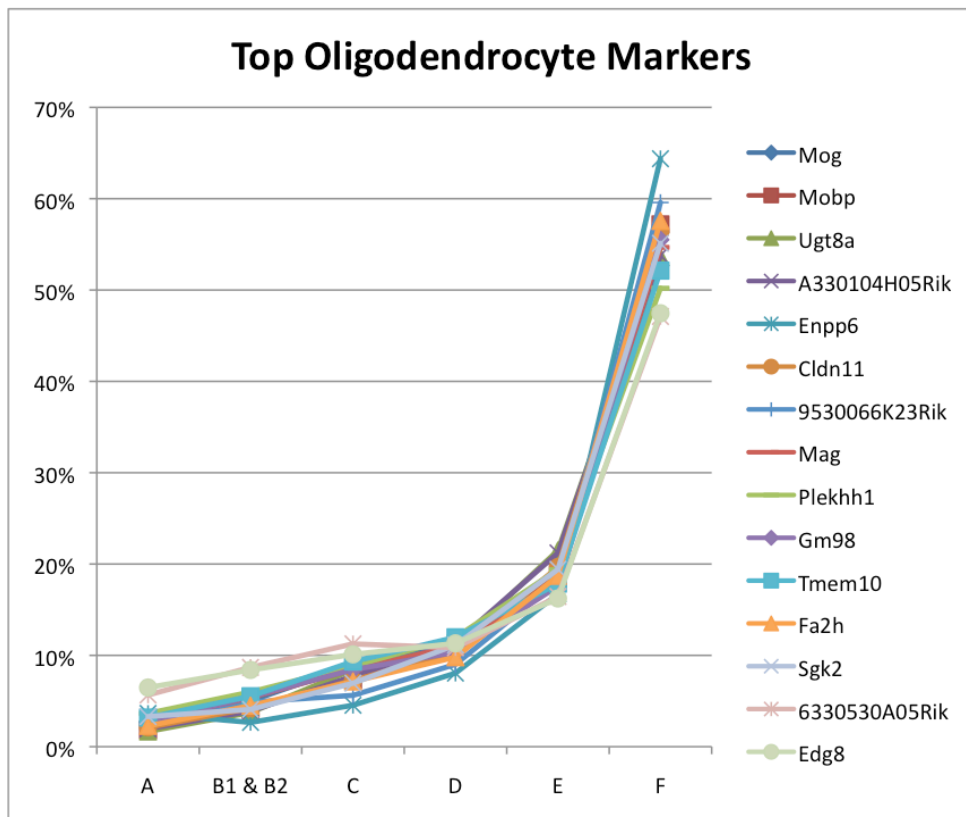
A



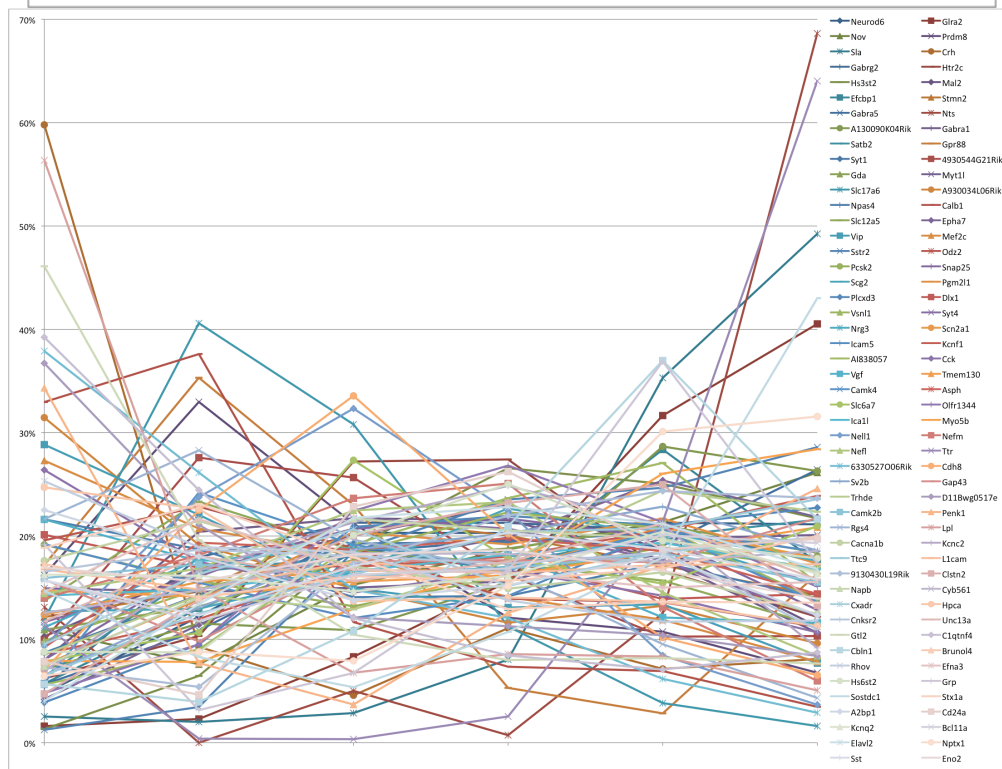
B



C



D



E

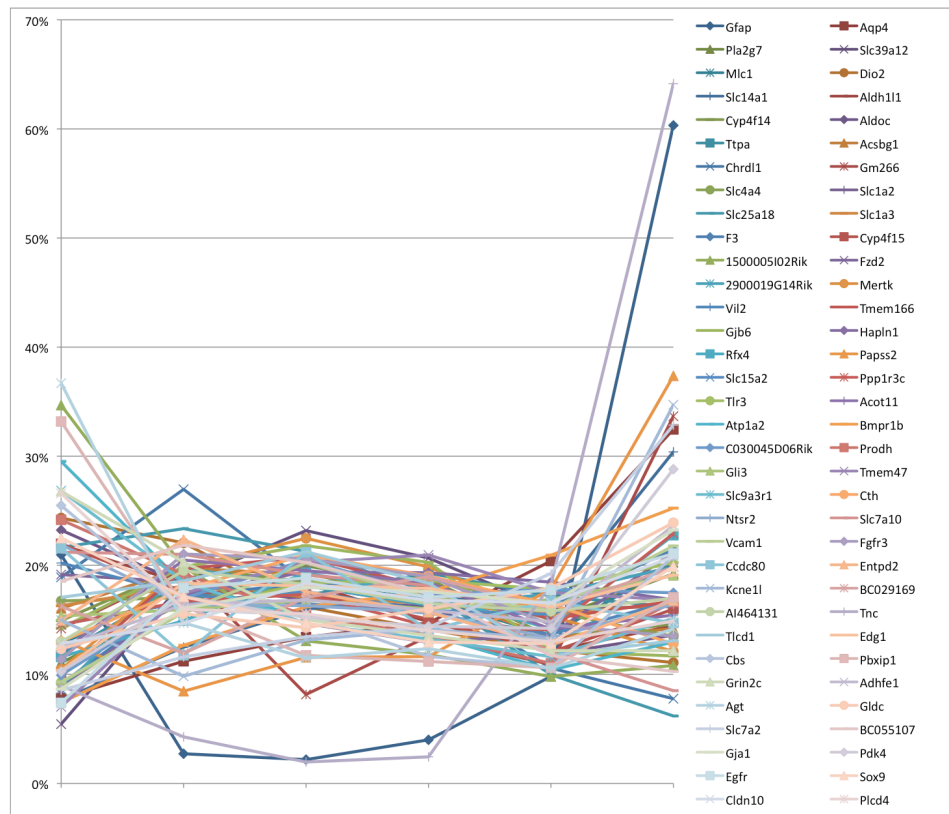


Figure B.3: Various neuron- and astrocyte-specific genes show different expression patterns across laminarly-dissected samples, suggesting laminar populations of these cell types. Relative percentage of total expression across samples A-F (B1 & B2 are averaged) for the Ensembl genes 50 fold or greater enriched in (A) neurons, (B) astrocytes and (C) oligodendrocytes (Cahoy et al. 2008). These samples were normalized against one another (Belgard et al. 2011). Distributions across these same samples for genes enriched 20-fold or greater are shown for (D) neurons and (E) astrocytes.

Table B.5: Numbers of genes with different types of annotations. ¹Union of primary somatosensory cortex, dorsal cortex and lateral cortex dissections. ²Mouse orthologs of the corresponding human genes. ³Including genes not assigned to a module.

Source of annotations	Genes with quantified expression ¹	Genes with layer enrichment probabilities ¹
all genes (Flicek et al. 2011)	24,130	14,160
GO biological process from Ensembl (Ashburner et al. 2000; Flicek et al. 2011)	13,566	9,314
GO cellular component from Ensembl (Ashburner et al. 2000; Flicek et al. 2011)	14,462	10,004
GO molecular function from Ensembl (Ashburner et al. 2000; Flicek et al. 2011)	14,766	10,084
loci associated with human disease ² (Chen et al. 2010)	6,688	4,848
MGI mouse knockout phenotypes (Blake et al. 2011)	5,505	3,653
Reactome protein complexes (Croft et al. 2011)	2,875	2,080
Reactome protein interactions (Croft et al. 2011)	3,389	2,432
Reactome protein pathways (Croft et al. 2011)	3,468	2,502
KEGG pathways (Kanehisa et al. 2004)	4,796	3,315
co-expressed in mouse brain (Miller et al. 2010) ³	4,578	4,578
co-expressed amongst mouse neurons (Windén et al. 2009) ³	11,769	8,721
genes enriched in neurons or astrocytes (Cahoy et al. 2008)	1,284	1,130
SFARI autism candidate genes ² (Basu et al. 2009)	143	110
annotation from any of the above sources	17,389	11,908

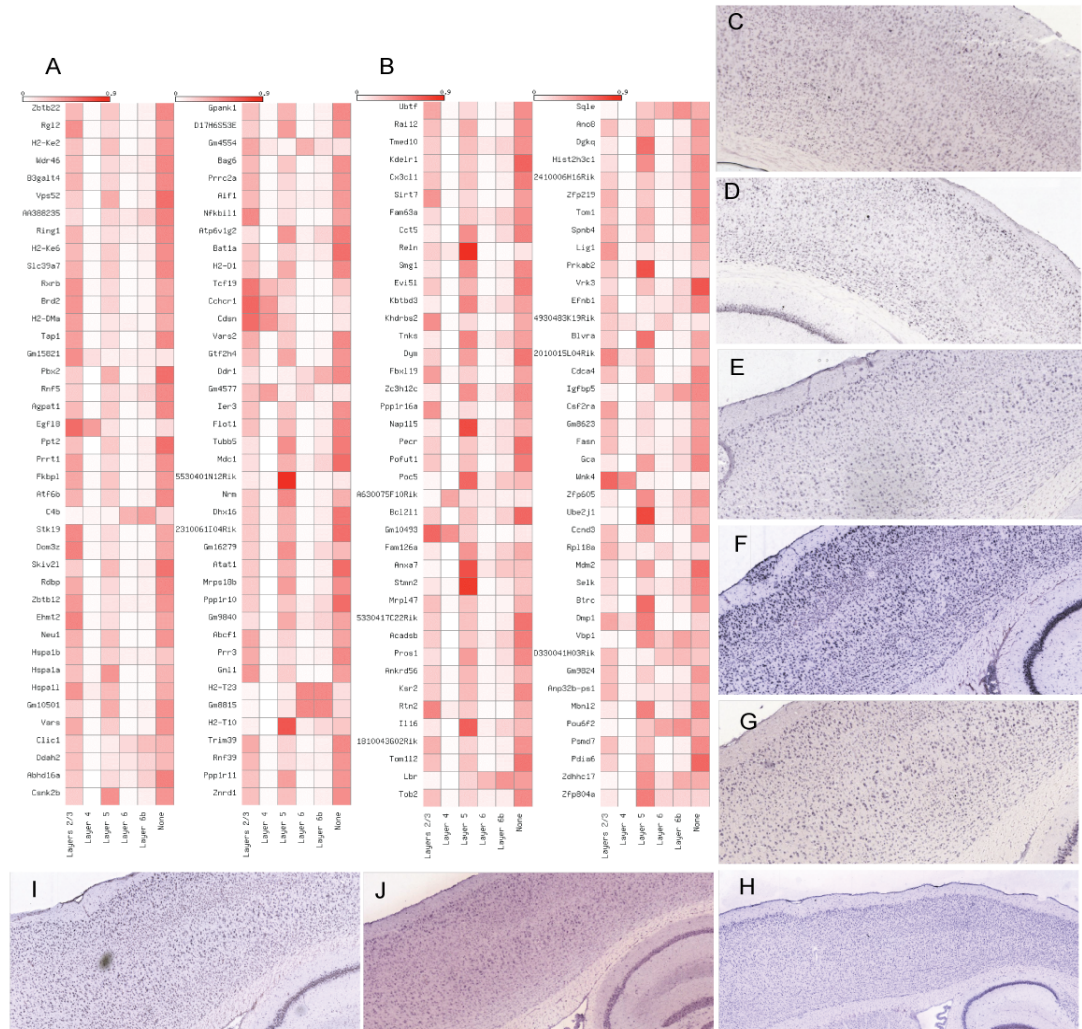


Figure B.4: Many MHC region genes show enriched expressed in layers 2/3. (A) Heatmaps (Pavlidis and Noble 2003) representing calibrated layer enrichment probabilities of MHC region genes in order of their leftmost coordinate, ordered from top to bottom, then left to right. Genes show complex patterns of enrichment across layers. (B) Heatmaps representing calibrated layer enrichment probabilities for 80 genes selected at random using a pseudorandom number generator from the list of classified genes. MHC region genes are 34% more likely than these randomly selected genes to be enriched in layers 2/3 (0.302 vs 0.226 average calibrated probabilities of enrichment; $p < 10^{-6}$, case resampling bootstrap). I

display *in situ* hybridisations from the Allen Mouse Brain Atlas (AMBA) of some example genes that appear correctly predicted by the classifiers as being enriched in at least layers 2/3 (they may also be enriched in other layers): (C) *Ier3* (AMBA curators confirm 2/3, also report 6), (D) *Hspa1l* (AMBA curators confirm 2/3, also report 6 and 6b), (E) *H2-DMa* (however AMBA curators report none), (F) *Dom3z*, (G) *Ehmt2*, (H) *Cchcr1*, (I) *Stk19* and (J) *Egfl8*. *Tcf19* is an example of a gene unresolved in most coronal sections (expression appears even across everything in brain), but at least one section has a clean stain (this stain shows a layers 2/3 enrichment). *Cdsn*, another gene with high likelihood of layers 2/3 enrichment, is untried at the time of this writing. Such high-level descriptive findings highlight the advantages of a genomics approach to neuroanatomy.

Table B.6: Genes in the Parkinson's Disease enrichment of layer 5. Nineteen genes with the Parkinson's Disease annotation would have been expected to be enriched in layer 5 by chance. Most of these thirty-four genes encode proteins active or associated with mitochondrial functions.

Symbol	Name
<i>Uqcrc2</i>	ubiquinol cytochrome c reductase core protein 2
<i>Lrrk2</i>	leucine-rich repeat kinase 2
<i>Atp5c1</i>	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, gamma polypeptide 1
<i>Cox6a2</i>	cytochrome c oxidase, subunit VI a, polypeptide 2
<i>Slc25a4</i>	solute carrier family 25 (mitochondrial carrier, adenine nucleotide translocator), member 4
<i>Ndufv2</i>	NADH dehydrogenase (ubiquinone) flavoprotein 2
<i>Uqcrc1</i>	ubiquinol-cytochrome c reductase core protein 1
<i>Ndufb8</i>	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 8
<i>Uqcrfs1</i>	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1
<i>Ndufa6</i>	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6 (B14)
<i>Ndufb6</i>	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6
<i>Cyc1</i>	cytochrome c-1
<i>Sdhb</i>	succinate dehydrogenase complex, subunit D, integral membrane protein
<i>Ndufs1</i>	NADH dehydrogenase (ubiquinone) Fe-S protein 1
<i>Cox5a</i>	cytochrome c oxidase, subunit Va
<i>Sdha</i>	succinate dehydrogenase complex, subunit A, flavoprotein (Fp)
<i>Vdac1</i>	voltage-dependent anion channel 1
<i>Casp3</i>	caspase 3
<i>Atp5f1</i>	ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit B1
<i>Vdac2</i>	voltage-dependent anion channel 2
<i>Ube2j1</i>	ubiquitin-conjugating enzyme E2, J1
<i>Ube2g1</i>	ubiquitin-conjugating enzyme E2G 1 (UBC7 homolog, C. elegans)
<i>Ube2g2</i>	ubiquitin-conjugating enzyme E2G 2
<i>Atp5b</i>	ATP synthase, H ⁺ transporting mitochondrial F1 complex, beta subunit
<i>Ppid</i>	peptidylprolyl isomerase D (cyclophilin D)
<i>Ndufa9</i>	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9
<i>Sdhb</i>	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
<i>Atp5j</i>	ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit F
<i>Atp5g3</i>	ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit C3 (subunit 9)
<i>Ndufa1</i>	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1
<i>Ndufc1</i>	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 1
<i>Ndufc2</i>	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 2
<i>Atp5a1</i>	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, alpha subunit 1
<i>Ndufb5</i>	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5

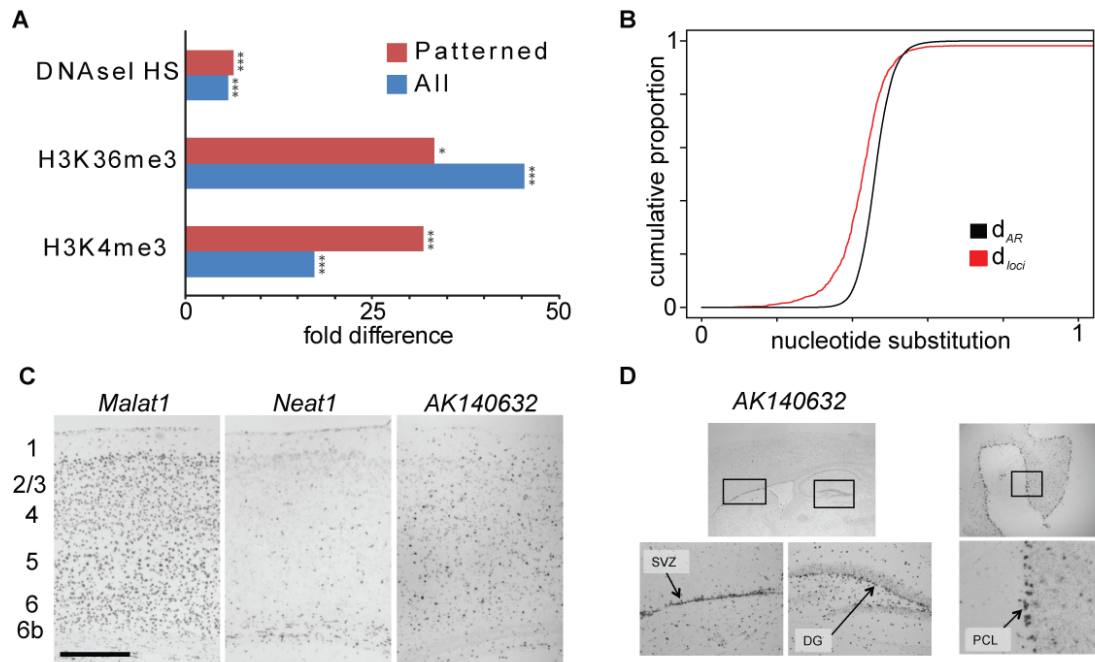


Figure B.5: LincRNAs are transcribed, evolutionarily constrained, and patterned in expression. (A) LincRNA loci significantly coincide with DNaseI hypersensitivity sites and histone methylation marks associated with active transcription in neuronal precursor cells. Plot, created by Dr. Marques, represents fold enrichments of observed over expected values for the overlap of layer patterned lincRNA loci and all lincRNA loci with DNaseI hypersensitivity sites, H3K36me3 sites and H3K4me3 sites (Meissner et al. 2008; Mikkelsen et al. 2007). $*p < 0.05$, $***p < 0.001$ (B) LincRNA loci show significant evolutionary constraint. Plot, created by Dr. Marques, represents the cumulative distributions of substitution rate for lincRNA loci (red) and putatively neutral sequence-AR (black). LincRNA loci accumulate significantly fewer nucleotide substitutions between mouse and human than neighbouring neutral sequence (0.428 and 0.460: median substitution rate lincRNA loci and neutral sequence respectively; $p < 10^{-16}$). (C) I predicted some cortical lincRNAs to be *bona fide* layer markers, and these were confirmed by Dr.

Peter Oliver. *In situ* hybridisation of three lincRNAs in the male mouse cortex at P56. *Malat1* is the among the most highly expressed transcripts identified from the RNA-seq dataset. *Neat1* expression is highly patterned in layer 6b, as predicted, although in these images *AK140632* is expressed at low levels throughout the cortex. Scale bar is 500 μ m. (D) Some cortical lincRNAs are expressed more highly elsewhere in the brain. *AK140632* is also expressed in a specific subset of neurons in adult brain as shown from parasagittal adult brain sections, including the dentate gyrus (DG), subventricular zone (SVZ) and the nuclei of neurons in the Purkinje cell layer (PCL) of the cerebellum.

Appendix C - Supplemental Figures and Tables

for Chapter 4

Table C.1: RNA concentration and quality metrics of mouse samples

Sample	Abbreviation	ng/ μ L	260:280	260:230
Dorsal cortex A (layers I-III)	MDC_A	51.8	2.03	0.87
Dorsal cortex B (layer IV)	MDC_B	128.7	1.96	1.11
Dorsal cortex C (layers IV-V)	MDC_C	84.6	1.92	0.61
Dorsal cortex D (layer V)	MDC_D	127.4	1.91	0.64
Dorsal cortex E (layer VI)	MDC_E	64.9	2.04	0.73
Dorsal cortex F (layers VI-VIb)	MDC_F	298.8	1.7	0.57
Lateral cortex A (layers I-III)	MLC_A	67.7	2.02	1.07
Lateral cortex B (layer IV)	MLC_B	128.8	2.06	1.23
Lateral cortex C (layers IV-V)	MLC_C	31.5	2.01	0.66
Lateral cortex D (layers V-VI)	MLC_D	79.4	2.04	0.76
Lateral cortex E (layer VI)	MLC_E	55.5	1.95	1.03
Lateral cortex F (layer VIb)	MLC_F	13.7	1.93	0.46
Hippocampus	MHipp	285.3	2.03	1.2
Clastrum/endopiriform	MCE	82	1.92	1.1
Pallial amygdala	MPA	125.8	1.92	1.14
Striatum	MS	99.8	2.05	1.41

Table C.2: RNA concentration, quality metrics and RIN of chicken samples

Sample	Abbreviation	ng/ μ L	260: 280	260: 230	RIN
arcopallium	CA	203.7	2.11	2	6.9
dorsolateral cortex	CDLC	255.1	2.11	2.01	8.3
hippocampus	CHipp	490.1	2.12	1.83	7.6
mesopallium	CM	266.5	2.1	1.97	8.3
nidopallium	CN	696.6	2.14	1.95	not called
hyperpallium	CH	232.2	2.13	1.76	5.3
striatum	CS	174.1	2.09	1.9	8.4

Table C.3: Insert sizes and mapping information for lanes from all libraries

Species	Region	Date YY- MM- DD	Sequencer	Flowcell	L n	IIS	ISS std dev	good reads (M)	alignment s (M)	properly paired	both mates mapped
Chicken	Arcopallium	091029	DOGFISH	42WDDAAXX	1	52	31	28.1	24.7	15.0 (61%)	22.4 (91%)
		091120	DOGFISH	42VFJAAXX	6	52	31	34.7	30.4	18.4 (61%)	27.3 (90%)
		091124	ALPACA	42UY8AAXX	1	55	33	30.2	26.8	17.2 (64%)	24.5 (91%)
	Dorsolateral cortex	091029	DOGFISH	42WDDAAXX	2	52	31	25.8	20.5	13.3 (65%)	18.4 (89%)
		091120	DOGFISH	42VFJAAXX	7	52	31	32.9	26.3	16.8 (64%)	23.2 (88%)
		091124	ALPACA	42UY8AAXX	2	53	32	29.0	23.5	15.5 (66%)	21.0 (89%)
	Hippocampus	091029	DOGFISH	42WDDAAXX	3	90	30	27.9	21.8	11.7 (54%)	19.1 (88%)
		091120	DOGFISH	42VFJAAXX	8	90	30	32.6	25.7	13.9 (54%)	22.4 (87%)
		091124	ALPACA	42UY8AAXX	3	90	30	29.7	23.2	12.5 (54%)	20.3 (88%)
	Mesopallium	091029	DOGFISH	42WDDAAXX	4	89	28	28.8	23.4	14.1	20.6

										(61%)	(89%)
		091124	ALPACA	42UY8AAXX	4	89	28	30.1	24.2	14.9 (61%)	21.5 (89%)
	Nidopallium	091120	DOGFISH	42VFJAAXX	3	120	31	37.8	31.4	18.1 (58%)	27.4 (87%)
		091124	ALPACA	42UY8AAXX	5	121	31	30.9	26.6	15.8 (60%)	23.8 (90%)
	Hyperpallium	091120	DOGFISH	42VFJAAXX	2	53	30	29.7	26.3	16.0 (61%)	23.5 (90%)
		091124	ALPACA	42UY8AAXX	7	53	31	26.4	23.7	14.7 (62%)	21.5 (91%)
	Striatum	091120	DOGFISH	42VFJAAXX	1	62	31	32.7	27.2	17.4 (64%)	24.5 (90%)
		091124	ALPACA	42UY8AAXX	6	63	32	27.2	23.0	15.2 (66%)	21.1 (92%)
Mouse	Dorsal cortex A (layers I-III)	100720	PLATYPUS	621GFAAXX	1	287	30	31.0	30.0	18.9 (63%)	27.4 (91%)
		100805	LEOPARD	6232KAAXX	1	287	31	60.4	59.8	38.9 (65%)	55.9 (93%)
	Dorsal cortex B (layer IV)	100720	PLATYPUS	621GFAAXX	2	290	28	58.3	55.3	34.5 (62%)	51.0 (92%)
		100805	LEOPARD	6232KAAXX	2	290	28	69.4	66.2	42.1 (64%)	61.5 (93%)
	Dorsal cortex C (layers IV-V)	100720	PLATYPUS	621GFAAXX	3	291	28	52.3	50.7	30.2 (59%)	47.3 (93%)
		100805	LEOPARD	6232KAAXX	3	291	28	62.3	60.7	36.6 (60%)	56.6 (93%)
	Dorsal cortex D (layer V)	100720	PLATYPUS	621GFAAXX	4	269	30	58.9	58.1	34.7 (60%)	54.9 (94%)

		100805	LEOPARD	6232KAAXX	4	268	30	66.6	65.4	38.9 (59%)	61.2 (93%)
	Dorsal cortex E (layer VI)	100720	PLATYPUS	621GFAAXX	5	284	31	56.8	56.1	33.2 (59%)	53.0 (95%)
	Dorsal cortex F (layers VI-VIb)	100720	PLATYPUS	621GFAAXX	6	279	21	53.6	52.8	29.9 (57%)	49.8 (94%)
		100810	MARMOSET	62328AAXX	5	279	21	61.6	60.6	34.4 (57%)	56.7 (94%)
	Lateral cortex A (layers I-III)	100720	PLATYPUS	621GFAAXX	7	281	40	40.5	39.0	25.3 (65%)	36.1 (93%)
		100810	MARMOSET	62328AAXX	6	281	40	50.7	48.9	32.2 (66%)	45.3 (93%)
	Lateral cortex B (layer IV)	100720	PLATYPUS	621GFAAXX	8	275	32	53.3	52.1	34.3 (66%)	48.7 (93%)
		100831	PLATYPUS	627RLAAXX	5	275	32	60.5	60.0	39.7 (67%)	56.0 (94%)
	Lateral cortex C (layers IV-V)	100720	MARMOSET	621UPAAXX	1	281	40	60.8	60.0	36.8 (61%)	55.7 (93%)
		100810	MARMOSET	62328AAXX	8	283	42	53.0	52.3	31.8 (61%)	48.6 (93%)
	Lateral cortex D (layers V- VI)	100720	MARMOSET	621UPAAXX	2	267	29	47.8	47.1	29.1 (62%)	44.0 (94%)
		100810	WALLABY	62344AAXX	5	268	29	46.8	46.4	29.0 (63%)	43.7 (94%)
	Lateral cortex E (layer VI)	100720	MARMOSET	621UPAAXX	3	268	38	42.3	40.8	23.7 (58%)	37.8 (93%)
		100810	WALLABY	62344AAXX	6	268	38	46.0	44.7	26.1 (58%)	41.6 (93%)
	Lateral cortex	100720	MARMOSET	621UPAAXX	4	273	34	47.2	46.2	26.8	43.3

	F (layer VIb)									(58%)	(94%)
		100810	WALLABY	62344AAXX	7	272	34	51.7	50.9	29.8 (59%)	48.0 (94%)
	Hippocampus	100720	MARMOSET	621UPAAXX	7	259	34	49.2	49.4	29.9 (61%)	45.8 (93%)
		100810	WALLABY	62344AAXX	4	259	34	52.4	53.1	32.6 (61%)	49.8 (94%)
	Caudate/ putamen	100720	MARMOSET	621UPAAXX	6	276	28	56.3	55.8	33.8 (61%)	52.1 (93%)
		100810	WALLABY	62344AAXX	3	276	28	55.6	55.7	34.2 (61%)	52.6 (95%)
	Pallidum medial	100720	MARMOSET	621UPAAXX	5	283	32	58.2	55.4	33.1 (60%)	50.6 (91%)
		100810	WALLABY	62344AAXX	2	284	33	51.6	49.5	29.7 (60)	45.7 (92%)
	Striatum	100720	MARMOSET	621UPAAXX	8	283	35	56.3	55.6	34.1 (61%)	51.2 (92%)
		100810	WALLABY	62344AAXX	8	284	35	47.2	47.3	29.2 (62%)	44.5 (94%)

Table C.4: Estimated pre-mRNA fraction in mouse and chick samples

Mouse sample	Median pre-mRNA fraction	Chicken sample	Median pre-mRNA fraction
Dorsal cortex A (layers I-III)	4.49% (N=2972)	arcopallium	6.95% (N=7237)
Dorsal cortex B (layer IV)	4.20% (N=3150)	dorsolateral cortex	8.16% (N=7608)
Dorsal cortex C (layers IV-V)	3.50% (N=3102)	hippocampus	8.19% (N=7761)
Dorsal cortex D (layer V)	3.39% (N=3148)	mesopallium	8.05% (N=7048)
Dorsal cortex E (layer VI)	2.92% (N=2705)	nidopallium	7.07% (N=7095)
Dorsal cortex F (layers VI-VIb)	2.55% (N=3150)	hyperpallium	7.31% (N=5911)
Lateral cortex A (layers I-III)	4.86% (N=2943)	striatum	9.24% (N=7144)
Lateral cortex B (layer IV)	4.11% (N=3093)		
Lateral cortex C (layers IV-V)	3.37% (N=3097)		
Lateral cortex D (layers V-VI)	3.22% (N=3023)		
Lateral cortex E (layer VI)	2.63% (N=2975)		
Lateral cortex F (layer VIb)	2.55% (N=3066)		
Hippocampus	2.35% (N=2980)		
Clastrum/endopiriform	3.13% (N=3119)		
Pallial amygdala	3.29% (N=3136)		
Striatum	2.31% (N=2993)		

Table C.5: Downsampling read fragments

Species	Region	# unique fragments mapped (% fragments sequenced)	# (%) unique mapped fragments in pc genes >200nt	Ratio to downsample	# unique fragments sampled
Chicken	Arcopallium	43,356,930 (93%)	29,563,619 (68%)	0.488717	21,189,264
	Dorsolateral cortex	37,306,331 (85%)	21,573,838 (58%)	0.537091	20,036,913
	Hippocampus	38,116,880 (85%)	22,922,907 (60%)	0.509703	19,428,271
	Mesopallium	25,280,073 (86%)	15,302,760 (61%)	0.724450	18,314,158
	Nidopallium	31,066,189 (90%)	19,310,314 (62%)	0.653362	20,297,472
	Hyperpallium	26,299,836 (94%)	17,071,439 (65%)	N/A	26,299,836
	Striatum	26,246,353 (88%)	15,125,692 (58%)	0.752200	19,742,508
Mouse	Dorsal cortex A	43,539,003 (96%)	28,760,754 (66%)	0.749822	32,646,495
	Dorsal cortex B	59,885,178 (94%)	41,072,917 (69%)	0.508724	30,464,999
	Dorsal cortex C	54,582,567 (95%)	41,863,139 (77%)	0.502602	27,433,302
	Dorsal cortex D	60,080,966 (96%)	46,445,743 (77%)	0.464065	27,881,490
	Dorsal cortex E	27,363,620 (96%)	20,950,956 (77%)	N/A	27,363,620
	Dorsal cortex F	56,037,833 (97%)	44,312,232 (79%)	0.454334	25,459,896
	Lateral cortex A	43,682,519 (96%)	26,523,474 (61%)	0.775756	33,886,992
	Lateral cortex B	54,747,441 (96%)	36,283,334 (66%)	0.564627	30,911,895
	Lateral cortex C	55,058,857 (97%)	41,437,719 (75%)	0.506372	27,880,256
	Lateral cortex D	45,732,069 (97%)	34,588,590 (76%)	0.605109	27,672,889
	Lateral cortex E	42,428,722 (96%)	33,439,586 (79%)	0.613748	26,040,543
	Lateral cortex F	48,260,801 (98%)	37,984,095 (79%)	0.522717	25,226,725
	Hippocampus	49,280,437 (97%)	39,544,659 (80%)	0.582436	28,702,719
	Clastrum/endopiriform	54,403,180 (97%)	41,710,672 (77%)	0.501311	27,272,906
	Pallial amygdala	52,006,334 (95%)	38,990,285 (75%)	0.519640	27,024,585
	Striatum	49,859,361 (96%)	39,701,019 (80%)	0.572858	28,562,311

Table C.6: Variance explained by the first 15 components of a lane-wise principal component analysis

principal component	mouse proportion of variance explained	mouse cumulative proportion of variance	chicken proportion of variance explained	chicken cumulative proportion of variance
1	32.8%	32.8%	65.8%	65.8%
2	15.9%	48.8%	10.7%	76.5%
3	11.2%	60.0%	6.3%	82.8%
4	9.2%	69.1%	5.5%	88.3%
5	6.5%	75.6%	5.1%	93.4%
6	5.3%	80.1%	3.1%	96.5%
7	2.7%	83.6%	0.7%	97.2%
8	2.5%	86.1%	0.4%	97.6%
9	1.9%	88.0%	0.4%	98.0%
10	1.8%	89.8%	0.3%	98.3%
11	1.5%	91.3%	0.3%	98.7%
12	1.1%	92.4%	0.3%	99.0%
13	1.0%	93.4%	0.3%	99.2%
14	0.8%	94.1%	0.3%	99.5%
15	0.7%	94.8%	0.3%	99.8%

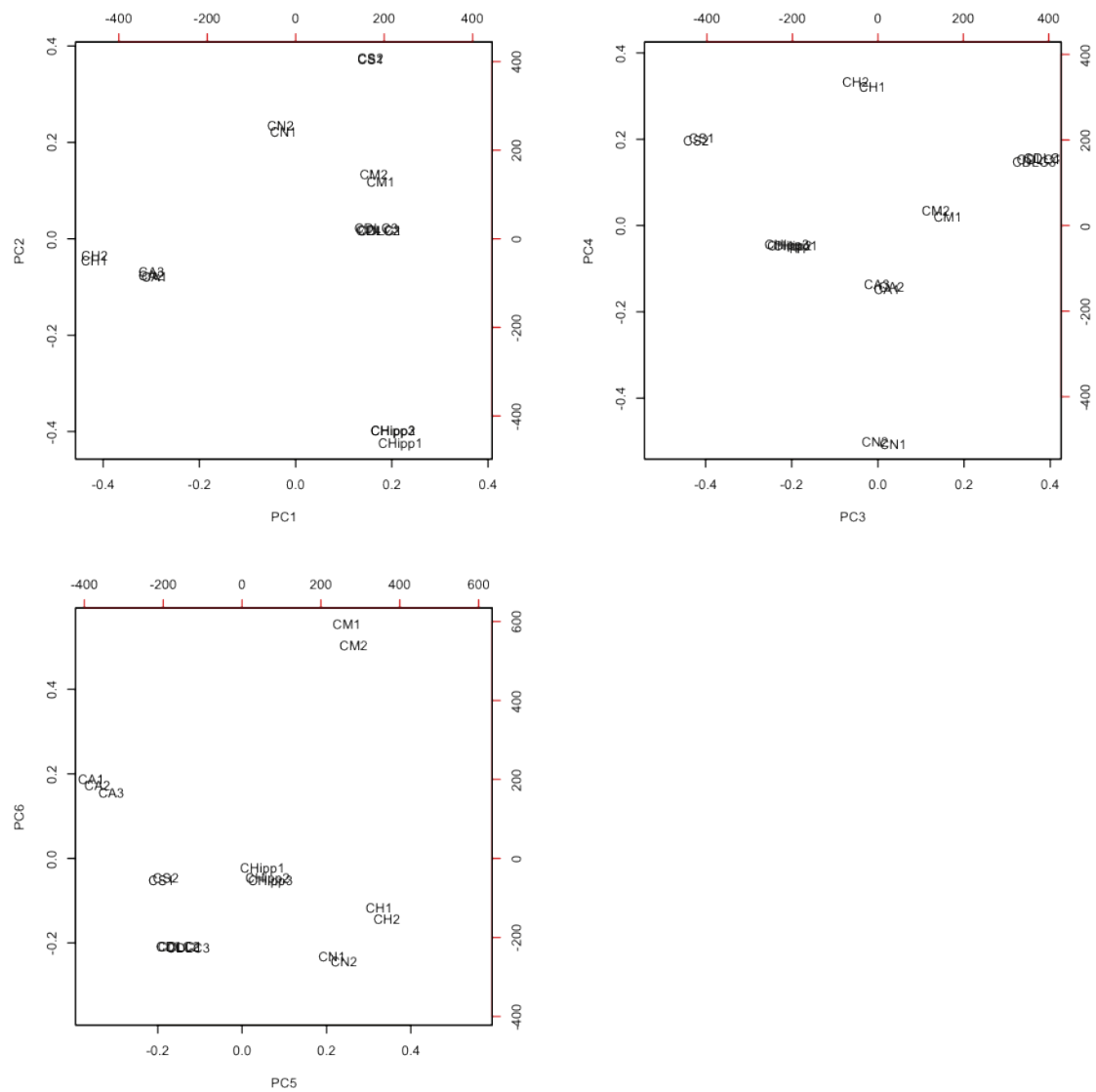


Figure C.1: Different lanes and flowcells of chicken libraries plotted on the first six principal components.

Table C.7: Variance explained by the first 15 components of a library-wise principal component analysis

principal component	mouse proportion of variance explained	mouse cumulative proportion of variance	chicken proportion of variance explained	chicken cumulative proportion of variance
1	34.7%	34.7%	68.8%	68.8%
2	22.6%	57.3%	13.3%	82.1%
3	15.8%	73.0%	7.4%	89.5%
4	7.6%	80.6%	4.6%	94.1%
5	5.8%	86.4%	3.3%	97.4%
6	3.9%	90.3%	2.6%	100%
7	2.2%	92.5%	0%	100%
8	1.7%	94.2%		
9	1.5%	95.7%		
10	1.1%	96.7%		
11	0.8%	97.6%		
12	0.7%	98.3%		
13	0.7%	99.0%		
14	0.5%	99.5%		
15	0.5%	100%		

Table C.8: Primers used for the chicken *in situ* hybridisations

Gene	Forward primer	Reverse primer
<i>RORβ</i>	GGCTTGGGCAATCTGAATAA	GAACATCTGCATCCCTCCAT
<i>RSPO3</i>	CCAAGGAGGGTGTGCTACAT	TTGTTTTCCCTCTGCTCTCG
<i>KCNH7</i>	AGCTGTCTGGGACTGGCTAA	GGATGGTCCAGAACTTGCAT
<i>TSHZ3</i>	CTCACGACACTTTGCAGGAA	GTTTTTGGCACAGGTGAGGT
<i>MGAT4C</i>	TGGGCTATCTTCCGTGAAAC	ATGTGCTGGAACAGTGATGG

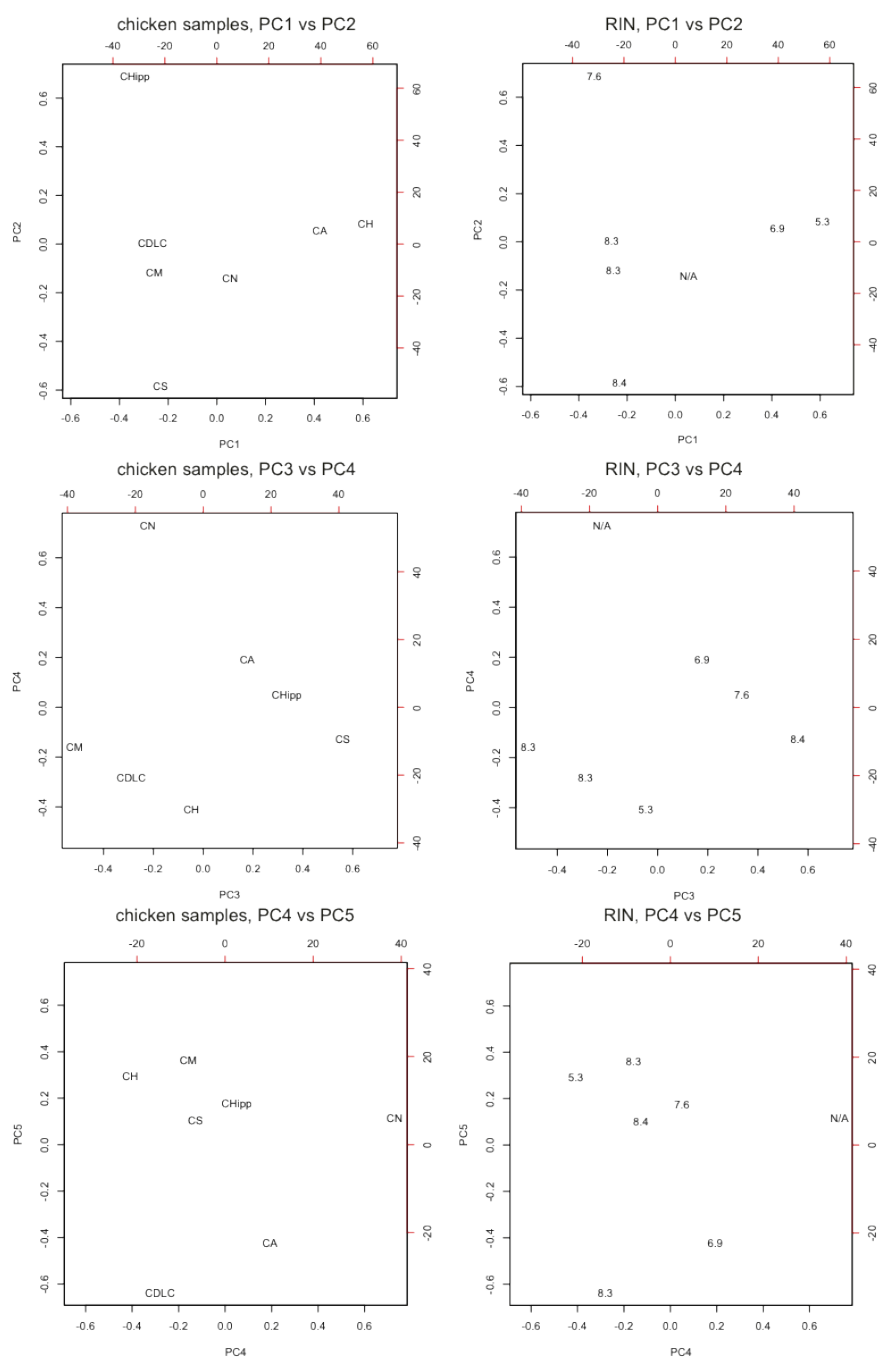


Figure C.3: RNA Integrity Numbers (RINs) of chicken samples plotted on chicken principal components.

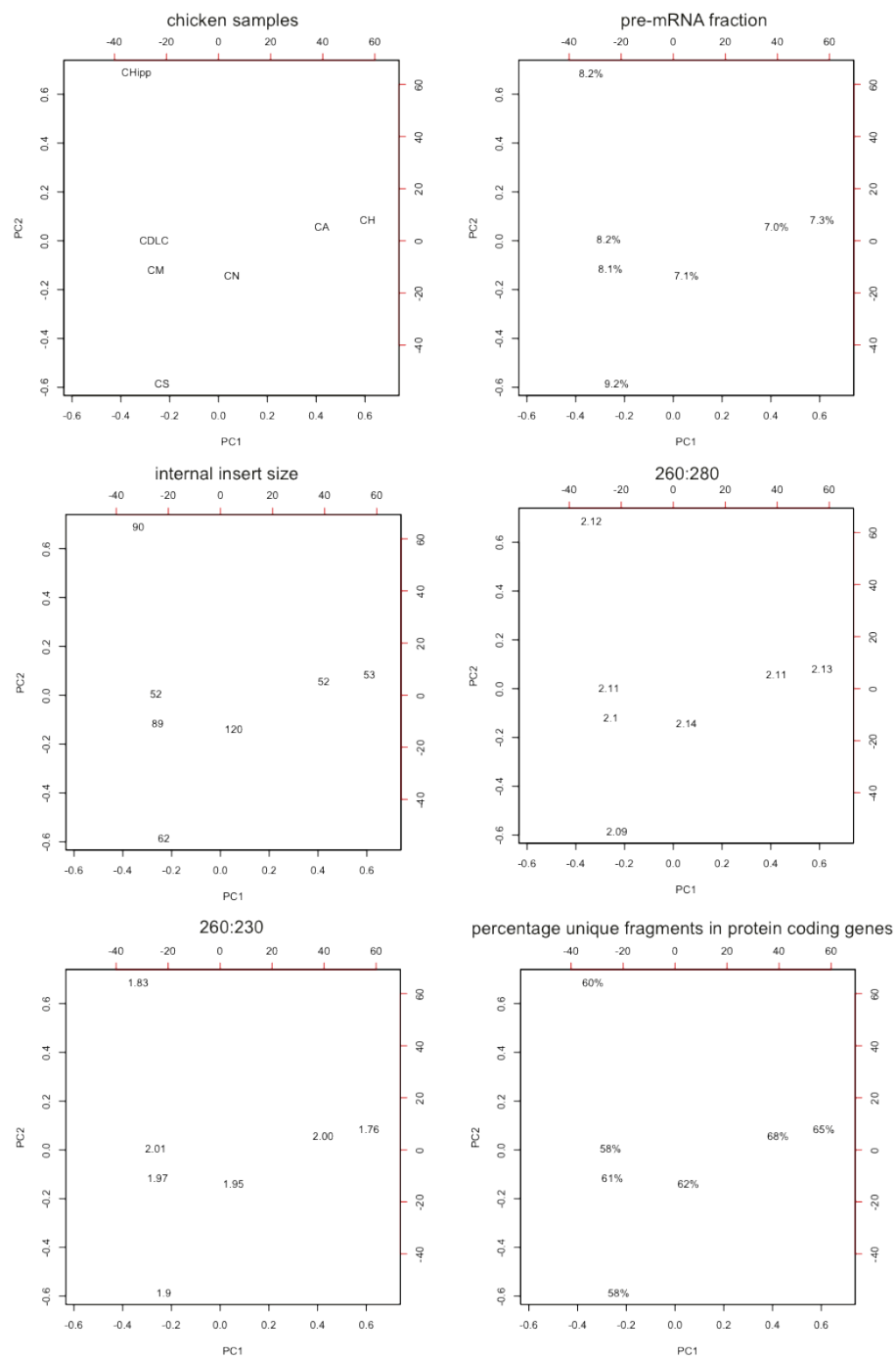


Figure C.4: Technical variables plotted on the first two chicken principal components.

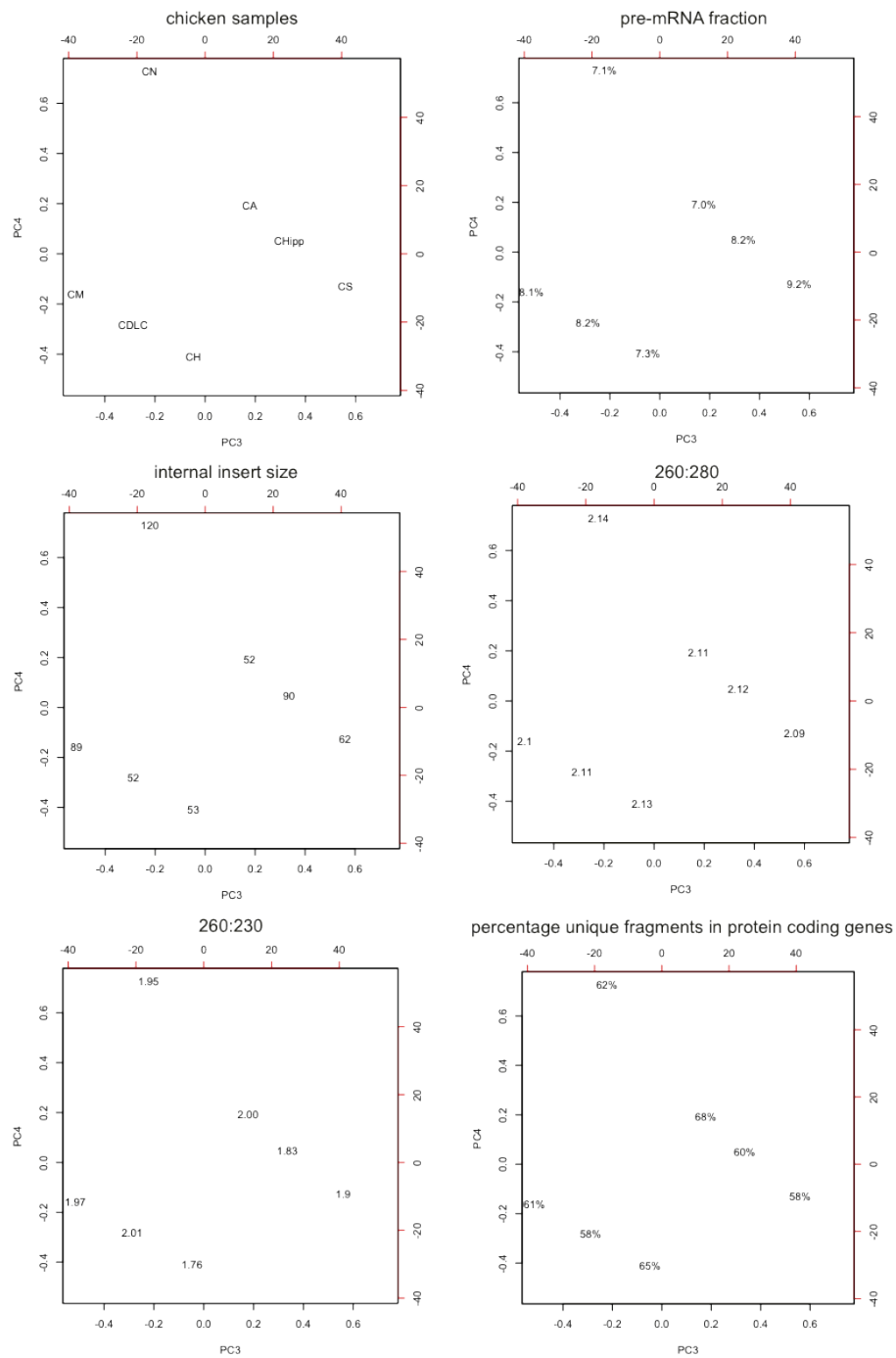


Figure C.5: Technical variables plotted on the third and fourth chicken principal components.

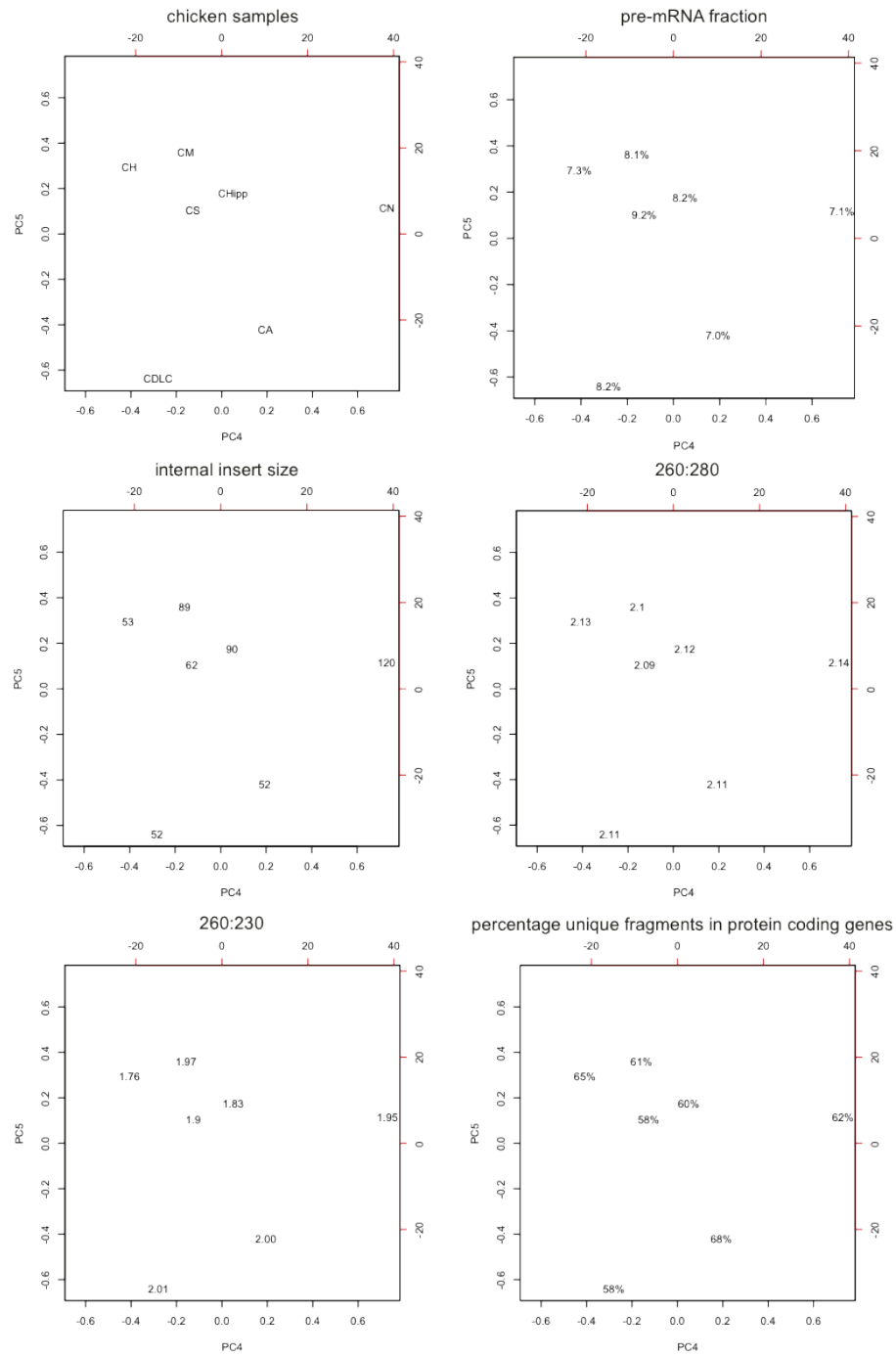


Figure C.6: Technical variables plotted on the first and fifth chicken principal components.

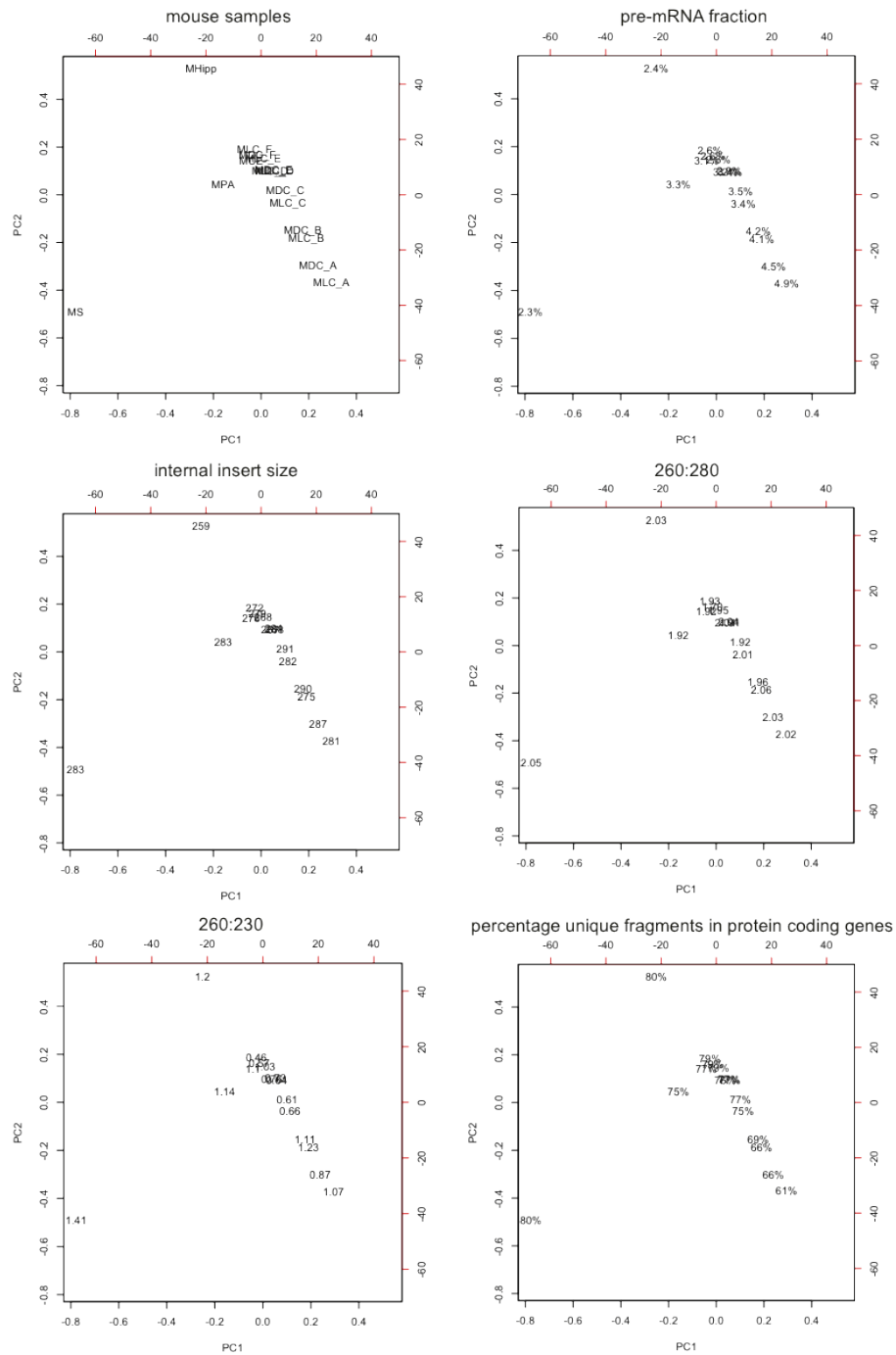


Figure C.7: Technical variables plotted on the first and second mouse principal components.

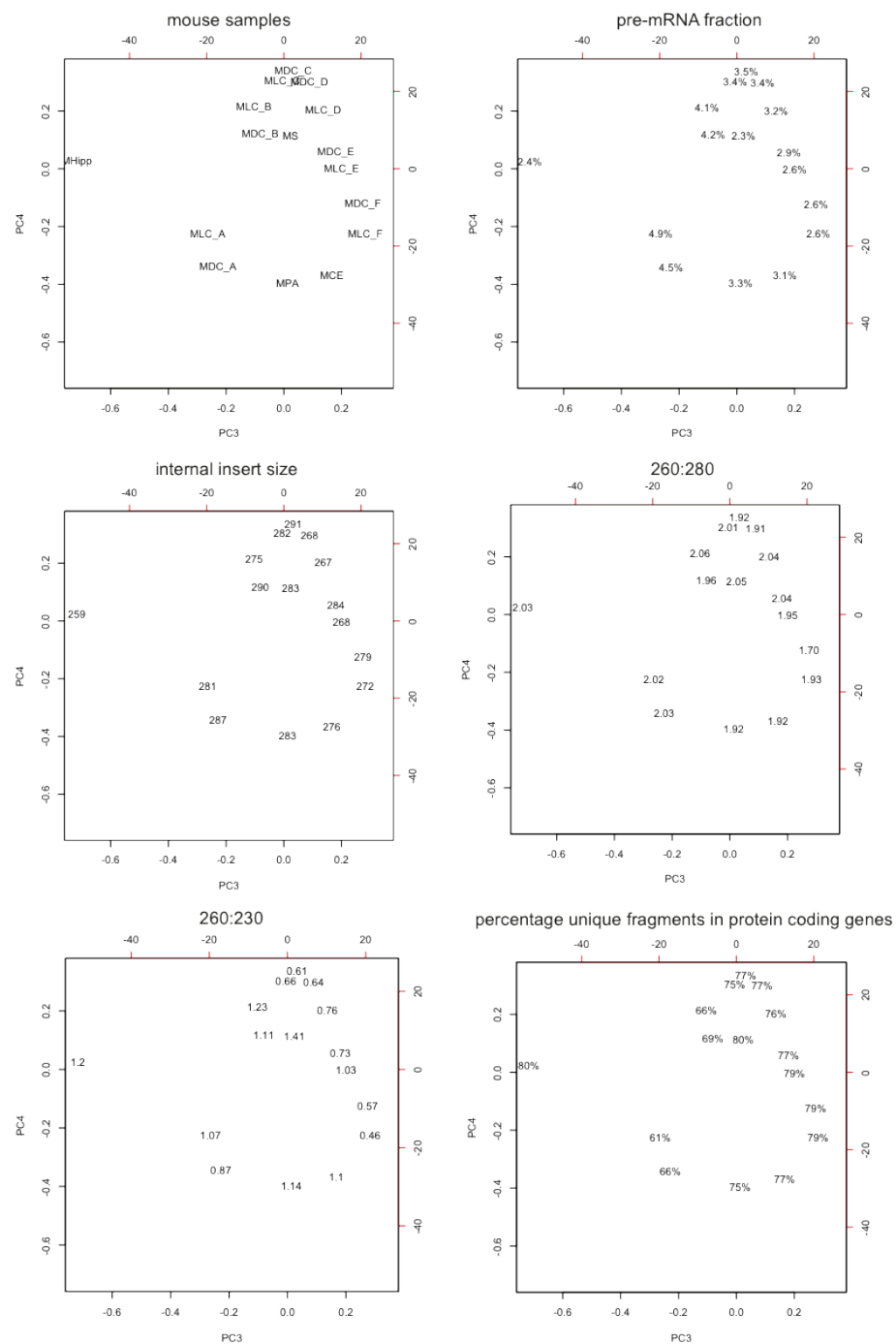


Figure C.8: Technical variables plotted on the third and fourth mouse principal components.

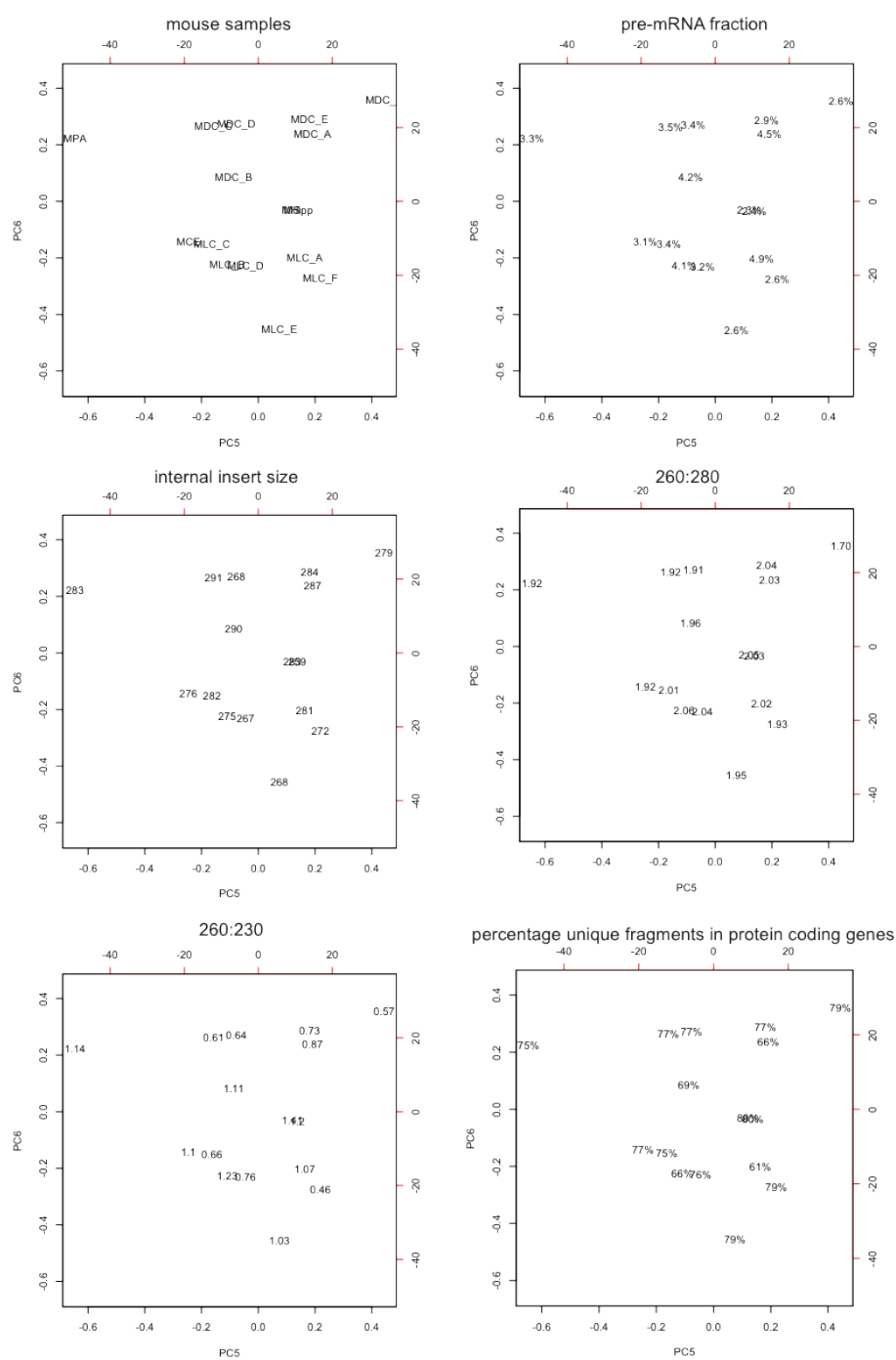


Figure C.9: Technical variables plotted on the fifth and sixth mouse principal components.

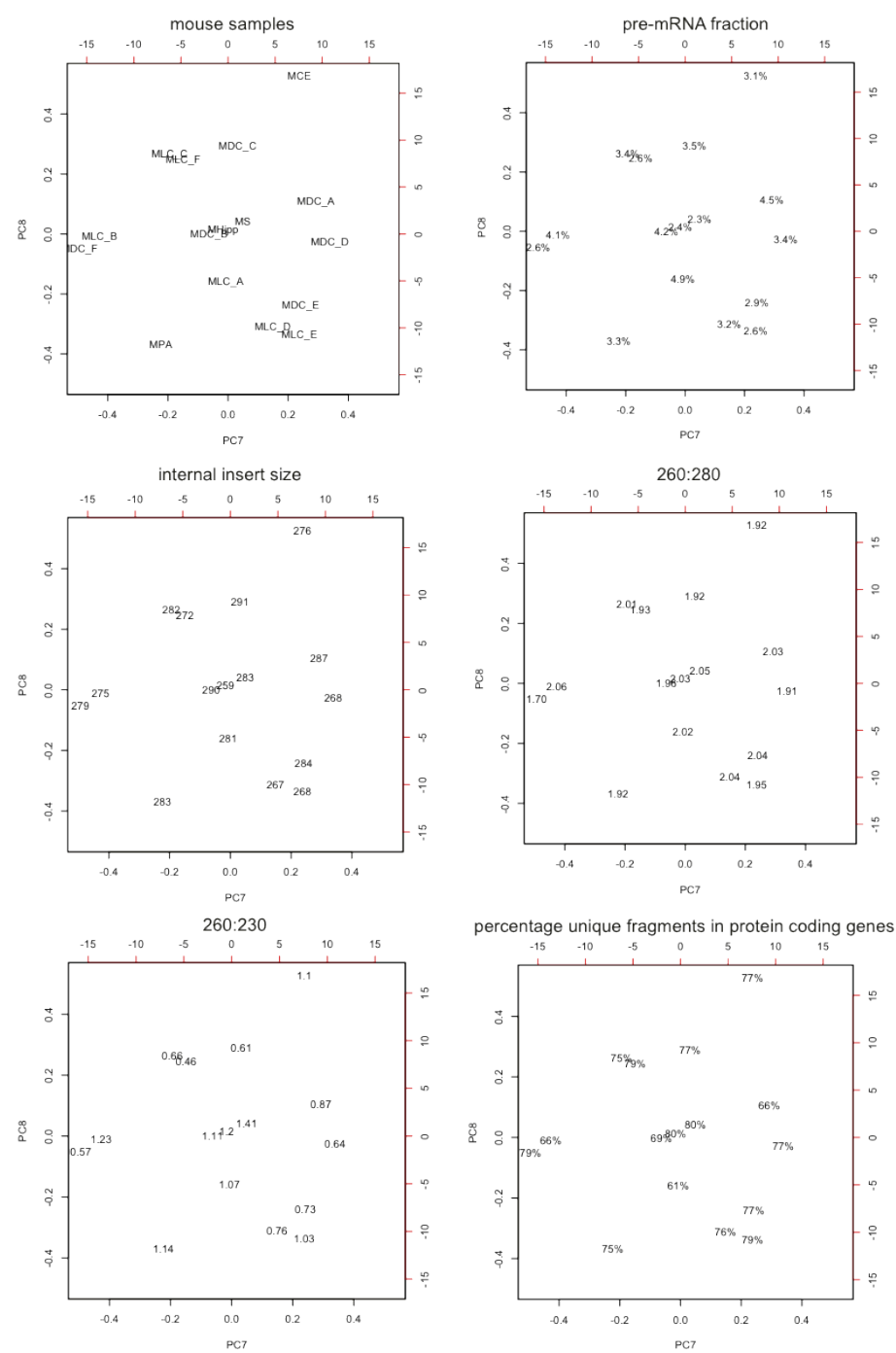


Figure C.10: Technical variables plotted on the seventh and eighth mouse principal components.

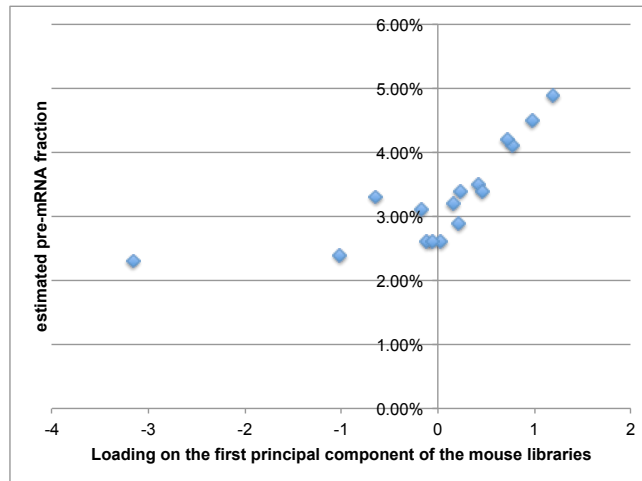


Figure C.12: The first principal component of the mouse libraries was correlated ($r = 0.73$; two-tailed $p = 0.0014$) with the pre-mRNA fraction.

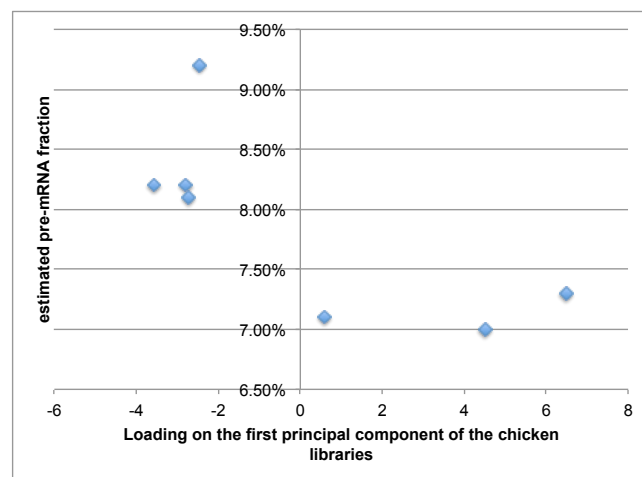


Figure C.13: The first principal component of the chicken libraries may also be correlated ($r = -0.74$; two-tailed $p = 0.056$) with the pre-mRNA fraction.

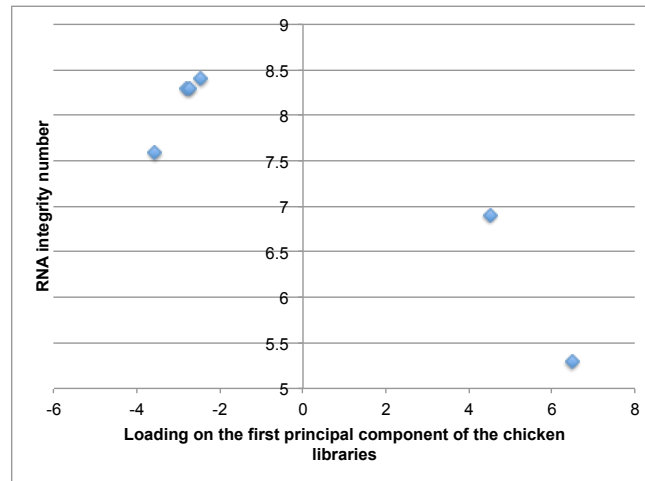


Figure C.14: The first principal component of the chicken libraries was correlated ($r = -0.90$; two-tailed $p = 0.013$) with RNA Integrity Number.

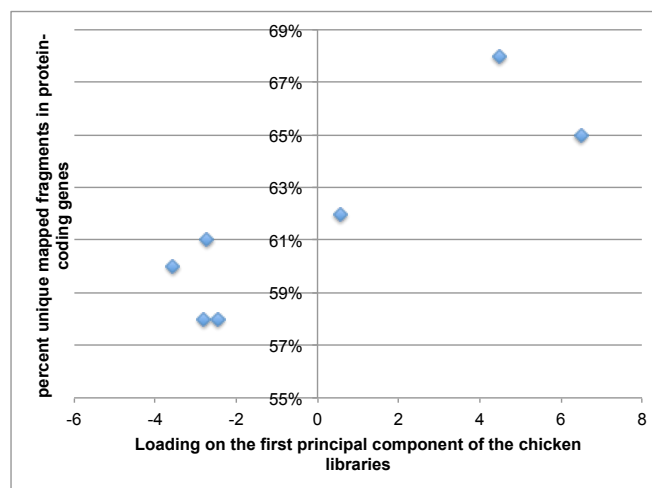


Figure C.15: The first principal component of the chicken libraries was correlated ($r = 0.87$; two-tailed $p = 0.011$) with the percentage of uniquely mapped fragments in protein-coding genes.

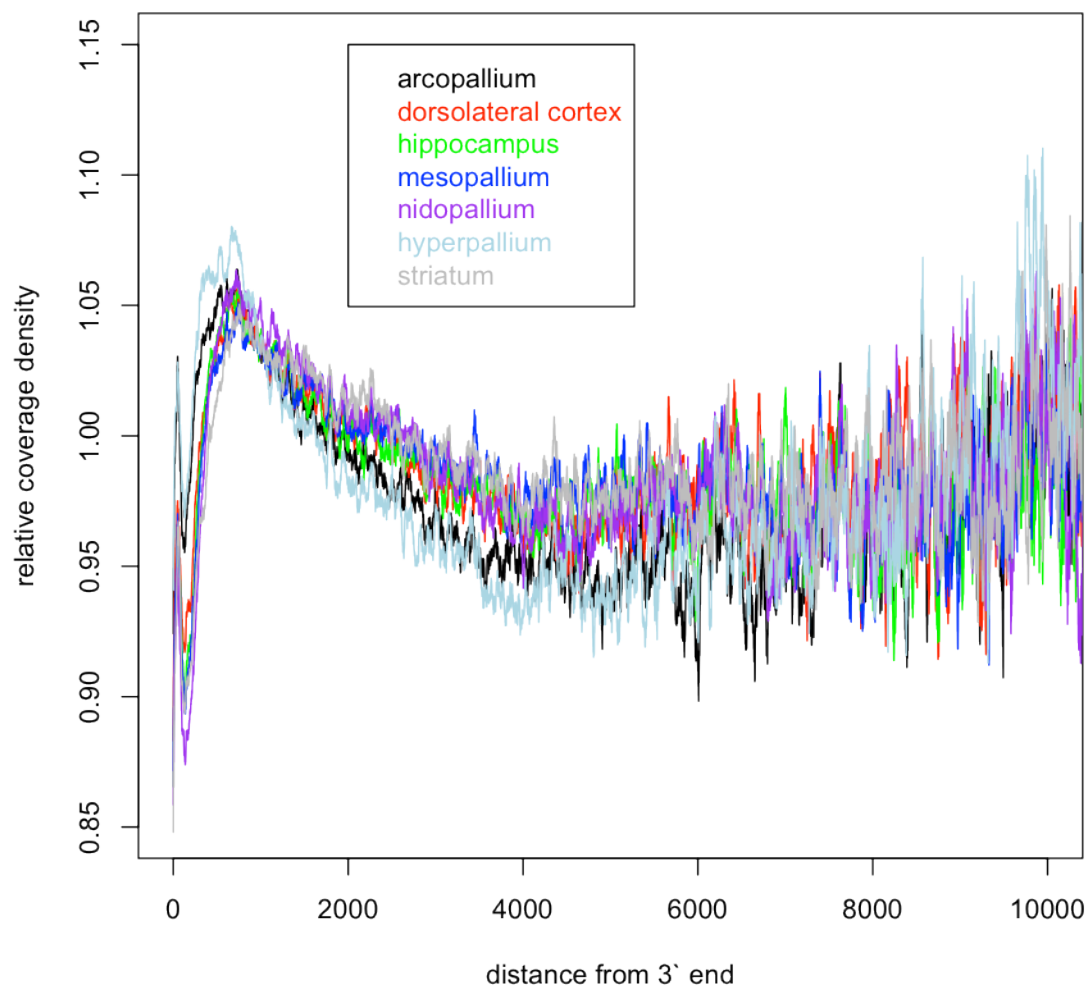


Figure C.16: Relative read coverage density in major chicken transcripts with respect to distance from the 3' end.

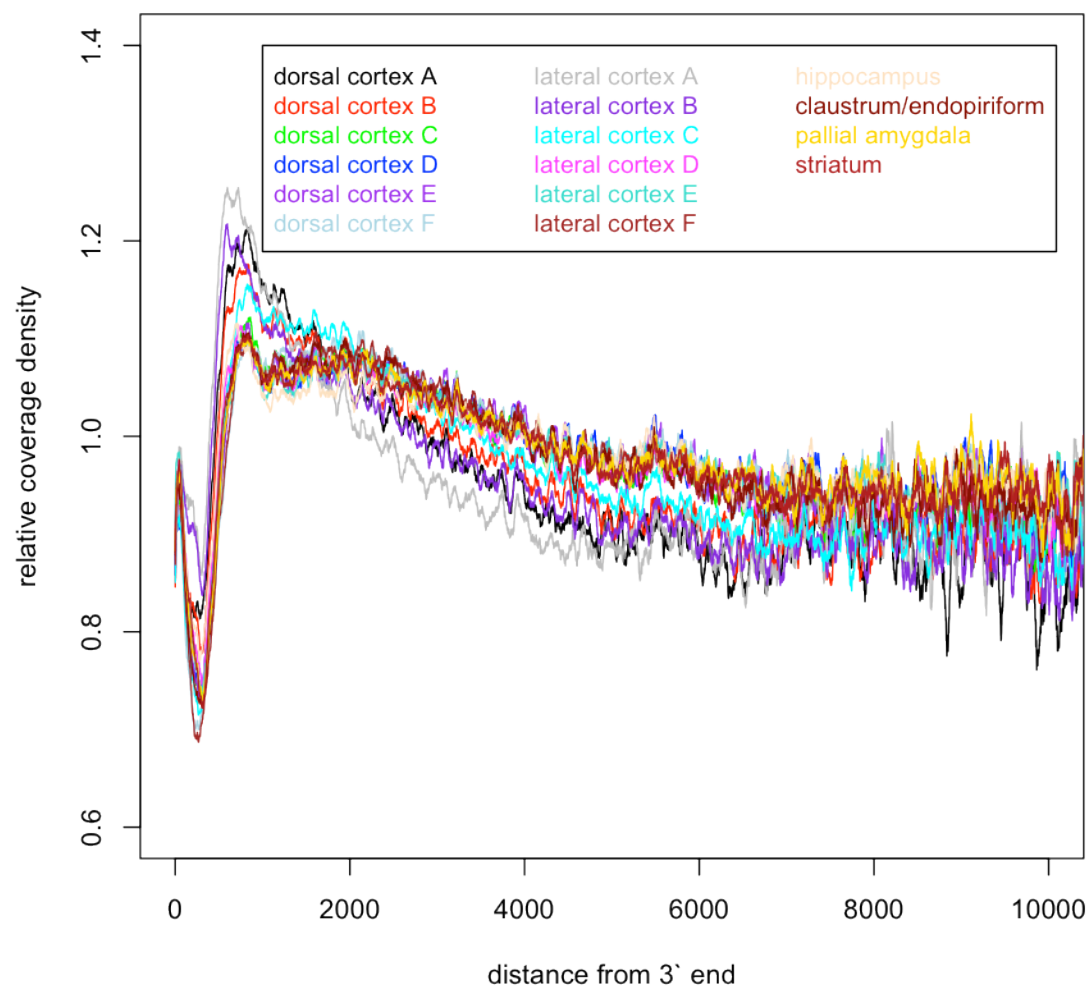


Figure C.17: Relative read coverage density in major mouse transcripts with respect to distance from the 3' end.

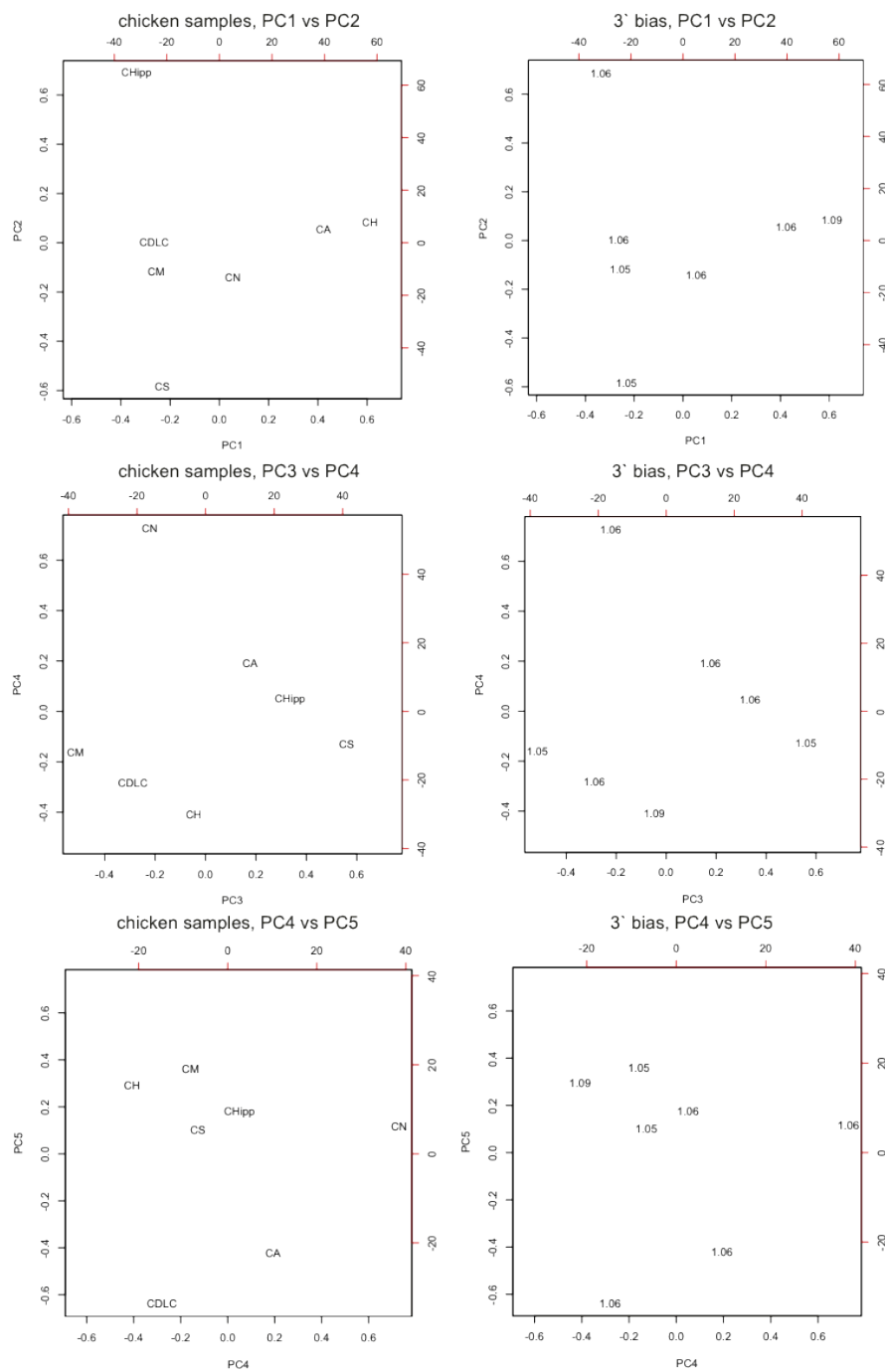


Figure C.18: Measures of 3' bias of chicken samples plotted on chicken principal components.

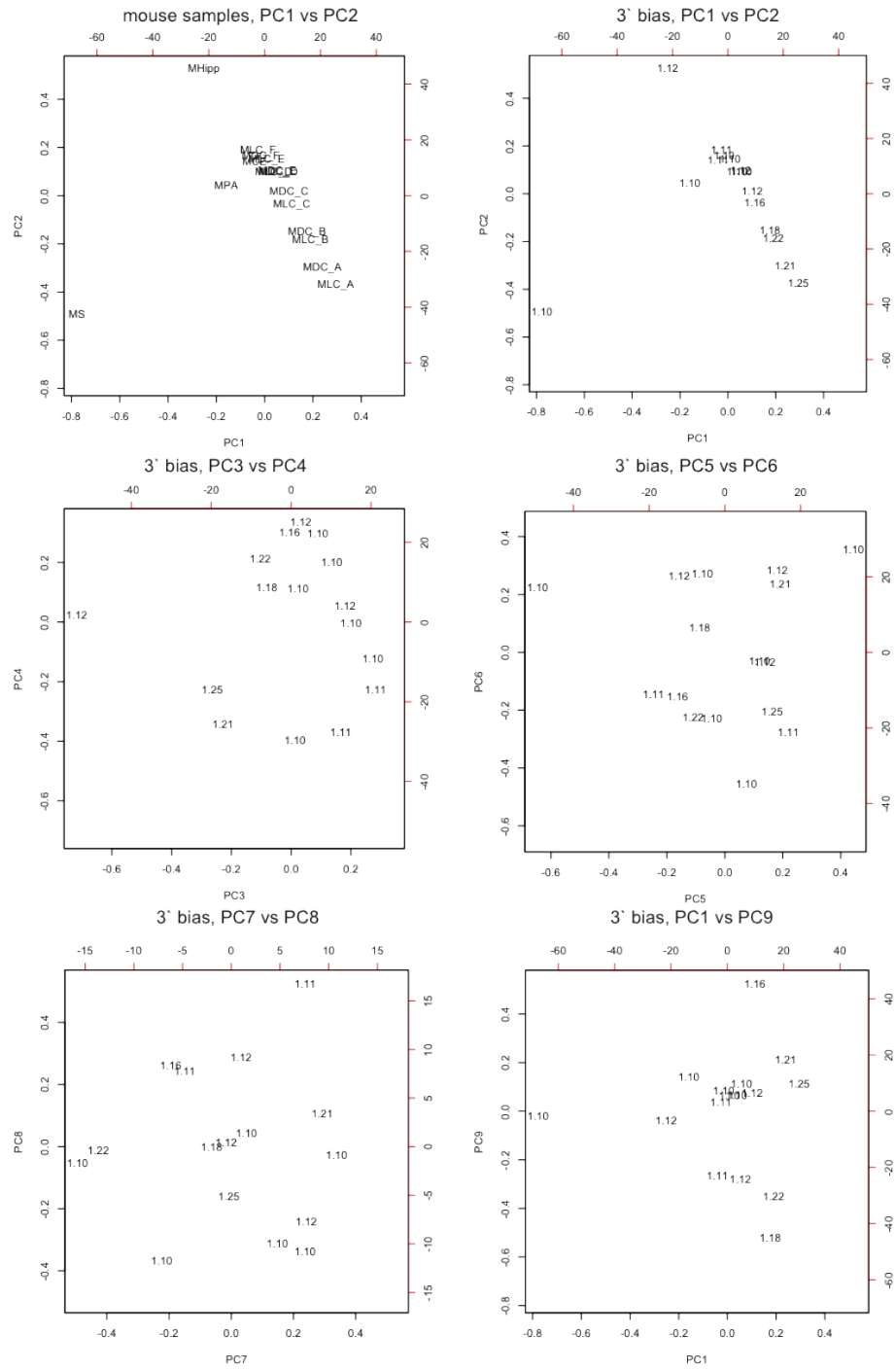


Figure C.19: Measures of 3' bias of mouse samples plotted on mouse principal components.

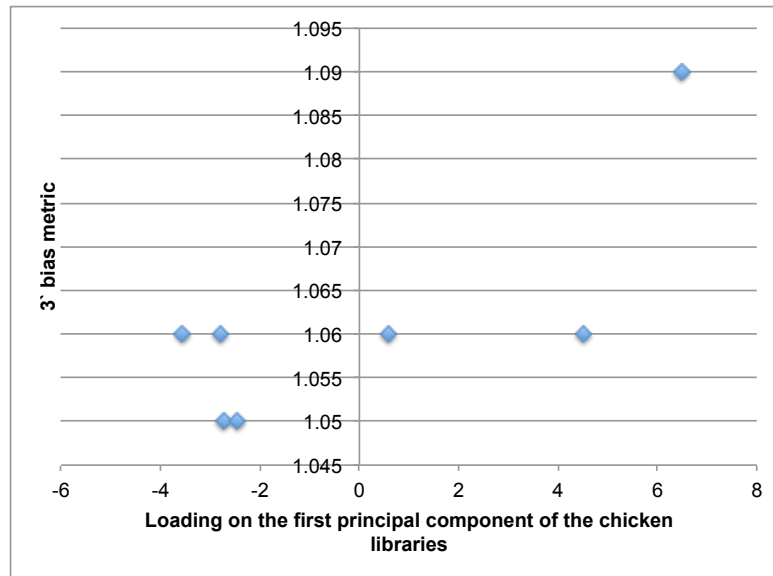


Figure C.20: The first principal component of the chicken libraries was correlated ($r = 0.76$; two-tailed $p = 0.047$) with a metric of 3' bias.

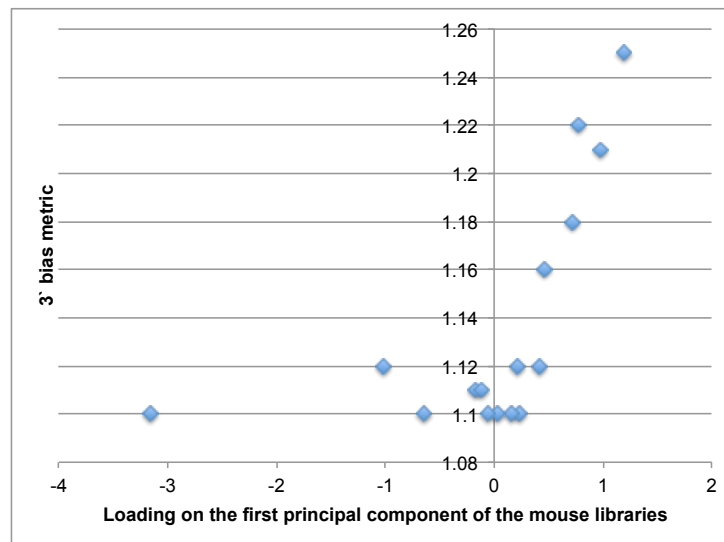


Figure C.21: The first principal component of the mouse libraries was correlated ($r = 0.59$; two-tailed $p = 0.016$) with a metric of 3' bias.

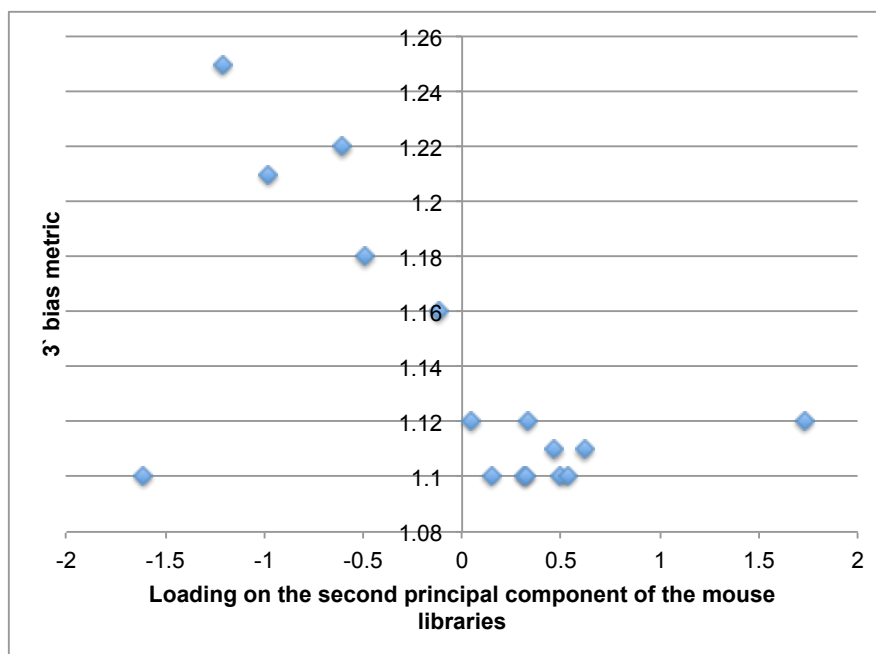


Figure C.22: The second principal component of the mouse libraries was correlated ($r = -0.57$; two-tailed $p = 0.020$) with a metric of 3' bias.

Table C.9: Scale-free topology (SFT) approximation and connectivity (k) in chicken for different values of the soft threshold.

Power	SFT R^2	slope	mean k	median k	max k
1	0.632	1.53	3170	3560	3960
2	0.36	0.589	2360	2760	3380
3	0.153	0.27	1890	2230	3010
4	0.0311	0.102	1580	1850	2730
5	0.000157	-0.00626	1360	1560	2510
6	0.0308	-0.0814	1190	1330	2330
7	0.103	-0.144	1050	1140	2170
8	0.204	-0.197	941	993	2040
9	0.332	-0.243	850	868	1930
10	0.425	-0.279	774	764	1820
12	0.599	-0.343	652	600	1650
14	0.712	-0.393	560	479	1510
16	0.797	-0.437	488	386	1390
18	0.846	-0.473	430	318	1290
20	0.878	-0.505	383	264	1200
22	0.907	-0.533	344	222	1120
24	0.932	-0.562	311	189	1060
26	0.943	-0.587	284	160	998
28	0.953	-0.605	260	137	949
30	0.961	-0.63	239	119	905

Table C.10: Scale-free topology (SFT) approximation and connectivity (k) in mouse for different values of the soft threshold.

Power	SFT R^2	slope	mean k	median k	max k
1	0.562	2.06	1860	1860	2710
2	0.0996	0.35	942	907	1790
3	0.159	-0.353	557	511	1300
4	0.511	-0.711	362	313	991
5	0.649	-0.941	249	203	786
6	0.721	-1.07	180	139	640
7	0.749	-1.19	134	97.8	530
8	0.781	-1.25	103	70.7	447
9	0.796	-1.3	80.3	52.1	381
10	0.805	-1.35	64.1	39.1	328
12	0.824	-1.42	42.5	22.8	249
14	0.835	-1.46	29.6	14.1	194
16	0.846	-1.48	21.4	9.1	155
18	0.856	-1.49	15.9	6.09	125
20	0.866	-1.49	12.1	4.2	102
22	0.874	-1.5	9.38	2.96	85
24	0.876	-1.49	7.42	2.14	71.1
26	0.885	-1.47	5.95	1.57	60
28	0.893	-1.46	4.85	1.18	51.1
30	0.919	-1.41	3.99	0.89	43.7

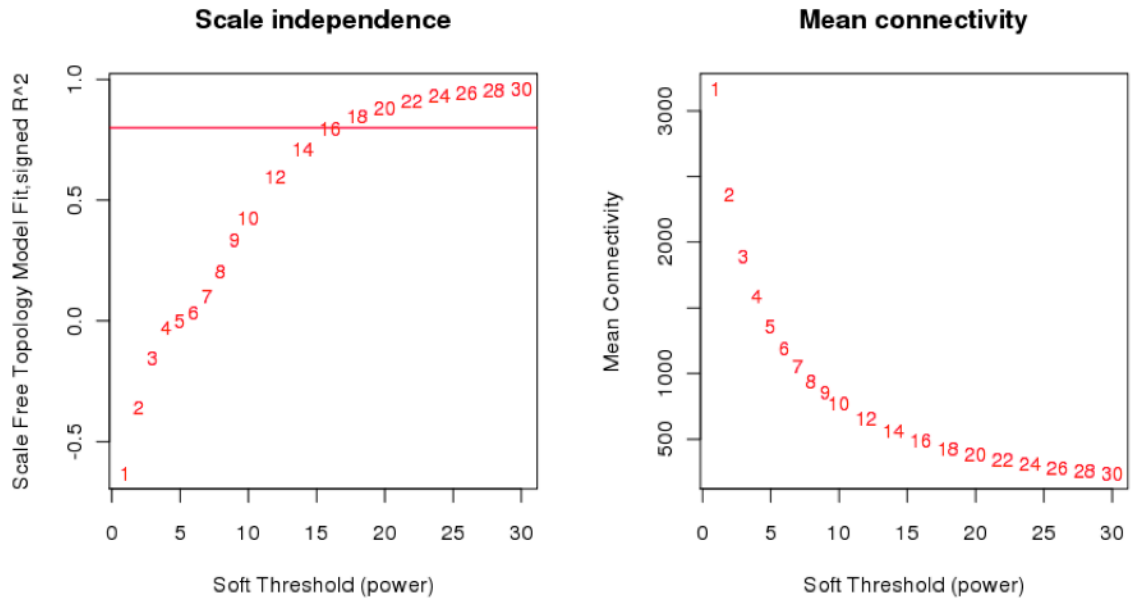


Figure C.23: Plots showing the relationships between chicken soft threshold and both fit to a scale free topology model and mean connectivity. The horizontal line on the left indicates an R^2 of 0.8.

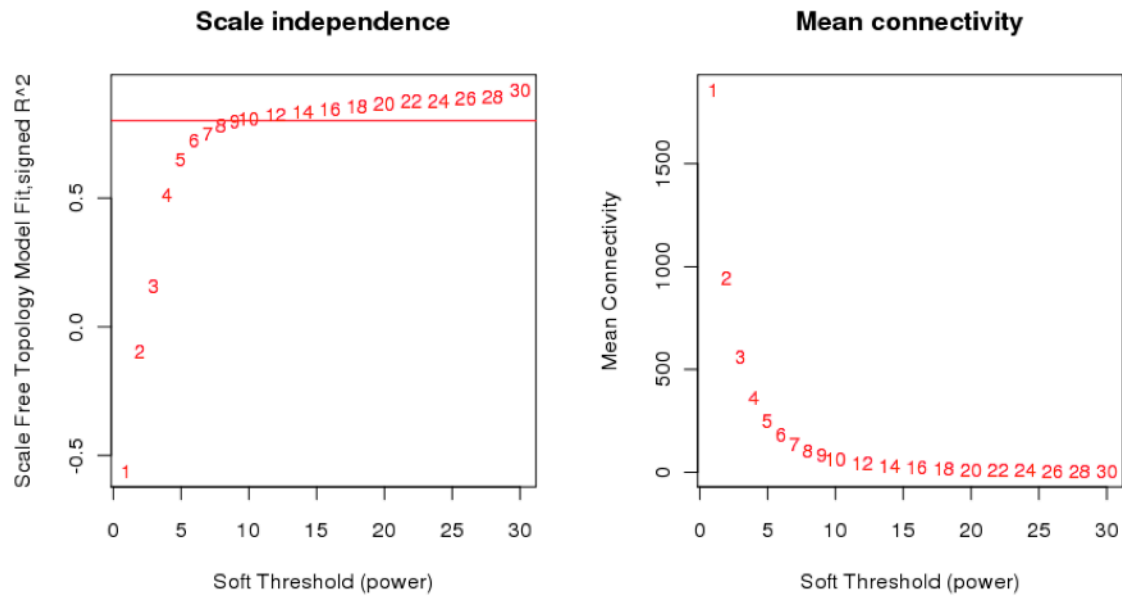


Figure C.24: Plots showing the relationships between mouse soft threshold and both fit to a scale free topology model and mean connectivity. The horizontal line on the left indicates an R^2 of 0.8.

Table C.11: Number of preserved modules ($Z_{\text{pres}} > 3$) with different soft thresholds and deep split parameters

Mouse Power	Chicken Power	Species	First split	Second split	Third split	Fourth split	Fifth split
10	24	Mouse	0	1	0	2	0
		Chicken	1	1	1	1	1
10	10	Mouse	0	1	0	2	0
		Chicken	1	1	1	2	0
10	15	Mouse	0	1	0	2	0
		Chicken	1	1	2	1	2
15	24	Mouse	0	1	1	2	1
		Chicken	1	1	1	1	1
15	15	Mouse	0	1	1	2	1
		Chicken	1	1	2	1	2
10	20	Mouse	0	1	0	2	0
		Chicken	1	1	3	1	1

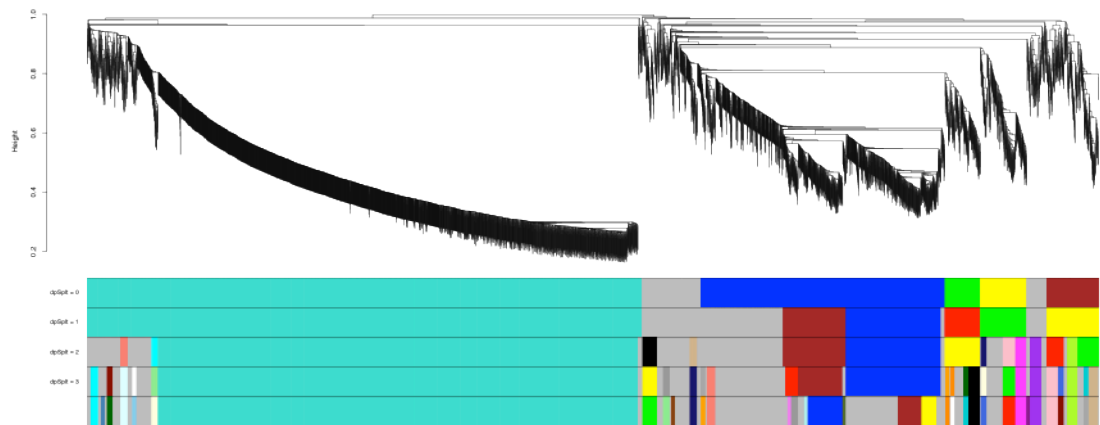


Figure C.25: The chicken dendrogram is shown at top with the module choices for different splitting parameters at bottom.

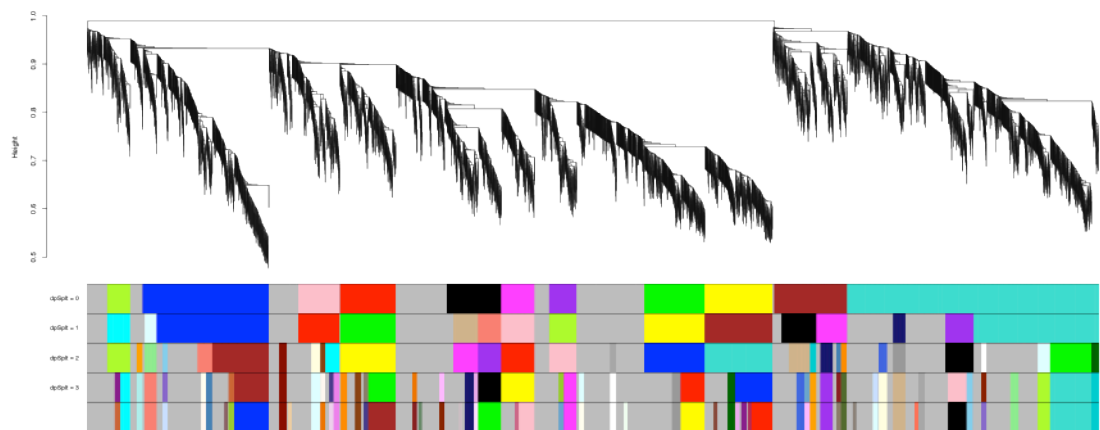


Figure C.26: The mouse dendrogram is shown at top with the module choices for different splitting parameters at bottom.

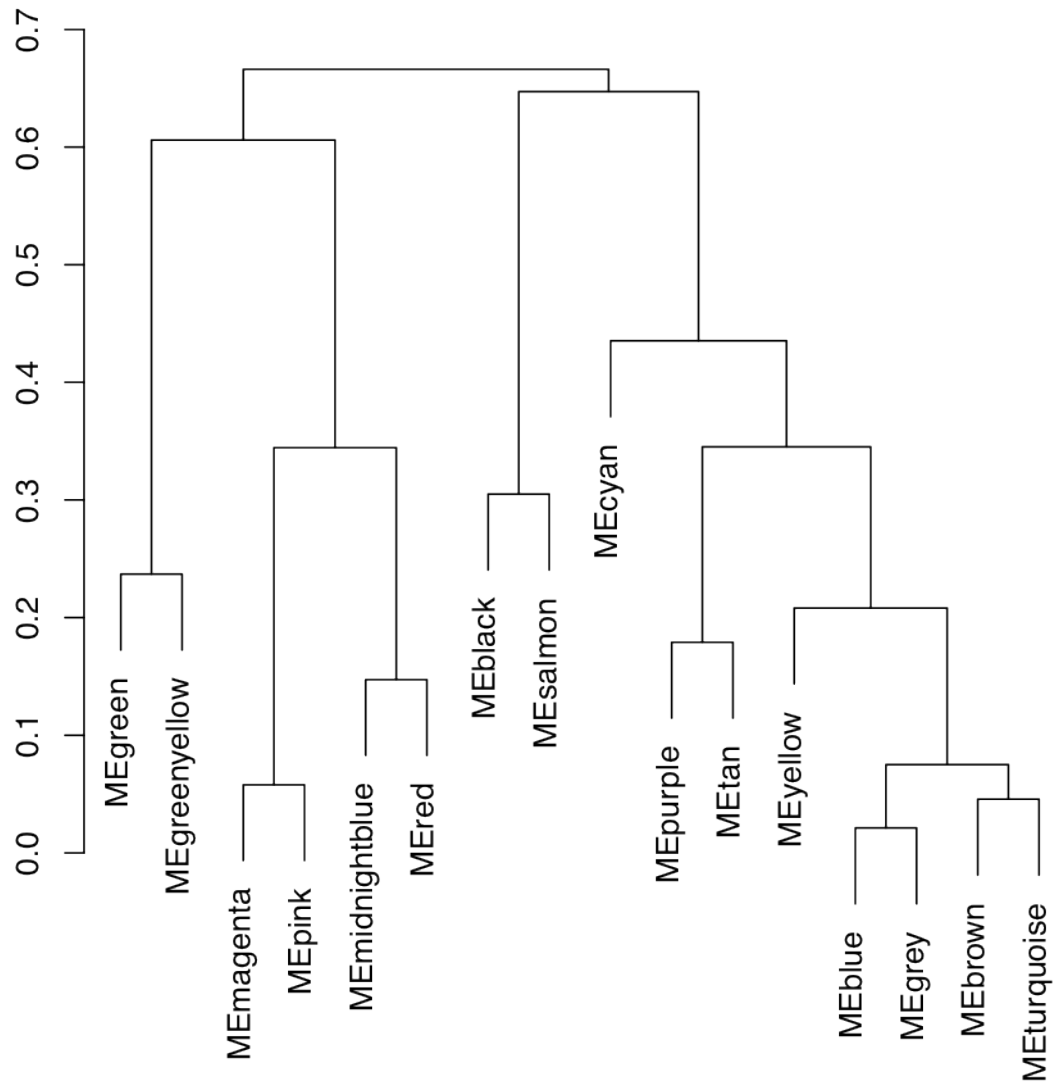


Figure C.27: Hierarchical clustering of the chicken module eigengenes using the hclust R function (agglomerating on the average).

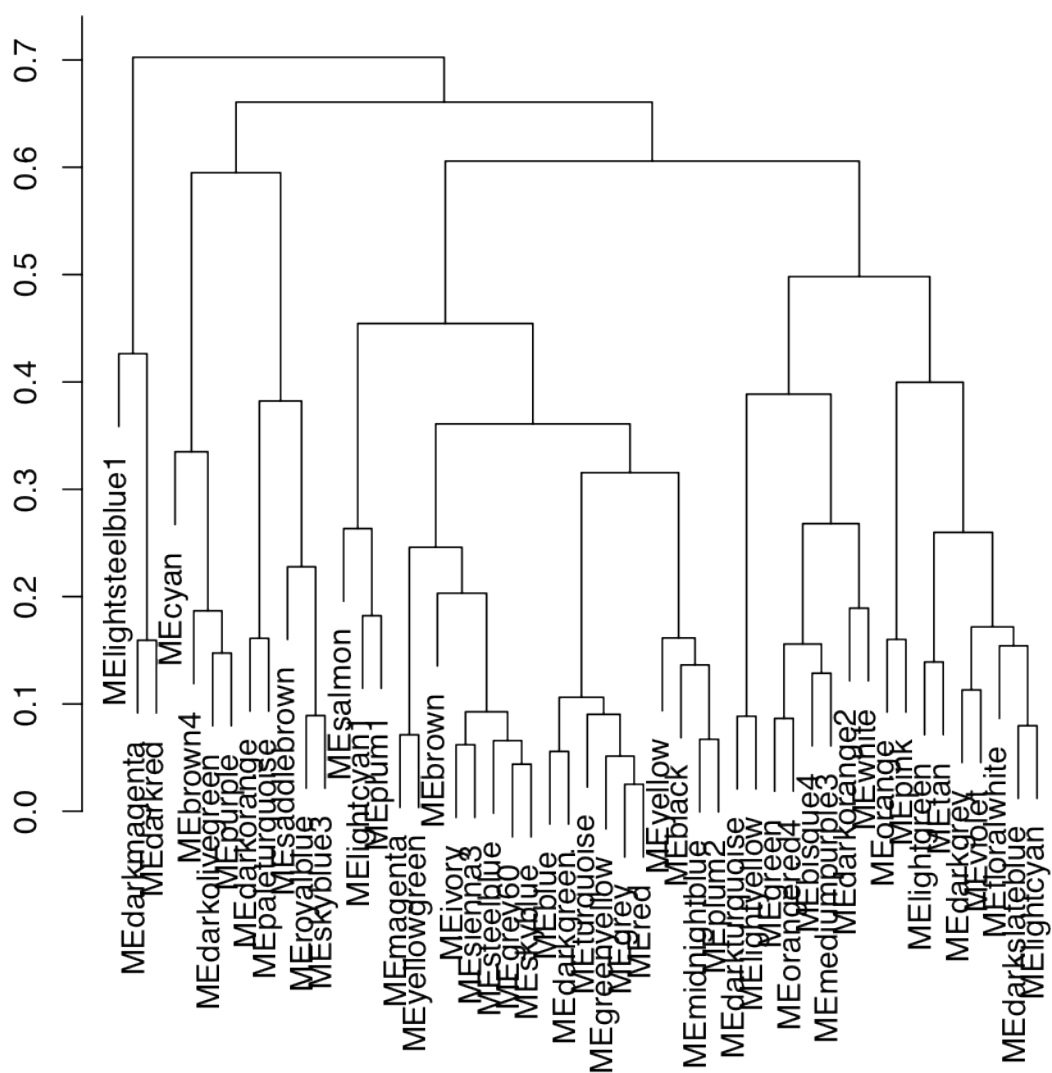


Figure C.28: Hierarchical clustering of the mouse module eigengenes using the hclust R function (agglomerating on the average).

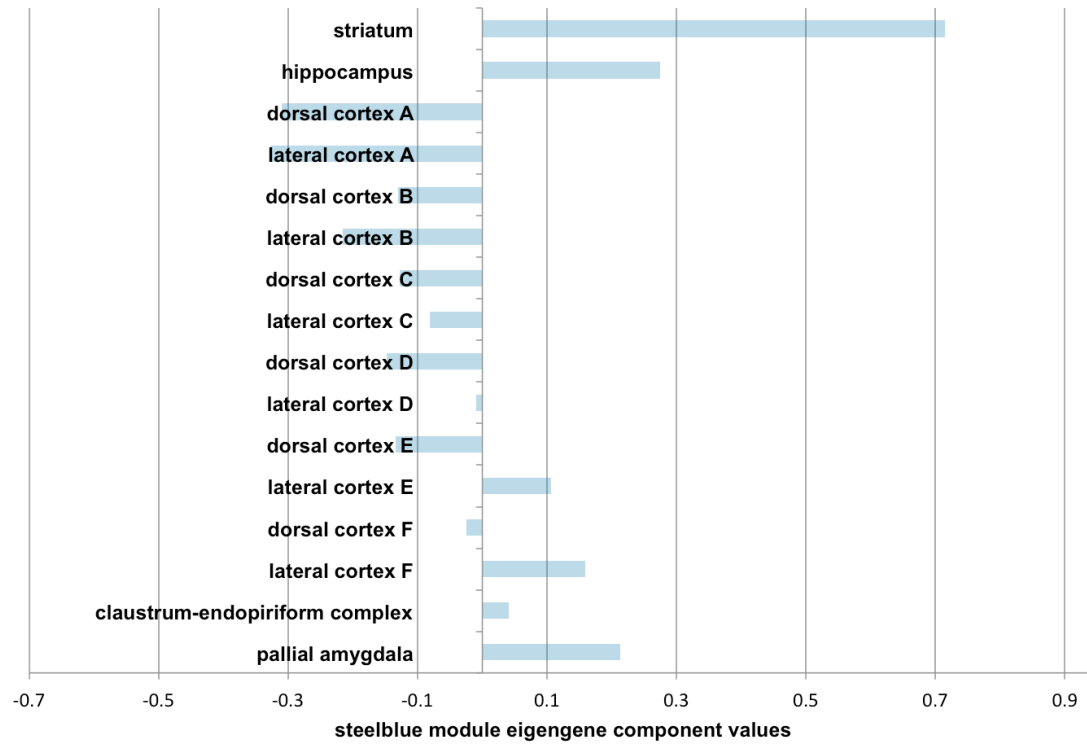
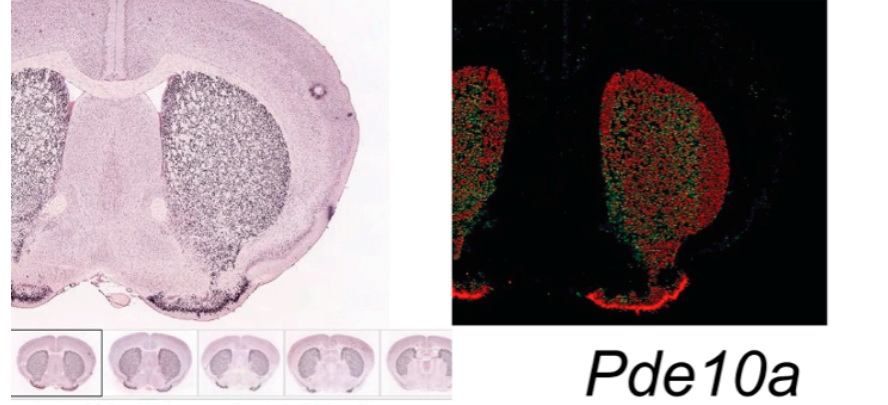
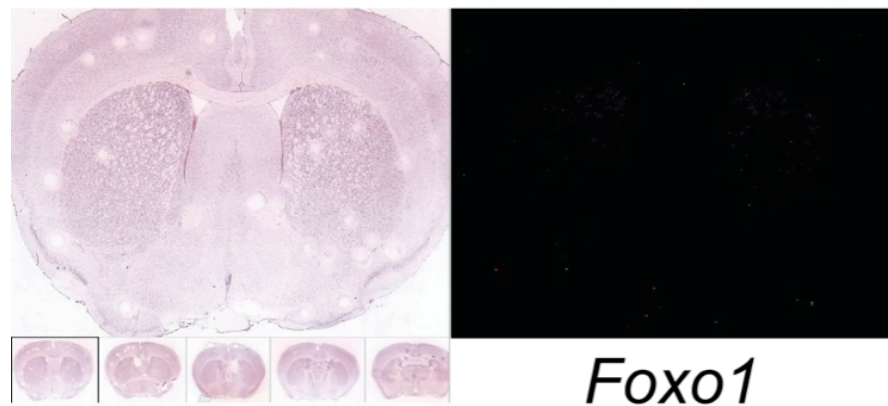
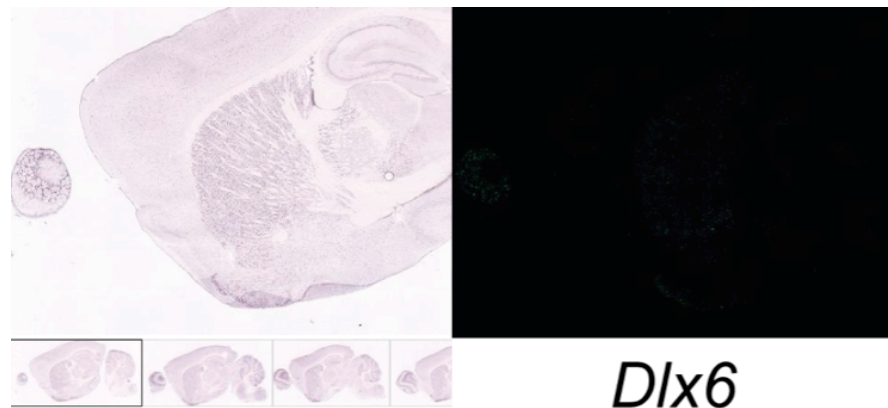
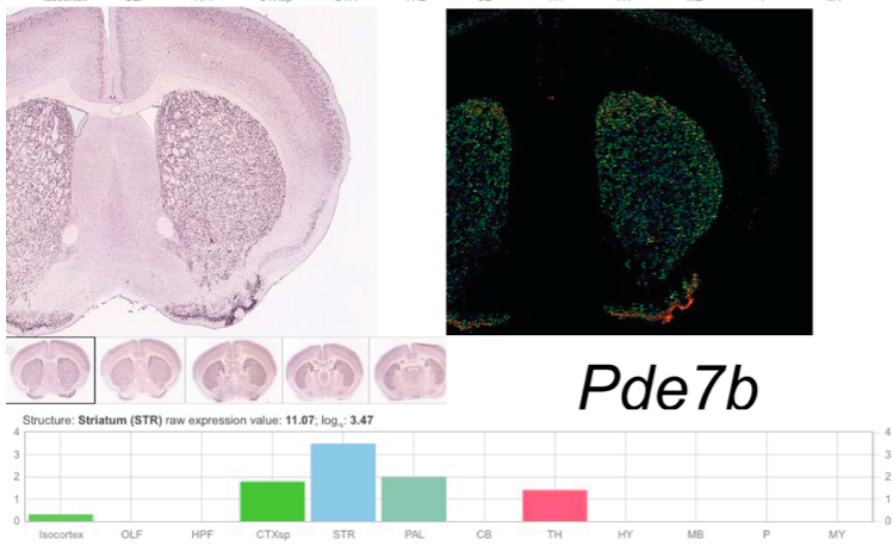
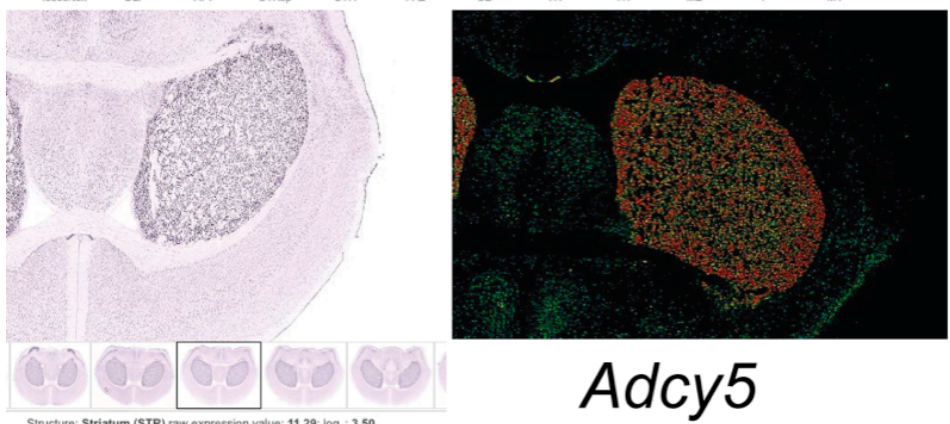
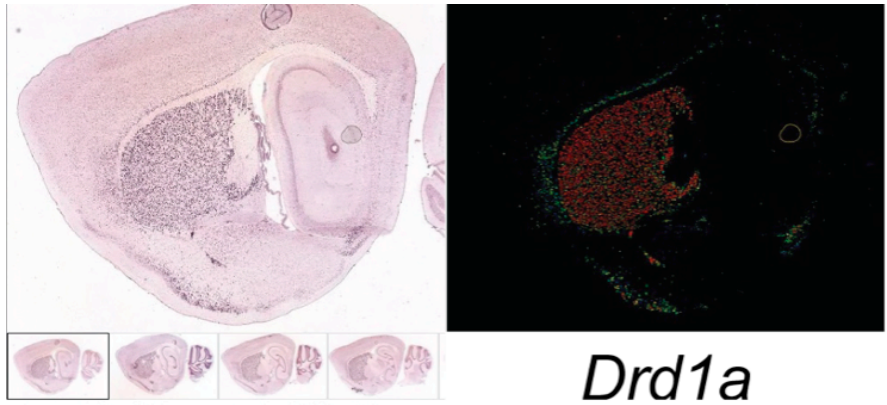


Figure C.29: The steelblue mouse module eigengene is highly expressed in striatum.

The x-axis represents the values of the components of the module eigengene.





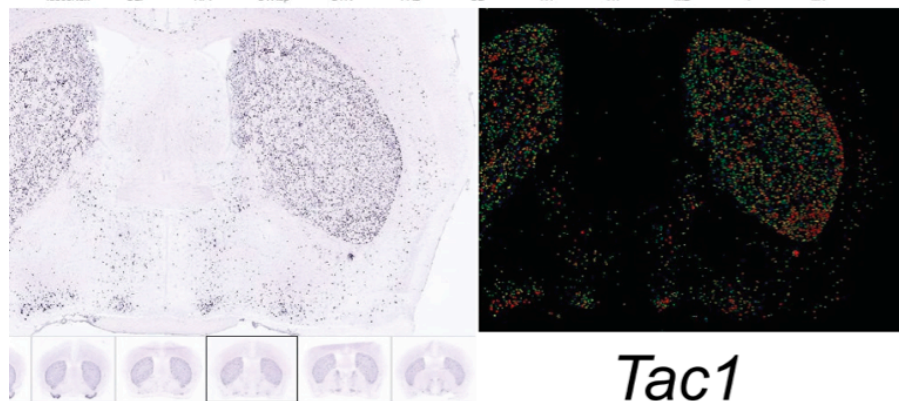
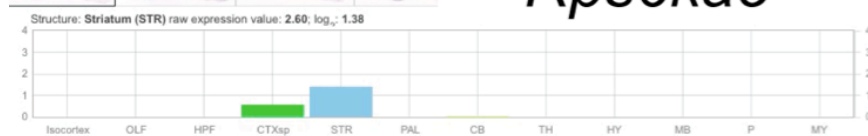
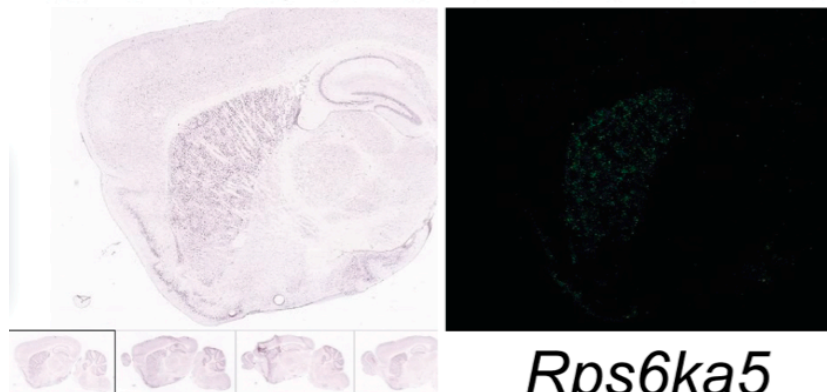
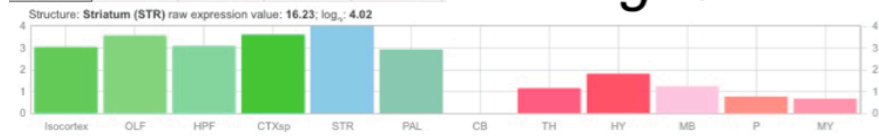
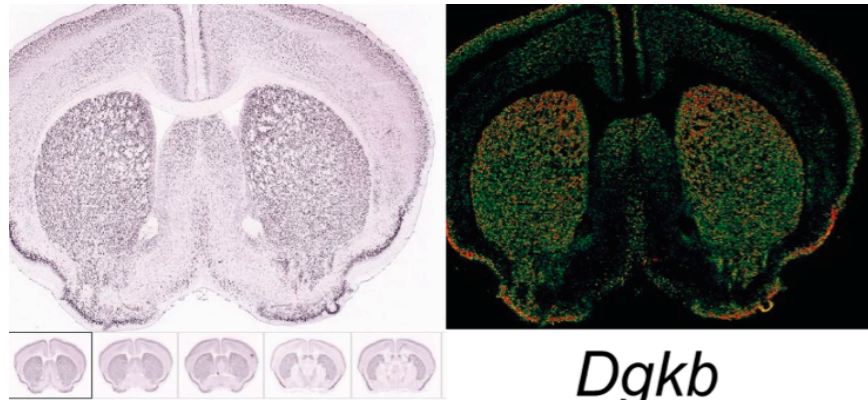
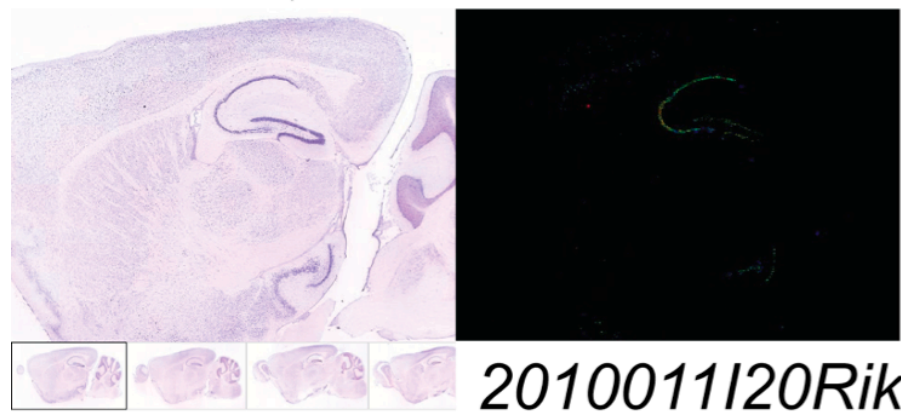
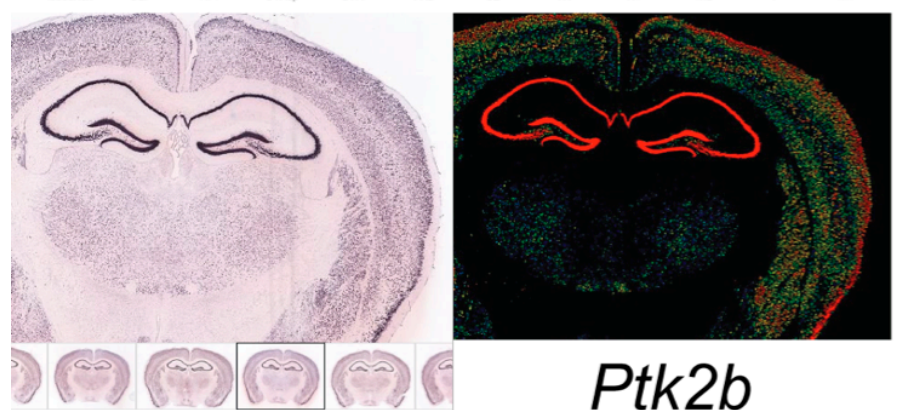
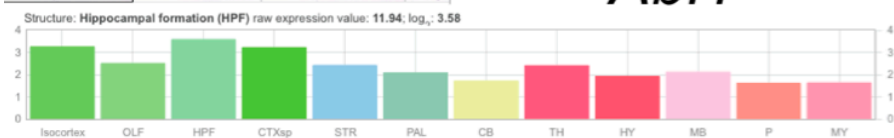
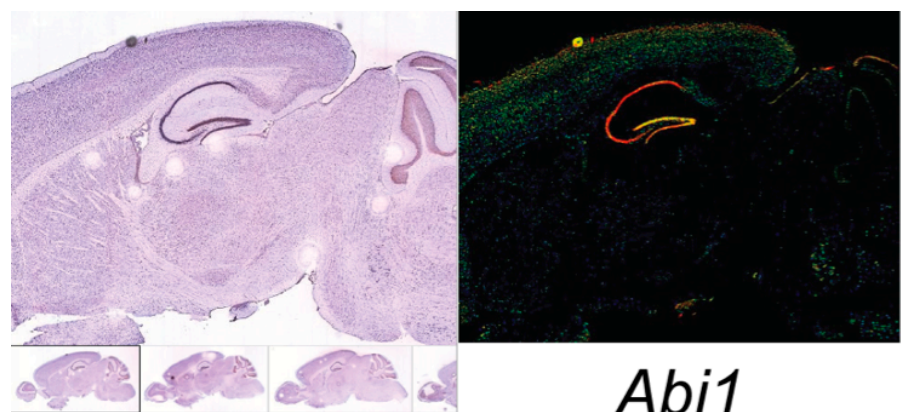


Figure C.30: *In situ* hybridisations, expression analyses of the hybridisations, and automated expression analyses across several major regions (STR corresponds to striatum) from the Allen Mouse Brain Atlas (Lein et al. 2007) showing the nine unique genes from the top five in either the chick green/mouse brown overlap or the chick greenyellow/mouse brown overlap.



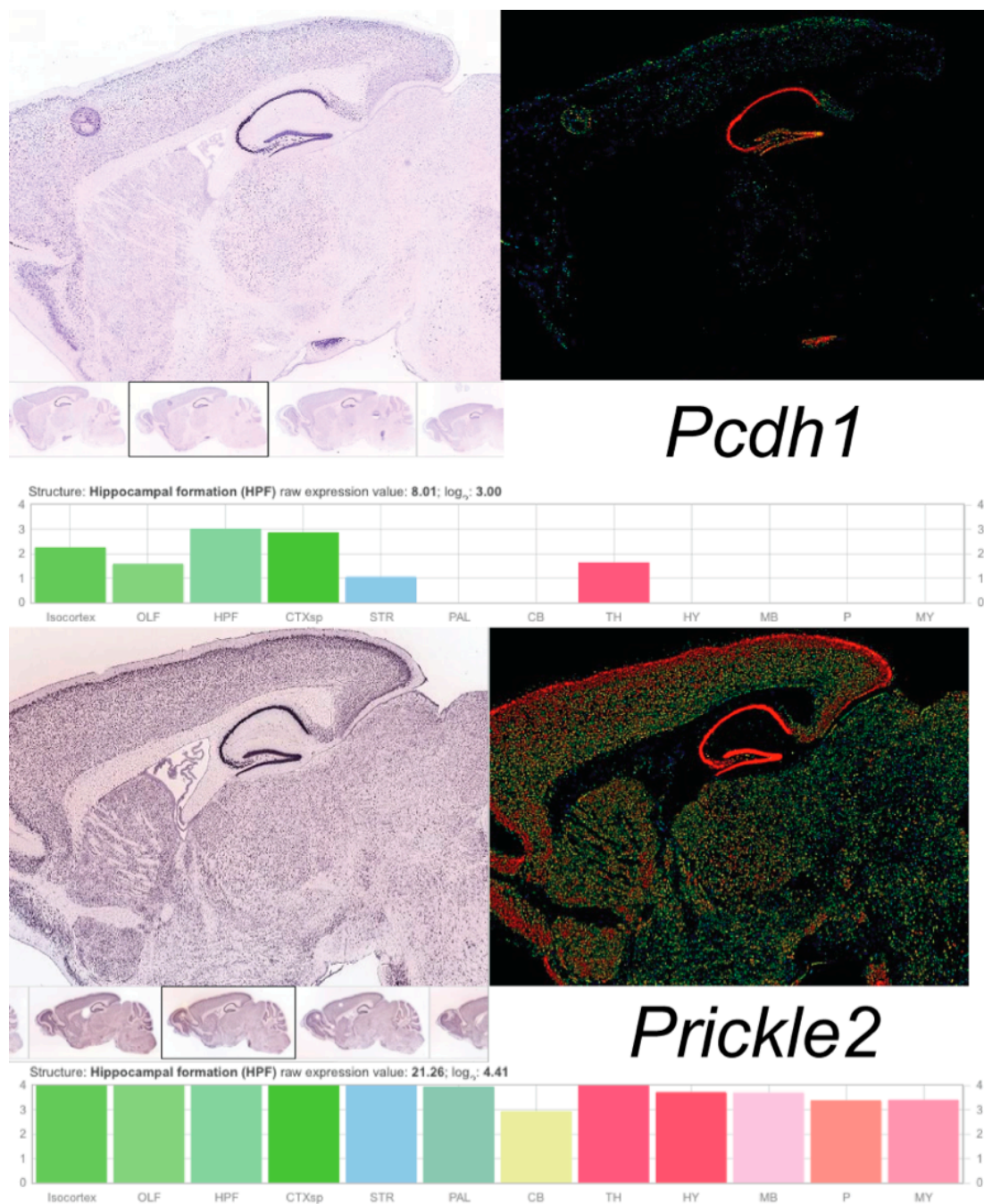


Figure C.31: *In situ* hybridisations, expression analyses of the hybridisations, and automated expression analyses across several major regions (HPF corresponds to the hippocampal formation) from the Allen Mouse Brain Atlas (Lein et al. 2007) showing the top five genes in the chick magenta/mouse black overlap.

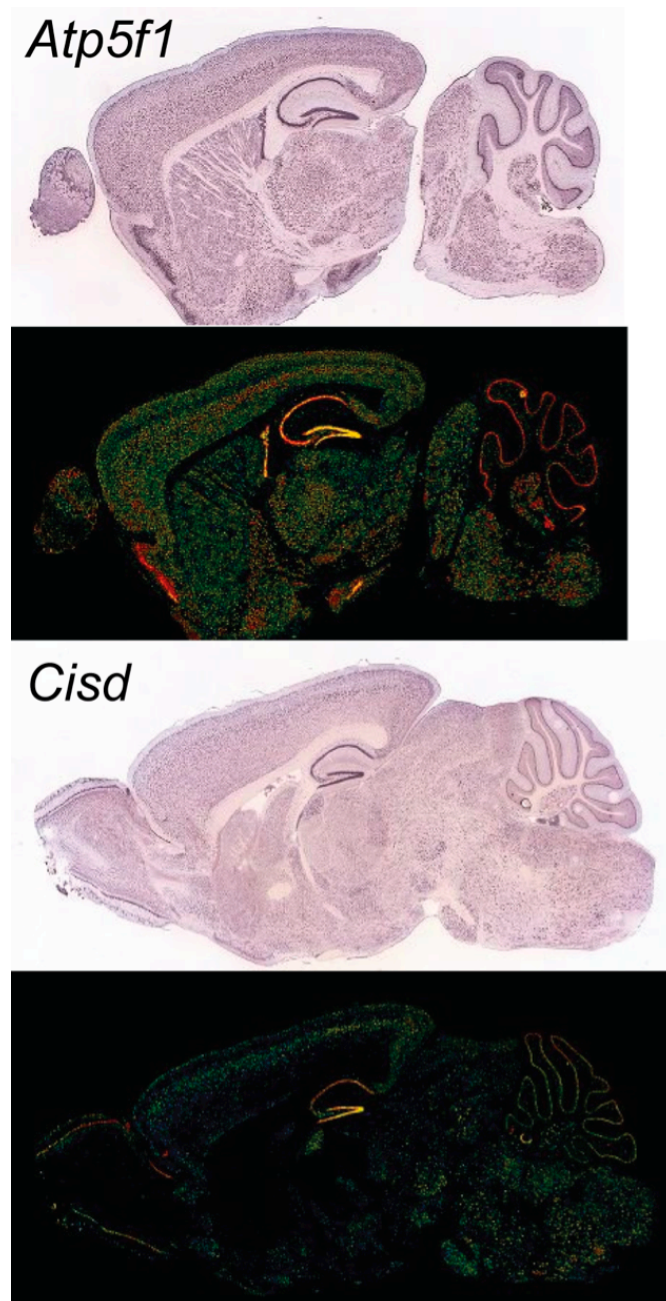


Figure C.32: Two genes correlated with both chick turquoise and mouse salmon – *Atp5f1* and *Cisd* – were differentially expressed across different brain regions in the Allen Mouse Brain Atlas (Lein et al. 2007).