Title
Genetic and Epigenetic Control of Gene Expression in Human and Non-Human Primates

Author
Zelaya, Ivette

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Genetic and Epigenetic Control of Gene Expression in Human and Non-Human Primates

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Ivette Maria Zelaya

2019

ABSTRACT OF THE DISSERTATION

Genetic and Epigenetic Control of Gene Expression

in Human and Non-Human Primates

by

Ivette Maria Zelaya

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2019

Professor Giovanni Coppola, Chair

A majority of the work presented in this dissertation focuses on identifying differences in transcriptome profiles across different phenotypes. The first project I present incorporates controls from different developmental time points, namely, prenatal and postnatal, to identify gene expression and splicing differences in SMA cases. Findings from this study report a large number of genes with prenatal expression patterns in iliopsoas from postnatal SMA samples. Similarly, differential splicing analyses uncovered prenatal splicing patterns in SMA cases in two muscle relevant genes: *TNNT3* and *MYBPC1*.

The next project characterizes the transcriptome profile of seven different tissues in the vervet monkey using RNA-seq data. Transcriptome profiles from two of the three brain tissues explored showed expression patterns correlated with developmental time point. Additionally, this project presents an eQTL study which

resulted in identifying eQTL SNPs within a region associated with hippocampal volume.

Building on the observation of developmental expression patterns in Brodmann's area 46 and caudate in the previous project, the next project I present focuses on the identification of age-related genes in vervet hippocampus. With the addition of younger samples, I also perform an eQTL analysis and report two additional genes, *CHMP1B* and *RAB31,* with associated SNPs within the hippocampal volume associated region.

Finally, the final project described focuses on improving the characterization of vervet chromatin modifications using human epigenomic datasets. Through the use of machine learning algorithms and prediction variables previously shown to correlate with conversion depth of histone marks across species, I show improved accuracy can be obtained while still maintaining biologically relevant peak signals.

The dissertation of Ivette Maria Zelaya is approved.

Rita Cantor

Eleazar Eskin

Nelson Freimer

Bogdan Pasaniuc

Giovanni Coppola, Committee Chair

University of California, Los Angeles

2019

Table of Contents

List of Figures and Tables

Acknowledgements

My graduate research would not have been possible without the help and support of my advisor, Dr. Giovanni Coppola. I am thankful for his patience and guidance, which allowed me to become a better researcher, and for the resources he provided, which enabled me to complete my graduate studies. I am also grateful for my committee members: Drs. Rita Cantor and Eleazar Eskin for the skills I learned in their courses which provided me a strong bioinformatics foundation; Dr. Bogdan Pasaniuc for his support during my rotation and his eagerness to help when I struggled with defining my project and Dr. Nelson Freimer for allowing me to work with the vervet datasets. I never imagined the huge role vervet datasets would play over the course of my graduate career.

I am thankful for the comradery I found in the Coppola lab, without which the long hours spent in the office would have been less enjoyable. Thank you to Alden Huang, Marisa Ramos, Doxa Chatzopoulou, Thomas Crisman and honorary "semelite" Deepika Dokuru for the many conversations over lunch, for the moral support and for providing me a mental reprieve from long days of coding.

Most importantly, I am eternally grateful for the wonderful support of my family. To my parents, my sister and my brother for taking an interest in my work and for encouraging me to pursue my graduate studies. I would also like to thank my husband Henry for all his patience, love and understanding throughout my graduate career.

The work presented in Chapter 2 was performed on samples obtained from Dr. Charlotte Sumner's group at John's Hopkins University. RNA library preparation and extraction were performed by Dr. Qing Wang. The remaining processing and analyses

presented in the chapter were performed by myself. Additionally, I am thankful to the people in the Sumner lab, specifically, Dr Sumner, Daniel Ramos and Christine Hatem, for their insight regarding the biological significance of my splicing results. This work is being prepared into a manuscript which will include additional muscle histological analyses and validation of observed splicing results performed by the Sumner lab.

The work described in Chapter 3 contains excerpts adapted from Jasinska, Anna J., et al. "Genetic variation and gene expression across multiple tissues and developmental stages in nonhuman primate." *Nature genetics* 49.12 (2017):1714. Nelson B. Freimer is the corresponding author of this study. I performed alignment and quantification of the RNA sequencing expression data from seven different tissues, as well as analyses of the RNA-seq data including expression quantitative trait analyses (eQTL), deconvolution analyses and comparisons with human and rhesus datasets that were described in this thesis. Susan Service performed hippocampal volume quantitative trait locus (QTL) analysis, expression quantitative trait analysis (eQTL) on blood microarrays, heritability analyses using SOLAR and principal component analyses. Anna Jasinska focused on analyzing the possible biological significance of our eQTL and QTL results.

The work presented in Chapter 4 is currently being prepared for submission, tentatively titled "Developmental patterns in the hippocampal transcriptome of the vervet monkey", with myself as the first author and Dr. Giovanni Coppola as the senior author. The analyses presented here makes use of 59 samples from a hippocampal dataset used in Chapter 3 with an additional 32 hippocampal samples. We thank Scott Fears and Anna Jasinska for the additional 32 samples from vervet

animals aged less than 1 year. Alignment and quantification of these samples was performed by myself, in addition to weighted gene coexpression (WGCNA) and expression quantitative trait analyses (WGCNA) on combined vervet samples. Fuying Gao assisted with the age-related differential expression analysis.

Chapter 5 makes use of vervet ChIP-seq data obtained from the manuscript Villar, et al. "Enhancer Evolution across 20 Mammalian Species". *Cell* (2015). The work presented in this chapter is unpublished.

Vita

Education

2009	Bachelor of Science in Mathematics/Applied Science University of California, Los Angeles, CA
2007	Associate of Arts in Biology Los Angeles Valley College, Los Angeles, CA

Employment

2014	Teaching Assistant Life Sciences Department, University of California, Los Angeles
2011-2012	Staff Research Associate I Physiology Department, University of California, Los Angeles
2005-2011	Operations Supervisor JP Morgan Chase Bank, North Hollywood, CA

Publications

Zelaya, I., Jasinska, A.J., Gao, F., Jorgensen, M.J., Fears, S.C., Woods, R., Freimer, N., Coppola, G. Developmental patterns in the hippocampal transcriptome of the vervet monkey. In preparation.

Auslander, N., Ramos, D.M., **Zelaya, I.**, Karathia, H., Crawford, T.O., Sumner, C.J., Ruppin, E. The GENDULF algorithm: prediction of modifier genes from gene expression data. Submitted.

Jasinska, A.J., **Zelaya, I.**, Service, S.K., Peterson, C.B., Cantor, R.M., Choi, O.W., DeYoung, J., Eskin, E., Fairbanks, L.A., Fears, S. and Furterer, A.E., 2017. Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. Nature genetics, 49(12), p.1714.

Huang, A.Y., Yu, D., Davis, L.K., Sul, J.H., Tsetsos, F., Ramensky, V., **Zelaya, I.**, Ramos, E.M., Osiecki, L., Chen, J.A. and McGrath, L.M., 2017. Rare copy number variants in NRXN1 and CNTN6 increase risk for Tourette syndrome. *Neuron*, *94*(6), pp.1101-1111.

Crisman, T.J., **Zelaya, I.**, Laks, D.R., Zhao, Y., Kawaguchi, R., Gao, F., Kornblum, H.I. and Coppola, G., 2016. Identification of an efficient gene expression panel for glioblastoma classification. *PloS one*, *11*(11), p.e0164649.

Peterson, C.B., Jasinska, A.J., Gao, F., **Zelaya, I.**, Teshiba, T.M., Bearden, C.E., Cantor, R.M., Reus, V.I., Macaya, G., López-Jaramillo, C. and Bogomolov, M., 2016. Characterization of Expression Quantitative Trait Loci in Pedigrees from Colombia and Costa Rica Ascertained for Bipolar Disorder. *PLoS Genet*, *12*(5), p.e1006046.

Zong, N., Ping, P., Lau, E., Choi, H.J., Ng, D., Meyer, D., Fang, C., Li, H., Wang, D., **Zelaya, I.M**. and Yates, J.R., 2014. Lysine ubiquitination and acetylation of human cardiac 20S proteasomes. *PROTEOMICS-Clinical Applications*, *8*(7-8), pp.590-594.

Zong, N.C., Li, H., Li, H., Lam, M.P., Jimenez, R.C., Kim, C.S., Deng, N., Kim, A.K., Choi, J.H., **Zelaya, I.** and Liem, D., 2013. Integration of cardiac proteome biology and medicine by a specialized knowledgebase. Circulation research, 113(9), pp.1043-1053.

Li, H., Zong, N.C., Liang, X., Kim, A.K., Choi, J.H., Deng, N., **Zelaya, I.**, Lam, M., Duan, H. and Ping, P., 2013. A novel spectral library workflow to enhance protein identifications. *Journal of proteomics*, *81*, pp.173-184.

Chapter 1

Introduction

Over the past decade, next generation sequencing technologies have revolutionized the field of genetic research (Metzker, 2010). Technological advances have afforded us improved transcriptome quantification methods without the need for dedicated microarray or other gene expression platforms (Ozsolak and Milos, 2011). This in turn has provided research avenues that would have been challenging to explore using microarray platforms, areas such as: exploration of alternative splice-sites, transcript-level quantification and detection of gene fusion events (Ozsolak and Milos, 2011). This is especially beneficial in studying brain disorders, where splicing has been found to play a role in disease pathogenesis in various neurological disorders (Dredge, Polydorides, and Darnell, 2001).

One such example occurs in spinal muscular atrophy (SMA), where splicing mutations in the *SMN1* gene splice out exon 7 which leads to a non-functional SMN protein (Lunn and Wang, 2008). SMA is a neurodegenerative disorder characterized by the degeneration of motor neurons in the spinal cord which leads to muscle atrophy. However, previous studies suggest a lack of maturation occurring during development in muscles from SMA type I cases (Martínez-Hernández et al., 2013). In an effort to identify developmental processes that may be affected in SMA cases, chapter 2 of this dissertation makes use of iliopsoas and diaphragm samples from 6-8 SMA cases, prenatal controls and postnatal controls to identify genes with prenatal expression and splicing patterns.

At the DNA level, genome-wide association studies have identified numerous genetic variants associated with disease (Welter et al., 2014), mostly located in non-coding regions of the genome, making it difficult to understand the functional impact of such variants. As a result, interest in understanding the functional role of noncoding variation has paved the way for studies exploring effects of genetic variants on gene expression (Nicolae et al., 2010) as well as identifying regulatory elements within the genome (Tak and Farnham, 2015).

Systematic studies aimed at the identification of expression quantitative loci (eQTL) in humans are limited by tissue sample availability across multiple developmental stages (Consortium et al., 2017; C.-H. Yu, Pal, and Moult, 2016). In addition, even an imperfect control of environmental conditions is impracticable in human studies. As such, model organisms have provided a feasible alternative. Mammalian model organisms, including invertebrates and rodents, are widely used in research studies and now multiple resources exist to facilitate gene expression studies (Blake et al., 2017; Shimoyama et al., 2015). However, the evolutionary distance of these models from humans limits their applicability, especially in the context of human disease. Nonhuman primate models, including the vervet monkey (*Chlorocebus aethiops sabaeus*) (Jasinska et al., 2013), constitute an attractive alternative for this type of studies, as considerable tissue resources have already been collected and the genome sequenced.

Chapter 3 describes the creation of an RNA-seq-based transcriptional resource across seven vervet tissues, including blood, fibroblasts, three brain tissues (caudate, Brodmann's area 46 [BA46], hippocampus), and two endocrine tissues (adrenal and pituitary gland). This resource provides an assessment of gene expression levels in

multiple vervet tissues across ten developmental time points, ranging from infant (7 days) to adult (9 years). Two of the three brain regions, caudate and BA46, are found to have developmental related expression patterns. Additionally, we perform characterization of eQTLs within these seven tissues and identify a hippocampal eQTLs located within a region associated with hippocampal volume. Chapter 4 builds on these results and expands our interrogation of developmental relevant genes in vervet hippocampus. In addition to uncovering genes associated with aging pathways, we also show many of these vervet genes are also developmentally regulated in the Allen Brain Atlas human and rhesus datasets.

Finally, significant progress has been made in understanding the contribution of noncoding variation to epigenetic marks regulating gene expression, both in rodents (Stamatoyannopoulos et al., 2012) and in humans (ENCODE Project Consortium, 2012). Through the use of these resources, it has been observed that regulatory regions, such as histone marks, vary across tissues and, importantly, that loci associated with specific diseases are enriched in tissue-specific histone marks in disease relevant tissues (Trynka et al., 2013). Chapter 5 describes a machine learning approach to classify vervet enhancer (H3K27ac) and promoter (H3K4me3) marks using human data from the Epigenomics Roadmap project (Kundaje et al., 2015). We show that factors such as distance to the transcription start site, GC content, and peak length can be used as predictive variables to classify true peak calls obtained by lifting over human epigenomic coordinates to the vervet genome.

Chapter 2

Prenatal expression patterns in Spinal Muscular Atrophy

2.1 Introduction

Spinal muscular atrophy (SMA) is an autosomal recessive neurodegenerative disease and the leading genetic cause of infant mortality (Lunn and Wang, 2008). Mutations in the survival motor neuron gene (*SMN1*) cause disease by reducing the amount of functional SMN protein. The *SMN1* paralog *SMN2* differs from *SMN1* by five nucleotides which consequently produces a non-functional, truncated SMN protein. Nonetheless, *SMN2* has been found to act as a disease modifier whose copy number is inversely related to disease severity. Lack of a functional SMN protein results in degeneration of α-motor neurons in the anterior horn of the spinal cord which leads to muscle atrophy (Hamilton and Gillingwater, 2013). Despite recent advances in the treatment of SMA (Finkel et al., 2017; Mendell et al., 2017), the molecular pathway by which muscle atrophy occurs is not fully understood and seems to be dependent on disease severity (Deguise et al., 2016). Muscle biopsies from SMA type I cases report a prenatal appearance (Fidziańska, Goebel, and Warlo, 1990), while additional studies suggest a lack of maturation in SMA muscle (Martínez-Hernández et al., 2009).

Prenatal patterns are not unique to SMA and have previously been reported in other muscular degenerative diseases such as Duchenne muscular dystrophy (DMD) (Fitzsimons and Hoh, 1981). By uncovering possible genes and pathways altered in SMA muscles in comparison to normal muscle development, a better understanding may be gained into mechanisms involved in SMA muscle pathology. Thus, we sought

Sample Name	Sample Type	Age	SMN1 Copies	SMN2 Copies	Cause of Death	Iliopsoas	Diaphragm
SMA_08_02	SMA	16m	0	2	Type 1 SMA	X	X
SMA_12_01		2.5m	0	2	Type 1 SMA	X	X
SMA_08_01		4.5m	0	2	Type 1 SMA	X	X
SMA_09_02		4m	0	2	Type 1 SMA	X	X
SMA_10_16		72m	0	2	Type 2 SMA		X
SMA_14_04		72m	0	2	Type 1 SMA	X	X
SMA_17_03		0.5m	0	2	Type 1 SMA	X	X
SMA_17_06		18w (prenatal)	0	2	Type 1 SMA	X	X
MBB_113	Prenatal Control	18w	2	1	Control	X	X
MBB_314		21w	3	1	Trisomy 1	X	X
MBB_684		22w	2	2	Tuberous Sclerosis	X	X
MBB_361		28w	2	1	Hydrops fetalis	X	X
CNTL_15_02		25w				X	
CNTL_15_03		22w				X	X
CNTL_15_04		18w					X
CNTL_15_07		34w			Congenital Heart Defect	X	X
CNTL_12_02	Postnatal Control	0.3m	2	2	Meconium Aspiration	X	X
UMB_86		1.9m	2	2	Congenital Heart Defect		X
UMB_1296		3.26m	2	2	Control	X	
UMB_1472		3.93m	2	2	Control		X
UMB_195		4.1m	2	2	Control	X	
CNTL_12_05		19m	2	1		X	X
CNTL_13_01		168m	2	1	Cardiac Arrest	X	X
CNTL_15_05		23w, 3 mon post-delivery				X	X
CNTL_15_06		144m				X	X
CNTL_17_01		9m			Trisomy 21	X	X
MBB_106		0.73m	2	1	Control		X
MBB_569		4.4m	2	1		X	

Table 2-1: Summary of samples used in differential expression and splicing analyses. Red "X" indicates outlier samples that were excluded from analyses.

to identify gene expression changes and splicing differences in diaphragm and iliopsoas muscle from SMA affected infants, compared to postnatal and prenatal controls.

2.2 Results

Comparison	Tissue	Up-regulated genes	Down-regulated genes	Total # of DE genes
SMA vs Postnatal Controls	Iliopsoas	540	496	1,036
Prenatal Controls vs SMA		2,314	2,073	4,387
Prenatal vs Postnatal Controls		3,775	3,447	7,222
SMA vs Postnatal Controls	Diaphragm	29	42	71
Prenatal Controls vs SMA		3,172	2,698	5,870
Prenatal vs Postnatal Controls		2,041	1,846	3,887

Table 2-2: Summary of differential expression results for each comparison group by tissue type.

We obtained RNA sequencing data from postnatal SMA cases, and prenatal and postnatal controls from diaphragm and iliopsoas tissues (Table 2-1). For our differential expression and splicing analyses we focused on three comparison groups for each tissue type: prenatal vs postnatal controls, SMA cases vs postnatal controls and prenatal controls vs SMA cases (Figure 2-1). To identify genes in SMA cases with prenatal patterns we focused on genes shared between prenatal controls and SMA cases vs postnatal comparisons, while excluding genes with similar patterns in prenatal controls vs SMA cases. By excluding genes shared with this third comparison group (prenatal controls vs SMA cases) we excluded differentially expressed SMA genes with expression values somewhere between those observed in prenatal and postnatal controls. Inferring biological significance of such

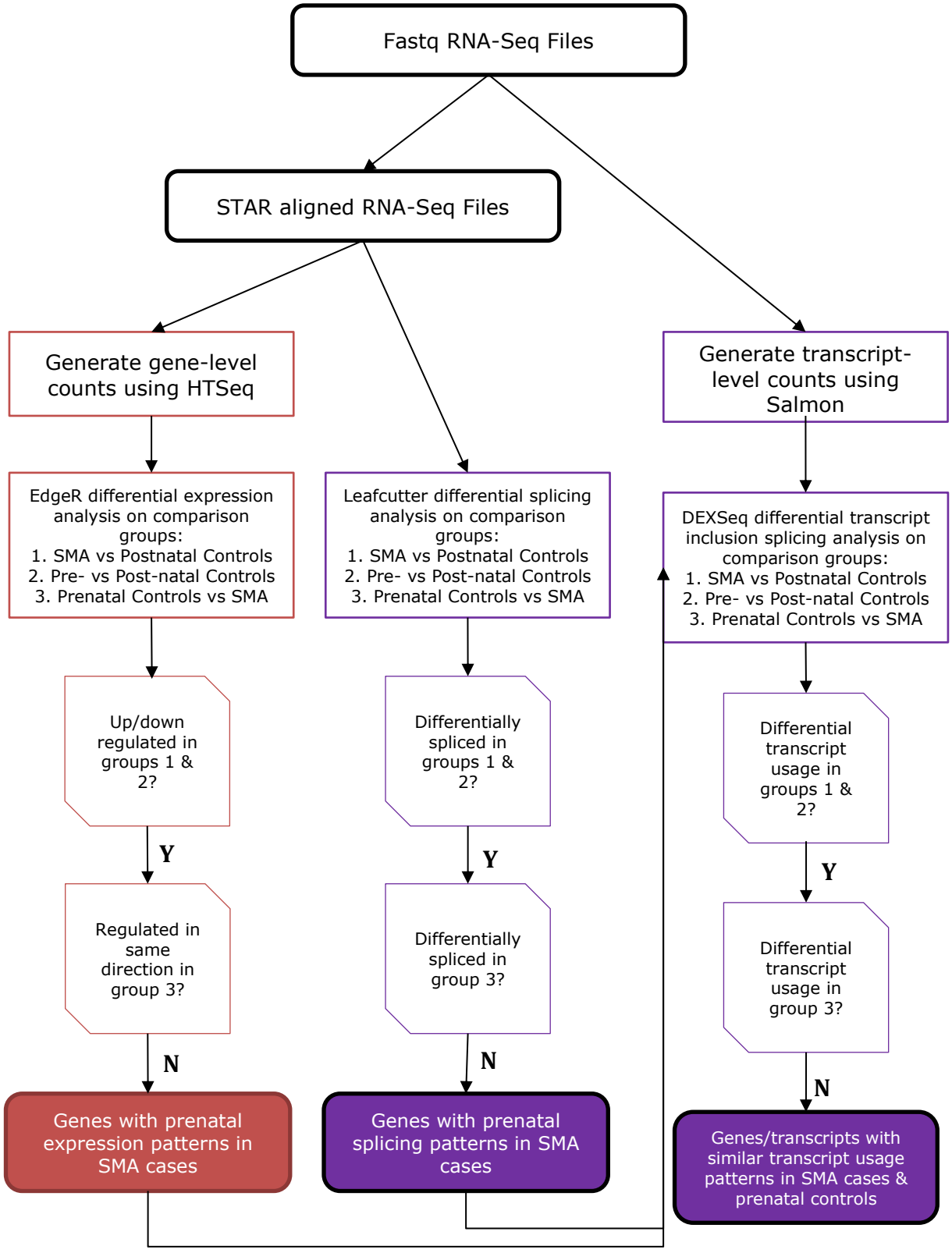


Figure 2-1: Schematic of differential expression and splicing analyses performed.

expression patterns is a more challenging task, thus we sought to focus our investigation on differentially expressed genes where SMA expression mirrored prenatal expression.

Differential expression analysis highlights neurodegeneration-associated genes

Differential expression analysis performed on diaphragm and iliopsoas muscle tissues resulted in the analysis of approximately 17,000 genes, after filtering. At an FDR threshold of 0.05, we observed the greatest number of differentially expressed genes in prenatal vs postnatal tissue comparisons in iliopsoas and in prenatal vs SMA comparisons in diaphragm (Table 2-2). First, we focused on the 500 genes up- or down-regulated in iliopsoas SMA and prenatal vs postnatal comparisons but not in prenatal vs SMA (Figure 2-2). We performed functional annotation using DAVID (D.

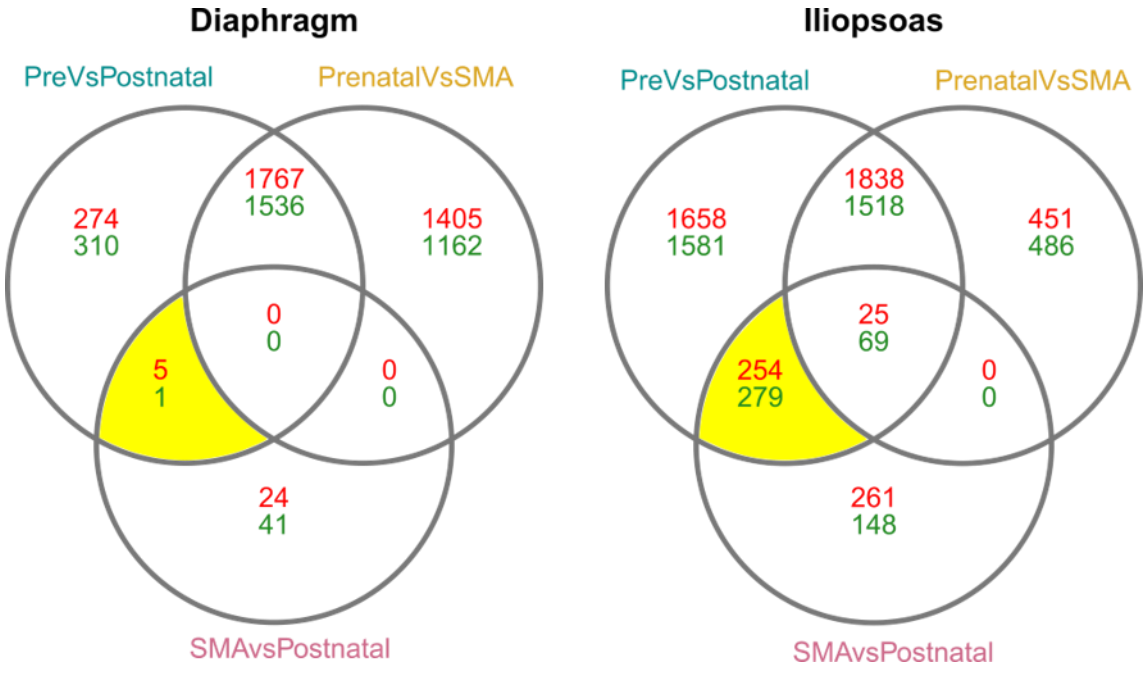


Figure 2-2. Venn diagrams of differential expression results. Venn diagrams show overlap of DE genes in each comparison group with the highlighted region indicating genes with potential prenatal expression patterns in SMA cases. Red and green numbers indicate up- and down-regulated genes, respectively.

W. Huang, Sherman, and Lempicki, 2009a, 2009b) on these genes and found an enrichment of SMA and prenatal down-regulated genes involved in Parkinson's, Huntington's and Alzheimer's disease as well as metabolic and mitochondrial translational pathways (FDR < 0.05; Figure A-1). Then, we explored if any of these 500 genes were DE in the opposite direction in prenatal vs SMA which would suggest

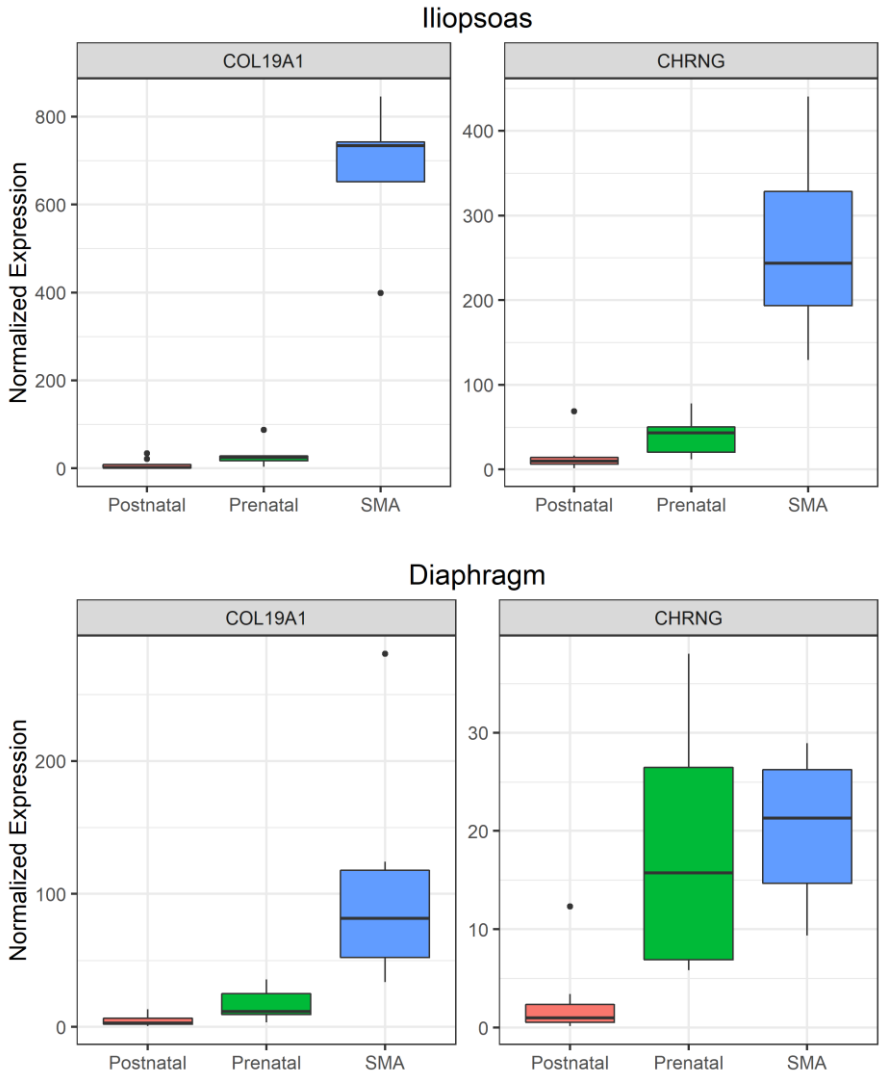


Figure 2-3. Normalized expression across sample types for *CHRNG* and *COL19A1*. Expression plots showing a greater divergence from postnatal expression in iliopsoas SMA cases vs prenatal controls. While similar expression trends are observed in iliopsoas and diaphragm for both genes, expression of these genes in diaphragm SMA cases more closely resemble diaphragm prenatal expression patterns.

a more significant departure from postnatal expression patterns in SMA cases than that observed in prenatal controls. We found 17 genes with such a pattern, including *COL19A1* and *CHRNG* (Figure 2-3)*,* both of which have been previously implicated in muscle-related disorders (Ana et al., 2018; Morgan et al., 2006)*. COL19A1* and *CHRNG* are upregulated in SMA iliopsoas, but not in pre- or post-natal control muscle. Next, we explored the six DE genes overlapping in SMA vs postnatal and pre- vs post-natal control comparisons in diaphragm. Of these six genes, one gene, *COL19A1*, recapitulated the increased expression signature observed in iliopsoas SMA cases in diaphragm SMA cases (Figure 2-3). Although *COL19A1* expression is considerably less in diaphragm SMA cases than iliopsoas SMA cases, the pattern of expression was similar in diaphragm and iliopsoas. Similarly, *CHRNG* expression also showed an increase in diaphragm SMA cases and prenatal controls, however, unlike our iliopsoas findings, *CHRNG* expression in diaphragm SMA cases better reflected *CHRNG* expression in prenatal controls (Figure 2-3).

Differential splicing analyses uncover prenatal splicing patterns in SMA cases

Comparison	Tissue	# of Splicing Events	# of Spliced Genes
SMA vs SMA	Diaphragm vs Iliopsoas	175	169
Prenatal vs Prenatal		12	11
Postnatal vs Postnatal		130	126
SMA vs Postnatal Controls	Iliopsoas	106	103
Prenatal Controls vs SMA		620	562
Prenatal vs Postnatal Controls		1,495	1,252
Prenatal Controls vs SMA	Diaphragm	1,121	1,016
SMA vs Postnatal Controls		24	24
Prenatal vs Postnatal Controls		465	427

Table 2-3: Summary of differential splicing results.

A

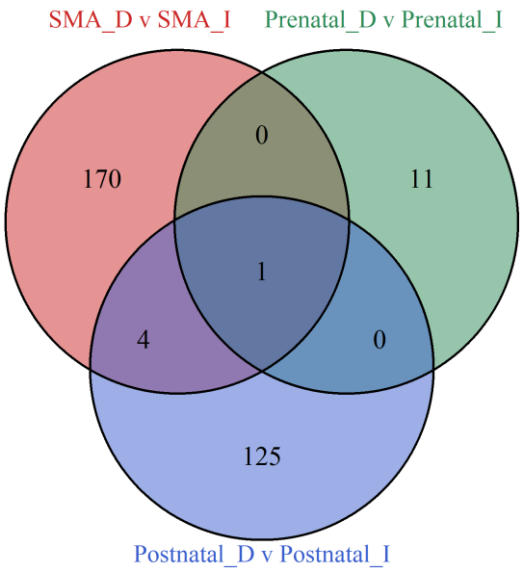
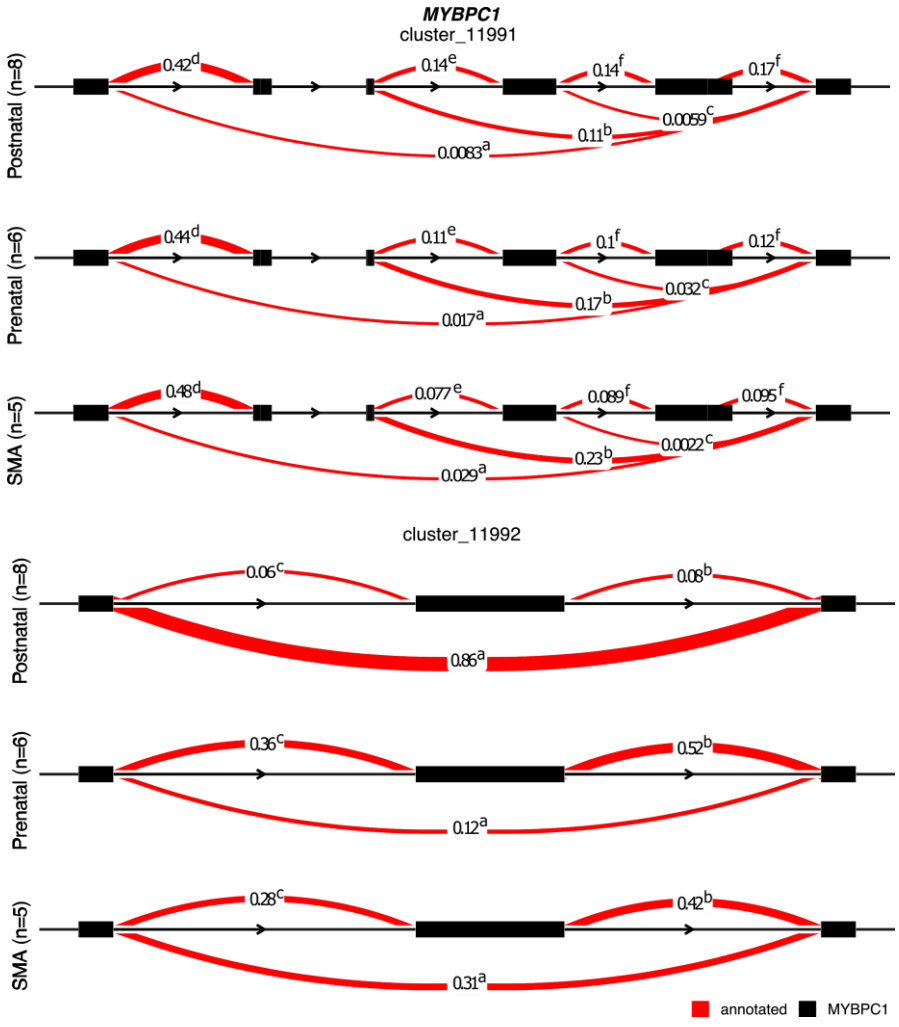


Figure 2-4. Overview of differential splicing results. (A) Overlap of iliopsoas and diaphragm splicing results by comparison group. (B) Splicing patterns in differentially spliced gene *MYBPC1.* Cluster 11991 represents exons 3-8 of the gene while cluster 11992 illustrates exons 26-28. Splicing clusters show increased exclusion of exons 6-7 (superscript b) and inclusion of exon 27 (superscript c) in SMA cases and prenatal controls.

B



Given the role splicing plays in various neurodevelopmental diseases (Mills and Janitz, 2012), we explored splicing differences between SMA cases and controls in iliopsoas and diaphragm. We first compared diaphragm and iliopsoas splicing differences within the same sample groups and then focused on comparisons between our three samples conditions within each tissue type.

Differential splicing analysis between conditions (i.e. diaphragm vs iliopsoas SMA cases) yielded the lowest number of splicing differences between prenatal diaphragm and iliopsoas samples and the highest number between SMA cases (Table 2-3). Functional annotation of these gene sets failed to uncover enriched pathways or gene ontology terms after correcting for multiple hypothesis testing (FDR < 0.05). Comparison of differentially spliced clusters across sample types resulted in very little overlap (Figure 2-4A), suggesting splicing differences are specific to sample type rather than tissue. Similar to what we observed in our differential expression analysis, differential splicing results across sample types within the same tissue yielded the greatest number of differential splicing events in diaphragm prenatal controls vs SMA cases and iliopsoas prenatal vs postnatal controls (Table 2-3). In addition, we observed a strong correlation between the number of differentially spliced events and the number of differentially expressed genes in each dataset (R=0.98, p=4e-4; Figure A-1B).

Next, we explored SMA vs postnatal differential splicing events that were present in prenatal and SMA vs postnatal comparisons but not in SMA vs prenatal dataset. We found 38 genes in iliopsoas (37% of the DS genes) that reported delta percent-spliced-in (dPSI) values in SMA cases similar to prenatal dPSI values. These results suggest that although all our SMA cases are postnatal, splicing patterns in

these genes better reflect what we observe in our prenatal controls. One notable example, *MYBPC1* (Figure 2-4B), encodes the slow skeletal isoform of the myosin-binding protein C (Geist and Kontrogianni-Konstantopoulos, 2016). Splicing results in *MYBPC1* suggest a higher proportion of exons 6-7 being spliced out in SMA cases and prenatal controls (cluster 11991), in addition to a higher spliced-in proportion of exon 27 (cluster_11992).

We sought to follow up on these 38 genes with prenatal splicing patterns in SMA cases by performing additional splicing analyses focusing on differential

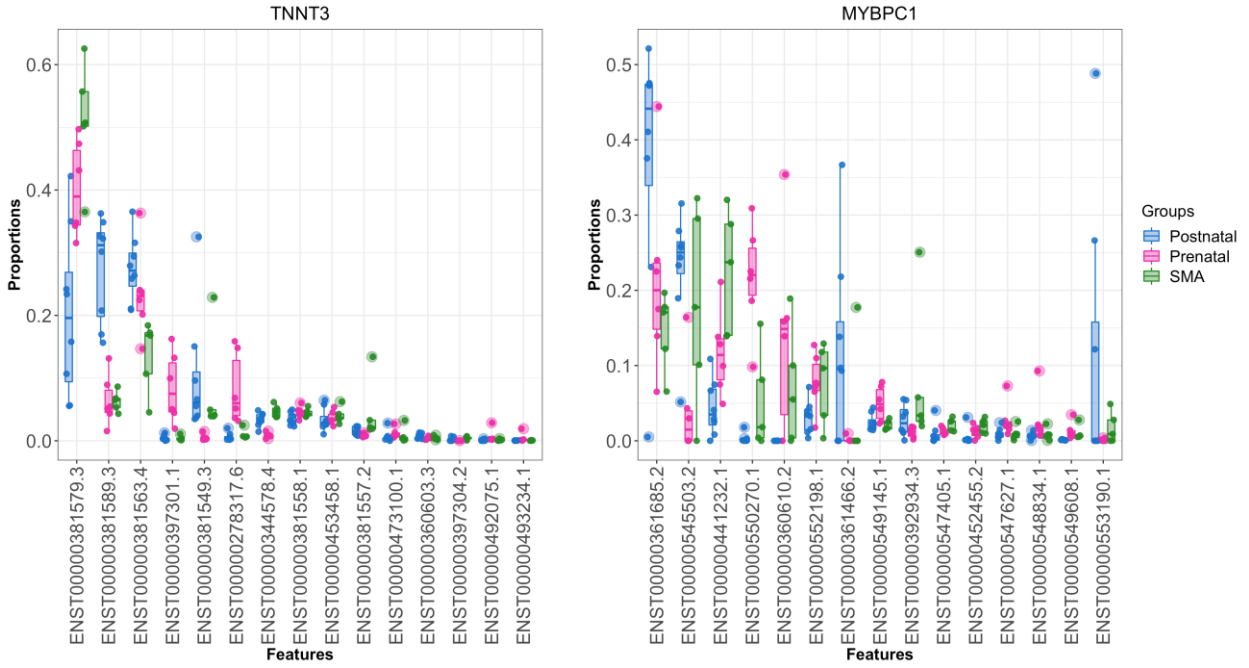


Figure 2-5. Differential isoform usage results for two validated differentially spliced genes: *TNNT3* and *MYBPC1*. SMA cases and prenatal controls both present decreased usage of ENST00000381589.3 *TNNT3* transcript and increased usage of ENST00000381579.3 isoform. Similarly, postnatal controls show a preference for *MYBPC1* transcript ENST00000361685.2, while SMA cases and prenatal controls show a preference for ENST00000441232.1.

transcript usage. Through this analysis, we replicated our differential splicing results in two of our 38 genes, *TNNT3* and *MYBPC1* (Figure 2-5). Lack of replication of the other 36 genes may be due to SMA- or prenatal-specific splicing events which result

in different transcripts being generated. Observed differential isoform usage in *MYBPC1* coincide with our original results which highlight a higher inclusion and exclusion percentage of exon 27 and exons 6-7, respectively, in SMA cases and prenatal controls when compared to postnatal controls (Figure 2-4B). These splicing events contribute to increased and decreased proportions of ENST00000441232.1 and ENST00000361685.2 transcripts, respectively, in SMA cases and prenatal controls (Figure 2-5).

2.3 Discussion

Our study interrogated the transcriptome in two relevant tissues from human postnatal SMA cases, and compared it to pre- and postnatal controls. Our analysis revealed a subset of gene expression changes in postnatal SMA cases which resembled the pattern of expression observed in prenatal controls.

Our differential expression analysis uncovered sets of interesting genes associated with various neurodegenerative disorders. Interestingly, one of the top genes, *CHRNG*, encodes the fetal acetylcholine receptor subunit gamma (Gu and Hall, 1988). During development, as muscle maturation occurs, the fetal gamma subunit of the acetylcholine receptor is switched to the adult epsilon subunit (Hesselmans, Jennekens, Van Den Oord, Veldman, and Vincent, 1993). The increased *CHRNG* expression observed in our iliopsoas SMA cases further supports previous findings in human and mouse studies reporting the presence of the fetal gamma subunit in postnatal SMA muscle (Kariya et al., 2008; Martínez-Hernández et al., 2013). Similarly, studies exploring *COL19A1* expression reported increased expression in fetal muscle compared to adult muscle, and decreased expression in fetal brain

compared to adult brain (Sumiyoshi, Inoguchi, Khaleduzzaman, Ninomiya, and Yoshioka, 1997). More importantly, similar to our observations in SMA cases, recent studies also found increased expression of *COL19A1* in amyotrophic lateral sclerosis (ALS) cases, even suggesting the use of *COL19A1* as a prognostic biomarker for the disease (Ana et al., 2018). Taken together, these results confirm a gene expression signature switched toward a prenatal state in postnatal in SMA muscle.

Our differential splicing analysis provided novel insights into splicing patterns in SMA cases when compared to pre- and postnatal controls. Our observation of splicing patterns in iliopsoas postnatal SMA cases that better reflect those of prenatal controls in genes such as *TNNT3 and MYBPC1* suggests either a lack of developmental progression, or reversal to a fetal state in muscle gene expression. *TNNT3* knockout studies in mice suggest troponin T3 is essential for growth and postnatal survival (Ju et al., 2013). More importantly, alternative splicing of *TNNT3* is developmentally regulated with several isoforms exclusively expressed in fetal or adult muscle tissues (Wei and Jin, 2016). The higher abundance of the ENST00000381579.3 transcript in SMA cases and prenatal controls contrasts isoform expression reported by the GTEx portal (www.gtexportal.org, data source v7) in adult skeletal muscle tissues where this transcript is only the third most abundant transcript. Similarly, splicing results in *MYBPC1* also suggest preferential expression of specific isoforms in prenatal controls and SMA cases. *MYBPC1* belongs to the myosin-binding protein family which plays a crucial role in muscle contraction (Lin et al., 2018). Myosin genes are known to have isoforms expressed exclusively during development (Schiaffino, Rossi, Smerdu, Leinwand, and Reggiani, 2015), however, no known fetal isoforms have been reported for *MYBPC1*. Despite the lack of known fetal isoforms, the higher abundance

of the ENST00000441232.1 transcript in SMA cases and prenatal controls, combined with low expression of this transcript in our postnatal controls as well as in GTEx adult muscle tissues (TPM = 4.04, www.gtexportal.org, data source v7), suggests this transcript may be developmentally regulated. While these observed transcripts identified in *TNNT3* and *MYBPC1* may not be exclusive to fetal muscle tissues, their increased abundance in prenatal controls suggest a function for them and their encoded proteins which is diminished in our postnatal controls.

Finally, our expression and splicing results highlighted genes known to be involved in muscle development and suggested a possible role for these genes in SMA muscle pathology. Functional studies are now needed to better understand how prenatal expression patterns of these genes affect normal development and if they contribute to muscle atrophy. The inclusion of prenatal control samples in our study, coupled with our differential expression and splicing results can provide a valuable resource in understanding developmental pathways that may be affected in spinal muscular atrophy and other neurodegenerative diseases.

2.4 Methods

Sample Collection

RNA sequencing was performed on diaphragm and iliopsoas tissues from SMA cases, prenatal & postnatal controls. Samples were run in two batches using different methods due to availability of technology at the time of extraction and sequencing. For batch one, human iliopsoas and diaphragm muscle tissues were disrupted and homogenized in Lysing Matrix A (MP Biomedicals, LLC, Santa Ana, CA), plus RLT plus buffer (QIAGEN, Valencia, CA) by FastPrep 5G (MP Biomedicals, LLC, Santa Ana, CA).

About 30 mg of tissues were used. The lysate was span down and the supernatant was saved. The total RNA was then extracted using RNeasy ® Plus Mini Kit (QIAGEN, Valencia, CA) according to manufacturer's instructions. The RNA integrity (RIN) was examined using Bioanalyzer 2100 (Agilent, Santa Clara, CA). The RNA with RIN 4.6 or above was used in the following RNA-seq prep.

The RNA-seq libraries were prepared using TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina, La Jolla, CA) following the manufacturer's instruction. The libraries were pooled for sequencing of pair-end 50-bp on HiSeq™2500 (Illumina, La Jolla, CA).

For samples in our second batch, about 30 mg of human iliopsoas and diaphragm muscle tissues were used in a 2-ml tube with a 5-mm stainless steel bead and RLT buffer (QIAGEN, Valencia, CA), then disrupted and homogenized in TissueLyser LT (QIAGEN, Valencia, CA), operated at 40 Hz for 2 min. The lysate was used for the further total RNA extraction using RNeasy Fibrous Tissue Mini Kit (QIAGEN, Valencia, CA) following manufacturer's instructions. The RNA integrity (RIN) was examined using Bioanalyzer 2100 (Agilent, Santa Clara, CA). The RNA with RIN 4.2 or above was used in the following RNA-seq prep.

The RNA-seq libraries were prepared using TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina, La Jolla, CA) following the manufacturer's instruction. The libraries were pooled for sequencing of pair-end 50-bp on HiSeq™4000 (Illumina, La Jolla, CA).

Sample Processing and Outlier Detection

RNA-seq fastq files were aligned to human reference hg19 using STAR aligner v2.5.0a (Dobin et al., 2013) and expression counts were obtained using HTSeq

(Anders, Pyl, and Huber, 2015). Outliers were determined through hierarchical clustering and removed from subsequent analyses (CNTL_15_03 in diaphragm and SMA14_04 and CNTL_15_07 in iliopsoas; Figure A-2A,B). Additional outlier detection was performed using principal component analysis (PCA), where two additional diaphragm outliers were observed: SMA_17_03 and CNTL_17_01 (Figure A-2C,D). Similarly, in iliopsoas and diaphragm, we observe our single prenatal SMA sample (SMA_17_06) and our only premature postnatal control sample (CNTL_15_05) cluster with prenatal controls. Due to the nature of these samples and the way they clustered in both of our hierarchical clustering and PCA approaches, they were excluded so as to eliminate any possible influence on our prenatal comparisons.

Differential Expression Analyses

Lowly expressed genes with a counts-per-million (cpm) values less than 0.5 in more than 25% of the samples were removed. Differential expression analyses were performed using the edgeR generalized linear model (GLM) approach (McCarthy, Chen, and Smyth, 2012) on trimmed mean normalized counts (Robinson and Oshlack, 2010). To account for possible batch effects due to differences in extraction methods and sequencing technology, batch was included as a factor in our model. Differential expression results were considered significant at an FDR threshold of 0.05 for each comparison group.

Differential Splicing Analysis

Alternative splicing analysis was performed using the leafcutter package v1.0 (Y. I. Li, Knowles, and Pritchard, 2016), an annotation-free splicing analysis software. Default parameters were used in the leafcutter script, except for minimum samples per intron which was reduced to three to account for the smaller number of iliopsoas

SMA cases. Batch was included as a covariate. FDR values were calculated using the p.adjust function in R and results were considered significant at an FDR threshold of 0.05.

Additional splicing analyses were limited to genes identified by differential gene expression and leafcutter differential splicing analyses as having prenatal patterns in SMA samples. Differential transcript usage isoform expression was performed using the DEXSeq v1.28 R package (Anders, Reyes, and Huber, 2012). Transcript counts were obtained using Salmon (Patro, Duggal, Love, Irizarry, and Kingsford, 2017) due to its fast computing time and strong correlation with alternative isoform quantification methods (C. Zhang, Zhang, Lin, and Zhao, 2017). Counts were imported and scaled using tximport (Soneson, Love, and Robinson, 2015) to account for differences in library size and transcript length. Filtering was performed to remove lowly expressed transcripts, defined as transcripts expressed in fewer than 4 samples with a minimum count value of 5. To ensure consistency between all our analyses, batch was also included in our linear model. Finally, two-stage FDR control was performed using the stageR package (Van den Berge, Soneson, Robinson, and Clement, 2017) to control the false-discovery rate at the gene and feature level. Leafcutter splicing events were considered validated if differential splicing was observed at an overall (two-stage) FDR threshold of 0.05.

Chapter 3

Expression quantitative trait loci in vervet tissues

3.1 Introduction

Efforts to understand how genetic variation contributes to common diseases and quantitative traits increasingly focus on the regulation of gene expression. Most loci identified through genome-wide association studies (GWAS) lie in noncoding genome regions (Hindorff et al., 2009) and are enriched for eQTLs, SNPs regulating transcript levels, primarily of nearby genes (Nicolae et al., 2010). This observation suggests that eQTL catalogs may signpost variants responsible for GWAS signals (Albert and Kruglyak, 2015). Normal functioning of complex organisms depends on tightly regulated gene expression at specific developmental stages in specific cell types. Existing human eQTL data sets are likely missing information relevant to understanding disease, as most known human eQTLs have been identified in adult individuals, largely from lymphocytes or lymphoblastoid cell lines (Gibson, Powell, and Marigorta, 2015; Gilad, Rifkin, and Pritchard, 2008). This lack is particularly striking for neuropsychiatric disorders, given the inaccessibility of brain tissues in living individuals and the enormous modifications occurring in the brain across development (H. J. Kang et al., 2011).

Databases of gene expression obtained in samples from post-mortem donors have begun to remedy the lack of human data connecting genotypic variation and multitissue transcriptome variation. The Genotype-Tissue Expression (GTEx) project eQTL catalog is the most extensive of such resources available (Mele et al., 2015). However, limitations of the GTEx project inherent to human research, namely the

lack of developmental data, the relatively low number and variable quality of samples, and their genetic heterogeneity, motivate the generation and investigation of equivalent resources from model organisms. The advantages of model systems include (i) the feasibility of controlling for interindividual heterogeneity in environmental exposures and minimizing the interval between death and tissue preservation; (ii) the practicability of obtaining sizable numbers of samples from multiple tissues across development; and (iii) the opportunity to systematically phenotype individuals carrying particular eQTL variants. The similarities between humans and nonhuman primates (NHPs) in behavior, neuroanatomy and brain circuitry (Jasinska et al., 2013; Jennings et al., 2016; Rogers and Gibbs, 2014) make NHP eQTLs particularly valuable for illuminating neuropsychiatric disorders.

We report here, in Caribbean vervets (Chlorocebus aethiops sabaeus) from the Vervet Research Colony (VRC) extended pedigree, the first NHP resource combining genotypes from whole-genome sequencing (WGS) (Y. S. Huang et al., 2015), multitissue expression data across postnatal development, controlled environmental exposures (see Methods), and quantitative phenotypes relevant to human brain and behavior. Caribbean vervets are Old World monkeys whose population expanded dramatically from a founding bottleneck occurring when West African vervets were introduced to the Caribbean in the seventeenth century (Jasinska et al., 2013); genetic variation has drastically declined in Caribbean vervet populations since then, resulting in enrichment for numerous deleterious, or otherwise rare alleles.

Through necropsies performed under uniform conditions, we obtained brain and peripheral tissue samples from captive VRC vervets. Using these resources, we have delineated cross-tissue RNA-seq-based expression profiles for seven of these

tissues across multiple developmental stages from birth to adulthood. We identified numerous local and distant eQTLs in each tissue and validated a locus associated with multiple distant eQTLs, observed previously using pedigree-wide analyses (Jasinska et al., 2009). Additionally, we demonstrated the relevance of vervet eQTLs to an example of higher-order traits: hippocampus-specific local eQTLs regulate a set of lncRNAs associated with hippocampal volume, a phenotype related to neuropsychiatric disorders (Stein et al., 2012).

3.2 Results

Tissue	Protein-Coding	Non-Coding	Pseudogene	Other/Unknown	Total Genes
		Dataset 1			
Blood	5436	59	89	2	5586
		Dataset 2			
Adrenal	18221	3898	3036	32	25187
BA46	18451	5656	3393	30	27530
Blood	20529	8112	5093	42	33776
Caudate	18695	5961	3559	34	28249
Fibroblast	16614	2913	2787	14	22328
Hippocampus	18290	5411	3223	33	26957
Pituitary	18879	4976	3344	37	27236

Table 3-1. Biotypes of genes analyzed in Datasets 1 and 2

We investigated two data sets. Data set 1, described previously (Jasinska et al., 2009), consists of gene expression levels obtained by hybridizing all available VRC whole blood–derived RNA samples (n = 347) to Illumina HumanRef-8 v2 microarrays, which we used because no vervet arrays are available. After filtering out probe sequences not represented in the vervet genome (Warren et al., 2015) or containing common vervet SNPs (Y. S. Huang et al., 2015), we estimated expression levels at 6,018 probes, corresponding to 5,586 unique genes (Table 3-1). Data set 2 consists of RNA-seq reads from seven tissues collected under identical conditions

from each of 58 VRC monkeys (representing ten developmental stages, from birth through adulthood; Methods). Five of these tissues have prominent roles in

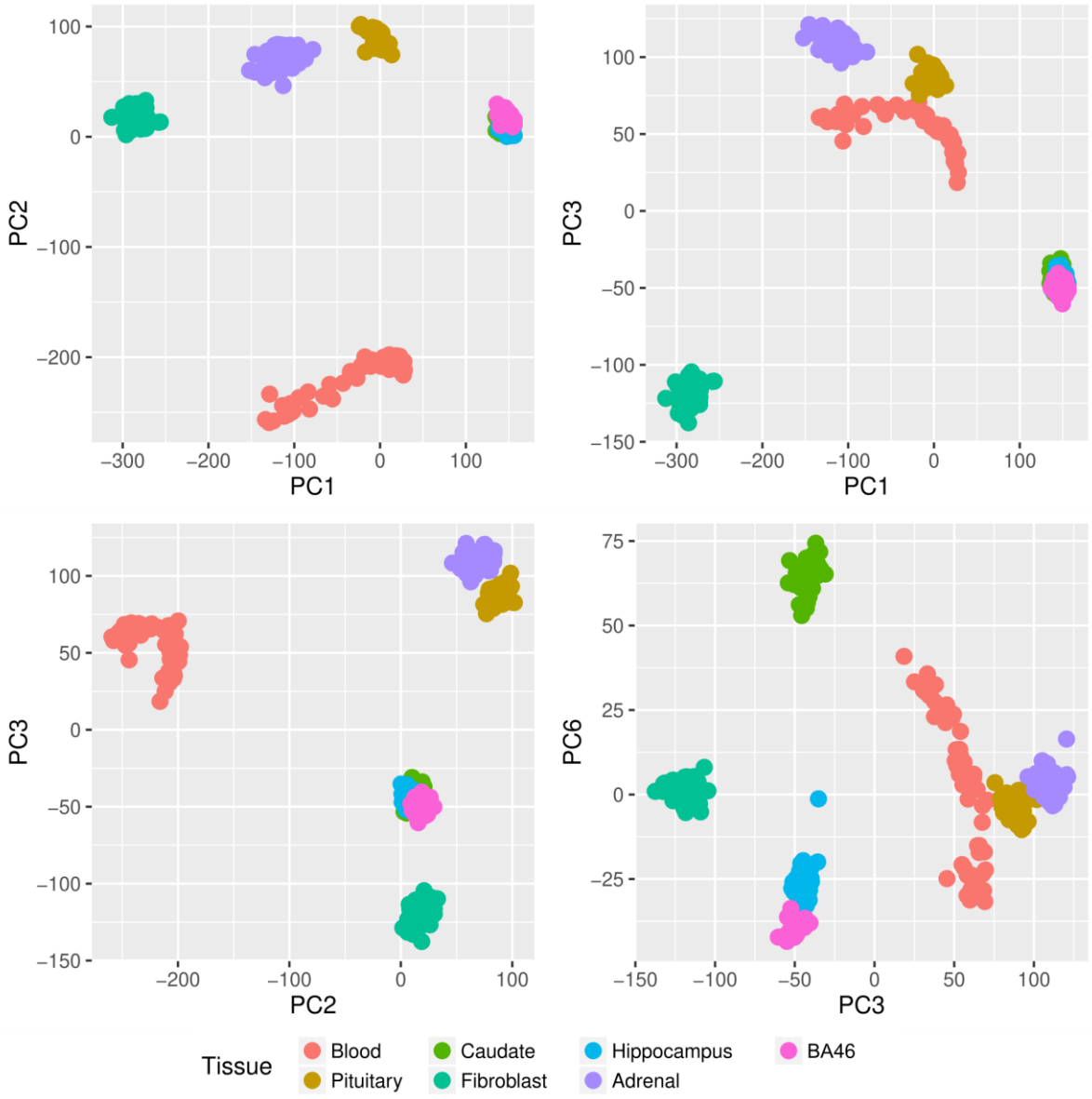


Figure 3-1. Principal components 1, 2, 3 and 6 from analysis of gene expression levels (RNA-seq) in seven tissues. PC1 (47.5% of total variance) separates fibroblast from brain tissues and PC2 (18.2% of variance) separates blood from all other tissues, while the three brain regions do not separate until PC6 (2% of variance).

cognitive and behavioral phenotypes (Arnett, Muglia, Laryea, and Muglia, 2016; McEwen, Gray, and Nasca, 2015; Nestler, E., Hyman, S., Holtzman, D. & Malenka,

2015): Brodmann area 46 (BA46), a cytoarchitectonically defined region encompassing most of the dorsolateral prefrontal cortex (DLPFC); hippocampus; caudate nucleus, a component of dorsal striatum; pituitary gland; and adrenal gland. The other two tissues (cultured skin fibroblasts and whole blood) are relatively accessible and are thus widely used in studies aimed at identifying biomarkers. We assessed expression of 33,994 annotated genes but minimized spurious signals by excluding genes expressed in <10% of individuals or at lower than one read per tissue (Table 3-1).

Principal-components analysis (PCA) of data set 2 showed that, overall, expression levels clustered more by tissue than by individual (Figure 3-1).

Sources of variation in multitissue expression data

The availability in data set 2 of multiple samples from both sexes at each age point enabled us to examine developmental trajectories and sex differences in gene expression. To maximize our ability to observe patterns, we conducted PCA on the expression of the 1,000 most variably expressed genes separately for each tissue (Figure 3-2). Comparison of the ranks of expression for the orthologs of these genes in matched tissues in humans and rhesus macaques yielded Spearman correlations of ~0.5–0.8 and ~0.3–0.4, respectively (Tables 3-2,3-3,3-4). Among the seven vervet tissues, the patterns in BA46 and caudate nucleus displayed the clearest association with development; PC1 (20.1% of BA46 variability and 18.5% of caudate nucleus variability) distinguished the vervets nearly linearly by age. All tissues except fibroblasts showed sharply demarcated expression patterns between males and females: on PC1 (hippocampus and pituitary gland, 19.3% and 16.2% of variability,

respectively), on PC2 (BA46, caudate nucleus and blood, 15.5%, 17.4% and 3.2% of variability, respectively) and on PC3 (adrenal gland, 8.2% of variability).

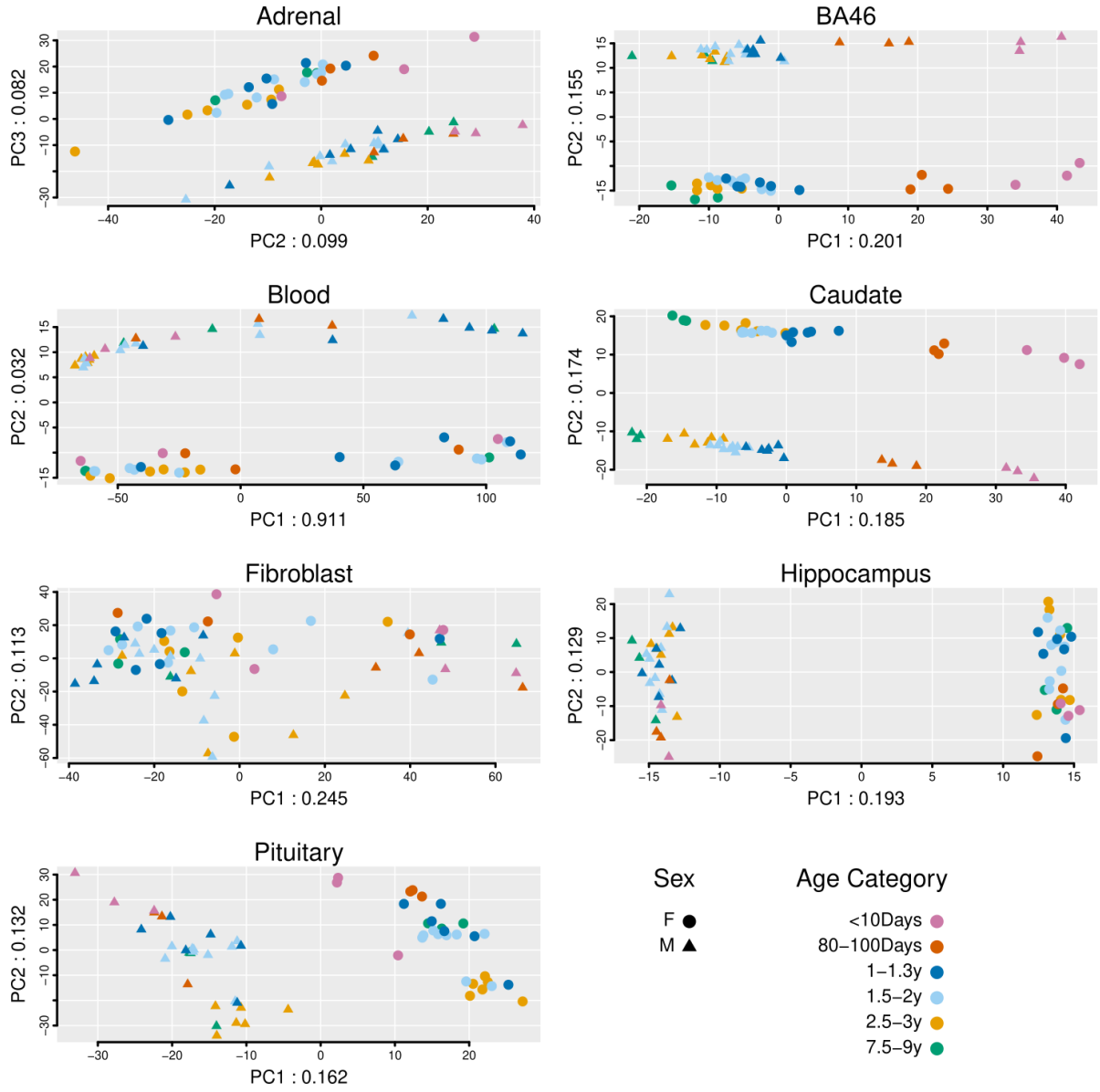


Figure 3-2. Principal-components analysis of the 1,000 genes with the most variable expression levels. Analysis was performed separately by tissue; sample size was 60 animals for adrenal gland, blood, fibroblasts and pituitary gland and 59 for BA46, caudate nucleus and hippocampus. Numbers in the labels for the x and y axes correspond to the proportion of total variance accounted for by that PC.

Age (V)	Age (H)	BA46 (V) vs DLPFC (H)			Caudate (V) vs Striatum (H)			Hippocampus (V) vs Hippocampus (H)		
		# of samples (V)	# of samples (H)	Rho	# of samples (V)	# of samples (H)	Rho	# of samples (V)	# of samples (H)	Rho
7 d	<=5 m	5	2	0.638	5	2	0.548	5	2	0.628
90 d	6-18 m	6	2	0.618	6	1	0.537	6	1	0.615
1-1.25 y	19m-5y	12	3	0.544	12	2	0.539	12	2	0.552
1.5-2.5 y	6-11y	22	3	0.599	22	1	0.567	23	3	0.66
3-4y	12-19y	6	3	0.603	6	2	0.512	6	3	0.601
>=5y	20-60+ y	6	5	0.591	6	5	0.549	6	5	0.622

Table 3-2. Rank correlation values (rho) for expression comparison between vervet and human data from ABA. V=Vervet; H=Human. d=days, y=years, m=months

Age(V)	Age(R)	BA46 (V) vs Medial Frontal Cortex (R)			Caudate (V) vs Basal ganglia (R)			Hippocampus (V) vs Hippocampal Cortex (R)		
		# of samples (V)	# of samples (R)	Rho	# of samples (V)	# of samples (R)	Rho	# of samples (V)	# of samples (R)	Rho
7 d	0 m	2	3	0.381	2	2	0.274	2	3	0.379
90 d	3 m	3	3	0.349	3	3	0.287	3	3	0.372
1-1.25y	12 m	6	3	0.326	6	3	0.303	6	3	0.371
>=4 y	48 m	3	3	0.339	3	3	0.288	3	3	0.376

Table 3-3. Rank correlation values (rho) for expression comparison between vervet and rhesus data from ABA. V=Vervet; R=Rhesus. d=days, m=months, y=years

Vervet-GTEx Comparison

Vervet		Human		Correlation (rho)
Tissue	# of Samples	Tissue	# of Samples	
Adrenal	58	Adrenal	126	0.794
Blood	58	Blood	338	0.78
Caudate	57	Caudate	100	0.683
Hippocampus	58	Hippocampus	81	0.717
Pituitary	58	Pituitary	87	0.795

Table 3-4. Rank correlation values (rho) for expression comparison between vervet and human GTEx

To evaluate whether cell type heterogeneity influences the interpretation of our expression and eQTL results for blood and brain tissues, we conducted a transcriptional deconvolution analysis of these tissues using published data (Gaujoux and Seoighe, 2013; Y. Zhang et al., 2014) (Figure 3-3). We estimated the diversity of cell types per sample in each tissue by calculating entropy, observing that blood had substantially higher diversity of cell types than the three brain tissues (Figure 3-3).

We also examined the relationship between the proportion of specific cell types and developmental stage. For BA46 and hippocampus, the proportion of oligodendrocyte precursor cells decreased as age increased, as observed previously

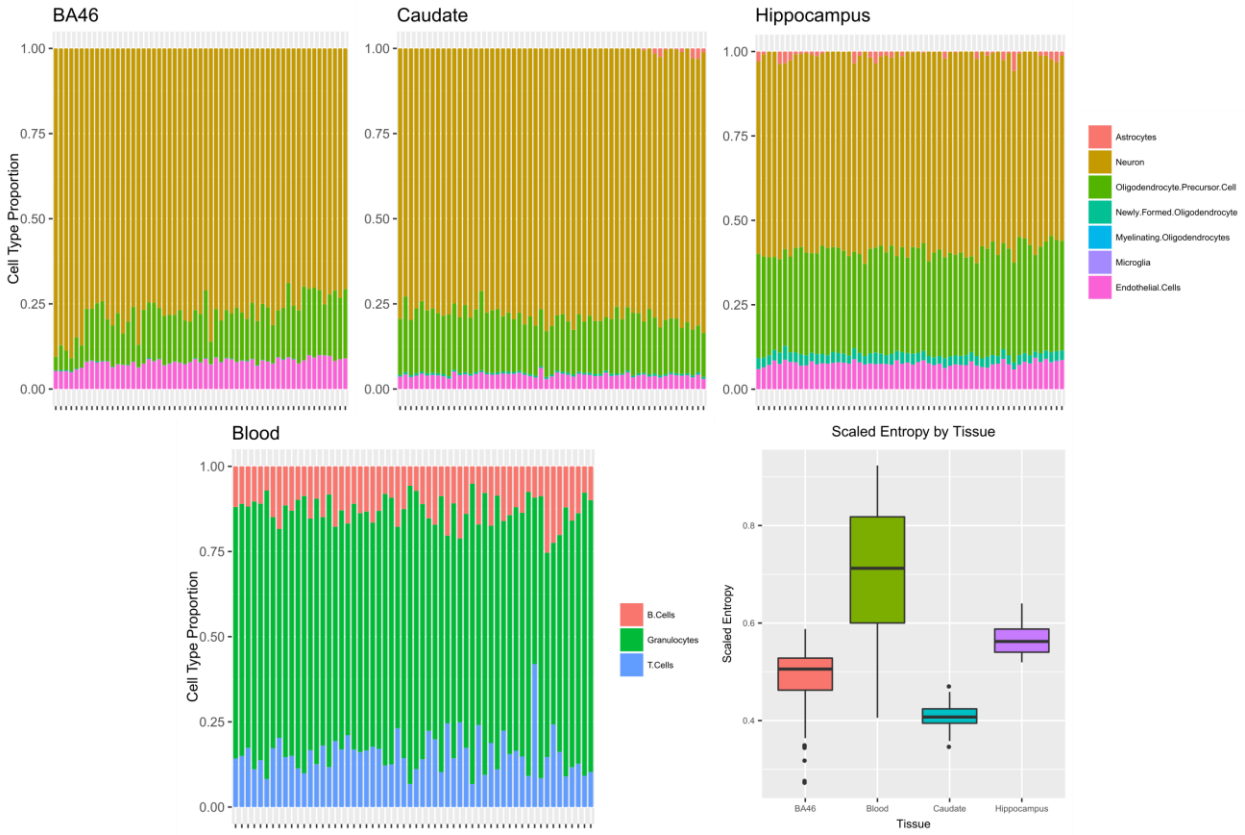


Figure 3-3. Cell type composition in each animal and distribution of scaled entropy of cell type. Deconvolution analysis was applied to vervet BA46, caudate, hippocampus and blood, and the proportion of cell types is presented for each animal, as well as the distribution, over 58 vervets, of scaled entropy for each tissue.

in human (Q. Yu and He, 2017); in contrast, in caudate nucleus, the proportion of this cell type increased with age. Similarly, the proportion of neurons increased with age in BA46 and hippocampus but decreased with age in caudate nucleus (Figures B-1,B-2,B-3). We found no correlation between estimated cell proportions and major PC axes in any tissue. These estimated proportions may not fully reflect in vivo cellular composition, but any bias would remain relatively systematic across animals and so would be unlikely to confound other analyses.

We evaluated the effect of RNA-seq sample batch on transcriptomic profiles and PC patterns. As batch showed association with expression profiles in pituitary gland and adrenal gland (PC2) and caudate nucleus and pituitary gland (PC3), we included it as a covariate in eQTL analyses.

Identification of eQTLs

WGS of 721 VRC monkeys provided the first NHP genome-wide, high-resolution genetic variant set (Y. S. Huang et al., 2015), which includes 497,163 WGS based SNPs that tag common variation across the genome. Using these SNPs, we conducted separate GWAS of data sets 1 and 2 to identify local (probe/gene <1 Mb from an associated SNP) and distant (all other probe/gene–SNP associations) eQTLs in each data set. The covariates in all eQTL analyses included age, sex and batch.

Using SOLAR (Almasy and Blangero, 1998), we identified significant estimated heritability for 3,417 probes in data set 1 (out of the 6,018 filtered probes that we evaluated, corresponding to 5,586 unique genes) at a false discovery rate (FDR) threshold of FDR < 0.01. A GWAS of each heritable probe identified one or more significant eQTLs at 461 local and 215 distant probes (Bonferroni-corrected threshold of 4.8×10^{-8} for local eQTLs and 1.5×10^{-11} for distant eQTLs; Table 3-5).

Approximately 35% of probes with a significant eQTL (173/498) displayed at least one local and one distant significant association.

Tissue	Probes/genes analyzed	Local eQTL	Distant eQTL	% Distant eQTL on same chr
Dataset 1: Microarray				
Blood	3,417	461	215	80.80%
Dataset 2: RNA-seq				
Adrenal	25,187	555	80	54.50%
BA46	27,530	307	30	81.80%
Blood	33,776	60	4	100%
Caudate	28,249	441	47	69.00%
Fibroblast	22,328	239	43	33.20%
Hippocampus	26,957	361	45	70.60%
Pituitary Gland	27,236	596	80	77.50%

Table 3-5. Gene expression data sets. The number of probes/genes with at least one significant local and distant eQTL (at Bonferroni corrected thresholds) are presented. We have 80% power to detect distant eQTLs accounting for 15% of the variability in expression in Dataset 1 and 66% of the variability in Dataset 2

In data set 2, we observed, for each of the five solid tissues, 361–596 genes with local eQTLs and 30–80 genes with distant eQTLs. For blood and fibroblasts, 60 and 239 genes showed local eQTLs and 4 and 43 genes showed distant eQTLs, respectively, all at Bonferroni corrected thresholds (6.5×10^{-10} (local) and 5.3×10^{-13} (distant); Table 3-5). The paucity of eQTLs in blood likely reflects heterogeneity in the proportions of different cell types in this tissue, as found in deconvolution analyses (Figures 3-1 and 3-3). The paucity of eQTLs in fibroblasts has no obvious explanation, although we analyzed fewer genes overall in fibroblasts than in tissues with cellular heterogeneity. For about 70% of Bonferroni-significant eQTLs (local and distant and in all tissues), the SNPs demonstrating association had minor allele frequency (MAF) >30%.

Comparison to human eQTLs

While the eQTLs summarized in Table 3-5 are genome-wide significant at Bonferroni thresholds, we also applied FDR-controlling procedures to expand the list of local eQTLs for more exploratory investigations and to make our results comparable to those of the GTEx project (Table 3-6). We controlled FDR at 0.05 for eGenes (genes with a significant eQTL; see Methods), accounting for multiple testing using a hierarchical error-controlling procedure developed for multitissue eQTL analysis (Bogomolov, Peterson, Benjamini, and Sabatti, 2017). We applied this same procedure to GTEx eQTLs to facilitate comparisons between the data sets.

Tissue	Vervet number of individuals	# Local eQTL Vervet Genes[a]	GTEx number of individuals	GTEx number of eGenes[a]	# Vervet Genes with Human Ortholog	# Orthologous Genes Tested in GTEx[b]	% Tested Genes p<0.05	% Tested Genes p <.05/# tested Genes[c]	% Tested Genes significant genome-wide in GTEx[d]
Adrenal	58	2932	126	2915	1828	1674	100%	28.70%	18.20%
Blood	58	574	338	5438	264	229	100%	70.70%	38.90%
Caudate	57	3140	100	2396	1737	1548	100%	24.60%	14.10%
Hippocampus	58	2437	81	1405	1436	1296	100%	18.40%	9.20%
Pituitary	58	3395	87	2222	1863	1743	100%	20.70%	13.00%

Table 3-6. Comparison of specific genes with local eQTL in Vervet Dataset 2 to GTEx. The number of genes with at least one significant local eQTL in Vervet (at FDR thresholds) are presented.

a The number of eGenes found in the multi-tissue hierarchical FDR procedure applied to vervet Dataset 2 and to GTEx.

b Vervet genes with a human ortholog that were not tested in GTEx were filtered by their QC procedures

c The threshold for significance corrected for the number of genes compared between Vervet and GTEx (column 7).

d Genes were declared significant by GTEx at an FDR of 0.05.

Despite having a smaller sample size than v6 (accession phs000424. v6.p1) of the GTEx project, we identified more local eQTLs (at FDR thresholds applied to both data sets; Methods) for the five solid tissues evaluated in both resources (Table 3-6). The larger number of local eQTLs in vervets likely reflects the more homogenous environment of colonied NHPs as compared to humans and the more uniform tissue

collection process in this study. Specific vervet and GTEx eQTLs overlapped substantially. All vervet genes with a genome-wide significant eQTL (FDR < 0.05) also displayed a human eQTL in the same tissue (P < 0.05), given that the gene had a known human ortholog and was tested in the GTEx project. Using instead the GTEx-defined significance threshold for orthologous genes (FDR < 0.05), an average of 19% of vervet eQTLs corresponded to a human eQTL (Table 3-6). Restricting the comparison to Bonferroni-significant local eQTLs, an average of 23% of vervet eQTLs also had an eQTL in the same tissue in the GTEx data set.

Tissue	# Local eQTL Vervet Genes	# Vervet Genes with Human Ortholog	# Genes Tested in CMC	% Tested Genes CMC FDR<0.20	% Tested Genes significant genome-wide in CMC
Vervet local eQTL at Bonferroni Thresholds					
BA46	307	183	130	100%	90.77%
Caudate	441	225	151	100%	87.42%
Hippocampus	361	187	137	99%	87.59%
Vervet eQTL at FDR thresholds					
BA46	2251	1346	1079	99%	88.60%
Caudate	3079	1712	1316	99%	87.61%
Hippocampus	2377	1391	1115	99%	88.25%

Table 3-7. Comparison of Vervet eQTL with Common Mind Consortium (CMC)

We additionally compared our local eQTL results for brain tissues to the open-access version of human eQTLs from DLPFC, available from the CommonMind Consortium (CMC) (Fromer et al., 2016). Almost 90% of vervet brain local eQTL genes with human orthologs in the CMC data set had a local eQTL at FDR < 0.05 in that data set (Table 3-7).

Hippocampus eQTLs in a region linked to hippocampal volume

As an initial investigation of the impact of vervet eQTLs on higher order traits, we focused on magnetic resonance imaging (MRI)-based hippocampal volume, a highly heritable trait in the VRC (h^2 = 0.95) (Fears et al., 2009) for which the strongest QTL signal across the genome (peak logarithm of odds (LOD) score = 3.42) lies in an ~8.3-Mb segment of CAE18. Power simulations (SOLAR) indicated that, in the VRC pedigree, quantitative trait data for 347 vervets (the number with hippocampal volume data) provided 80% power to detect a locus with LOD = 2 when locus-specific heritability was >45%.

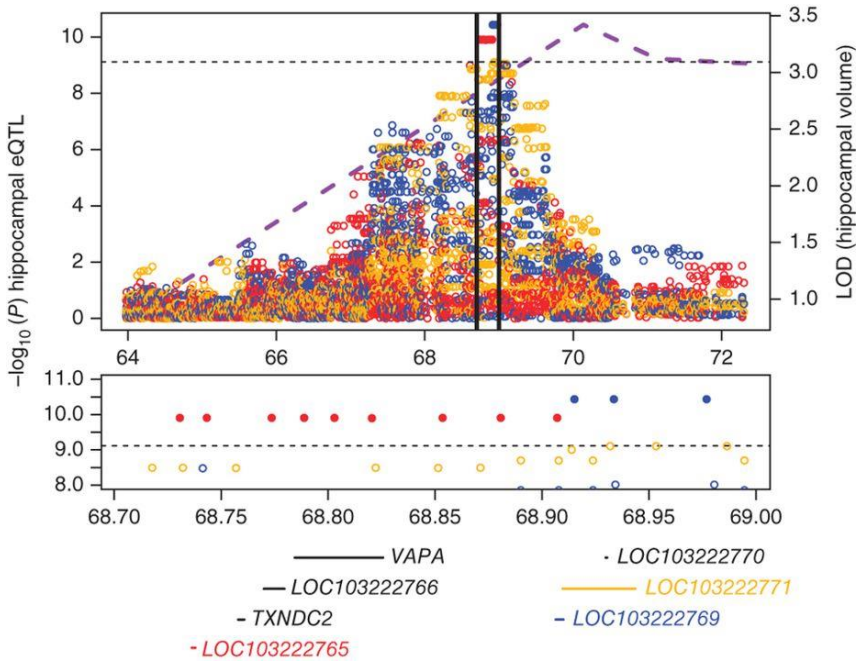


Figure 3-4. Hippocampal volume QTL and local hippocampal eQTLs in RNA-seq analysis. Top, the dashed purple line is the multipoint LOD score for hippocampal volume (measured in 347 animals). Circles correspond to evidence for association in 58 animals of SNPs with hippocampal expression of three genes: LOC103222765 (red), LOC103222769 (blue) and LOC103222771 (gold). Filled circles correspond to genomewide significant associations. The region between the black vertical lines is expanded in the middle and bottom panels. The dashed horizontal line represents the genome-wide significance threshold for local eQTLs. Middle, SNPs with −log10 P > 8 for association with expression in hippocampus; color codes are as in the top panel. Bottom, genes located between 68.7 and 69 Mb (the eQTL region); color codes are as in the top panel. The Pearson correlations between expression of these three genes are as follows: LOC103222765–LOC103222769, r = −0.16; LOC103222765–LOC103222771, r = 0.32; LOC103222769–LOC103222771, r = 0.60.

In the center of the broad region around this linkage peak, two hippocampus-specific local eQTLs were genome-wide significant (Bonferroni threshold; Figure 3-4). These SNPs reside in and regulate expression of two lncRNAs located 168 kb apart: LOC103222765 (nine associated SNPs) and LOC103222769 (three associated SNPs). An additional lncRNA, LOC103222771, situated 2 bp from LOC103222769, showed hippocampus-specific association with six SNPs at a significance level ($P < 1 \times 10^{-9}$) just above the genomewide threshold. While all three genes displayed hippocampus-specific eQTLs, the genes themselves were expressed across all seven tissues that we analyzed and showed no significant sex- or age specific differences in expression patterns (data not shown). The incomplete database annotation for lncRNAs (Mattick and Rinn, 2015) limits comparative analyses of such genes among primates; however, a BLAST search found a homolog for LOC103222765 in the white-tufted-ear marmoset (*Callithrix jacchus*) and one for LOC103222771 in the crab-eating macaque (*Macaca fascicularis*). While LOC103222765 overlaps a coding gene (RAB31), LOC103222769 and LOC103222771 do not overlap the exons of any coding genes (Ulitsky and Bartel, 2013).

3.3 Discussion

We describe here the first NHP resource for investigating the genetic contribution to interindividual variation in multitissue gene expression across development. This resource complements the GTEx project (Ardlie et al., 2015; Wang et al., 2016) but is differentiated from it by a study design that is infeasible in human research. Notably, age-based sampling enabled delineation of tissue-specific expression profiles in relation to developmental trajectories. These profiles illuminate

biological processes associated with the expression patterns of particular genes. For example, several genes critical in synapse formation and postnatal myelination of the central nervous system (Bergoffen et al., 1993; Bond et al., 2002; Sargiannidou et al., 2009; Tang et al., 2005) contribute to the nearly linear age-related pattern observed in BA46 and caudate nucleus, suggesting that the observed expression pattern reflects this process. Conversely, the lack of such a developmentally specific pattern in the hippocampus may be related to the lifelong generation of functional neurons in this tissue, underpinning its functions in learning and memory (Eriksson et al., 1998; van Praag et al., 2002).

Three factors increase the signal-to-noise ratio of vervet eQTL analyses relative to human studies: (i) the homogeneity of environmental exposures; (ii) the greater control over necropsy conditions; and (iii) the restricted genetic background of the population. These factors enabled us to identify 385 genes with genome-wide significant distant eQTLs, including the MRL at IFIT1B.

Just as GTEx data help refine signals from human GWAS of complex traits (Gibson et al., 2015), we used vervet hippocampal eQTLs to identify a set of lncRNAs as candidate genes for hippocampal volume. The genetic and environmental homogeneity of the relatively small vervet study sample likely facilitated these findings and support multitissue vervet eQTL studies as a strategy for identifying loci with a large impact on higher-order phenotypes generally. The tissues examined thus far are a fraction of those available from the same vervets; the investigations reported here can be extended to an additional 60 brain regions and 20 peripheral tissues.

Expanding tissue resources in NHPs, generally, will create additional opportunities to identify biomedically relevant eQTLs (Bakken et al., 2016; Rogers and Gibbs, 2014). The abundance of wild Caribbean vervet populations, and their almost complete identity genetically to the samples we analyzed, make them uniquely valuable for maximizing the value of our eQTL resource (Jasinska et al., 2013, 2009). Each lead SNP for the eQTLs associated with hippocampal volume in the VRC is common in the Caribbean vervet population. We anticipate that our eQTL database will enhance interpretation of well-powered GWAS that can be conducted in these populations for a wide range of complex traits.

3.4 Methods

Study Sample

The monkeys in this study were from the VRC, established by UCLA during the 1970s to 1980s from 57 founder animals captured in the wild on St. Kitts and Nevis (Jasinska et al., 2013). MRI phenotypes were obtained before the VRC moved to the Wake Forest School of Medicine in 2008. All vervets in this study were born in captivity, reared by a mother and socially housed in large indoor–outdoor enclosures, in matrilineal groups that approximated the social structure of wild vervet populations. They had uniform exposure to light and darkness and were fed a standardized diet.

Gene expression

Two gene expression data sets were collected. Data set 1 consisted of microarray (Illumina HumanRef-8 v2) assays of whole-blood RNA in 347 vervets. Data set 2 consisted of RNA-seq data from seven tissues assayed in 60 animals. Six

vervets were in both data sets. No randomization was applied in allocating animals to data sets, and investigators were not blinded to the allocation of animals to data sets.

Data set 1: microarrays from whole blood

The microarray data set has been described previously (GSE15301) (Jasinska et al., 2009). To obtain a set of probes usable in vervet from the Illumina HumanRef-8 v2 microarray, we used the vervet reference sequence to select probes containing no vervet indels and demonstrating ≤5 mismatches, with a maximum of one mismatch in the 16-nt central portion of the probe. To prevent bias in expression measurement due to SNP interference with hybridization, we excluded probes targeting sequences with common SNPs identified in the VRC. A total of 11,001 probes passed these filters. Illumina provides a 'detection P value' for detection of a given probe in a specific individual (with $P < 0.05$ considered significant). We analyzed 6,018 probes with detection P values of $P < 0.05$ in at least 5% of vervets and tested 3,417 significantly heritable probes for eQTL association. Expression data were inverse normal transformed before analysis.

Data set 2: RNA-seq data from seven tissues

Tissues collected during experimental necropsies (Wake Forest School of Medicine IACUC protocol A09-512) were obtained from 60 vervets representing ten developmental stages, ranging from neonates (7 d) through infants (90 d and 1 year), young juveniles (1.25, 1.5, 1.75 and 2 years), subadults (2.5 and 3 years) to adults (4+ years), with 6 vervets (3 male and 3 female) from each developmental time point. Two vervets (a 1.75-year-old female and a 7-d-old male) for which we did not have WGS data were excluded from this study. Altogether, we included 11

vervets less than 1 year old, 23 vervets between 1 and 2 years old, and 24 vervets between 2 and 4 years old, 29 males and 29 females.

For all vervets, we conducted RNA-seq in seven tissues: three brain tissues (BA46, caudate nucleus and hippocampus), two neuroendocrine tissues (adrenal gland and pituitary gland) and two peripheral tissues (blood and fibroblasts). From purified RNA, we created two types of cDNA libraries; poly(A)+ RNA (fibroblasts, adrenal gland and pituitary gland) and total RNA (blood, caudate nucleus, hippocampus and BA46) libraries. For one vervet in which the RNA-seq data indicated a mix-up between the caudate nucleus and BA46 samples, we excluded data from these two tissues in all analyses.

RNA-seq reads were aligned to the vervet genomic assembly Chlorocebus_sabaeus 1.1 by the ultrafast STAR aligner (Dobin et al., 2013) using our standardized pipeline. STAR was run using default parameters, which allow up to ten mismatches. Gene expression was measured as total read counts per gene. For paired-end experiments, we considered total fragments. Fragment counts aligning to known exonic regions (based on NCBI Chlorocebus sabaeus Annotation Release 100) were quantified using the HTSeq package (Anders et al., 2015). The counts for all 33,994 genes were then combined; weakly expressed genes (mean in raw counts of <1 across all samples) and genes detected in <10% of individuals were filtered out. The calcNormFactors function in the edgeR package (Robinson, McCarthy, and Smyth, 2010) was applied to normalize counts. Finally, an inverse normal transform was applied to counts per million, before analysis.

Deconvolution analysis was performed in vervet brain and blood tissues using available references for these tissues. For brain tissues, gene signatures were

obtained from Zhang et al. (Y. Zhang et al., 2014); for blood, cell-type-specific markers were taken from data sets built into the CellMix package (Gaujoux and Seoighe, 2013). Cell type composition for each tissue was evaluated using the CellMix R package.

Data sets for comparative expression analysis between species

We performed comparative analysis of gene expression between vervet brain samples, GTEx and age-matched samples from Allen Brain Atlas (ABA) data sets; BrainSpan (human RNA-seq data) and the NIH Blueprint NHP Atlas (rhesus macaque microarray data) (Bakken et al., 2016; H. J. Kang et al., 2011). Matching the three vervet brain tissues to the most closely corresponding available tissues in the other species, we compared overall expression profiles between these species and inspected developmental expression patterns for selected genes.

Overall mean levels of expression were compared between species using a rank correlation. GTEx and BrainSpan were compared to vervet independently. For the GTEx comparison, vervet tissues were matched to the five available corresponding tissues: adrenal gland, blood, caudate nucleus, hippocampus and pituitary gland. Analyses involving the two ABA data sets were limited to the three brain regions most closely related to the brain tissues analyzed in vervets. As the rhesus macaque data set included only males, we limited comparisons to male vervets.

For each of the three data set comparisons, vervet raw counts were first converted to RPKM values using the edgeR R package (Robinson et al., 2010). GTEx and human ABA counts were already normalized to RPKM values; rhesus macaque counts had been normalized using an RMA approach (Bakken et al., 2016). Mean

expression was then calculated by tissue for each data set. For ABA data sets, mean expression was calculated by tissue type and time point, according to matched age groups. Vervet gene names were converted to their corresponding human orthologs to ensure gene names matched between vervet and comparison data sets; genes with no human ortholog were excluded. Additionally, genes not present in both vervet and the comparison species data set were also removed. Variances were then calculated for each gene across the five or three different vervet tissues, for GTEx and ABA comparisons, respectively. The top 1,000 genes with the highest variances were then selected for rank–rank correlation testing. The base R function cor.test was used to perform correlation testing.

Hippocampal volume

Estimates of hippocampal volume were obtained in 347 vervets >2 years of age using MRI. Details of the image acquisition and processing protocol were described previously (Fears et al., 2009). Prior to genetic analysis, hippocampal volume was log transformed and regressed on sex and age (SOLAR (Almasy and Blangero, 1998)); residuals were used as the final phenotype.

Genotype data

Genotypes were generated through WGS, as described previously (ERP008917) (Y. S. Huang et al., 2015). Genotypes from 721 VRC vervets that passed quality control procedures can be queried via the EVA at EBI. Two genotype data sets were used (Y. S. Huang et al., 2015): (i) the Association Mapping Set consists of 497,163 SNPs on the 29 vervet autosomes. This set has, on average, 198 SNPs per megabase of vervet sequence, with a maximal gap of 5 kb between adjacent SNPs. (ii) The Linkage Mapping Set consists of 147,967 SNPs on the 29 vervet

autosomes. This set has, on average, 58.2 SNPs per megabase of vervet sequence, with an average gap of 17.5 kb between adjacent SNPs.

The software package Loki (Heath, Snow, Thompson, Tseng, and Wijsman, 1997), which implements Markov chain Monte Carlo methods, was used to estimate multipoint identical by decent (MIBD) allele sharing among all vervet family members from the genotype data. As long stretches of IBD were evident among these closely related animals, a reduced marker density (9,752 SNPs of the 148,000 set) was sufficient to evaluate MIBD at 1-cM intervals. The correspondence between the physical and genetic positions of vervet SNPs was established by interpolation using 360 markers from the vervet STR linkage map (Jasinska et al., 2007), for which physical and genetic positions were known.

Principal-component analysis.

The top 1,000 genes with the most variable expression were selected for each tissue (data set 2), and PCA was applied to log2-transformed counts per million, using the singular value decomposition and prcomp function in R. Expression was mean-centered before analysis. We examined genes in the top and bottom 10% of the distribution of PC loadings on PC1, PC2 or PC3 (200 genes per tissue, per PC) where these loadings are taken from the eigen decomposition of the expression matrix. The gene loadings represent the amount that gene contributes to the PC value for that sample on the axis in question.

Mapping gene expression and hippocampal volume phenotypes

For the higher-order phenotype hippocampal volume, we anticipated having power only to detect loci with a strong effect and therefore evaluated it using linkage

analysis. For gene expression traits, we expected to have power to identify relatively small effects and therefore applied genome-wide association analyses.

Heritability and multipoint linkage analysis

We estimated the familial aggregation (heritability) of traits using SOLAR, which implements a variance components method to estimate the proportion of phenotypic variance due to additive genetic factors. This model partitions total variability into polygenic and environmental components. The environmental component is unique to individuals while the polygenic component is shared between individuals as a function of their pedigree kinship. Genome-wide multipoint linkage analysis of hippocampal volume was also implemented in SOLAR, which further partitions the genetic covariance between relatives for each trait into locus-specific heritability and residual genetic heritability. Linkage analysis was performed at 1-cM intervals using the likelihood-ratio statistic.

Association analysis

Association between specific SNPs and gene expression phenotypes was evaluated using EMMAX (H. M. Kang et al., 2010). EMMAX employs a linear mixed model approach, where SNP genotype is a fixed effect, and correlation of phenotype values among individuals is accounted for using an identity-by-state approximation to kinship. Association analyses used 497,163 SNP markers and for both data set 1 and data set 2 included age (in data set 2, age corresponds to developmental stage), sex and sample batch as covariates. It is common to try to account for unmeasured factors influencing global gene expression by including probabilistic estimation of expression residuals (PEER) factors as covariates (Stegle, Parts, Piipari, Winn, and

Durbin, 2012). We considered the controlled nature of the study environment and experimental design to preclude the need for this adjustment.

Multiple-testing considerations in eQTLs

As our primary error-controlling strategy for eQTL discovery, we used a Bonferroni correction to account for multiple testing across genes, SNPs and tissues. Thresholds for data set 2 were more stringent, as it included analysis of multiple tissues and tested more genes than in data set 1 (~25,000 versus ~3,000). In data set 1, we analyzed association with 3,417 heritable probes. The local eQTL significance threshold (4.8×10^{-8}) was corrected for testing of SNPs within 1 Mb of 3,417 probes. The distant eQTL significance threshold (1.5×10^{-11}) accounted for genome-wide testing of 3,417 probes. Data set 2 significance thresholds were constructed in a similar fashion but also accounted for testing of 191,263 gene–tissue combinations (Table 3-5). The RNA-seq local eQTL threshold was 6.5×10^{-10}, and the distant eQTL threshold was 5.3×10^{-13}.

To identify multitissue eGenes, the tissues in which they are active and the associated SNPs in each of these tissues, we used TreeBH, a hierarchical testing approach (Bogomolov et al., 2017) that extends the error-controlling procedure characterized in Peterson et al. (C. B. Peterson, Bogomolov, Benjamini, and Sabatti, 2016) to multitissue eQTLs.

We compared the number of eGenes identified in each tissue using the above procedure with the results of GTEx (Analysis Release V6; dbGaP accession phs000424.v6.p1). We downloaded all eQTL association results for tissues in common with our study and applied this same hierarchical procedure to the GTEx results to identify eGenes.

Chapter 4

Exploring gene expression changes across developmental time points in vervet hippocampus

4.1 Introduction

Over the last decade, various genetic and genomic resources have been compiled to further our understanding of brain phenotypes across healthy and diseased individuals (Negi and Guda, 2017; Ramasamy et al., 2014; van Erp et al., 2016). More importantly, many studies have focused on identifying genomic regions which may be involved in regulating gene expression and thus contributing to quantitative or disease phenotypes (Majewski and Pastinen, 2011). These regions, or expression quantitative trait loci (eQTL), are enriched in genomic regions associated with disease phenotypes in genotype-wide association (GWAS) studies (Welter et al., 2014). Published resources have revealed a high level of tissue specificity in eQTL results (Consortium et al., 2017), thus highlighting the importance of studies of brain regions to further our understanding of the genetic contributions to brain disorders. Interestingly, even within the same tissue, eQTLs have been found to vary across specific cell types (Ackermann, Sikora-Wohlfeld, and Beyer, 2013; Gerrits et al., 2009), proportions of which have been shown to differ across development (Q. Yu and He, 2017). While existing resources, such as the Genotype-Tissue Expression (GTEx) project (Ardlie et al., 2015; Consortium et al., 2017), provide expression and genetic data across various brain regions, donor ages range from 18-70 years and thus fail to provide data across developmental time points; such data are crucial to our understanding of neurodevelopmental disorders. Developmental expression data

are available through the Allen Brain Atlas (Hawrylycz et al., 2015; Miller et al., 2014) (ABA) for human and rhesus macaque, however no genetic marker information is available for the ABA dataset and, perhaps more importantly, a limited number of samples are available at each developmental time point, rendering the heterogeneity in expression signatures difficult to resolve.

Non-human primate models provide numerous advantages over other model organisms for the study of neuropsychiatric disorders, due to their large genetic similarities and strong resemblance in brain circuitry and anatomy with humans (Warren et al., 2015). The Caribbean-origin vervet monkey (*Chlorocebus aethiops sabaeus*) is an Old World monkey species frequently used in biomedical research (Jasinska, 2019; Jasinska et al., 2013) that has served as a model for studies in Alzheimer's disease and aging (J. A. Chen et al., 2018; Kalinin et al., 2013; Postupna et al., 2017), the role of insulin in increasing the risk of Alzheimer's disease in diabetic individuals (Morales-Corraliza et al., 2016), and the effects of fetal alcohol exposure on hippocampal neurons (Burke, Ptito, Ervin, and Palmour, 2015). The Caribbean-origin vervets, which have been previously described in detail (Y. S. Huang et al., 2015), provide a unique opportunity for genetic trait mapping for multiple reasons. First, the Caribbean populations, which were founded from a small number of West African vervets, are characterized by reduced genetic variability due to their rapid expansion from an extreme bottleneck; because of this demographic history, many highly deleterious alleles are present in relatively high frequency in these populations. Second, the large genetically and phenotypically characterized vervet pedigree established from these founder populations, the Vervet Research Colony (VRC), facilitates studies under a controlled environment thus increasing the power to

observe effects of other variables. Third, hundreds of specimens are available from VRC brain and peripheral tissues from various developmental time points.

We previously characterized the role of genetic variation in determining gene expression differences across seven different tissues spanning ten time points (ranging between 7 days and 9 years) in VRC monkeys, and identified numerous eQTL genes (eGenes), as well as a region on chromosome 18 associated with hippocampal volume (Jasinska et al., 2017), although no age-related genes were identified in hippocampal tissues. Much of hippocampal development occurs during embryonic stages (Khalaf-Nazzal, R; Francis, 2013) and therefore identifying age-related genes in the hippocampus might require younger animals. To address this possibility, we obtained hippocampal tissue from 32 animals under 1 year of age, extending the range of ages from 0 to 9 years, and explored expression differences across developmental time points using RNA sequencing. We identified age-related groups of transcripts, which correlated well with existing human and primate resources. We also confirmed and expanded the hippocampal eQTL catalog, and identified an additional two genes correlated with hippocampal volume.

4.2 Results

Hippocampal samples were obtained at six time points ranging between 0 and 270 days of age (6-7 per group, 32 animals in total). RNA sequencing was performed and analyzed as previously described (Jasinska et al., 2017). This novel dataset was then combined with the previous hippocampal dataset that included 59 animals studied at ten developmental time points ranging between 7 days and 9 years. Thus, a combined analysis was performed on samples from 91 animals at 15 developmental

time points, ranging between 0 days and 9 years (Figure 4-1, Table C-1), with roughly the same number of male and female animals at each time point.

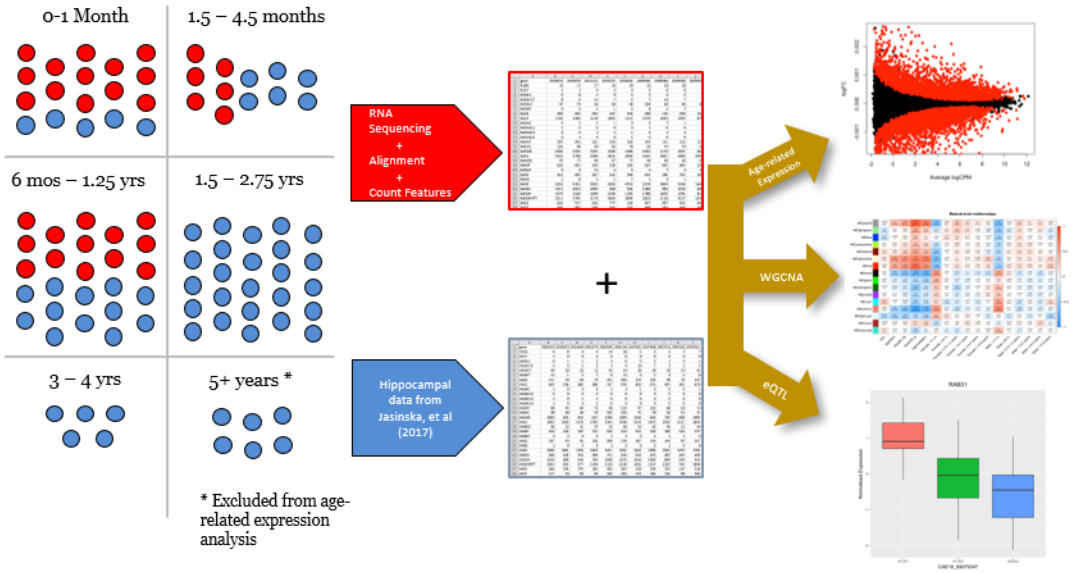


Figure 4-1. Schematic summarizing hippocampal samples and analyses performed. Newly collected hippocampal samples are shown in red while previously published hippocampal samples (Jasinska, et al.) are in blue under their corresponding time points. After processing new samples, expression data was combined with existing hippocampal data. Age related gene expression, WGCNA and eQTL analyses were then performed on the combined dataset.

We examined the relationship between quantitative phenotypes (such as body and brain weight) and age in days. Body weight was linearly correlated with age in days (Figure C-1A), while brain weight showed a steep increase in animals younger than 100 days and a plateau shortly after (Figure C-1B). These trends are consistent with what has been previously reported in humans (Dekaban and Sadowsky, 1978), highlighting the importance of studying younger animals when exploring genes involved in developmental processes in brain regions such as hippocampus.

Correlation and network analysis identify age-related transcripts in the developing hippocampus.

Over 2,000 transcripts were correlated with age in days (FDR < 0.05) with a similar number of directly and inversely correlated transcripts (Figure C-2A). Gene ontology analysis identified biological processes involving regulation of signal transduction and cell communication, as well as various protein kinase activities (Figure C-2B). Because our six oldest animals were significantly older than the rest, we wanted to determine whether these samples might be driving our observed age-

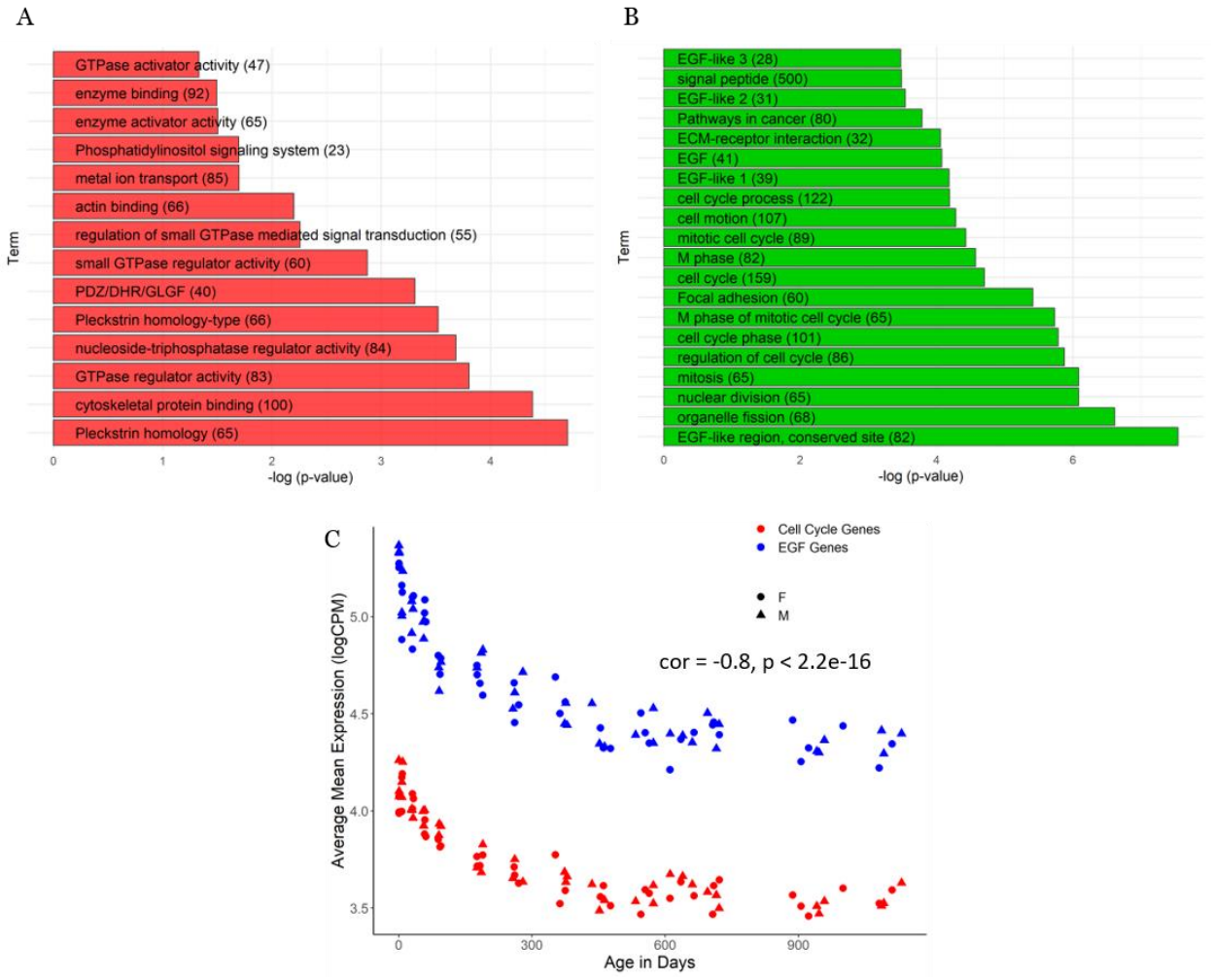


Figure 4-2. DAVID functional analysis results of up- and down-regulated age-related genes. (A) Functional analysis results of positively correlated age-related genes (Animals aged < 1500 days; FDR < 0.05). (B) Functional analysis terms enriched in list of negatively correlated age-related genes (Animals aged < 1500 days; FDR < 0.05). (C) Average expression trend of epidermal growth factor and cell cycle genes with age.

correlations. We calculated correlation values for all 2,767 age-related genes, both including and excluding our six oldest samples, and determined that the age effect was being driven by these six samples for 15% of our age-related genes (n=404). Thus, in order to identify with confidence developmentally regulated transcripts, we repeated our analysis after excluding our six oldest animals.

Age-related expression analysis limited to animals aged between 0 and 1,500 days identified age-related expression in more than 6,000 genes (FDR<0.05; Figure C-3). Comparison of these results to the results that we obtained when including older animals identified significant overlap of genes with negative and positive correlation values ($p < 2.2e-16$, Fisher exact test; Figure C-3B). We performed functional annotation analyses on our age-related genes using DAVID (D. W. Huang et al., 2009b, 2009a) and observed a significant overlap of actin binding genes with our positively correlated age-related transcripts (FDR < 0.05, Figure 4-2A). Similarly, terms significantly associated with our negatively correlated transcripts included mitotic nuclear division, cell cycle and epidermal growth factor-like conserved regions (FDR < 0.05, Figure 4-2B,C).

In order to refine the list of age-associated transcripts, and to identify groups of co-expressed transcripts during development, we performed weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008). We identified 16 WGCNA modules and correlated their eigengenes (see Methods) with phenotypic information, including age in days (or categorized in 6 age groups), sex, and brain weight (Figure 4-3). Modules correlated with body weight were also significantly correlated with age in days, as expected given the linear relationship between body weight and age (Dekaban and Sadowsky, 1978) (Figure C-1A). Modules with

eigengenes correlated with brain weight (g) were also correlated with the "<= 1 month" age category, independent of sex, confirming the relationship we had observed between animals under 100 days old and brain weight (Figure C-1B).

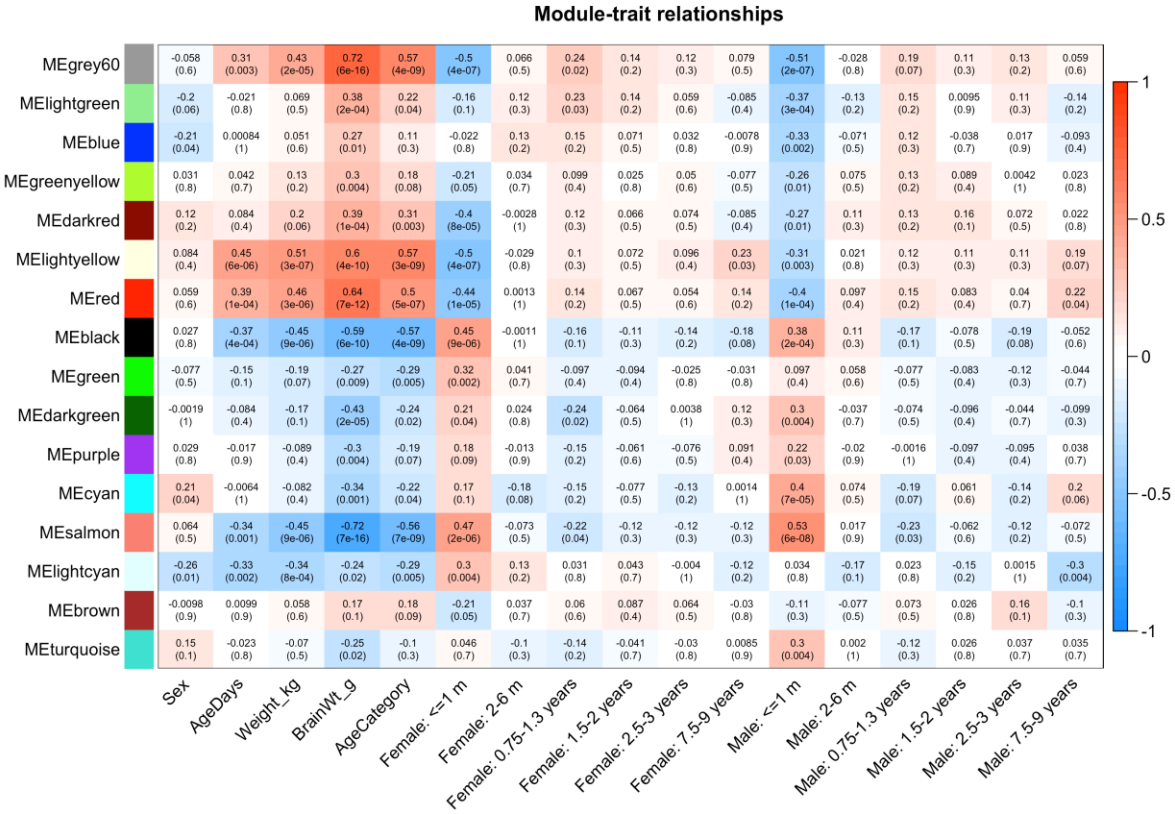


Figure 4-3. WGCNA module correlation coefficients and p-values for multiple traits (x-axis). P-values, listed in parenthesis, have been FDR corrected to account for multiple hypothesis testing. Modules with eigengenes positively correlated with specific traits are shown in red, while those with negatively correlated eigengenes are shown in blue.

We used enrichR (E. Y. Chen et al., 2013; Kuleshov et al., 2016) to functionally annotate the top modules positively (grey60 module, including 1,067 transcripts, Figure 4-4A) and negatively (salmon module, 1,608 transcripts, Figure 4-4B) correlated with brain weight, which also presented a marginally significant correlation to age in days (Figures 4-4C, 4-4D). Enrichment analysis showed an

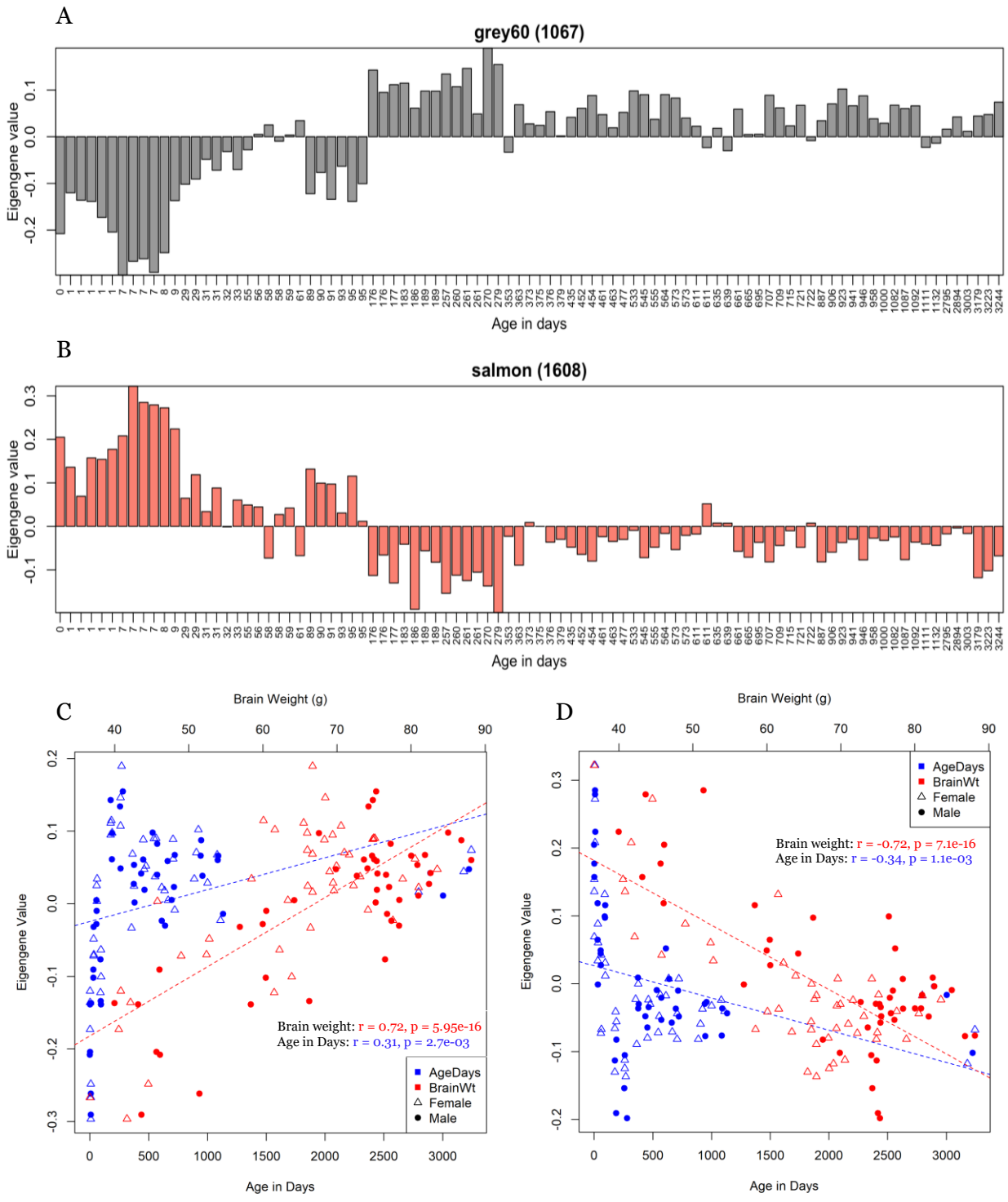


Figure 4-4. Bar and scatterplots of eigengene values for Grey60 and Salmon modules. Eigengene values of (A) grey60 and (B) salmon module by increasing age in days. (C) Grey60 and (D) salmon module eigengene values plotted against age in days (bottom x-axis) and brain weight (top x-axis).

overrepresentation of apoptosis modulation and signaling genes in the grey60

module, while cell cycle genes were enriched in the salmon module (FDR < 0.05).

Comparisons to other developmental datasets

We compared our hippocampal age-related genes with age-related hippocampal genes from humans and other non-human primates, using developmental datasets available in public repositories. We first analyzed human hippocampal data from the BrainSeq project (including 286 samples from individuals ranging from 3.5 weeks-84 years of age)(Collado-Torres et al., 2019). Using the same approach applied to our vervet data (see Chapter 4-4, pg 60), we identified 8,567 age-related human hippocampal genes of which 4,652 had known vervet orthologs. Of those with known vervet orthologs, 2,401 overlapped our age-related

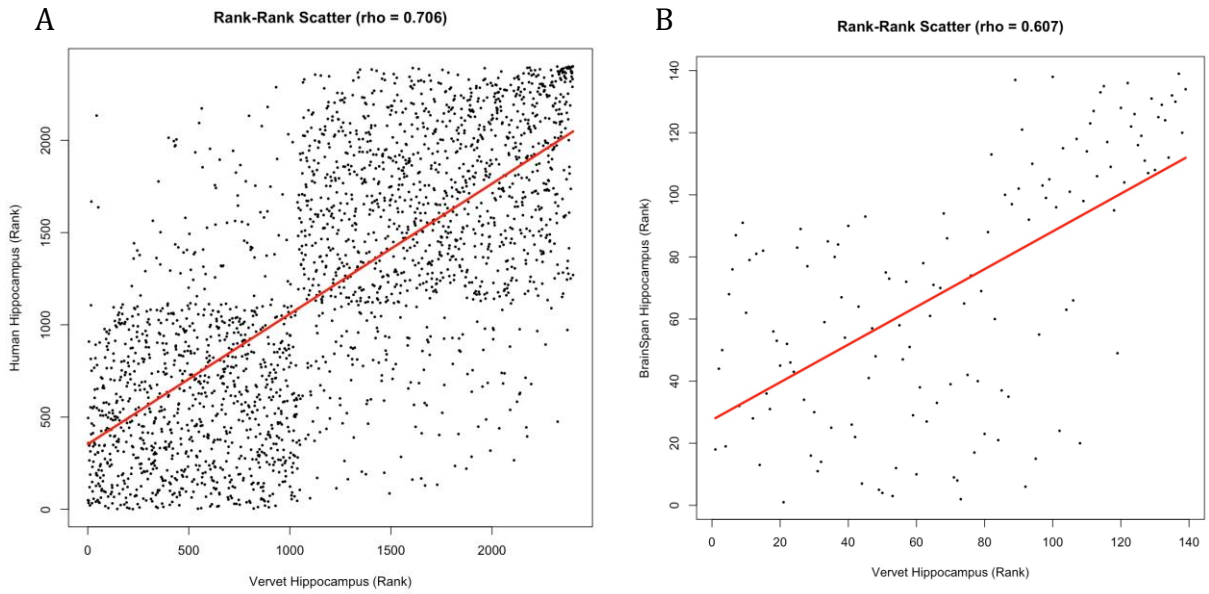


Figure 4-5. Rank-rank plots using signed log P-values. Plots comparing signed log p-values of shared age-related genes in vervet hippocampus versus human (A) BrainSeq and (B) BrainSpan signed log P-values.

vervet genes with the age-effect occurring in the same direction in 2,140 of these (89%). We then explored the ranking of our shared age-related genes using a Spearman rank correlation test on signed log p-values and found a strong correlation

between vervet and human hippocampal results (Figure 4-5A, rho=0.706, p<0.05). Additionally, we plotted rank mean expression values of the top 1000 vervet age-related genes at matched human and vervet developmental timepoints (Table C-2) and observed an even higher degree of correlation (rho = 0.873-0.908, Fig. C-4).

We performed a similar analysis on the BrainSpan human developmental dataset (comprising 17 postnatal samples across 15 time points) (H. J. Kang et al., 2011) and Allen Brain Atlas (ABA) rhesus developmental dataset (including 12 samples from 4 timepoints)(Bakken et al., 2016). The Spearman rank correlation value of signed log p-values was lower in human BrainSpan comparisons (rho=0.607, p=2.21e-15, Fig. 4-5B), most likely due to the smaller sample size of the BrainSpan

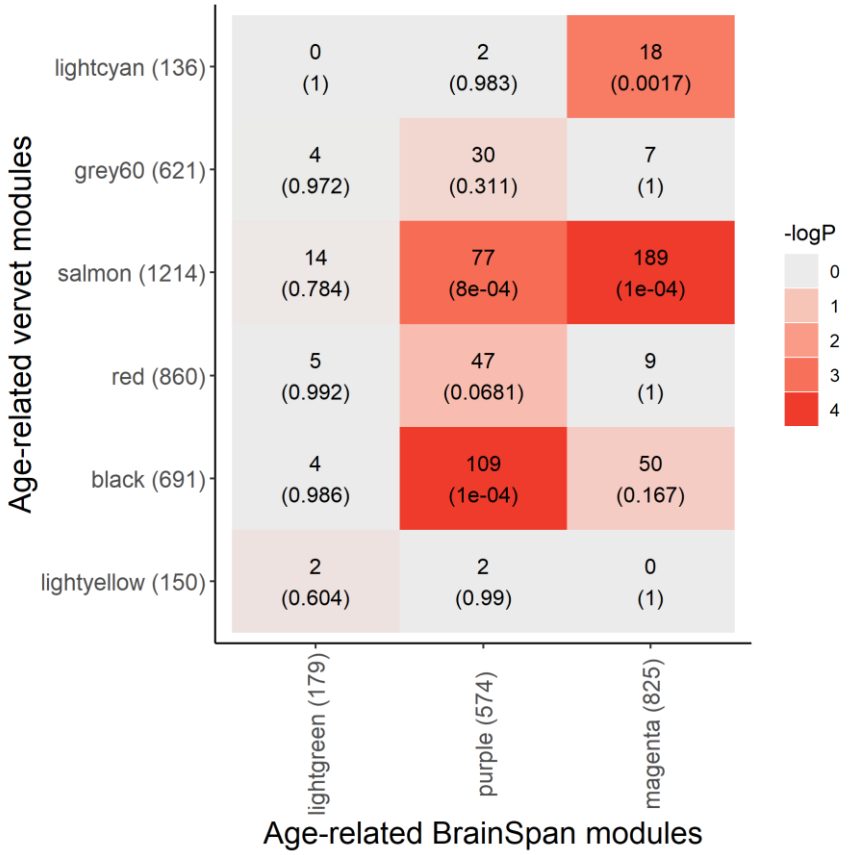


Figure 4-6. Overlap between vervet and human BrainSpan WGCNA age-related modules. Table shows number of overlapping genes between modules and permutation p-values in parenthesis.

dataset, which in turn led to the identification of a smaller number of age- related genes (n=736). While we observe higher correlation values of ranked expression values at almost all six timepoints (rho = 0.593-0.756, Fig. C-5) they are lower than BrainSeq ranked expression correlations, most likely a result of the different normalization methods used on vervet and human BrainSpan expression data. Age-related analysis on the ABA rhesus dataset failed to uncover age-related genes at $p < 0.05$. Moreover, correlation analysis of mean expression data of top vervet age-related genes at age-matched timepoints (Table C-3) yielded low correlation values (rho=0.407-0.436, data not shown).

Finally, we performed WGCNA on the human developmental BrainSpan dataset to identify age-related modules. While no modules were significantly correlated with age, we identified three modules significantly correlated with younger age categories in male and female samples (FDR<0.05, Figure C-6). We then tested for enrichment of genes from age-related vervet modules in these three human age category-related modules. We observed genes from three vervet modules were enriched in at least one human BrainSpan module ($p < 0.05$, permutation test, Figure 4-6). More importantly, similar to BrainSpan modules, these age-related vervet modules were correlated with our youngest age group (< 5 mos) in both females and males. In addition, these enriched vervet modules were correlated in the same direction as BrainSpan modules, while modules without significant correlation were inversely correlated.

Expression Quantitative Trait Locus Analysis

We previously identified (Chapter 3.2, pg 32) eQTLs in multiple tissues in vervets, and comparison of our identified eQTL genes (eGenes) with published

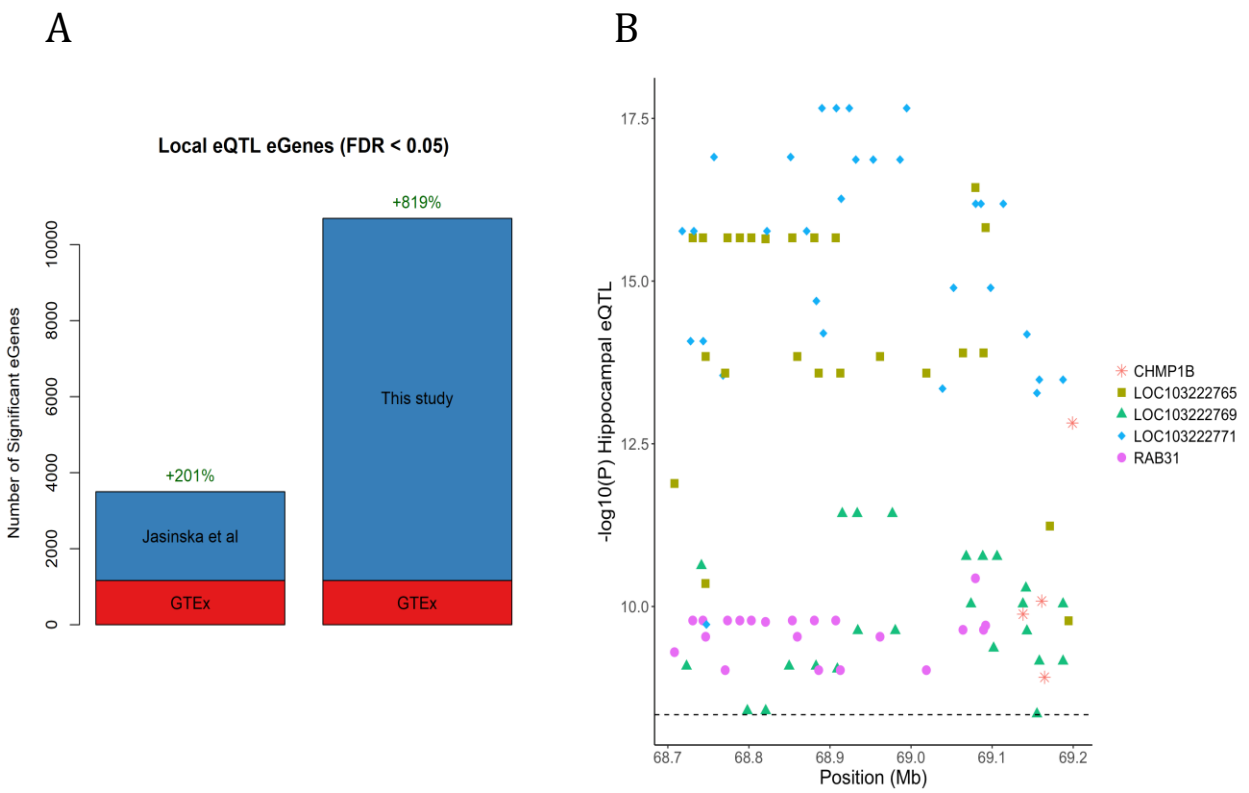


Figure 4-7. eQTL results. (A) eQTL results comparing number of local eGenes identified in Jasinska et al, GTEx and this study at an FDR threshold of 0.05. (B) Significant associations located within hippocampal volume associated locus.

resources, such as GTEx, revealed an increased number of local eGenes identified despite our smaller sample size (Figure 4-7A)(Jasinska et al., 2017). More importantly, we performed linkage analysis in this same study and identified a region of chromosome 18 that was significantly associated with hippocampal volume (LOD score = 3.42). Incorporating our hippocampal eQTL results with this linkage finding, we observed three long noncoding RNA (lncRNA) eGenes associated to SNPs within this locus, one of which was marginally significant (*LOC103222771*). Using qRT-PCR we confirmed expression of these lncRNAs was significantly correlated with hippocampal volume. Thus, to follow up on our eQTL findings and possibly identify additional eGenes within this hippocampal volume associated region, we performed

an eQTL analysis on our combined dataset. We identified 1,567 genes with local eQTLs at a Bonferroni threshold of $4.6x10^{-09}$, which is four times more than identified in our previous dataset (Jasinska et al., 2017) (n=361, Figure 4-7A) denoting an increased power due to a larger sample size. In addition to confirming SNPs at the locus previously linked to hippocampal volume as being associated with gene expression of the three nearby lncRNAs: *LOC103222765*, *LOC103222769*, and *LOC103222771*, we identified two additional genes, *RAB31* and *CHMP1B*, which were also associated to SNPs within that locus (Figure 4-7B) expanding the list of possible candidates to investigate.

We then compared our results to hippocampal eQTL results from GTEx. All of our local eGenes replicated in GTEx at a threshold of $p < 0.05$ (Table 4-1). More importantly, in comparison with GTEx we identified more than eight times the number of local eGenes (Figure 4-7A). At a Bonferroni threshold of $3.67x10^{-12}$, we identified 262 genes with distant associations, defined as locus/transcript associations in which the locus was located further than +/- 1Mb from the associated transcript, or on a different chromosome. Of these 262 genes, 26 were found to have eQTLs on a different chromosome. These 26 genes were distributed across the genome with the largest number associated to SNPs clustered on chromosomes 16, 9 and 25 (Table C-4). No locus was found to be associated to more than 1 gene from a different

Multiple hypothesis correction method	Vervet number of individuals	# Local eQTL Vervet Genes	GTEx number of individuals	GTEx number of eGenes	# Vervet Genes with Human Ortholog	# Orthologous Genes Tested in GTEx	% Tested Genes p<0.05	% Tested Genes significant genome-wide in GTEx
Bonferroni	87	1,567	81	310	1,045	968	100%	4.40%
FDR	87	9,530	81	1,164	6,306	5,908	100%	8.10%

Table 4-1. GTEx comparison results using Bonferroni and FDR corrected thresholds.

chromosome. Functional analysis of these 26 genes was not possible due to the fact that 25 have no known human ortholog.

Next, following up on our differential expression analysis, we sought to determine whether our age-related differentially expressed genes were over- or underrepresented in our eQTL results. Transcripts positively correlated with age were neither over- nor under-represented in our eQTL results. In contrast, genes with expression inversely correlated with age (either including and excluding older animals) were significantly underrepresented in our eQTL results ($p <= 4.198x10^{-07}$). This result aligns with our previously reported finding that age-related genes were less likely to be eGenes than non-age-related genes (Jasinska et al., 2017).

4.3 Discussion

Ethical limitations restrict the availability of developmental samples to investigate differences in gene expression across development in humans, a limitation that has obvious relevance for our understanding of brain-related traits and diseases (Glass et al., 2013). Here, we provide a survey of transcriptional activity in the vervet hippocampus across various developmental time points, corresponding to human time points ranging from birth to old age. The correlation between brain weight and earlier developmental timepoints highlights the importance of studying younger samples as they contribute to this phenotype.

We observe genes with significant correlation with age driven by older animals included several classes of genes essential for neurodevelopment (indicated in Supplementary Table 2). Among them were genes implicated in autism, such as *NLGN3, MARK1, SETD5, NLGN4X*, and other genes involved in synaptic functions. For

example, *DVL1* is implicated in modulating of APP processing (Mudher et al., 2001) and is a key player in aging and Alzheimer disease-related Wnt signaling pathway (Palomer, Buechler, and Salinas, 2019; Tapia-Rojas and Inestrosa, 2018).

Age-related transcripts identified after removing our six oldest animals show an overrepresentation of cell cycle and epidermal growth factor genes. Previous studies have reported epidermal growth factor genes play a role in healthy aging and longevity in C. elegans (S. Yu and Driscoll, 2011). Moreover, a study performed in mice reported increased neurogenesis in the hippocampal dentate subgranular zone and the subventricular zone when treated with heparin-binding epidermal growth factor-like growth factor (HB-EGF) (Jin et al., 2003). Just as neurogenesis decreases with aging (Apple, Solano-Fonseca, and Kokovay, 2017), expression of EGF genes also decreases which further suggests a shared mechanism for these biological processes.

Additionally, six of our 16 modules identified by WGCNA were correlated to the same multiple phenotypes, namely, age in days, body weight, brain weight, age category and "< = 1 month" in both sexes. We hypothesized that such modules contain genes involved in general aging pathways and thus their being enriched in cell cycle processes and apoptosis is consistent with previously reported findings regarding aging (Chandler and Peters, 2013; Cooper, 2012). The directionality of our observations reinforces the idea that apoptosis increases during the aging process, since the genes driving this enrichment show a positive correlation with age (Figure 4-3).

We have shown that expression patterns for our age-related genes are comparable across vervet and human datasets. More importantly, we found that

about half of the genes correlated with age in human and with known vervet orthologs were also identified as age-related genes in our vervet dataset, with the majority of these correlations occurring in the same direction. These findings support our identification of developmental genes in vervet hippocampus and suggest that relative expression across timepoints are conserved across species.

Finally, the increase in sample size in our eQTL analysis allowed us to identify a greater number of local and distant associations. We validated our previous finding implicating two lncRNAs within a locus associated with hippocampal volume (Jasinska et al., 2017), and identified two additional genes, *CHMP1B* and *RAB31*, whose expression levels are associated to SNPs within the hippocampal volume-linked region. These additional candidates can contribute to our understanding of this phenotype, as they have known human orthologs. Specifically, *RAB31* has been reported to play a role in the differentiation of neural progenitor cells into astrocytes, with overexpression resulting in enhanced differentiation, and silencing in a reduction of differentiation (Chua, Goh, and Tang, 2014). Reduced glial density as well as reduced hippocampal volume have been previously implicated in chronic stress (Rahman, Callaghan, Kerskens, Chattarji, and O'Mara, 2016) and major depressive disorder (Cotter, Mackay, Landau, Kerwin, and Everall, 2001). This finding suggests a possible shared pathway between these brain phenotypes, in which *RAB31* might play a role. Finally, our eQTL findings coupled with our identified developmental genes can provide additional insight into underlying mechanisms contributing to hippocampal developmental phenotypes.

4.4 METHODS

Data Collection

Total RNA was processed with Ribo-Zero Gold kit (Epicentre, WI) to remove ribosomal RNAs. Sequencing libraries were prepared using Illumina TruSeq RNA sample prep kit following manufacturer's protocol. After library preparation, amplified double-stranded cDNA was fragmented into 125 bp (Covaris-S2, Woburn, MA) DNA fragments, which were (200 ng) end-repaired to generate blunt ends with 5'-phosphates and 3'- hydroxyls and adapters ligated. The purified cDNA library products were evaluated using the Agilent Bioanalyzer (Santa Rosa, CA) and diluted to 10 nM for cluster generation in situ on the HiSeq paired-end flow cell using the CBot automated cluster generation system. All samples were multiplexed into a single pool in order to avoid batch effects (Auer and Doerge, 2010) and sequenced using an Illumina HiSeq 2500 sequencer (Illumina, San Diego, CA) across 2 lanes of 69bp-paired-end sequencing, corresponding to 3 samples per lane and yielding between 52 and 65 million reads per sample. Quality control was performed on base qualities and nucleotide composition of sequences.

RNA data processing

Alignment to the Chlorocebus sabeus reference annotation (NCBI release 100) was performed using the STAR (Dobin et al., 2013) spliced read aligner with default parameters. Additional QC was performed after the alignment to examine: the level of mismatch rate, mapping rate to the whole genome, repeats, chromosomes, key transcriptomic regions (exons, introns, UTRs, genes), insert sizes, AT/GC dropout, transcript coverage and GC bias. Between 83 and 91% (average 89.9%) of the reads mapped uniquely to the vervet genome. Total counts of read-fragments aligned to candidate gene regions were derived using HTSeq (Anders et al., 2015) program with

Chl. Sab (May 2014) NCBI annotation as a reference and used as a basis for the quantification of gene expression. Only uniquely mapped reads were used for subsequent analyses.

To identify the main sources of variation in our dataset, we obtained principal components (PC) from the top 1000 most variable genes across all samples. We found that PC1 (24.9% variance) differentiated animals by batch while PC2 (16.5% variance) separated animals by sex (Figure C-7A). We then corrected for batch effect and once again performed principal component analysis on the corrected data (Figure C-7B,C). Finally, we performed correlation analysis on the top 15 principal components and known covariates (Table C-5) and found that the top three PCs were strongly correlated with age category or sex ($p<0.05$), but no significant correlation occurred with age in days.

Age-related gene expression analysis

Age-related gene expression analysis was performed using the edgeR R package (Robinson et al., 2010), treating age as a continuous variable. Genes with cpm counts < 0.5 in less than 25% of the samples were removed. Counts were normalized using a trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010). Batch and sex were used in the generalized linear model with age in days as the variable of interest. Finally, a likelihood ratio test was performed to identify age-related genes. Significance threshold was set at FDR <0.05. Reported log2FC represents the log2 of the multiplicative effect of a single unit increase in age.

Weighted gene co-expression analysis

Weighted co-expression analysis was performed using the WGCNA R package. Read counts were normalized and log-transformed using the edgeR R package

(Robinson et al., 2010). Genes with less than 0.5 counts-per-million (cpm) in less than twenty-two samples were removed to reduce the likelihood of spurious results. After filtering, WGCNA was performed on a total of 19,994 genes. We then corrected for batch using the removeBatchEffect function from the limma R package (Gentleman et al., 2004). A minimum threshold of 30 genes was used during module construction. Similar modules were merged at a correlation threshold of r=0.75. Finally, module eigengenes were correlated with brain weight (g), age in days, body weight (kg), sex, age category and age category by sex and p-values were calculated using the corPvalueStudent function from the WGCNA package. Enrichment analysis was performed on the grey60 module, made up of 1,067 genes and the salmon module, comprised of 1,608 genes.

Comparison to other datasets

Age-related gene comparisons were performed using downloaded expression values from the BrainSeq Phase II project, human BrainSpan developmental data, and rhesus Allen Brain Atlas database. Processed expression counts were downloaded from BrainSeq and genes with cpm < 0.5 in less than 25% of samples were removed. Normalization was then performed using edgeR to maintain consistency with our vervet age-related analysis. For human and rhesus developmental datasets, age-related genes were identified using the edgeR generalized linear model approach (McCarthy et al., 2012), with sex included in the model for human datasets. Additionally, for the BrainSeq dataset the first five principal components (obtained from genotype information) were also included to account for ethnicity. The intersection of vervet and human/rhesus age-related genes was obtained, and rank correlation values were calculated and plotted using signed log10 p-values.

Spearman rank correlation comparisons of normalized expression was also performed using the top 1000 vervet age-related genes with known human orthologs at corresponding time points (Tables C-2,C-3). Vervet expression data was normalized using the edgeR package and RPKM counts were calculated. Finally, since rhesus data included only males, the vervet comparison was also limited to males.

Weighted gene co-expression analysis was performed on the human BrainSpan dataset using the same module construction thresholds as described above. Analysis was limited to human genes with known vervet orthologs. After module construction, module eigengenes were correlated with age, age category, sex and a combination of sex + age category. Overlap significance of human BrainSpan and vervet age-related modules was performed using a permutation test (n=10,000). Overlap was defined as significant at a threshold of $p < 0.05$.

eQTL Analysis

Genotype information was available for 29 out of the 32 new samples, thus eQTL analysis was performed on 87 hippocampal samples. Lowly expressed genes, defined as reads with a zero count in more than 10% of the samples and a combined mean less than 1, were excluded from our analysis. Counts per million (cpm) were then obtained using the edgeR package and a quantile transformation was applied across all 27,425 remaining genes. Association analysis between expression and genotype data was performed using the linear mixed model package EMMAX (H. M. Kang et al., 2010). The first fifteen principal components were included in our model as covariates, which accounted for 60% of observed total variance. A kinship matrix was also calculated and included to account for the high degree of relatedness in our samples. Strict Bonferroni thresholds for local and distant eQTL associations were

calculated as $4.6x10^{-09}$ and $3.67x10^{-12}$, respectively. Finally, FDR results were obtained separately for local and distant associations using the hierarchical error control R package, TreeQTL. The distance parameter for local associations was set within 1MB from gene start and stop positions and anything greater than that was classified as distant. Comparison of local eQTL results with GTEx was performed using our Bonferroni threshold and an FDR threshold of 0.05, as calculated by TreeQTL (C. Peterson, Bogomolov, Benjamini, and Sabatti, 2015). For both thresholds, we observed all of our eGenes had a p-value < 0.05 in the corresponding GTEx results.

Chapter 5

Characterization of epigenetic marks in non-human primates

5.1 Introduction

The development of high-throughput technologies has facilitated the sequencing of thousands of individuals (Metzker, 2010). This has led to the identification of numerous genetic markers associated with various disease phenotypes (Welter et al., 2014). Despite the implications of such findings, functionally annotating these variants has been challenging due to their location within noncoding regions of the genome. Expression quantitative trait locus (eQTL) studies have been successful in finding association of some of these variants to changes in gene expression (Nicolae et al., 2010), however, a better understanding of the underlying mechanism by which these variants regulate gene expression is required. This sparked an interest in identifying regulatory regions of the genome and their impact on gene expression (Maurano et al., 2012).

Projects like ENCODE (ENCODE Project Consortium, 2012) and Roadmap Epigenomics (Bernstein et al., 2010) have succeeded in identifying chromatin modifications across various tissues and cell types. Studies incorporating regulatory features have identified tissue specificity in epigenomic marks overlapping risk associated variants (Trynka et al., 2013) and regulating gene expression (Heintzman et al., 2009). These studies not only reinforce the overall relevance of epigenomic resources, but the importance of incorporating in a tissue-specific manner.

The usefulness of model organisms such as the green African vervet monkey (*Chlorocebus sabeus*) has been previously established (Jasinska et al., 2013),

however, limited genomic information about epigenetic markers poses a challenge to their use in the characterization and refining of GWAS loci. While tools allowing for the conversion of genetic coordinates across species exist, corresponding genomic regions at the sequence level do not always correspond to corresponding chromatin markers. For example, we generated vervet liver H3K27ac peaks through liftover of human H3K27ac liver peak calls and compared them with a known H3K27ac vervet liver dataset (data not shown), and observed that only 49% of known peaks overlapped with liftover results. More importantly, 38% of peaks generated by liftover were identified as false peaks. The low true positive rate can perhaps be addressed by pooling results from multiple conversion algorithms, or by imputation methods (Ernst and Kellis, 2015). However, the introduction of large numbers of false positives can undermine our ability to draw relevant biological conclusions and has not previously been addressed.

Thus, we set out to improve the prediction of vervet histone marks obtained by the application of genome coordinate conversion tools like liftover. Through the use of machine learning algorithms and predictive variables compiled from factors such as distance from transcription start site, average GC percent and peak length, we attempt to differentiate actual peaks from false peaks. We then evaluate the efficacy of our model using known enhancer (H3K27ac) and promoter (H3K4me3) marks obtained from a vervet liver dataset.

5.2 Results

Numerous machine learning algorithms have been implemented in the field of genetic research (Larrañaga et al., 2006). We selected random forest (RF) and support vector machine (SVM) algorithms due to their run-time efficiency and ability for parallelization.

Model building and selection of best predictive features

To determine whether our predictive variables share similar patterns across species, we decided to test the use of human and vervet datasets as training sets in our machine learning models (Figure 5-1). For our human training set, human liver H3K27ac and H3K4me3 peak calls were downloaded from the Roadmap Epigenomics database (Bernstein et al., 2010). Simulated peaks were generated to serve as false peak calls. Simulated and known peak calls were then combined and a balanced subset of this dataset was then selected to use as our training set.

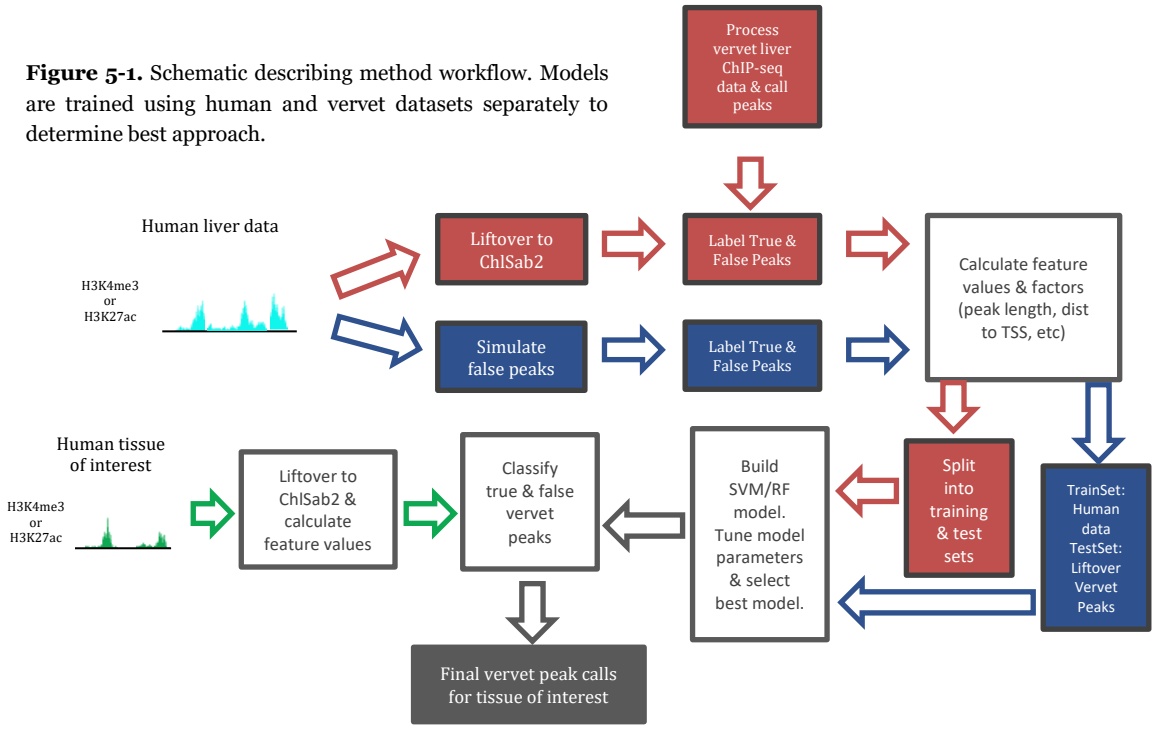


Figure 5-1. Schematic describing method workflow. Models are trained using human and vervet datasets separately to determine best approach.

Our vervet dataset was generated by applying the liftover algorithm as implemented by the rtracklayer R package (M. Lawrence, Gentleman, and Carey, 2009), to human liver H3K27ac and H3K4me3 peak calls. Human liver datasets were selected due to the availability of vervet liver ChIP-seq data which allow validation of liftover results and thus provide labelled vervet peaks to be used in our supervised learning approach. To obtain vervet liver H3K27ac and H3K4me3 peak calls we used published vervet liver ChIP-seq data included in Villar, et al (2015) (Villar et al., 2015). Peaks were called using the MACS2 peak caller (Yong Zhang et al., 2008) using the same p-value and q-value thresholds utilized by the Roadmap Epigenomics consortium (Kundaje et al., 2015). These peak calls were then used to differentiate between positive and negative peak calls in our generated vervet liftover peaks.

Once true and false peaks were labelled in our human and vervet datasets, we calculated features to be included in our model. Model features were selected based on evidence suggesting correlation of genetic features to conservation of enhancer (H3K27ac) or promoter (H3K4me3) marks across species (Villar et al., 2015). Histone mark features calculated included distance to transcription start site (TSS), GC content percentage and peak length. Additionally, due to the common practice of discretizing continuous variables in machine learning (Chmielewski and Grzymala-Busse, 1996), we converted our continuous variables into categorical variables by binning them according to value (see Methods). SVM and RF models were then trained and tested using one of three values: actual parameter values (values), discretized parameter values (factors), and combined values and factors (all).

Random forest and svm-linear algorithms were run using default parameter values, with the exception of the number of trees which was set to 100 (Probst and

Boulesteix, 2018). The use of vervet or human training datasets were tested along with varying feature values described above (i.e. values, factors, all). Regardless of training set used, our vervet dataset was split into a training and test sets. RF and SVM functions were then trained using either vervet training set or a subset of our human dataset and tested using the vervet test set.

Training Data	Algorithm	Feature	Accuracy	Sensitivity	Specificity
Human	SVM	Values	60.32%	0.55%	98.84%
		Factors	59.68%	0.05%	98.10%
		All	59.68%	0.05%	98.10%
	RF	Values	59.41%	1.37%	96.80%
		Factors	59.56%	0.28%	97.76%
		All	59.53%	0.26%	97.73%
Vervet	SVM	Values	69.58%	54.62%	78.21%
		Factors	66.08%	66.76%	65.65%
		All	69.12%	58.85%	75.74%
	RF	Values	67.04%	53.13%	76.00%
		Factors	65.80%	30.15%	88.78%
		All	70.05%	56.93%	78.59%

Table 5-1. H3K27ac model results. Results are broken down by training data type and data features included in the model. Red indicates predictive variables selected for further model tuning.

As expected, models trained on the vervet data outperformed human trained models for each histone mark and algorithm type (Tables 5-1, 5-2). Surprisingly, our SVM models yielded the same results when using factor and combined data values suggesting factors played a bigger role in building the hyperplane despite the inclusion of continuous variables. Overall, the random forest algorithm yielded higher accuracy for most features tested and thus was the selected machine learning approach for our final model. For feature selection we prioritized specificity to avoid losing relevant peak calls, while also ensuring minimal loss of accuracy and

sensitivity. Thus, for H3K27ac and H3K4me3 histone marks, we selected combined factor + values (all) and values only (values), respectively (Tables 5-1, 5-2).

Training Data	Algorithm	Feature	Accuracy	Sensitivity	Specificity
Human	SVM	Values	27.66%	0.04%	99.68%
		Factors	27.61%	0.09%	99.35%
		All	27.61%	0.09%	99.35%
	RF	Values	27.65%	0.39%	98.71%
		Factors	27.61%	0.09%	99.35%
		All	27.62%	0.20%	99.12%
Vervet	SVM	Values	78.61%	92.70%	41.90%
		Factors	79.55%	98.03%	31.39%
		All	79.55%	98.03%	31.39%
	RF	Values	81.71%	90.42%	59.01%
		Factors	78.73%	95.72%	34.42%
		All	82.90%	92.49%	57.89%

Table 5-2. H3K4me3 model results. Results are broken down by training data type and data features included in the model. Red indicates predictive variables selected for further model tuning.

Model parameter tuning

After determining the best features and machine learning algorithm for each histone mark, we attempted to tune our model parameters in an effort to improve model prediction. Additionally, since our current implementation of the SVM algorithm uses a linear kernel, which works best on data that can be separated linearly, we also decided to test an SVM model with a non-linear (radial) kernel. Given the computational time it takes to train such a model, using it to identify preferential features was not feasible.

Our random forest model was run on mtry values of 1 to 20, while maintaining the number of trees at 100. Meanwhile, the SVM radial algorithm was implemented using a random grid search to determine best cost (C) value. The accuracy metric was used to select optimal values for random forest and SVM-radial models (Figure

5-2). After tuning, our SVM-radial model failed to offer any improvement over our random forest model (Table 5-3).

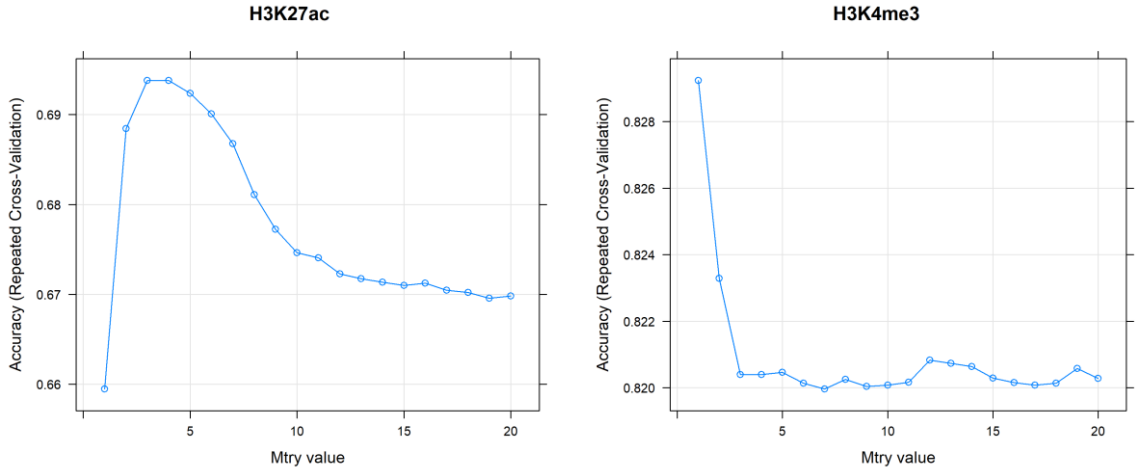


Figure 5-2. Tuning of random forest mtry parameter. Tuning was performed using best performing predictive variables. For H3K27ac random forest model was optimized at an mtry value of 3, while H3K4me3 model demonstrated the highest accuracy at an mtry value of 1.

Histone Mark	Algorithm	Accuracy	Sensitivity	Specificity
H3K27ac	RF	70.11%	56.79%	78.69%
	SVM-radial	70.85%	55.90%	78.70%
H3K4me3	RF	83.02%	91.20%	61.69%
	SVM-radial	82.58%	91.38%	59.65%

Table 5-3. Performance summary of the best RF and SVM-radial models after parameter tuning. RF outperforms SVM-radial by a slight margin for both histone marks.

Classification of vervet peaks in three brain tissues

We applied our optimized RF models to vervet peaks lifted over from three human brain datasets: anterior caudate, dorsolateral prefrontal cortex (DLPFC) and hippocampus middle. We then used the predicted true peaks to test whether our multi-tissue eQTL results, at Bonferroni or FDR thresholds, from chapter 3 were enriched in tissue or brain specific peak calls for each histone mark. For each brain region we classified peaks as tissue specific if it did not overlap any peak calls in

neither of the other two brain regions nor vervet liver. Additionally, due to the possibility of tissue specific peaks being false positives, we defined brain specific peaks as peaks present in at least two brain regions but not in vervet liver. We tested for enrichment of eQTLs from each brain tissue individually and found significant enrichment of caudate FDR-corrected eQTLs in caudate specific peaks for both

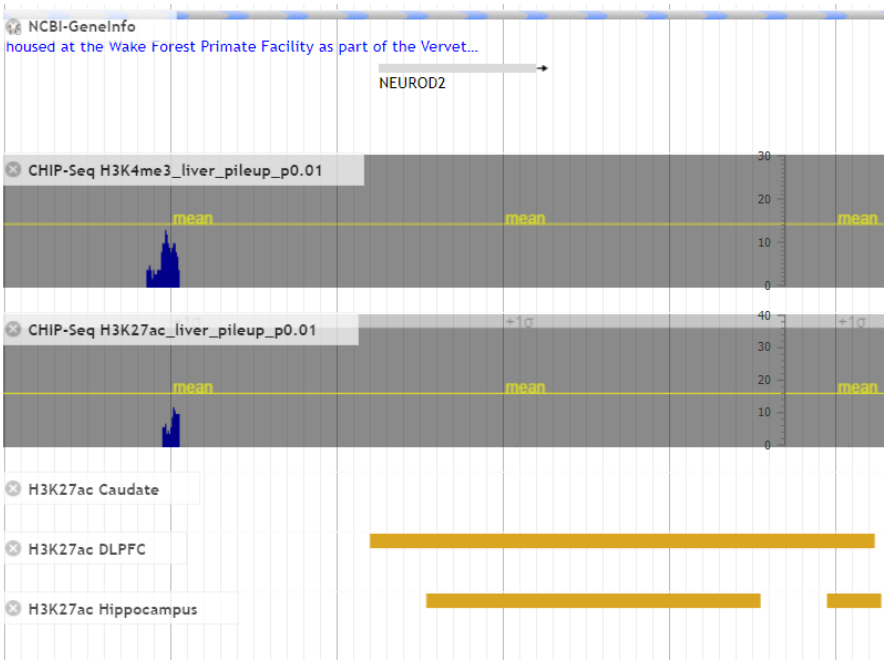


Figure 5-3. Example of a brain specific H3K27ac peak within *NEUROD2* gene. Top section illustrates location of *NEUROD2*, a gene enriched in brain tissues. Blue bars illustrate H3K27ac peaks in liver dataset. Orange bars at the bottom represent the peak span in dorsolateral prefrontal cortex (DLPFC) and hippocampus. Image was generated using the vervet genome browser (Ramensky, et al. Unpublished; https://coppolalab.ucla.edu/vgb/home).

histone mark ($p < 0.05$, hypergeometric test). Next, we analyzed whether the combined set of brain eQTLs was enriched in brain specific peaks and found nominal enrichment of our brain eQTLs (FDR< 0.05) in H3K27ac brain specific peaks ($p = 0.0558$, hypergeometric test).

Finally, we explored whether brain specific peaks fell near genes highly expressed in brain tissues using the top 12 genes with the highest level of enriched expression as listed at the Human Protein Atlas: *OPALIN*, *GFAP*, *OMG*, *OLIG2*, *GRIN1*, *NEUROD6*, *SLC17A7*, *CREG2*, *NEUROD2*, *C1orf61*, *ZDHHC22* and *KCNJ6* (Uhlen et al., 2015) (https://www.proteinatlas.org/humanproteome/tissue/brain). We observed H3K27ac and H3K4me3 brain specific peaks occurring within 7 and 8 of these genes, respectively. An example, using H3K27ac, is presented in Figure 5-3 for *NEUROD2*, a neuronal differentiation gene.

5.3 Discussion

We presented a method to improve the accuracy of vervet histone peak predictions generated by lifting over histone peaks from other species such as human. Our approach makes use of the reported relationship of certain features in relation to conservation of histone marks (Villar et al., 2015) and uses them as predictive variables to identify true peak calls. Limitations of our method exist, as evidenced by the lower accuracy in predicting H3K27ac peaks. We hypothesize this may be due to the lower contribution of our selected features in relation to conservation depth H3K4me3 marks when compared to H3K27ac (Villar et al., 2015). However, despite the lower accuracy observed, we still retain biologically relevant tissue specific information as observed in our brain specific peak calls.

Additionally, we demonstrated the effectiveness of random forest models in differentiating true peaks from false peaks. Random forest models have been used in gene expression studies (Kursa, 2014), while machine learning approaches characterizing chromatin modifications have focused on the use of hidden markov

models (HMM) (Ernst and Kellis, 2010; Larson and Yuan, 2010; Won, Chepelev, Ren, and Wang, 2008). However, we found no studies focused on improving the conversion of genomic features across species despite the large number of studies incorporating current methods (Kuhn, Haussler, and Kent, 2013; Zhao et al., 2014). Our method can be expanded to take advantage of the multiple conversion tools available, plus it can build on these tools through the use of imputation methods (Ernst and Kellis, 2015).

However, in order to introduce additional predicted peak calls without introducing a large number of false positives, we would need to ensure high specificity and sensitivity values. Thus, additional prediction variables will be needed to improve upon the accuracy of our model. One such variable, peak intensity, might be able to be carried over from original human histone values, however, it would not contribute much when differentiating between two peak calls corresponding to the same original human peak. Incorporating additional information such as gene expression values corresponding to the nearest genes may be a feasible predictive variable worth testing due to the high degree of correlation between gene expression and histone modification levels (Karlić, Chung, Lasserre, Vlahovicek, and Vingron, 2010).

Finally, like most predictive methods, our method would greatly benefit from additional vervet ChIP-seq data sets to further improve our model and validate our results. Despite the improvements that can still be made, our approach provides a good starting point to ensure accuracy of predicted peak calls generated from liftover. Due to the availability of epigenomic resources from a number of species, our approach may prove beneficial to researchers without the necessary resources to generate their own epigenomic datasets.

5.4 Methods

Chromatin modification peak calls from vervet liver data

ChIP-Seq files for vervet liver H3K27ac and H3K4me3 histone modifications were downloaded from the Villar, et al manuscript (Villar et al., 2015). Vervet peak calls generated by Villar, et al were not used due to the method by which they were obtained. Namely, Villar, et al aligned vervet ChIP-seq data to the rhesus macaque genome and applied liftover to obtain corresponding vervet coordinates. We observe that through the use of this method, vervet peaks were limited to autosomal chromosomes 1-20 and sex chromosome X, effectively ignoring vervet chromosomes 21-29.

ChIP-seq files were aligned to the Chlorocebus sabaeus v1.0 reference genome using the bwa aligner with default parameters (Heng Li and Durbin, 2010). Low quality reads or non-uniquely mapped reads were removed using samtools option -q 1 (H. Li et al., 2009). Peaks were called using MACS2 peak caller (Yong Zhang et al., 2008) using the -nomodel and -broad options, with -p threshold set to 0.01. Peaks with a p-value less than 0.01 within gapped peak files were retained as valid peak calls. Gapped peak files contain broad peak calls ($p < 0.1$) with at least one overlapping strong narrow peak call ($p < 0.01$).

Vervet training dataset

Human liver H3K27ac and H3K4me3 gapped peak calls were downloaded from the epigenomics roadmap data repository (Bernstein et al., 2010). Liftover of human peak calls was implemented using the rtracklayer R package (M. Lawrence et al., 2009) and hg19ToChlSab2 chain file downloaded from the UCSC browser

(Rosenbloom et al., 2015). The rtracklayer function was selected over the command line script due to an observed increase in true H3K27ac vervet liver peak calls from human data (data not shown), though the reason for this difference is not known. The liftOver function implemented in rtracklayer produced smaller blocks often a few base pairs apart; thus, blocks were combined using a gap value of 100 added to start and end coordinates, and overlapping regions were consolidated. Once all overlapping blocks were combined, start and end coordinates were adjusted to account for gap value. Feature values were then generated and true peaks were classified based on overlap of known vervet liver peaks.

Human training dataset

The human training set was composed of known human peak calls and simulated human peak calls. Simulation was performed by randomly sampling chromosome values and chromosome start position. We ensured randomly sampled start positions were within chromosome boundaries by including known lengths of human chromosomes. Peak end coordinates were determined by random sampling the width ranges observed in human peaks at a qvalues > 1.30 corresponding to an FDR > 0.05. Simulated false peaks were then combined with actual peak calls and feature values were calculated as described. Simulated peaks with missing GC content information were excluded from our final human dataset. Finally, our human dataset was subsampled to better match the size of the vervet training set.

Model parameters

Human and vervet transcription start site (TSS) and GC percent data was downloaded from the UCSC browser (Rosenbloom et al., 2015), using genome assembly hg19 and chlSab2 for human and vervet, respectively. GC content

information was provided in 5-bp regions combined across 1024 total regions. For each human or vervet peak, average GC percent values were calculated for the entire region spanning each histone mark. Distance to the nearest transcription start site was determined using the distanceToNearest function from the GenomicRanges R-package (Michael Lawrence et al., 2013). Length of peaks was determined using difference between peak start and end coordinates.

Feature values were converted to factors by binning actual values for GC percentage, peak length and distance to nearest TSS. For TSS distance, values were binned into the following ranges: distance = 0, 0 < distance <= 10000, 10000 < distance <= 30000, and distance > 30000. Similarly, peak lengths were categorized as length <= 2000, 2000 < length <= 4000, 4000 < length <= 7000, 7000 < length <= 15000, and length > 15000. Finally, average GC percent values were factored into the following groups: less than 20, 20 to 30, 30 to 40, 40 to 50, 50 to 60 and greater than 60.

Machine learning implementation

Machine learning algorithms were implemented using the caret R-package. Random forest and svm algorithms were run using default parameters, except for nTree which was reduced to 100. For SVM implementation, default cost (C) value is set to 1, while for RF, default mtry value for classification is the sqrt(p), where p is the number of variables. Additionally, scale and center pre-processing options were applied to feature variables. Finally, accuracy calculations were determined through repeated cross-validation with repeats set to a value of three within the trainControl function from the caret package.

Model parameter tuning

Random forest parameter tuning was implemented using the tuneGrid parameter to include mtry values from 1 to 20. For our svm-radial model, random tuning was permitted and tuneLength was set to ten to allow testing of ten different cost values (C).

Peak prediction using best model

For vervet peak prediction, epigenomic roadmap human peak calls from anterior caudate, dorsolateral prefrontal cortex and hippocampal mid were used to correspond with vervet eQTL results obtained from caudate, Brodmann's area 46 and hippocampus. We applied our best model, identified as our random forest model with mtry values of 3 and 1 for H3K27ac and H3K4me3 marks, respectively.

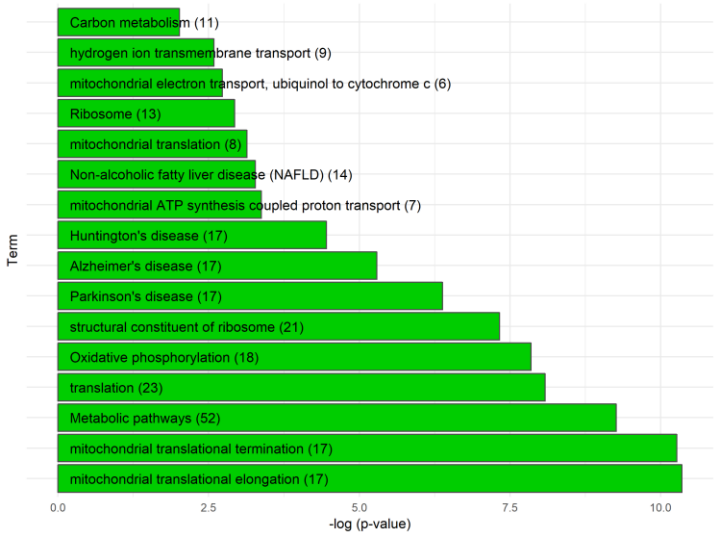
Classification of vervet peaks and eQTL enrichment in three brain regions

Enrichment of eQTLs, at Bonferroni or FDR thresholds, was determined by focusing on two different peak types: tissue specific and brain specific peak calls. To obtain tissue specific peak calls, we defined a factor variable for each tissue comparison, with values of 1 and 0 representing overlap and no overlap in comparison tissue, respectively. Hypergeometric tests were then performed to test for enrichment of tissue specific eQTL SNPs in tissue specific histone peaks from the same tissue type. Next, we labelled peaks with present in at least one other brain region but not present in liver data as brain specific, and performed enrichment analysis using the combined set of eQTLs identified in brain tissues.

Appendix A

Supplementary Figures for Chapter 2

A



B

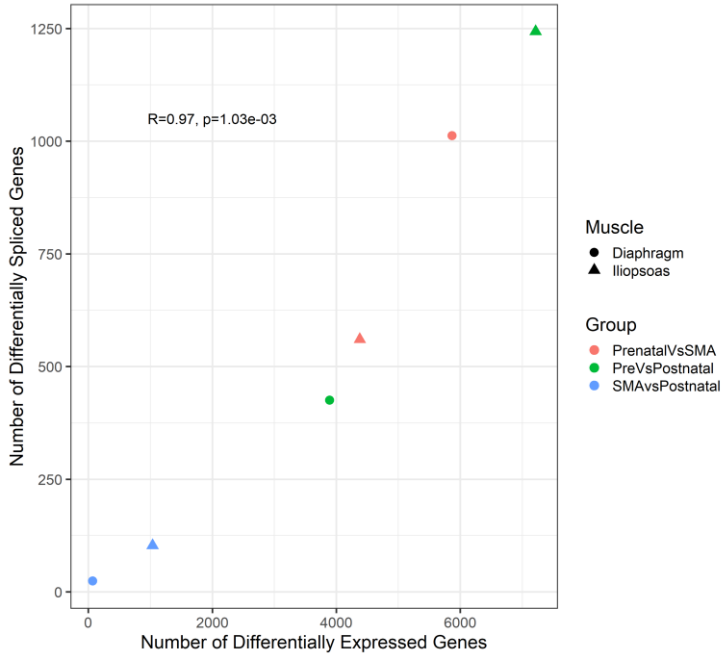


Figure A-1. (A) Functional enrichment results for down-regulated DE genes in iliopsoas SMA cases and prenatal controls vs postnatal controls. (B) Relationship between number of differentially expressed genes and differentially spliced events in each comparison group and tissue type.

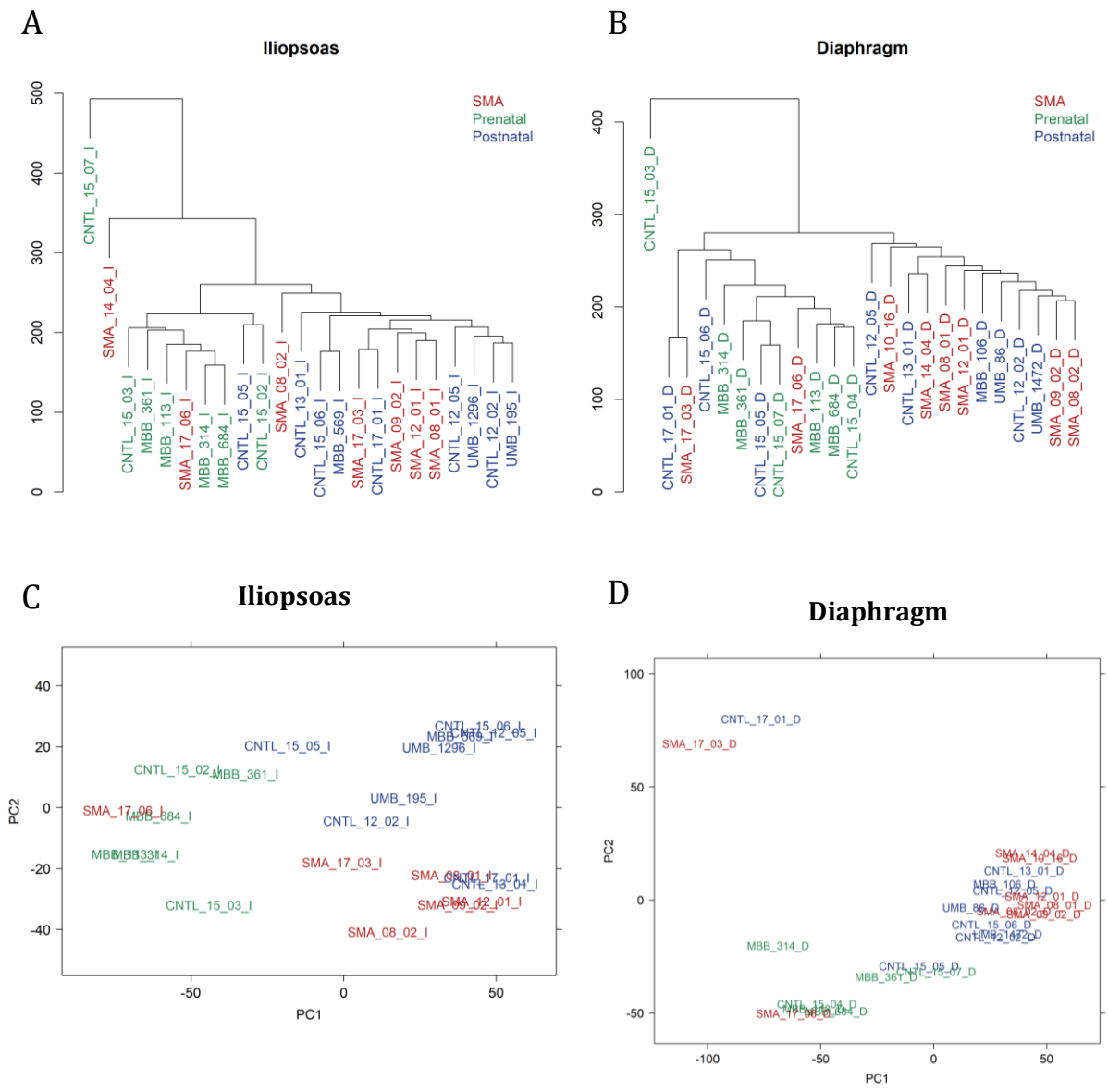


Figure A-2. Hierarchical clustering and principal component analyses of diaphragm and iliopsoas samples. Samples in plots are color coded by sample type with SMA cases, prenatal and postnatal controls displayed in red, green and blue, respectively. (A) Hierarchical clustering of Iliopsoas identified two outlier samples: SMA_14_04 and CNTL_15_07. Similarly, clustering shows the only prenatal SMA case (SMA_17_06) and prematurely born postnatal control (CNTL_15_05) clustering with prenatal controls. (B) Hierarchical clustering of diaphragm samples identified one outlier: CNTL_15_03. Similar to iliopsoas, prenatal SMA case (SMA_17_06) and premature postnatal control (CNTL_15_05) once again cluster with prenatal controls. (C) PC2 as a function of PC1 for iliopsoas, showing clustering of prenatal SMA sample (SMA_17_06) and premature postnatal control (CNTL_15_05) clustering with prenatal controls. (D) PCA identified two additional outliers in diaphragm: CNTL_17_01 and SMA_17_03. Additionally, PCA once again showed clustering of prenatal SMA case and prematurely born postnatal control (SMA_17_06 and CNTL_15_05) clustering with prenatal controls.

Appendix B

Supplementary Figures for Chapter 3

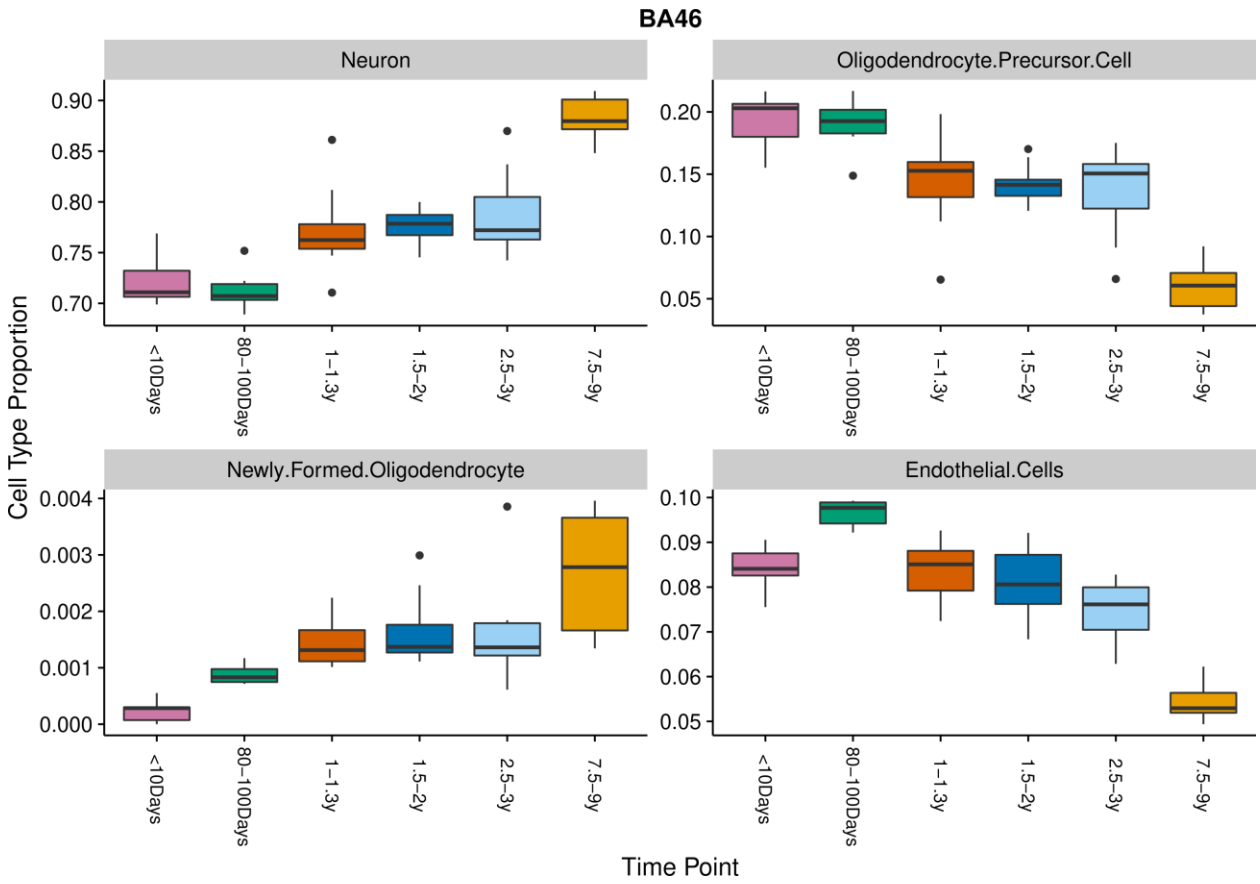


Figure B-1. Distribution of cell type composition by age for vervet BA46. Deconvolution analysis was applied to vervet BA46. The distribution of cell type proportions is plotted for six vervet age groups.

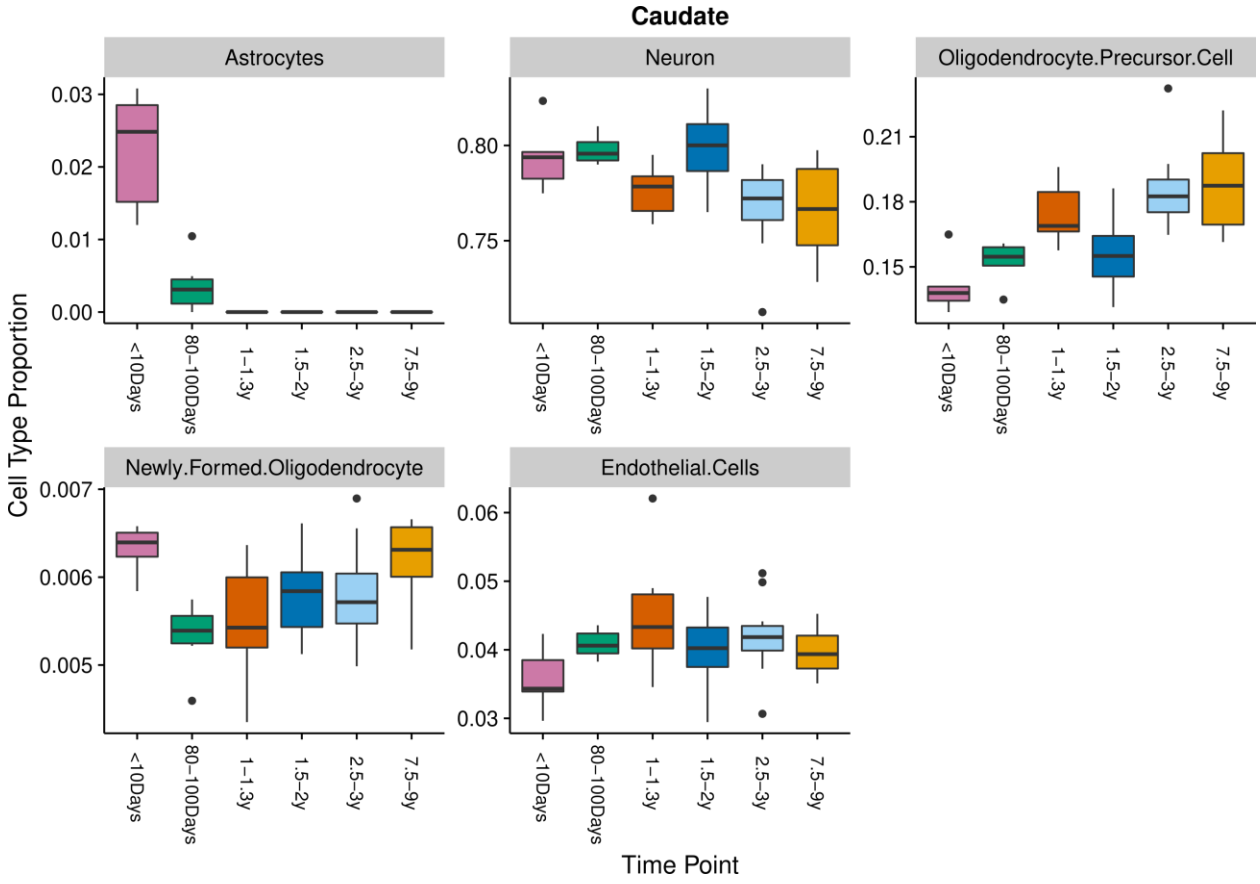


Figure B-2. Distribution of cell type composition by age for vervet caudate. Deconvolution analysis was applied to vervet caudate. The distribution of cell type proportions is plotted for six vervet age groups.

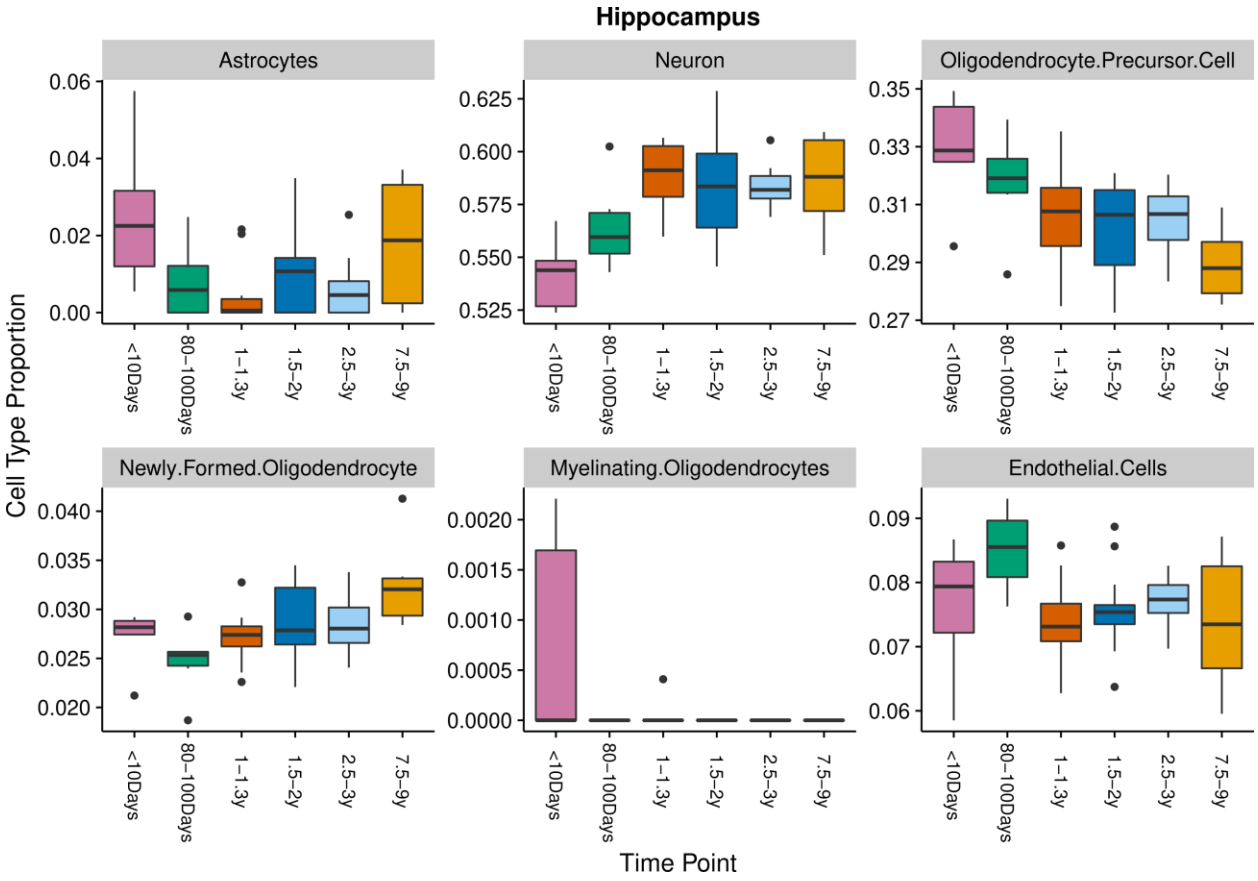


Figure B-3. Distribution of cell type composition by age for vervet hippocampus. Deconvolution analysis was applied to vervet hippocampus. The distribution of cell type proportions is plotted for six vervet age groups.

Appendix C

Supplementary Tables and Figures for Chapter 4

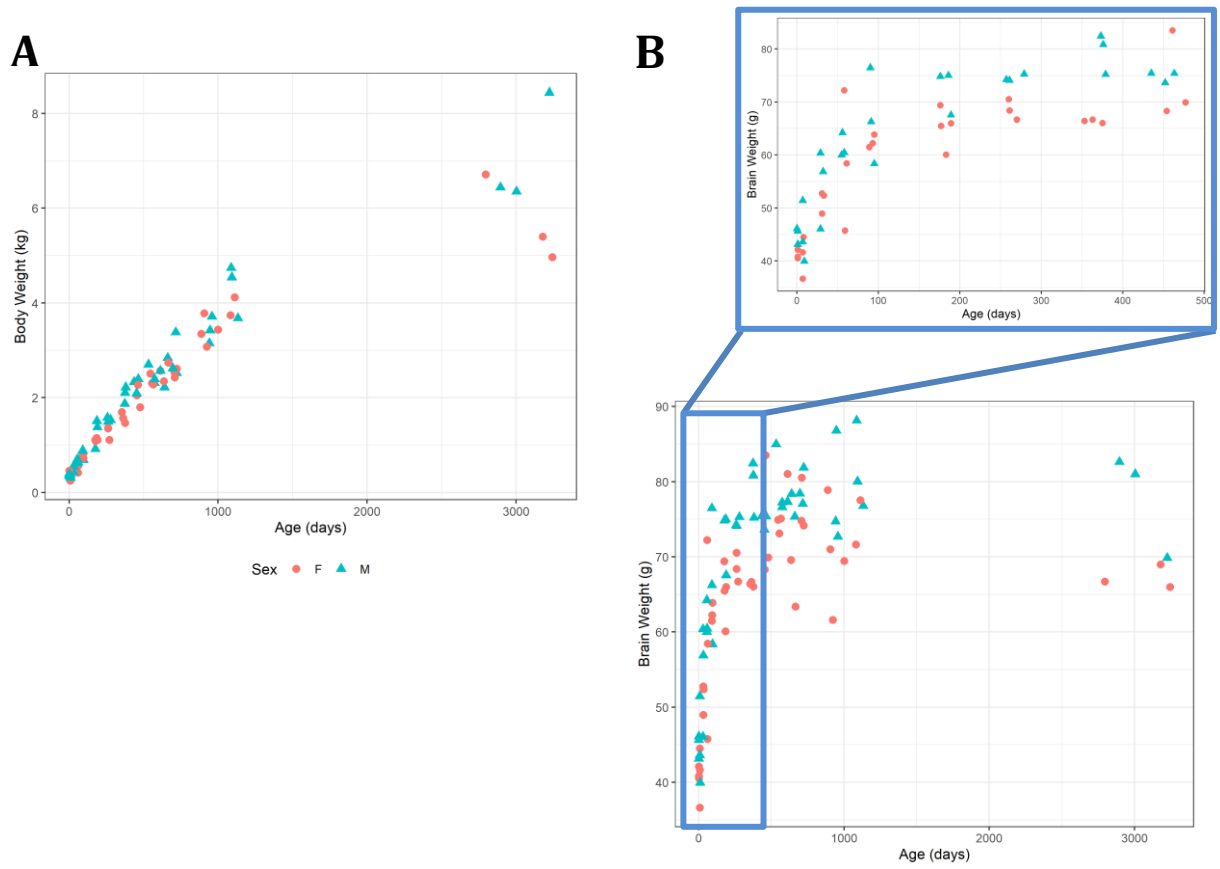


Figure C-1. Brain and body weight as a function of age in days. (A) Body weight (in kg) as a function of age (in days). (B) Brain weight (in g) as a function of age (in days). Inset: in animals aged under 500 days brain weight rises rapidly and begins to plateau after the 100-day mark. Males and females are indicated by blue triangles and red circles, respectively.

Sample ID	Sex	Age (days)	Weight (kg)	BrainWt (g)	Batch
2009098	F	7	0.34	41.66	1
2010056	F	7	0.25	36.65	1
2010069	M	7	0.43	51.44	1
2010083	M	7	0.36	43.61	1
2010014	F	8	0.34	44.53	1
2009055	F	89	0.77	61.51	1
2010023	M	90	0.89	76.47	1
2010024	M	91	0.88	66.26	1
2010018	F	93	0.81	62.23	1
2010016	M	95	0.69	58.37	1
2010019	F	95	0.72	63.88	1
2008138	F	353	1.7	66.43	1
2009052	F	363	1.57	66.69	1
2008148	M	373	1.87	82.44	1
2008143	F	375	1.47	66.04	1
2008122	M	376	2.1	80.83	1
2009057	M	379	2.22	75.2	1
2008036	M	435	2.33	75.44	1
2008102	M	452	2.09	73.62	1
2008064	F	454	2.05	68.3	1
2008093	F	461	2.27	83.54	1
2008089	M	463	2.39	75.4	1
2008060	F	477	1.8	69.93	1
2008010	M	533	2.7	84.99	1
2008012	F	545	2.51	74.93	1
2008090	F	555	2.31	73.11	1
2008066	F	564	2.28	75.1	1
2008054	M	573	2.31	77.23	1
2008101	M	573	2.39	76.62	1
2008014	F	611	2.57	81.02	1
2008017	M	611	2.57	77.34	1
2008147	F	635	2.35	69.59	1
2008023	M	639	2.22	78.37	1
2008007	M	661	2.84	75.35	1
2008141	F	665	2.74	63.38	1
2008095	M	695	2.62	78.4	1
2008074	F	707	2.54	74.82	1
2008114	F	709	2.43	80.53	1
2008021	M	715	3.38	77.05	1
2008022	M	721	2.52	81.87	1
2008052	F	722	2.61	74.18	1
2007047	F	887	3.35	78.87	1
2007043	F	906	3.78	71.01	1
2007044	F	923	3.08	61.61	1
2007035	M	941	3.15	74.72	1
2007041	M	946	3.42	86.8	1
2007023	M	958	3.71	72.67	1
2007006	F	1000	3.44	69.47	1
2007032	F	1082	3.74	71.65	1
2007016	M	1087	4.74	88.14	1
2007002	M	1092	4.54	80.03	1
2007031	F	1111	4.12	77.56	1
2007020	M	1132	3.68	76.77	1

2003104	F	2795	6.71	66.71	1
2002090	M	2894	6.44	82.62	1
2002024	M	3003	6.35	81	1
2002079	F	3179	5.4	69	1
2002060	M	3223	8.44	69.87	1
2002053	F	3244	4.97	66	1
2010084	M	0	0.38	46.1	2
2010004	F	1	0.36	40.83	2
2010055	F	1	0.36	42.13	2
2010059	M	1	0.39	43.14	2
2010072	F	1	0.46	40.54	2
2011020	M	1	0.34	45.62	2
2011017	M	9	0.31	39.95	2
2010052	M	29	0.55	60.38	2
2010071	M	29	0.43	46.05	2
2010009	F	31	0.42	52.75	2
2010045	F	31	0.44	48.97	2
2009088	M	32	0.51	56.87	2
2010051	F	33	0.47	52.39	2
2010042	M	55	0.7	59.99	2
2010040	M	56	0.62	64.23	2
2010032	F	58	0.69	72.23	2
2010053	M	58	0.64	60.46	2
2009078	F	59	0.42	45.77	2
2010033	F	61	0.57	58.46	2
2009077	M	176	0.92	74.83	2
2009092	F	176	1.11	69.42	2
2009086	F	177	1.09	65.5	2
2009085	F	183	1.15	60.08	2
2009007	M	186	1.51	75.01	2
2009075	F	189	1.11	65.99	2
2010102	M	189	1.38	67.56	2
2009050	M	257	1.58	74.24	2
2009091	F	260	1.36	70.54	2
2009048	F	261	1.35	68.42	2
2009064	M	261	1.49	74.11	2
2009060	F	270	1.11	66.71	2
2009084	M	279	1.53	75.29	2

Table C-1. Summary of vervet hippocampus samples

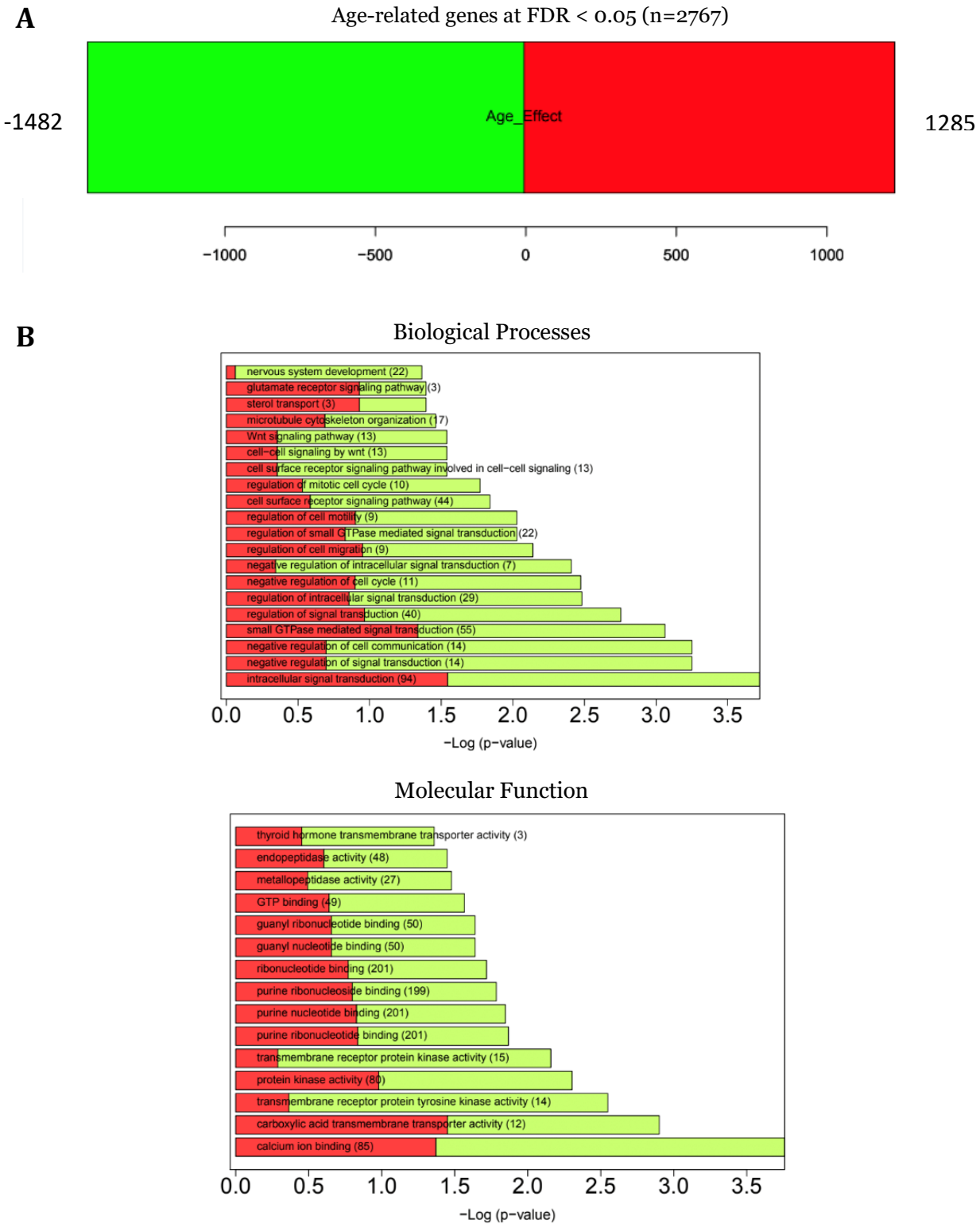


Figure C-2. (A) Number of differentially expressed age-related genes using N=91 animals. (B) Gene ontology results for genes positively and negatively correlated with age, represented by red and green, respectively.

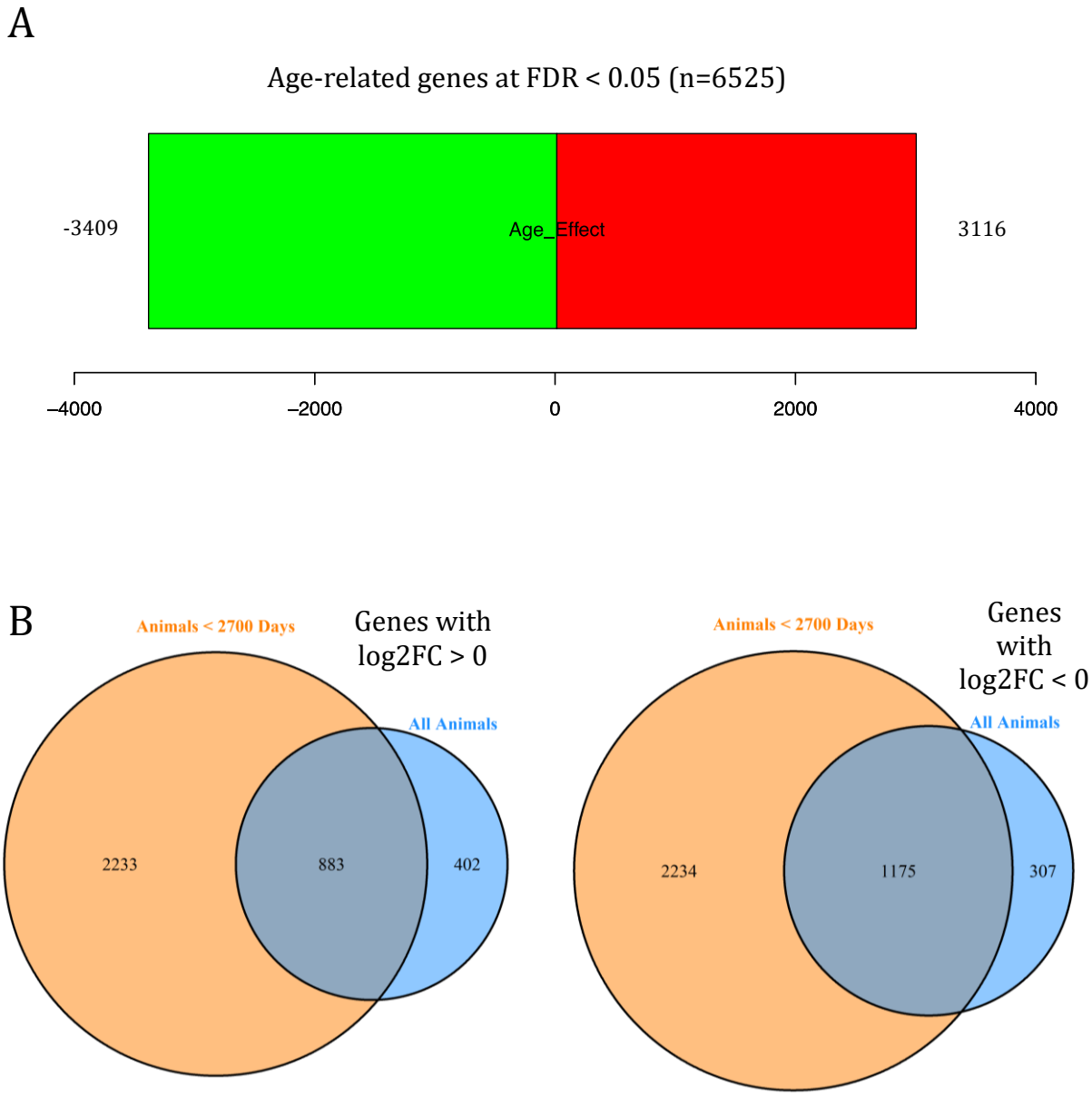


Figure C-3. (A) Number of positively correlated (red) and negatively correlated (green) age-related genes after excluding six oldest animals. (B) Comparison of age-related expression results with and without six oldest animals.

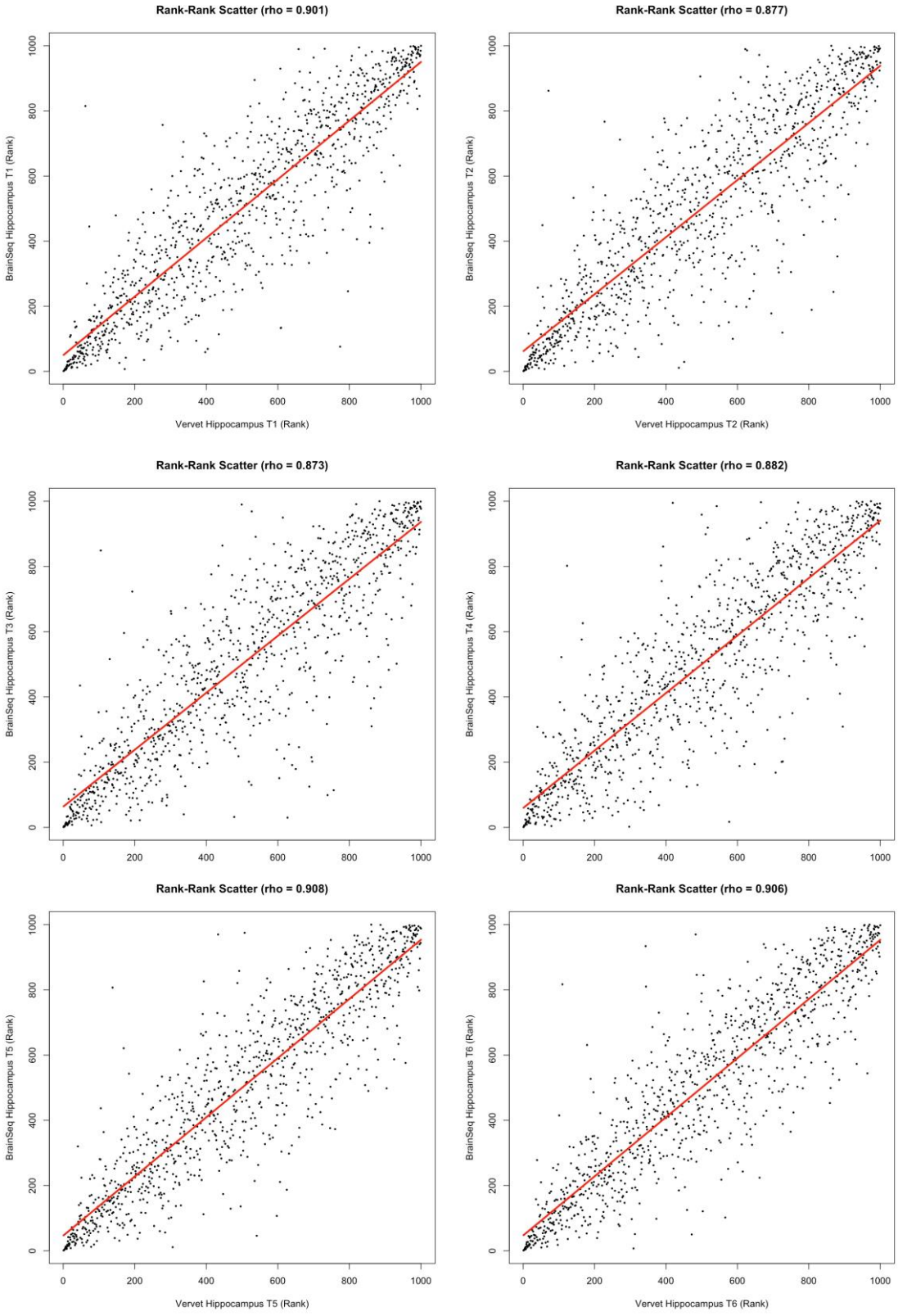


Figure C-4. Rank-Rank plots for human BrainSeq and vervet hippocampal gene expression comparison across six time points, using the top 1,000 vervet age-related genes.

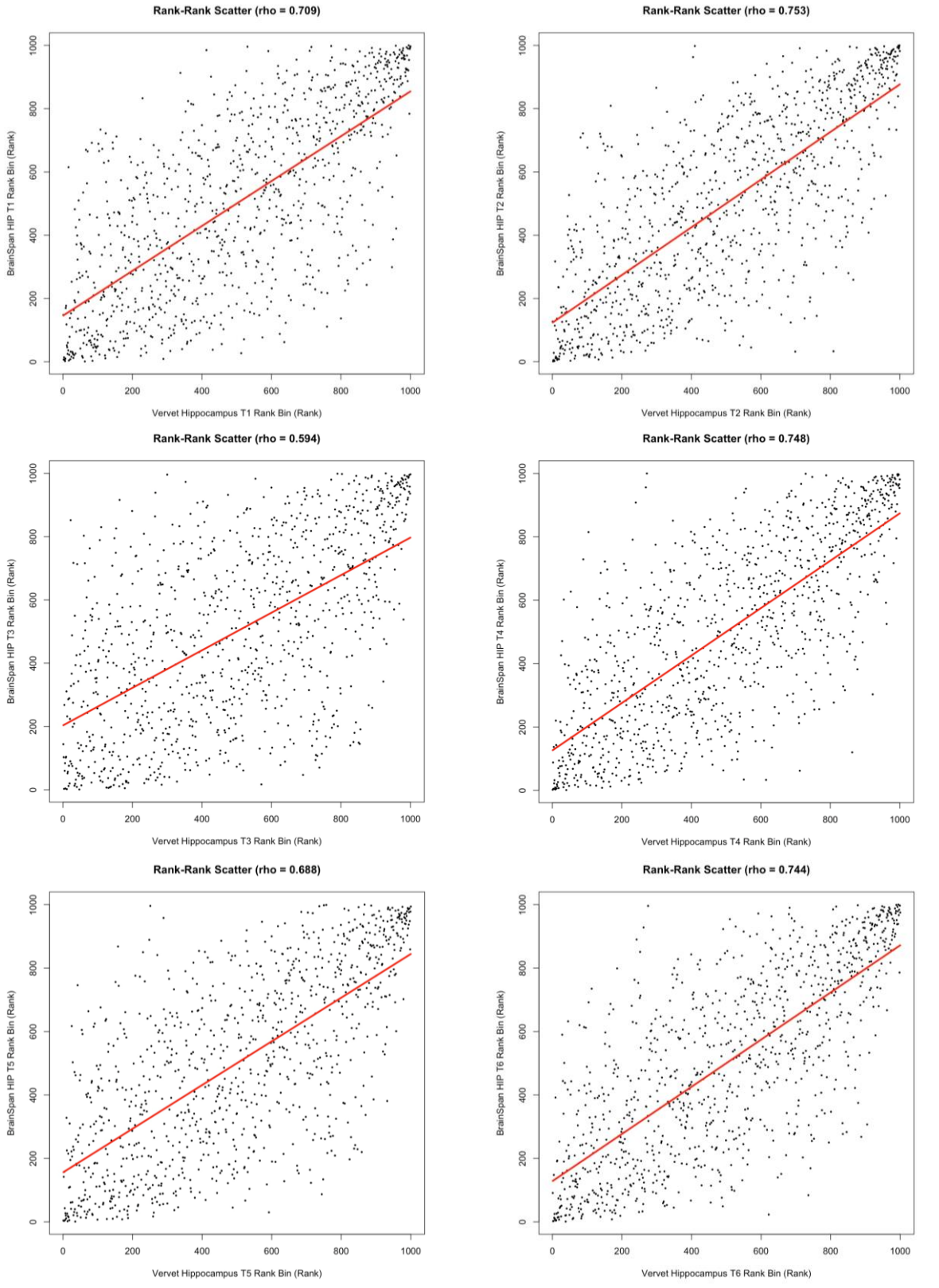


Figure C-5. Rank-Rank plots for human BrainSpan and vervet gene expression comparison across six time points, using the top 1,000 vervet age-related genes.

Time Period	Human Age	Human Developmental Period	Vervet Age	Vervet Category	Vervet Transcriptome Age Categories
T1	Birth - 5 months	Early Infancy	0 to 1 month	Neonates	1d, 7d, 30d
T2	6-18 months	Late Infancy	1.5 - 4.5 months	Young Infants	60d, 90d
T3	19 months - 5 years	Early Childhood	6 months - 1.25 years	Older infants to young juveniles	180d, 270d, 1y, 1.25y
T4	6-11 years	Late Childhood	1.5 - 2.75 years	Older Juveniles	1.5y, 1.75y, 2y, 2.5y
T5	12-19 years	Adolescence	3 - 4.75 years	Adolescents	3y, 4y
T6	20-60+ years	Adulthood	5+ years	Adults	5y or older

Table C-2. Corresponding age categories between vervet and human datasets [Adapted from Jasinkska, et al (2017)]

Time Period	Age rhesus macaques	Age Vervet
T1	0m	Infants 0, 1 and 7 days
T2	3m	Infants 90 days
T3	12m	Infants 1 and 1.25 years old
T4	48m	Adults 4+ years

Table C-3. Corresponding age categories between vervet and rhesus macaque [Adapted from Jasinska, et al (2017)]

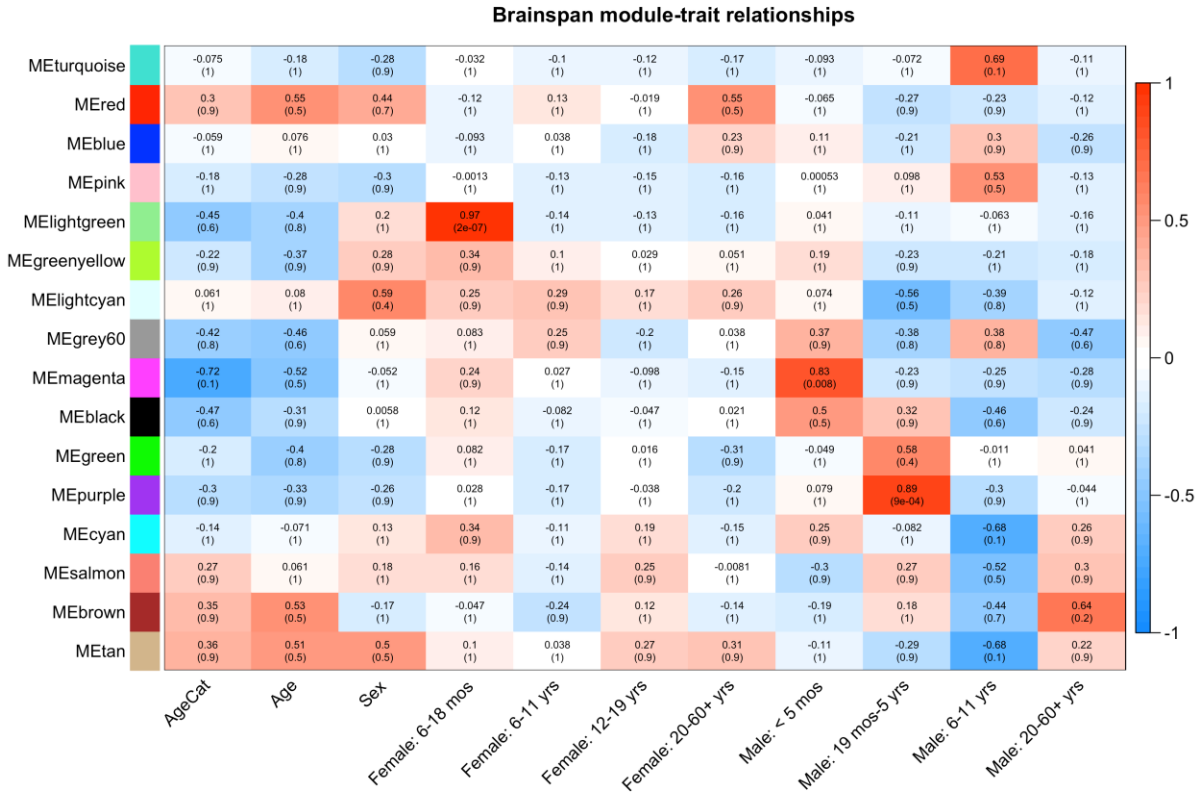


Figure C-6. Correlation of BrainSpan WGCNA module eigengenes with sample traits. P-values have been corrected for multiple hypothesis testing.

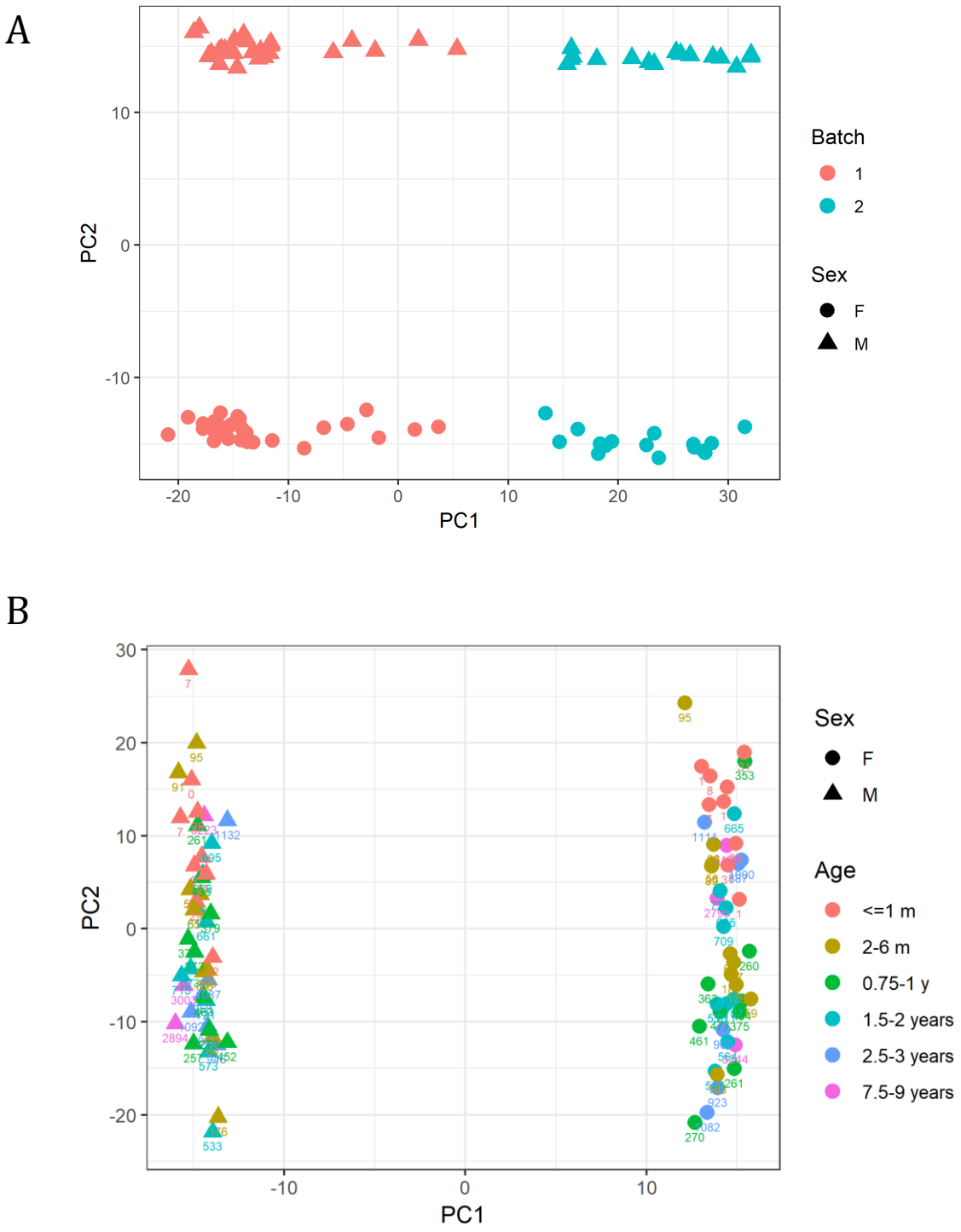


Figure C-7. (A) PCA plot of combined hippocampal samples (n=91). (B) PCA plot of PC1 vs PC2 after removing batch effect, with labels specifying age in days.

Gene Name	Gene Chromosome	Number of Distant eQTLs	eQTL Chromosome(s)
LOC103217586	CAE10	33	CAE9
LOC103218775	CAE11	10	CAE16
LOC103238280	CAE11	28	CAE1
LOC103219243	CAE12	1	CAE25
LOC103239854	CAE12	25	CAE29
LOC103240737	CAE13	29	CAE5
LOC103241453	CAE15	3	CAE14
LOC103242588	CAE16	1	CAE14
DPPA5	CAE17	1	CAE18
LOC103222806	CAE18	12	CAE12
LOC103244222	CAE18	2	CAE28
LOC103246126	CAE2	97	CAE18
LOC103226062	CAE21	11	CAE9
LOC103244257	CAE23	4	CAE6
LOC103244881	CAE23	1	CAE6
LOC103229920	CAE24	26	CAE26
LOC103230363	CAE25	4	CAE9
LOC103246473	CAE27	10	CAE23
LOC103214808	CAE3	28	CAE16
LOC103233076	CAE5	9	CAE16
LOC103236151	CAE7	7	CAE2
LOC103236677	CAE7	10	CAE16
LOC103237663	CAE8	3	CAE25
LOC103216138	CAE9	51	CAE28
LOC103232262	CAEX	4	CAE25
LOC103232591	CAEX	11	CAE3

Table C-4. eGenes with associated SNPs located on a different chromosome.

Principal Component	Age Category	Age in Days	Sex	Concentration	Volume	RIN
PC1	0.426	0.106	0.999	0.0778	0.106	0.0764
PC2	0.678	0.142	0.106	0.133	0.106	0.0909
PC3	0.641	0.354	0.106	0.105	0.106	0.296
PC4	0.267	0.284	0.106	0.0662	0.105	0.121
PC5	0.353	0.099	0.104	0.0912	0.106	0.0712
PC6	0.176	0.062	0.104	0.0465	0.106	0.0202
PC7	0.641	0.409	0.106	0.101	0.106	0.126
PC8	0.141	0.0998	0.106	0.0549	0.106	0.19
PC9	0.222	0.0627	0.106	0.051	0.106	0.015
PC10	0.11	0.104	0.106	0.114	0.106	0.091

Table C-5. Correlation between first 10 principal components and known covariates.

References

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genetics*, *9*(6), e1003514. https://doi.org/10.1371/journal.pgen.1003514

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891

Almasy, L., and Blangero, J. (1998). Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees. *The American Journal of Human Genetics*, *62*(5), 1198–1211. https://doi.org/10.1086/301844

Ana, C. C., Gabriela, A. C., Paz, T. M., Juan, F. R., Adrián, G., Alexandra, J. R., … García, R. A. (2018). Collagen XIX Alpha 1 Improves Prognosis in Amyotrophic Lateral Sclerosis. *Aging and Disease*, *10*(2), 278–292. https://doi.org/10.14336/AD.2018.0917

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. https://doi.org/10.1101/gr.133744.111

Apple, D. M., Solano-Fonseca, R., and Kokovay, E. (2017). Neurogenesis in the aging brain. *Biochemical Pharmacology*, *141*, 77–85. https://doi.org/10.1016/J.BCP.2017.06.116

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Lockhart, N. C. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Arnett, M. G., Muglia, L. M., Laryea, G., and Muglia, L. J. (2016). Genetic Approaches to Hypothalamic-Pituitary-Adrenal Axis Regulation. *Neuropsychopharmacology*, *41*(1), 245–260. https://doi.org/10.1038/npp.2015.215

Auer, P. L., and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, *185*(2), 405–416. https://doi.org/10.1534/genetics.110.114983

Bakken, T. E., Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., … Lein, E. S. (2016). A comprehensive transcriptional map of primate brain development. *Nature*, *535*(7612), 367–375. https://doi.org/10.1038/nature18637

Bergoffen, J., Scherer, S., Wang, S., Scott, M., Bone, L., Paul, D., … Fischbeck, K. (1993). Connexin mutations in X-linked Charcot-Marie-Tooth disease. *Science*, *262*(5142), 2039–2042. https://doi.org/10.1126/science.8266101

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., … Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium.

Nature Biotechnology, *28*(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

Blake, J. A., Eppig, J. T., Kadin, J. A., Richardson, J. E., Smith, C. L., and Bult, C. J. (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, *45*(D1), D723–D729. https://doi.org/10.1093/nar/gkw1040

Bogomolov, M., Peterson, C. B., Benjamini, Y., and Sabatti, C. (2017). *Testing hypotheses on a tree: new error rates and controlling strategies*. Retrieved from http://arxiv.org/abs/1705.07529

Bond, J., Roberts, E., Mochida, G. H., Hampshire, D. J., Scott, S., Askham, J. M., … Woods, C. G. (2002). ASPM is a major determinant of cerebral cortical size. *Nature Genetics*, *32*(2), 316–320. https://doi.org/10.1038/ng995

Burke, M. W., Ptito, M., Ervin, F. R., and Palmour, R. M. (2015). Hippocampal neuron populations are reduced in vervet monkeys with fetal alcohol exposure. *Developmental Psychobiology*, *57*(4), 470–485. https://doi.org/10.1002/dev.21311

Chandler, H., and Peters, G. (2013). Stressing the cell cycle in senescence and aging. *Current Opinion in Cell Biology*, *25*(6), 765–771. https://doi.org/10.1016/J.CEB.2013.07.005

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., … Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128. https://doi.org/10.1186/1471-2105-14-128

Chen, J. A., Fears, S. C., Jasinska, A. J., Huang, A., Al-Sharif, N. B., Scheibel, K. E., … Coppola, G. (2018). Neurodegenerative disease biomarkers $A\beta_{1-40}$, $A\beta_{1-42}$, tau, and p-tau$_{181}$ in the vervet monkey cerebrospinal fluid: Relation to normal aging, genetic influences, and cerebral amyloid angiopathy. *Brain and Behavior*, *8*(2), e00903. https://doi.org/10.1002/brb3.903

Chmielewski, M. R., and Grzymala-Busse, J. W. (1996). Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, *15*(4), 319–331. https://doi.org/10.1016/S0888-613X(96)00074-6

Chua, C. E. L., Goh, E. L. K., and Tang, B. L. (2014). Rab31 is expressed in neural progenitor cells and plays a role in their differentiation. *FEBS Letters*, *588*(17), 3186–3194. https://doi.org/10.1016/J.FEBSLET.2014.06.060

Collado-Torres, L., Burke, E. E., Peterson, A., Shin, J., Straub, R. E., Rajpurohit, A., … Jaffe, A. E. (2019). Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron*. https://doi.org/10.1016/J.NEURON.2019.05.013

Consortium, Gte., analysts:, L., Laboratory, D. A. & C. C. (LDACC):, management:, N. I. H. program, collection:, B., Pathology:, … Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*, 204. Retrieved from http://dx.doi.org/10.1038/nature24277

Cooper, D. M. (2012). The Balance between Life and Death: Defining a Role for Apoptosis in

Aging. *J Clin Exp Pathol*. https://doi.org/10.4172/2161-0681.S4-001

Cotter, D., Mackay, D., Landau, S., Kerwin, R., and Everall, I. (2001). Reduced Glial Cell Density and Neuronal Size in the Anterior Cingulate Cortex in Major Depressive Disorder. *Archives of General Psychiatry*, *58*(6), 545. https://doi.org/10.1001/archpsyc.58.6.545

Deguise, M.-O., Boyer, J. G., McFall, E. R., Yazdani, A., De Repentigny, Y., and Kothary, R. (2016). Differential induction of muscle atrophy pathways in two mouse models of spinal muscular atrophy. *Scientific Reports*, *6*(1), 28846. https://doi.org/10.1038/srep28846

Dekaban, A. S., and Sadowsky, D. (1978). Changes in brain weights during the span of human life: Relation of brain weights to body heights and body weights. *Annals of Neurology*, *4*(4), 345–356. https://doi.org/10.1002/ana.410040410

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dredge, B. K., Polydorides, A. D., and Darnell, R. B. (2001). The splice of life: Alternative splicing and neurological disease. *Nature Reviews Neuroscience*, *2*(1), 43–50. https://doi.org/10.1038/35049061

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Eriksson, P. S., Perfilieva, E., Björk-Eriksson, T., Alborn, A.-M., Nordborg, C., Peterson, D. A., and Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine*, *4*(11), 1313–1317. https://doi.org/10.1038/3305

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, *28*(8), 817–825. https://doi.org/10.1038/nbt.1662

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, *33*(4), 364–376. https://doi.org/10.1038/nbt.3157

Fears, S. C., Melega, W. P., Service, S. K., Lee, C., Chen, K., Tu, Z., … Woods, R. P. (2009). Identifying Heritable Brain Phenotypes in an Extended Pedigree of Vervet Monkeys. *Journal of Neuroscience*, *29*(9), 2867–2875. https://doi.org/10.1523/JNEUROSCI.5153-08.2009

Fidziańska, A., Goebel, H. H., and Warlo, I. (1990). Acute Infantile Spinal Muscular Atrophy. *Brain*, *113*(2), 433–445. https://doi.org/10.1093/brain/113.2.433

Finkel, R. S., Mercuri, E., Darras, B. T., Connolly, A. M., Kuntz, N. L., Kirschner, J., … De Vivo, D. C. (2017). Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *New England Journal of Medicine*, *377*(18), 1723–1732. https://doi.org/10.1056/NEJMoa1702752

Fitzsimons, R. B., and Hoh, J. F. Y. (1981). Embryonic and foetal myosins in human skeletal

muscle: The presence of foetal myosins in duchenne muscular dystrophy and infantile spinal muscular atrophy. *Journal of the Neurological Sciences*, *52*(2–3), 367–384. https://doi.org/10.1016/0022-510X(81)90018-6

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453. https://doi.org/10.1038/nn.4399

Gaujoux, R., and Seoighe, C. (2013). CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, *29*(17), 2211–2212. https://doi.org/10.1093/bioinformatics/btt351

Geist, J., and Kontrogianni-Konstantopoulos, A. (2016). MYBPC1, an Emerging Myopathic Gene: What We Know and What We Need to Learn. *Frontiers in Physiology*, *7*, 410. https://doi.org/10.3389/fphys.2016.00410

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Wolber, P. (2004). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Gerrits, A., Li, Y., Tesson, B. M., Bystrykh, L. V., Weersing, E., Ausema, A., … de Haan, G. (2009). Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. *PLoS Genetics*, *5*(10), e1000692. https://doi.org/10.1371/journal.pgen.1000692

Gibson, G., Powell, J. E., and Marigorta, U. M. (2015). Expression quantitative trait locus analysis for translational medicine. *Genome Medicine*, *7*(1), 60. https://doi.org/10.1186/s13073-015-0186-7

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, *24*(8), 408–415. https://doi.org/10.1016/j.tig.2008.06.001

Glass, D., Viñuela, A., Davies, M. N., Ramasamy, A., Parts, L., Knowles, D., … Spector, T. D. (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biology*, *14*(7), R75. https://doi.org/10.1186/gb-2013-14-7-r75

Gu, Y., and Hall, Z. W. (1988). Immunological evidence for a change in subunits of the acetylcholine receptor in developing and denervated rat muscle. *Neuron*, *1*(2), 117–125. https://doi.org/10.1016/0896-6273(88)90195-X

Hamilton, G., and Gillingwater, T. H. (2013). Spinal muscular atrophy: going beyond the motor neuron. *Trends in Molecular Medicine*, *19*(1), 40–50. https://doi.org/10.1016/J.MOLMED.2012.11.002

Hawrylycz, M., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., … Lein, E. (2015). Canonical genetic signatures of the adult human brain. *Nature Neuroscience*, *18*(12), 1832–1844. https://doi.org/10.1038/nn.4171

Heath, S. C., Snow, G. L., Thompson, E. A., Tseng, C., and Wijsman, E. M. (1997). MCMC segregation and linkage analysis. *Genetic Epidemiology*, *14*(6), 1011–1016. https://doi.org/10.1002/(SICI)1098-2272(1997)14:6<1011::AID-GEPI75>3.0.CO;2-L

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112. https://doi.org/10.1038/nature07829

Hesselmans, L. F. G. M., Jennekens, F. G. I., Van Den Oord, C. J. M., Veldman, H., and Vincent, A. (1993). Development of innervation of skeletal muscle fibers in man: Relation to acetylcholine receptors. *The Anatomical Record*, *236*(3), 553–562. https://doi.org/10.1002/ar.1092360315

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, Y. S., Ramensky, V., Service, S. K., Jasinska, A. J., Jung, Y., Choi, O.-W., … Freimer, N. B. (2015). Sequencing strategies and characterization of 721 vervet monkey genomes for future genetic analyses of medically relevant traits. *BMC Biology*, *13*(1), 41. https://doi.org/10.1186/s12915-015-0152-2

Jasinska, A. J. (2019). Biological Resources for Genomic Investigation in the Vervet Monkey (Chlorocebus). In *Savanna Monkeys* (pp. 16–28). https://doi.org/10.1017/9781139019941.002

Jasinska, A. J., Schmitt, C. A., Service, S. K., Cantor, R. M., Dewar, K., Jentsch, J. D., … Freimer, N. B. (2013). Systems biology of the vervet monkey. *ILAR Journal / National Research Council, Institute of Laboratory Animal Resources*, *54*(2), 122–143. https://doi.org/10.1093/ilar/ilt049

Jasinska, A. J., Service, S., Choi, O., DeYoung, J., Grujic, O., Kong, S., … Freimer, N. B. (2009). Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Human Molecular Genetics*, *18*(22), 4415–4427. https://doi.org/10.1093/hmg/ddp397

Jasinska, A. J., Service, S., Levinson, M., Slaten, E., Lee, O., Sobel, E., … Ophoff, R. A. (2007). A genetic linkage map of the vervet monkey (Chlorocebus aethiops sabaeus). *Mammalian Genome*, *18*(5), 347–360. https://doi.org/10.1007/s00335-007-9026-4

Jasinska, A. J., Zelaya, I., Service, S. K., Peterson, C. B., Cantor, R. M., Choi, O.-W., … Freimer, N. B. (2017). Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. *Nature Genetics*, ng.3959. https://doi.org/10.1038/ng.3959

Jennings, C. G., Landman, R., Zhou, Y., Sharma, J., Hyman, J., Movshon, J. A., … Feng, G.

(2016). Opportunities and challenges in modeling human brain disorders in transgenic primates. *Nature Neuroscience*, *19*(9), 1123–1130. https://doi.org/10.1038/nn.4362

Jin, K., Sun, Y., Xie, L., Batteur, S., Mao, X. O., Smelick, C., … Greenberg, D. A. (2003). Neurogenesis and aging: FGF-2 and HB-EGF restore neurogenesis in hippocampus and subventricular zone of aged mice. *Aging Cell*, *2*(3), 175–183. https://doi.org/10.1046/j.1474-9728.2003.00046.x

Ju, Y., Li, J., Xie, C., Ritchlin, C. T., Xing, L., Hilton, M. J., and Schwarz, E. M. (2013). Troponin T3 expression in skeletal and smooth muscle is required for growth and postnatal survival: characterization of Tnnt3(tm2a(KOMP)Wtsi) mice. *Genesis (New York, N.Y. : 2000)*, *51*(9), 667–675. https://doi.org/10.1002/dvg.22407

Kalinin, S., Willard, S. L., Shively, C. A., Kaplan, J. R., Register, T. C., Jorgensen, M. J., … Feinstein, D. L. (2013). Development of amyloid burden in African Green monkeys. *Neurobiology of Aging*, *34*(10), 2361–2369. https://doi.org/10.1016/j.neurobiolaging.2013.03.023

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., … Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483–489. https://doi.org/10.1038/nature10523

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kariya, S., Park, G.-H., Maeno-Hikichi, Y., Leykekhman, O., Lutz, C., Arkovitz, M. S., … Monani, U. R. (2008). Reduced SMN protein impairs maturation of the neuromuscular junctions in mouse models of spinal muscular atrophy. *Human Molecular Genetics*, *17*(16), 2552–2569. https://doi.org/10.1093/hmg/ddn156

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(7), 2926–2931. https://doi.org/10.1073/pnas.0909344107

Khalaf-Nazzal, R; Francis, F. (2013). Hippocampal development – Old and new findings. *Neuroscience*, *248*, 225–242. https://doi.org/10.1016/J.NEUROSCIENCE.2013.05.061

Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1). https://doi.org/10.1093/nar/gkw377

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC*

Bioinformatics, *15*(1), 8. https://doi.org/10.1186/1471-2105-15-8

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., … Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Larson, J. L., and Yuan, G.-C. (2010). Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. *BMC Bioinformatics*, *11*(1), 557. https://doi.org/10.1186/1471-2105-11-557

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, *25*(14), 1841–1842. https://doi.org/10.1093/bioinformatics/btp328

Lawrence, Michael, Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Heng, and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Y. I., Knowles, D. A., and Pritchard, J. K. (2016). LeafCutter: Annotation-free quantification of RNA splicing. *BioRxiv*. Retrieved from http://www.biorxiv.org/content/early/2016/03/16/044107

Lin, B. L., Li, A., Mun, J. Y., Previs, M. J., Previs, S. B., Campbell, S. G., … Sadayappan, S. (2018). Skeletal myosin binding protein-C isoforms regulate thin filament activity in a Ca2+-dependent manner. *Scientific Reports*, *8*(1), 2604. https://doi.org/10.1038/s41598-018-21053-1

Lunn, M. R., and Wang, C. H. (2008). Spinal muscular atrophy. *The Lancet*, *371*(9630), 2120–2133. https://doi.org/10.1016/S0140-6736(08)60921-6

Majewski, J., and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics : TIG*, *27*(2), 72–79. https://doi.org/10.1016/j.tig.2010.10.006

Martínez-Hernández, R., Bernal, S., Also-Rallo, E., Alías, L., Barceló, M., Hereu, M., … Tizzano, E. F. (2013). Synaptic defects in type I spinal muscular atrophy in human development. *The Journal of Pathology*, *229*(1), 49–61. https://doi.org/10.1002/path.4080

Martínez-Hernández, R., Soler-Botija, C., Also, E., Alias, L., Caselles, L., Gich, I., … Tizzano, E. F. (2009). The Developmental Pattern of Myotubes in Spinal Muscular Atrophy Indicates

Prenatal Delay of Muscle Maturation. *Journal of Neuropathology & Experimental Neurology*, *68*(5), 474–481. https://doi.org/10.1097/NEN.0b013e3181a10ea1

Mattick, J. S., and Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. *Nature Structural & Molecular Biology*, *22*(1), 5–7. https://doi.org/10.1038/nsmb.2942

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

McEwen, B. S., Gray, J. D., and Nasca, C. (2015). 60 YEARS OF NEUROENDOCRINOLOGY: Redefining neuroendocrinology: stress, sex and cognitive and emotional regulation. *Journal of Endocrinology*, *226*(2), T67–T83. https://doi.org/10.1530/JOE-15-0121

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., … Guigo, R. (2015). The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Mendell, J. R., Al-Zaidy, S., Shell, R., Arnold, W. D., Rodino-Klapac, L. R., Prior, T. W., … Kaspar, B. K. (2017). Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *New England Journal of Medicine*, *377*(18), 1713–1722. https://doi.org/10.1056/NEJMoa1706198

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., … Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), 199–206. https://doi.org/10.1038/nature13185

Mills, J. D., and Janitz, M. (2012). Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. *Neurobiology of Aging*, *33*(5), 1012.e11-1012.e24. https://doi.org/10.1016/J.NEUROBIOLAGING.2011.10.030

Morales-Corraliza, J., Wong, H., Mazzella, M. J., Che, S., Lee, S. H., Petkova, E., … Mathews, P. M. (2016). Brain-Wide Insulin Resistance, Tau Phosphorylation Changes, and Hippocampal Neprilysin and Amyloid-β Alterations in a Monkey Model of Type 1 Diabetes. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(15), 4248–4258. https://doi.org/10.1523/JNEUROSCI.4640-14.2016

Morgan, N. V., Brueton, L. A., Cox, P., Greally, M. T., Tolmie, J., Pasha, S., … Maher, E. R. (2006). Mutations in the Embryonal Subunit of the Acetylcholine Receptor (CHRNG) Cause Lethal and Escobar Variants of Multiple Pterygium Syndrome. *The American Journal of Human Genetics*, *79*(2), 390–395. https://doi.org/10.1086/506256

Mudher, A., Chapman, S., Richardson, J., Asuni, A., Gibb, G., Pollard, C., … Lovestone, S.

(2001). Dishevelled Regulates the Metabolism of Amyloid Precursor Protein via Protein Kinase C/Mitogen-Activated Protein Kinase and c-Jun Terminal Kinase. *Journal of Neuroscience*, *21*(14), 4987–4995. https://doi.org/10.1523/JNEUROSCI.21-14-04987.2001

Negi, S. K., and Guda, C. (2017). Global gene expression profiling of healthy human brain and its application in studying neurological disorders. *Scientific Reports*, *7*(1), 897. https://doi.org/10.1038/s41598-017-00952-9

Nestler, E., Hyman, S., Holtzman, D. & Malenka, R. (2015). *Molecular Neuropharmacology: A Foundation for Clinical Neuroscience*. McGraw-Hill Education/Medical.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. https://doi.org/10.1371/journal.pgen.1000888

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Palomer, E., Buechler, J., and Salinas, P. C. (2019). Wnt Signaling Deregulation in the Aging and Alzheimer's Brain. *Frontiers in Cellular Neuroscience*, *13*, 227. https://doi.org/10.3389/fncel.2019.00227

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Peterson, C. B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. *Genetic Epidemiology*, *40*(1), 45–56. https://doi.org/10.1002/gepi.21942

Peterson, C., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2015). TreeQTL: hierarchical error control for eQTL findings. In *bioRxiv*. https://doi.org/10.1101/021170

Postupna, N., Latimer, C. S., Larson, E. B., Sherfield, E., Paladin, J., Shively, C. A., … Montine, T. J. (2017). Human Striatal Dopaminergic and Regional Serotonergic Synaptic Degeneration with Lewy Body Disease and Inheritance of APOE ε4. *The American Journal of Pathology*, *187*(4), 884–895. https://doi.org/10.1016/j.ajpath.2016.12.010

Probst, P., and Boulesteix, A.-L. (2018). To Tune or Not to Tune the Number of Trees in Random Forest. In *Journal of Machine Learning Research* (Vol. 18). Retrieved from http://jmlr.org/papers/v18/17-269.html.

Rahman, M. M., Callaghan, C. K., Kerskens, C. M., Chattarji, S., and O'Mara, S. M. (2016). Early hippocampal volume loss as a marker of eventual memory deficits caused by repeated stress. *Scientific Reports*, *6*(1), 29127. https://doi.org/10.1038/srep29127

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., … Weale, M. E. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience*, *17*(10), 1418–1428. https://doi.org/10.1038/nn.3801

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rogers, J., and Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, *15*(5), 347–359. https://doi.org/10.1038/nrg3707

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., … Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, *43*(D1), D670–D681. https://doi.org/10.1093/nar/gku1177

Sargiannidou, I., Vavlitou, N., Aristodemou, S., Hadjisavvas, A., Kyriacou, K., Scherer, S. S., and Kleopa, K. A. (2009). Connexin32 Mutations Cause Loss of Function in Schwann Cells and Oligodendrocytes Leading to PNS and CNS Myelination Defects. *Journal of Neuroscience*, *29*(15), 4736–4749. https://doi.org/10.1523/JNEUROSCI.0325-09.2009

Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A., and Reggiani, C. (2015). Developmental myosins: expression patterns and functional significance. *Skeletal Muscle*, *5*, 22. https://doi.org/10.1186/s13395-015-0046-6

Shimoyama, M., De Pons, J., Hayman, G. T., Laulederkind, S. J. F., Liu, W., Nigam, R., … Jacob, H. (2015). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Research*, *43*(D1), D743–D750. https://doi.org/10.1093/nar/gku1026

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.2

Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., … Dekker, J. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, *13*(8), 418. https://doi.org/10.1186/gb-2012-13-8-418

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Stein, J. L., Medland, S. E., Vasquez, A. A., Hibar, D. P., Senstad, R. E., Winkler, A. M., … Thompson, P. M. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nature Genetics*, *44*(5), 552–561. https://doi.org/10.1038/ng.2250

Sumiyoshi, H., Inoguchi, K., Khaleduzzaman, M., Ninomiya, Y., and Yoshioka, H. (1997). Ubiquitous expression of the alpha1(XIX) collagen gene (Col19a1) during mouse embryogenesis becomes restricted to a few tissues in the adult organism. *The Journal of*

Biological Chemistry, *272*(27), 17104–17111. https://doi.org/10.1074/JBC.272.27.17104

Tak, Y. G., and Farnham, P. J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin*, *8*, 57. https://doi.org/10.1186/s13072-015-0050-4

Tang, B., Zhao, G., Luo, W., Xia, K., Cai, F., Pan, Q., … Dai, H. (2005). Small heat-shock protein 22 mutated in autosomal dominant Charcot-Marie-Tooth disease type 2L. *Human Genetics*, *116*(3), 222–224. https://doi.org/10.1007/s00439-004-1218-3

Tapia-Rojas, C., and Inestrosa, N. C. (2018). Loss of canonical Wnt signaling is involved in the pathogenesis of Alzheimer's disease. *Neural Regeneration Research*, *13*(10), 1705–1710. https://doi.org/10.4103/1673-5374.238606

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, *45*(2), 124–130. https://doi.org/10.1038/ng.2504

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419–1260419. https://doi.org/10.1126/science.1260419

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, *154*(1), 26–46. https://doi.org/10.1016/j.cell.2013.06.020

Van den Berge, K., Soneson, C., Robinson, M. D., and Clement, L. (2017). stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology*, *18*(1), 151. https://doi.org/10.1186/s13059-017-1277-0

van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., … Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553. https://doi.org/10.1038/mp.2015.63

van Praag, H., Schinder, A. F., Christie, B. R., Toni, N., Palmer, T. D., and Gage, F. H. (2002). Functional neurogenesis in the adult hippocampus. *Nature*, *415*(6875), 1030–1034. https://doi.org/10.1038/4151030a

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Liu, X. S. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., … Chen, L. S. (2016). Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *The American Journal of Human Genetics*, *98*(4), 697–708. https://doi.org/10.1016/j.ajhg.2016.02.020

Warren, W. C., Jasinska, A. J., García-Pérez, R., Svardal, H., Tomlinson, C., Rocchi, M., … Freimer, N. B. (2015). The genome of the vervet (Chlorocebus aethiops sabaeus). *Genome*

Research, *25*(12), 1921–1933. https://doi.org/10.1101/gr.192922.115

Wei, B., and Jin, J.-P. (2016). TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure–function relationships. *Gene*, *582*(1), 1–13. https://doi.org/10.1016/J.GENE.2016.01.006

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006. https://doi.org/10.1093/nar/gkt1229

Won, K.-J., Chepelev, I., Ren, B., and Wang, W. (2008). Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, *9*(1), 547. https://doi.org/10.1186/1471-2105-9-547

Yu, C.-H., Pal, L. R., and Moult, J. (2016). Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease. *OMICS: A Journal of Integrative Biology*, *20*(7), 400–414. https://doi.org/10.1089/omi.2016.0063

Yu, Q., and He, Z. (2017). Comprehensive investigation of temporal and autism-associated cell type composition-dependent and independent gene expression changes in human brains. *Scientific Reports*, *7*(1), 4121. https://doi.org/10.1038/s41598-017-04356-7

Yu, S., and Driscoll, M. (2011). EGF signaling comes of age: promotion of healthy aging in C. elegans. *Experimental Gerontology*, *46*(2–3), 129–134. https://doi.org/10.1016/j.exger.2010.10.010

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, *18*(1), 583. https://doi.org/10.1186/s12864-017-4002-1

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., … Wu, J. Q. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *Journal of Neuroscience*, *34*(36), 11929–11947. https://doi.org/10.1523/JNEUROSCI.1860-14.2014

Zhang, Yong, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730