Cell type marker enrichment across brain regions and experimental conditions

by

Powell Patrick Cheng Tan

B. Sc. (Honours), Simon Fraser University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES
(Bioinformatics)

The University Of British Columbia
(Vancouver)

November 2012

Abstract

The first chapter of this thesis explored the dominant gene expression pattern in the adult human brain. We discovered that the largest source of variation can be explained by cell type marker expression. Across brain regions, expression of neuron cell type markers are anti-correlated with the expression of oligodendrocyte cell type markers. Next, we explored gene function convergence and divergence in the adult mouse brain. Our contributions are as follows. First, we provide candidate cell type markers for investigating specific cell type populations. Second, we highlight orthologous genes that show functional divergence between human and mouse brains.

In the second chapter, we present our preliminary work on the effects of tissue types and experimental conditions on human microarray studies. First, we measured the expression and differential expression levels of tissue-enriched genes. Next, we identified modules with similar expression levels and differential expression p-values. Our results show that expression levels reflect tissue type variation. In contrast, differential expression levels are more complex, owing to the large diversity of experimental conditions in the data. In summary, our work provides a different perspective on the functional roles of genes in human microarray studies.

Table of Contents

List of Tables

List of Figures

Glossary

AIBS Allen Institute for Brain Science, a nonprofit organization that makes publicly available large-scale data that pertains to neuroscience which includes *in situ* images of the mouse brain and human brain microarray

ANOVA Analysis of Variance, a set of statistical techniques to identify sources of variability between groups

AUC Area Under a Receiver Operating Characteristic Curve, the area under the curve that shows the true positive rate against the false positive rate at different cutoffs, an area of 1.0 shows perfect enrichment while an area of 0.5 indicates no enrichment

DE Differential Expression, the difference in expression levels between sample groups represented by a p-value

EE Expression, the relative mean expression level of a gene across all samples within a dataset ranging from 0.0 (no expression) to 1.0 (high expression)

GEO Gene Expression Omnibus, a public data repository of functional genomics studies

GO Gene Ontology, is a set of controlled vocabularies describing gene products in terms of biological processes, cellular components and molecular functions

H0351.2001 Allen Human Brain Atlas donor profile of a 24 year old African American male

H0351.2002 Allen Human Brain Atlas donor profile of a 39 year old African American male

ISA Iterative Signature Algorithm, a biclustering algorithm that iteratively selects genes and samples that are significantly different based on a threshold

ISH *In situ* Hybridization, is an experimental technique where labelled RNA strands hybridize to complementary strands localized in a specific tissue location

PC1 First Principal Component, the principal component with the largest variance

PCA Principal Component Analysis, a statistical technique that projects high dimensional data to lower dimensions in terms of orthogonal variables called principal components

RS Result Set, the pair of sample groups within a dataset from which ANOVA was used

WM/GM White Matter to Grey Matter Transcript Ratio, the gene expression ratio between gray matter samples and adjacent white matter samples

Acknowledgments

First and foremost, my appreciation goes to Dr. Paul Pavlidis who has been a kind and patient mentor. To Paul, thanks for taking me under your wings and always encouraging me to pursue my intuition. To Dr. Leon French, thanks for giving me a head start towards a productive and exciting thesis project and for providing insightful discussions and feedback on Chapter 1. I would also like to acknowledge Raymond Lim, Tyler Funnell, the Gemma developers and curators for their contributions toward the development of the differential expression matrix in Chapter 2. My thanks go to Elodie Portales-Casamar for providing feedback and suggestions toward this manuscript. To everyone in the Pavlidis Lab, thanks for all the support and camaraderie during my brief time in the lab. I would also like to thank my committee members, Dr. Joerg Gsponer and Dr. Ann-Marie Craig for their helpful feedback toward my work.

I also like to thank Dr. Frederic Pio and Dr. Jack Chen for providing my initial bioinformatics training, Dr. Paula Lario and Dr. Anders Ohrn for providing me with the practical experience to work in the industry, and, Dr. Ryan Brinkman, Dr. Wyeth Wasserman, Dr. Virginie Bernard, Dr. Matthew Farrer, and Dr. Carles Vilariño-Güell for the rotation opportunities and mentorship. Also, my thanks go to Sharon Ruschkowski for keeping me on track and to my fellow bioinformatics students for sharing this experience with me.

Finally, my research would not have been possible without the generous funding from the Canadian Institute for Health Research Bioinformatics Training Program and from Dr. Pavlidis. My data analysis would have not gone forward without the datasets provided by the Allen Institute for Brain Science and the many researchers and donors who donated their data towards the improvement of our scientific understanding.

Dedication

To my fellow friends and colleagues for their camaraderie.
To my Mom, Pacita, thanks for helping me stay healthy and keeping me out of the hospital. To my Dad, Lamberto, for providing us with a warm home. And to my Sister, Maria, for filling in the missing gaps and for being the best storyteller I have ever met.

Chapter 1

Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain

1.1 Introduction

Gene expression in the adult mammalian brain is highly complex and poorly understood. Over 80% of all genes are expressed in the central nervous system, often with patterns that vary in time and space [19, 22, 24]. Many genes show patterns that correspond to classical neuroanatomical subdivisions [24]. Others reflect neurotransmitter systems, and yet others appear to reflect patterns laid down during development [10, 22, 44]. The functional significance of many other patterns is not clear. As the neuroscience community increasingly integrates data across modalities, gaining a deeper understanding of expression patterns is important. One way to gain insight into these patterns is to examine their conservation in evolution. Another is to dissect them into sub-patterns that reflect different cell types. Progress on both of these fronts is enabled by the availability of large-scale data sets. In this paper we focus on expression patterns in the normal adult human brain, comparing them to expression in the normal adult mouse, extending our recent work [17].

There is a broad expectation that gene expression in the mouse and human should be similar, and the brain is no exception. It is well known that the fundamental anatomical structure and function of the nervous system is common across mammals. This is exemplified by the similarities observed in the gene expression patterns in the subcortical regions of the brain [25, 41]. Gene expression in the cortical regions on the other hand show greater gene expression diversity between mouse and human [45]. Differences in gene expression may be due to the increased number of cortical neurons in primates compared to rodents [20]. However, none of these studies is comprehensive in terms of brain regions or genes and insights into studies that look at cell type compositions have been limited. Within specific brain regions, inverse relationships between

cell type expression patterns have been observed in human [33]. However, it is unclear whether expression patterns are also anti-correlated between brain regions. Recently we reported that gene expression across adult mouse brain regions is dominated by patterns associated with neuron and oligodendrocyte marker expression levels [17]. These patterns were identified by seeking strong anti-correlated patterns of gene expression and also by principal component analysis (PCA). PCA captures the dominant patterns in the data in orthogonal variables termed principal components [35]. In the adult mouse brain, higher levels of expression of genes with a neuron-enriched pattern tended to be associated with anterior regions and regions with higher macroconnectivity [17]. The opposite was observed for the oligodendrocyte-enriched pattern. We hypothesized that similar relationships exist in the human brain.

To investigate the gene expression patterns in the human brain, we applied PCA to the regional transcriptomes of two adult human brains. Based on the first principal component (PC1) scores, we identified two groups of genes that were enriched for neuron cell type markers (the "neuron-enriched" pattern) and oligodendrocyte cell type markers (the "oligodendrocyte-enriched" pattern) respectively. Our results show that the significant portion of the transcriptome can be explained by the expression of neuron and oligodendrocyte cell type markers which are anti-correlated across brain regions. Moreover, in comparison to mouse subcortical regions, we report homologous genes with similar expression patterns which are also enriched for neuron and oligodendrocyte markers but not astrocyte markers. We also observed homologous genes with differences in expression patterns, the details of these patterns could provide additional insights into functional similarities and differences among mammalian brain lineages.

1.2 Methods

We used publicly available datasets and performed two independent analyses to study cell type expression patterns within the human brain and between the mouse and human brain. The overview of the materials and methods used are shown in Figure 1.1.

1.2.1 Human brain gene expression

We analyzed the normalized gene expression data from two healthy adult human post-mortem brains downloaded from the publicly available dataset called the "Allen Human Brain Atlas" provided by the AIBS (Allen Institute for Brain Science) (http://www.brain-map.org/) [19]. Briefly, donor H0351.2001 was a 24 year old African American male and donor H0351.2002 was a 39 year old African American male. For both brains, larger regions were manually macrodissected whereas smaller regions were laser captured microdissected. There are 896 brain region samples in the H0351.2001 dataset while the H0351.2002 dataset had 946 samples. The two human datasets were processed and analyzed separately. Sample replicates with the same "structure_name" column annotation were averaged, yielding 323 columns for H0351.2001 and 346 columns for H0351.2002. Samples from the left and right hemispheres were kept separate. Samples of white matter tracts (corpus callosum and cingulum bundle) were excluded from both matrices which resulted in 320 columns in the H0351.2001 dataset and 345 columns in the H0351.2002 dataset. Each normalized gene expression matrix contained data for 58,691 probes. We combined multiple probes for the same gene by taking the mean, yielding expression levels for 29,191 genes.

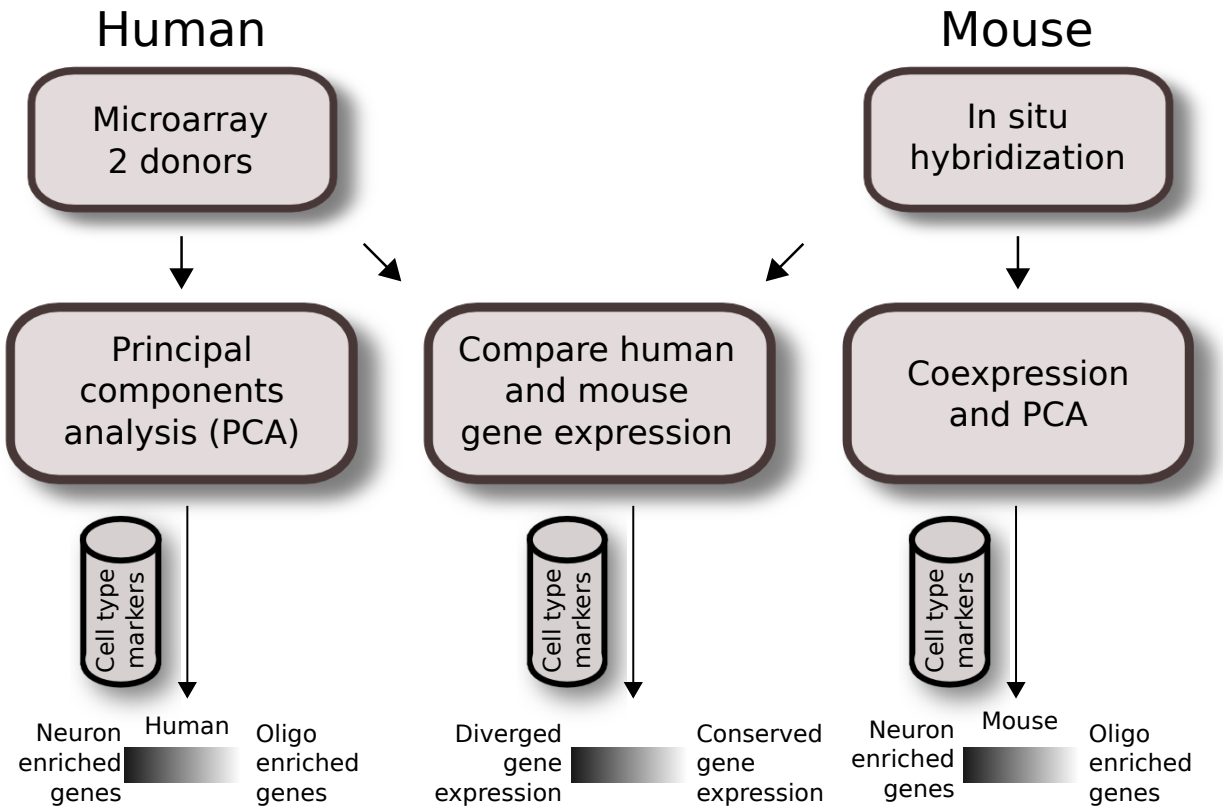


Figure 1.1: Analysis workflow of human and mouse gene expression across brain regions. Quality control for expressing genes and grey matter tissue samples were applied prior to analysis. Regional gene expression patterns were defined using PCA for two human brain microarray data in the first analysis (left). A similar method was applied to mouse ISH data previously as described in French et al. (2011) (right) [17]. The second analysis compares homologous data matrices of human H0351.2001 and mouse (middle). Cahoy et al. (2008) cell type markers were used to define neuron and oligodendrocyte-enriched patterns [9].

1.2.2 Mouse brain gene expression

We used the mouse gene expression data from the "Allen Mouse Brain Atlas" as described in our previous study [16]. Briefly, colorimetric *in situ hybridization* (ISH) images were collected from adult male, 56-day-old C57BL/6J normal mouse brains [24]. The ISH images were previously quantified and registered to a 3D reference atlas by Ng et al. [30]. The resulting brain region level expression energy (hereafter referred to as gene expression) is defined as the product of the expression area and the expression intensity [31]. Missing values are reported as NAs. The resulting mouse expression matrix has 20,444 genes and 207 brain regions.

1.2.3 Human brain analysis

For the analysis of the human data (independent of the mouse data), we focused our analysis on regionally variable grey matter expressed genes by discarding genes with standard deviation or mean expression below the 25^{th} percentile. After filtering, the H0351.2001 dataset had 14,595 genes and 320 brain regions while the H0351.2002 dataset had 14,615 genes and 345 brain regions.

We mean-centered and scaled the expression of each gene by its standard deviation across brain regions by using the "scale" function in R [38]. The "prcomp" function in R was used to calculate the principal components of the scaled gene expression matrix. PC scores for each gene correspond to the "x" value while PC loadings for each brain region correspond to the "rotation" value of the "prcomp" result. For consistency, we use the convention that the oligodendrocyte marker *MOBP* has a positive PC1 score.

We measured the cell type enrichment in PC1 scores by measuring the area under the curve (AUC) of the receiver operating characteristic curves, in a manner similar to the "wilcox.test" function in R. First, we ranked genes by their PC1 scores. Second, we divided the ranked list of genes into the positive and negative gene sets. This condition depends on the cell type of interest. For example, when we calculate the AUC for neuron markers, those genes that are found in the Cahoy neuron marker list are included in the positive gene set and all other genes are included in the negative gene set. Afterwards, we compare the positive and negative gene sets by calculating the AUC. To maintain positive AUC scores, the signs of the glia PC1 scores and loadings were reversed before AUCs are calculated.

1.2.4 Human-mouse comparisons

Human H0351.2001 and mouse brain region names were manually matched using the sample annotations and ontologies provided by the AIBS. Human genes were converted to mouse genes using HomoloGene build 66 [43].

We manually compared each brain region name in the AIBS mouse and human structure ontology files. For this analysis, we averaged the gene expression of both left and right human brain hemispheres with a matching structure name. However, there are many brain regions with structure names that do not match between species. To circumvent this, for each species, each brain region was manually annotated with a parent structure that is common to both species. Gene expression of multiple brain regions with the same parent structure were averaged. For example, the human regions "CA1"-"CA4" were averaged to match the parent structure "Ammon's horn". Likewise, the mouse regions "Lateral group of the dorsal thalamus", "Lateral posterior nucleus of the thalamus", and "Supragenicular nucleus" were averaged to match the parent structure "Lateral group of Nuclei, Dorsal Division".

Gene expression values of both matrices were then quantile normalized. Finally, genes with expression levels below the 25^{th} percentile in both species were removed. The resulting matched human and mouse matrices represent expression values of 7,911 genes across 58 subcortical brain regions.

We calculated the Spearman rank correlation for each homologous gene. Cell type enrichment of the homologous gene correlation was quantified as AUC in a similar manner to how AUC was calculated from PC1 scores.

1.2.5 Statistical analysis

We used the "cor.test" function in R to calculate Spearman rank correlations together with matching p-values. P-values were corrected for multiple testing by controlling for the false discovery rate, which are reported as q-values [4]. The distribution of orthologous gene expression pattern correlations was compared to 20 random distributions where human gene labels were shuffled without replacement. Correlations for data with missing values were calculated by using the "pairwise" method of the "cor" function in R [38].

Hierarchical clustering was performed with the "hclust" function in R [38], using Euclidean distances and Ward's minimum variance method as parameters [42].

Gene ontology analysis for the 100 most positively and negatively correlated expression patterns were performed using DAVID [11].

1.2.6 Additional data sources

Cell type markers were obtained from Cahoy et al. (2008) [9]. Only those marker genes that have at least 10x fold enrichment were used. In H0351.2001, there are a total of 267 neuron, 103 oligodendrocyte and 143 astrocyte cell type markers that are homologous to the mouse study. Similarly, the H0351.2002 dataset has 270 neuron, 104 oligodendrocyte and 145 astrocyte markers.

White matter to grey matter (WM/GM) transcript ratios within the anterior cingulate gyrus were obtained from Sibille et al. (2008) [40]. Sibille et al. defined WM/GM transcript ratio for each gene in each brain area as the ratio between the average expression of using all samples in the gray matter area and the average expression of using all samples in the adjacent white matter area. Ratios of multiple probe sets for the same gene were averaged. Glia to neuron cell ratios for the human cerebellum, cerebral cortex and the rest of the brain were obtained from Azevedo et al. (2009) who applied a chemomechanical dissociation technique to purify cells which were labelled by immunohistochemistry [2].

In relation with mouse and human expression pattern differences, the list of 73 genes that show differential expression pattern between mouse and human visual and temporal cortices was obtained from Zeng et al. (2012) [45]. Genes with discordant expression patterns between species were obtained from the list of 49 human-specific markers (genes that are correlated with modules enriched for cell types in human but not in mouse) in the meta-analysis of brain expression performed by Miller and colleagues (2010) [28]. These brain regions include both cortex and subcortical regions.

1.3 Results

1.3.1 Neuron-enriched and oligodendrocyte-enriched patterns are conserved

We characterized gene expression profiling data from two adult human brains (identified by the AIBS as donors H0351.2001 and H0351.2002) in a manner comparable to our previous analysis of the adult mouse brain (Figure 1.1). After filtering (see Methods), the H0351.2001 dataset had 14,595 genes while the H0351.2002 dataset had 14,615 genes, 13,250 of which were found in both datasets. For H0351.2001, we obtained 320 brain region samples. Telencephalon accounts for most of the brain region samples (53%),

metencephalon (22.1%), diencephalon (11%), myelencephalon (8.1%), and mesencephalon the least (5.9%). The H0351.2002 dataset had 345 samples with similar proportions of major brain divisions as H0351.2001. In H0351.2001, cerebellar samples clustered more closely compared to other brain regions (Figure 1.2), in line with previous observations that cerebellum gene expression is the most unique compared to other major brain divisions [26, 34, 39]. This was less apparent in H0351.2002 (data not shown). Hereafter, we report results based on these filtered datasets.

Next, we tested whether genes that express anti-correlated cell-type enriched patterns in mouse are also anti-correlated in humans [17]. We averaged the expression of all human homologs with the mouse neuron-enriched pattern. Similarly, we also averaged the expression pattern of all human homologs with the mouse oligodendrocyte-enriched pattern. As in mouse, the averaged neuron-enriched pattern is anti-correlated with the averaged oligodendrocyte-enriched pattern (H0351.2001 rho = -0.40, $P < 0.0001$ and H0351.2002 rho = -0.61, $P < 0.0001$) (Figure 1.2). Genes that show neuron-enriched patterns are predominantly expressed in metencephalon and telencephalon regions while genes in the oligodendrocyte-enriched patterns are not restricted to any major brain division.

This conservation of cell type marker enriched patterns is also evident in a PCA of the human data. The first three principal components of H0351.2001 accounted for 15.6%, 11.6%, and 8.31% of the total variance respectively whereas we see a slight decrease in the case of H0351.2002 with 15.2%, 8.07%, and 5.98% of the total variance respectively. The first principal component (PC1) gene scores of the two human datasets are strongly positively correlated (rho = 0.72, $P < 0.0001$), indicating that overall, the two brains have similar dominant expression patterns, consistent with the findings of Hawrylycz et al. (2012) [19]. We observed that these oligodendrocyte and neuron marker genes tend to have PC1 scores with opposite signs, consistent with our previous study in mouse. We term these as "oligodendrocyte-enriched" and "neuron-enriched" respectively. The top 25 genes in the "oligodendrocyte-enriched" and "neuron-enriched" gene sets are shown in Table 1.1 and Table 1.2 respectively. For each cell type, we measured the cell type enrichment by comparing the PC1 ranks of those cell type marker genes (as determined by Cahoy) against the PC1 ranks of the remaining genes (see Methods). In H0351.2001, neuronal markers showed the highest enrichment (AUC = 0.77), followed by oligodendrocyte markers (AUC = 0.73), and astrocyte markers the least (AUC = 0.66). We found evidence for comparable cell type marker enrichment in H0351.2002 PC1 loadings as well (neuron markers AUC = 0.82, oligodendrocyte markers AUC = 0.81, astrocyte markers AUC = 0.63). By way of comparison, in mouse we had found that PC2 gene loadings showed the highest enrichment for oligodendrocyte markers (AUC = 0.77) and neuron markers (AUC = 0.63) and no enrichment for astrocyte markers (AUC = 0.52) [17].

1.3.2 Principal component loadings partly reflect varying cell-type proportions

The PC1 gene loadings could either be explained by variations in expression levels within cells, or by variations in the ratio of different sub-populations of cells (or some combination of these). To further investigate this, we calculated the correlation between the H0351.2001 PC1 gene loadings and the white matter to gray matter transcript ratio (WM/GM) for 8,088 genes with data for both [40]. The correlation is statistically significant (rho = 0.59, $P < 0.0001$). Since white matter regions have been excluded from the

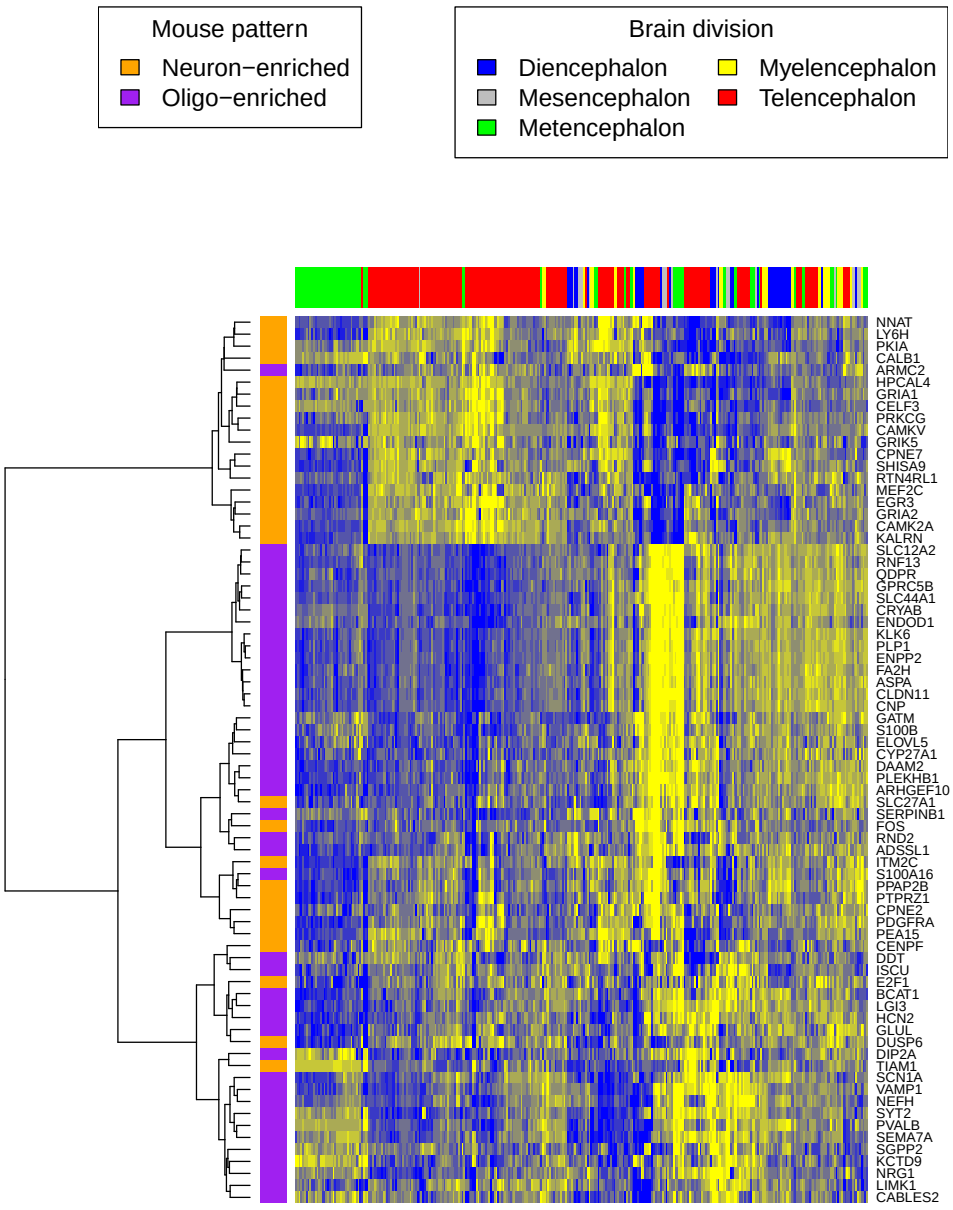


Figure 1.2: Human H0351.2001 gene expression profile of orthologous genes reported in the mouse neuron and oligodendrocyte-enriched patterns (French et al., 2011) [17]. High and low expression levels are colored in yellow and blue respectively. Rows are genes colored by their homolog cell type enrichment. Columns are brain region samples colored by major brain divisions. Hierarchical clustering was performed using the Ward's minimum variance method in R [42].

Gene symbol	Gene description	Entrez ID	PC1	Mean	StdDev
SLC27A1	solute carrier family 27 (fatty acid transporter), member 1	376497	12.67	10.92	0.33
REST	RE1-silencing transcription factor	5978	12.46	4.65	0.43
PPARA	peroxisome proliferator-activated receptor alpha	5465	12.41	3.65	0.37
ARHGEF10	Rho guanine nucleotide exchange factor (GEF) 10	9639	12.33	5.19	0.42
TRIM56	tripartite motif-containing 56	81844	12.14	3.93	0.42
EGFR	epidermal growth factor receptor	1956	12.12	4.21	0.55
A_24_P943258	AGILENT probe A_24_P943258 (non-RefSeq)	NA	12.11	4.92	0.43
A_23_P129258	AGILENT probe A_23_P129258 (non-RefSeq)	NA	12.01	13.75	0.51
RBMS2	RNA binding motif, single stranded interacting protein 2	5939	12.00	6.32	0.39
A_24_P316059	AGILENT probe A_24_P316059 (non-RefSeq)	NA	11.98	4.83	0.39
GPR75	G protein-coupled receptor 75	10936	11.96	6.83	0.39
PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	5209	11.95	7.15	0.39
C12orf39	chromosome 12 open reading frame 39	80763	11.95	4.65	0.47
MAFIP	MAFF interacting protein	727764	11.84	5.13	0.36
CXorf36	chromosome X open reading frame 36	79742	11.82	2.58	0.42
NPAS3	neuronal PAS domain protein 3	64067	11.75	8.08	0.35
SDPR	serum deprivation response	8436	11.74	3.43	0.44
LIMS1	LIM and senescent cell antigen-like domains 1	3987	11.74	3.45	0.35
A_24_P475689	AGILENT probe A_24_P475689 (non-RefSeq)	NA	11.67	3.67	0.43
BMP7	bone morphogenetic protein 7	655	11.59	7.21	0.38
CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	1495	11.58	6.69	0.41
TJP1	tight junction protein 1 (zona occludens 1)	7082	11.53	8.61	0.34
KIF19	kinesin family member 19	124602	11.53	3.02	0.50
A_23_P134887	AGILENT probe A_23_P134887 (non-RefSeq)	NA	11.53	5.72	0.49
F11	coagulation factor XI	2160	11.50	3.10	0.41

Table 1.1: Top 25 genes in the oligodendrocyte-enriched gene set of human H0351.2001 sorted by PC1 score.

human data we used, we interpret the WM/GM transcript ratios as variations in cell type proportions within grey matter regions.

We visualized the PC1 loadings on the schematic image of the brain using the Allen Brain Explorer 2 (see Methods) (Figure 1.3). Regions where there is high neuron marker expression include inferior frontal gyrus, CA2 and temporal pole. Regions where there is high oligodendrocyte marker expression include globus pallidus, putamen and head of caudate nucleus, in agreement with the enrichment of these regions in myelinated axons [14].

In addition, we calculated the ratio between "oligodendrocyte-enriched" PC1 markers and "neuron-enriched PC1" markers and compared it to the glia to neuron ratio measurements performed by Azevedo et

Gene symbol	Gene description	Entrez ID	PC1	Mean	StdDev
RNF41	ring finger protein 41	10193	-18.90	6.48	0.39
ARF5	ADP-ribosylation factor 5	381	-18.87	7.80	0.47
A_32_P86533	AGILENT probe A_32_P86533 (non-RefSeq)	NA	-18.49	8.19	0.64
GSTA4	glutathione S-transferase alpha 4	2941	-18.49	7.76	0.34
MMS19	MMS19 nucleotide excision repair homolog ...	64210	-18.26	6.51	0.40
TMEM59L	transmembrane protein 59-like	25789	-18.26	6.55	0.71
CLTA	clathrin, light chain A	1211	-18.26	9.45	0.38
UBE2K	ubiquitin-conjugating enzyme E2K ...	3093	-18.24	9.58	0.36
AP2A2	adaptor-related protein complex 2, alpha 2 subunit	161	-18.22	8.50	0.36
NMNAT2	nicotinamide nucleotide adenylyltransferase 2	23057	-18.11	8.75	0.53
LCMT1	leucine carboxyl methyltransferase 1	51451	-18.07	8.87	0.34
PDCD2L	programmed cell death 2-like	84306	-17.99	6.54	0.41
LOC727967	similar to block of proliferation 1	727967	-17.92	6.31	0.44
HAGH	hydroxyacylglutathione hydrolase	3029	-17.89	8.14	0.50
DHX30	DEAH (Asp-Glu-Ala-His) box polypeptide 30	22907	-17.88	6.94	0.42
RTN1	reticulon 1	6252	-17.87	9.77	0.61
CCT2	chaperonin containing TCP1, subunit 2 (beta)	10576	-17.81	10.38	0.37
PI4KA	phosphatidylinositol 4-kinase, catalytic, alpha	5297	-17.80	8.38	0.41
IARS	isoleucyl-tRNA synthetase	3376	-17.78	7.50	0.38
ABHD14A	abhydrolase domain containing 14A	25864	-17.76	7.74	0.42
PLD3	phospholipase D family, member 3	23646	-17.76	8.49	0.61
ATP6AP1	ATPase, H+ transporting, lysosomal accessory protein 1	537	-17.74	8.99	0.45
C19orf62	chromosome 19 open reading frame 62	29086	-17.66	7.36	0.43
RAB24	RAB24, member RAS oncogene family	53917	-17.64	8.06	0.38
KLHDC3	kelch domain containing 3	116138	-17.63	6.60	0.43

Table 1.2: Top 25 genes in the neuron-enriched gene set of human H0351.2001 sorted by PC1 score.

Brain division	H0351.2001	H0351.2002	Azevedo et al. 2009
Cerebellum	-0.032 $\pm$ 0.062	0.013 $\pm$ 0.045	0.23
Cerebral grey matter	-0.00069 $\pm$ 0.058	-0.012 $\pm$ 0.059	1.48
Rest of the brain	0.016 $\pm$ 0.042	0.015 $\pm$ 0.042	11.35

Table 1.3: PC1 brain loadings (mean $\pm$ standard deviation) of the two AIBS human datasets and measured glia to neuron ratio from Azevedo et al. (2009) [2] in cerebellum, cerebral grey matter and remaining brain regions.

al (2009) [2]. In agreement, in H0351.2001, we find that the human cerebellum, cerebral grey matter and the rest of the brain samples show increasing glia to neuron ratio respectively (Table 1.3). In H0351.2002, the cerebellum shows higher glia to neuron ratio than cerebral grey matter which may be due to individual variability or technical artifacts.

Together, these results suggest that gene expression variance in the human brain can partly be explained by variations in cell type composition, though we cannot exclude contributions from changes in expression

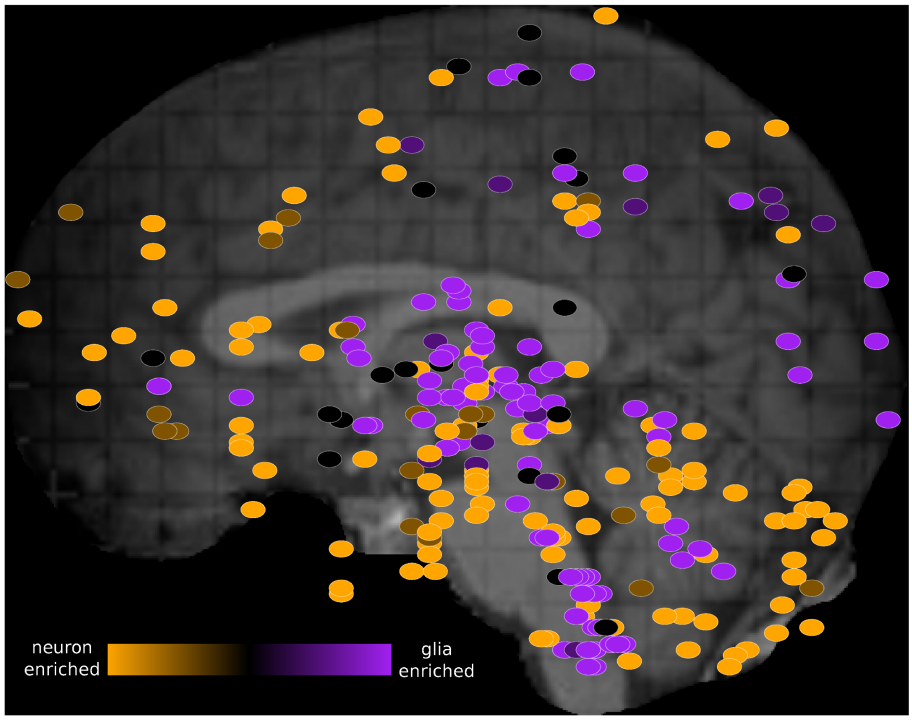


Figure 1.3: Schematic view of the H0351.2001 human brain showing oligodendrocyte-neuron PC1 marker ratio within each brain region sample. The brain PC1 loadings were obtained from the rotation attribute result object of the "prcomp" function in R. PC1 brain loadings range from 0.03 (orange) to -0.04 (purple) which suggest increasing glia-neuron ratio. Primary and secondary axes correspond to the mri_z and mri_y coordinates respectively. These dots were manually overlaid onto a brain image from the Allen Brain Explorer 2 software (http://mouse.brain-map.org/static/brainexplorer). In order to visualize subcortical region samples, we have hidden the visualization of the left cerebral hemisphere which causes some cortical samples (such as part of the left temporal cortex) to appear outside of the brain.

within cell types.

1.3.3 Orthologous genes with positively correlated expression patterns are enriched in cell type markers

In addition to identifying dominant gene expression patterns within each species, we also performed a comparison of gene expression patterns between orthologous gene and brain region samples in mouse and in human AIBS data, focusing on the H0351.2001 dataset which we deem to be the higher quality of the two data sets based on the clustering of cerebellar regions described above. Within data sets, we found that regional expression patterns show greater homogeneity in human (mean Spearman rho = 0.98 $\pm$ 0.0079)

than in mouse (mean Spearman rho = 0.86 $\pm$ 0.022). That is, expression patterns across mouse brain regions were apparently more variable than across human brain regions, possibly for technical reasons. Next, we measured the conservation of gene expression patterns by measuring the correlation for each homologous gene across matched brain regions. Finally, we compared our results with those of other studies by performing enrichment analyses on genes ranked by the strength of their correlation between species.

To prepare gene expression matrices of the same size, we limited the analysis to genes expressed above the 25^{th} percentile in both species and brain regions which could be matched between mouse and human, resulting in 7,911 genes and 58 subcortical brain regions (see Methods). Major brain regions include the hippocampal formation, cerebral nuclei, thalamus, epithalamus, hypothalamus, midbrain regions, pons, medulla and cerebellum. In this filtered data set, we saw consistent cell type marker enrichment in the PC scores in both mouse and human which indicates that the filtering process did not have a large effect on the data with respect to the patterns described in the previous section (data not shown).

We calculated the Spearman rank correlation between pairs of homologous brain regions and found statistically significant positive correlations (mean Spearman rho = 0.31 $\pm$ 0.031, $P < 0.0001$). The three most similar brain regions include Ammon's horn (rho = 0.40), dentate gyrus (rho = 0.38), and subiculum (rho = 0.35). Brain regions with the poorest correlation include nucleus raphe pontis (rho = 0.21), gracile nucleus (rho = 0.25), and pallidum (rho = 0.25).

In terms of genes, we measured the Spearman rank correlation of each homologous gene's expression levels across matched brain regions. We used these correlation values to rank homologous genes, such that those genes with conserved expression patterns are positively correlated while genes with discordant patterns have either no correlation or are anti-correlated across matched regions. We observed a positive skew in the correlation distribution (mean rho = 0.074, min rho = -0.57, max rho = 0.73; Figure 1.4). To verify whether this skew is significant or not, we compared this correlation distribution with a random distribution obtained by shuffling gene labels (see Methods). There are 53 fewer genes with correlation below -0.30 when compared to random while there are 645 more genes with correlation above 0.30 when compared to random. Together, this indicates that there are more genes with similar expression patterns than not. The top 25 genes with the most positively and negatively correlated gene expression between mouse and human are shown in Table 1.4 and Table 1.5 respectively. Figure 1.5 shows examples of genes with positively and negatively correlated expression levels across brain regions. We note that when only a few ($\sim$10) especially highly correlated brain regions were selected, the distribution became more positively skewed (data not shown), suggesting that more focused comparisons might provide higher resolution results, but it was not obvious how to choose such regions *a priori*.

Since we observed cell-type marker enrichment in the gene expression patterns for each species independently, we hypothesize that homologous genes that show conserved expression patterns are also enriched for cell type markers. We measured the cell-type marker enrichment using the ranked list of homologus genes, annotated by cell type in Cahoy et al. (2008) [9] (see Methods). In line with our hypothesis, our results show that expression patterns of homologous genes are enriched for neuronal (AUC = 0.74) and oligodendrocyte (AUC = 0.71) markers, but not astrocyte (AUC = 0.53) markers (see Methods). We interpret this as suggesting that neuronal markers and oligodendrocyte markers are generally more conserved in expression

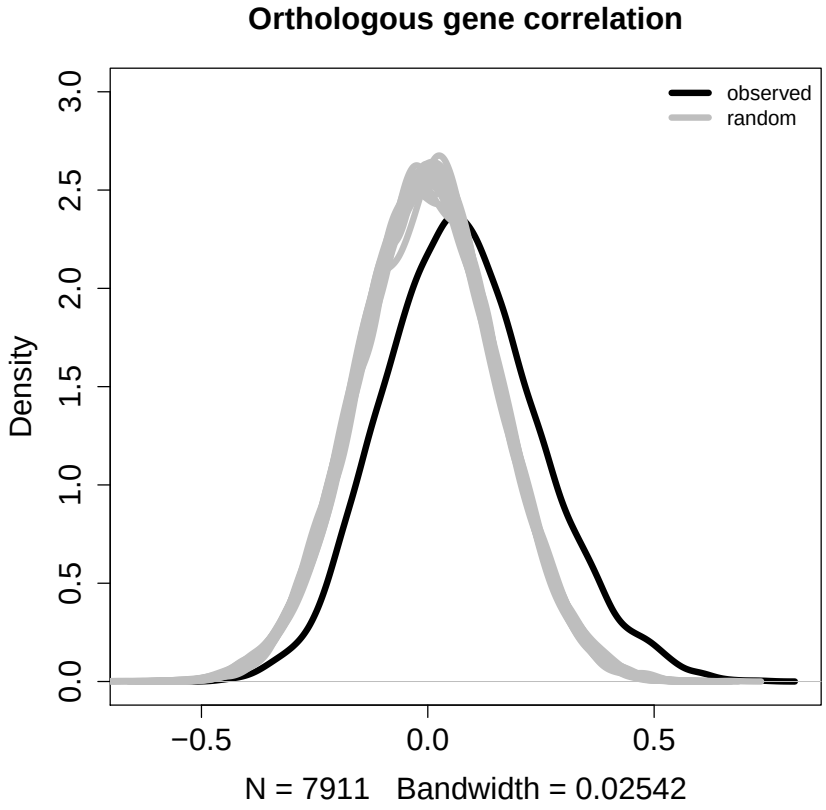


Figure 1.4: Correlation distribution between orthologous genes that are expressed. Correlation distribution is skewed towards the positive compared to random where human gene labels were shuffled without replacement.

patterns in comparison to the non-cell type markers (Table 1.4), consistent with the findings of Miller et al. [28]. In contrast, astrocyte markers were relatively poorly conserved overall, some astrocyte markers show positively correlated patterns (e.g. *AGT*, *GFAP*) while others show negatively correlated patterns (e.g. *SLC27A1*, *SCARA3*) between mouse and human (Table 1.4 and Table 1.5). We found similar results when using less stringent criterion for selecting genes from the Cahoy data (at least 5x enrichment instead of 10x, data not shown). We performed a Gene Ontology (GO) enrichment analysis for the top 100 genes with the most positively and negatively correlated patterns. Those genes with similar expression patterns are significantly enriched for Gene Ontology (GO) biological processes such as ion transport (GO:0006811), transmission of nerve impulse (GO:0019226), and synaptic transmission (GO:0007268). The top 100 genes with the most negatively correlated patterns are enriched in biological processes such as negative regulation of homeostatic process (GO:0032845), fatty acid oxidation (GO:0019395), and macromolecule catabolic process (GO:0009057).

Discordant expression patterns between mouse and human orthologs might indicate interesting functional divergences. We identified only 78 genes with reasonably strong negative correlations between mouse

Gene symbol	Gene description	rho	Mm Expr	Hs Expr
KCNC1	potassium voltage-gated channel, Shaw-related subfamily, member 1	0.73	9.93	3.84
SLC17A6	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 6	0.73	7.62	4.65
ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)	0.69	3.71	5.61
PCP4	Purkinje cell protein 4	0.66	8.16	4.69
GABBR2	gamma-aminobutyric acid (GABA) B receptor, 2	0.66	13.81	4.38
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit	0.65	2.63	1.81
CAMK2D	calcium/calmodulin-dependent protein kinase II delta	0.63	12.37	3.42
OSBPL5	oxysterol binding protein-like 5	0.63	3.17	2.67
VAT1	vesicle amine transport protein 1 homolog (T. californica)	0.62	1.76	1.73
SLC8A1	solute carrier family 8 (sodium/calcium exchanger), member 1	0.62	7.77	3.70
PLCB4	phospholipase C, beta 4	0.61	7.81	4.30
FOXP2	forkhead box P2	0.61	1.90	1.97
SPOCK1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 1	0.61	12.21	3.36
C20orf103	chromosome 20 open reading frame 103	0.61	3.47	3.38
KCNQ3	potassium voltage-gated channel, KQT-like subfamily, member 3	0.61	4.16	3.82
GNG4	guanine nucleotide binding protein (G protein), gamma 4	0.60	0.19	0.24
HTR1A	5-hydroxytryptamine (serotonin) receptor 1A	0.60	1.27	0.85
ZMAT4	zinc finger, matrin type 4	0.60	2.62	2.80
ADAM11	ADAM metallopeptidase domain 11	0.60	10.30	3.41
NRN1	neuritin 1	0.60	9.97	5.40
NTNG1	netrin G1	0.59	6.73	6.15
ST8SIA5	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 5	0.59	5.81	3.90
HCN1	hyperpolarization activated cyclic nucleotide-gated potassium channel 1	0.59	2.62	2.39
PCDH11X	protocadherin 11 X-linked	0.59	0.93	0.62
SYN2	synapsin II	0.59	10.29	4.76

Table 1.4: Top 25 genes with similar expression patterns between mouse (Mm) and human H0351.2001 (Hs) sorted by Spearman rank correlation (rho) with $q < 0.01$. Expr corresponds to the mean expression level across brain regions.

and human (rho $<$ -0.3). To seek supporting evidence for these and other negative correlations, we compared our findings to two previous mouse-human comparisons. Zeng et al. (2012) identified 73 genes with patterns considered discordant in the neocortex, including differences in laminar distribution [45]. Of these, 12 are negatively correlated in our study, including one of the 78 meeting a threshold of -0.3 (*SLC6A12* rho = -0.31; Table 1.6). Miller and colleagues identified 49 "human-specific" cell-type markers using meta-analysis of microarray data, of which fourteen are negatively correlated in our analysis, of which two are below -0.3 (*KIAA0174* rho = -0.36 and *ADK* rho = -0.32; Table 1.7). Thus despite major differences in methodology and brain regions considered, some previous reports of mouse-human differences are supported by our analysis.

1.4 Discussion

We studied the dominant gene expression patterns across the human brain and observed similar complementarity between "neuron/oligodendrocyte" enriched patterns as we previously identified in the mouse [17].

Gene symbol	Gene description	rho	q-value	Mm Expr	Hs Expr
TMEM2	transmembrane protein 2	-0.57	0.001	0.28	0.95
ABCA8	ATP-binding cassette, sub-family A (ABC1), member 8	-0.45	0.020	0.40	0.50
RPIA	ribose 5-phosphate isomerase A	-0.44	0.026	0.26	0.57
USP28	ubiquitin specific peptidase 28	-0.44	0.029	2.26	1.55
WARS2	tryptophanyl tRNA synthetase 2, mitochondrial	-0.43	0.030	0.46	0.38
SMOC2	SPARC related modular calcium binding 2	-0.42	0.037	0.58	1.16
CROT	carnitine O-octanoyltransferase	-0.41	0.039	2.20	1.54
KIAA1279	KIAA1279	-0.41	0.044	7.40	3.66
ACOX2	acyl-CoA oxidase 2, branched chain	-0.40	0.068	0.49	0.42
AGGF1	angiogenic factor with G patch and FHA domains 1	-0.39	0.079	0.28	0.11
MRPS34	mitochondrial ribosomal protein S34	-0.39	0.062	2.17	1.20
TMLHE	trimethyllysine hydroxylase, epsilon	-0.39	0.057	2.62	1.44
CTR9	Ctr9, Paf1/RNA polymerase II complex component, ...	-0.38	0.072	2.86	1.60
TNFRSF11B	tumor necrosis factor receptor superfamily, member 11b	-0.38	0.097	0.25	0.60
C2orf29	chromosome 2 open reading frame 29	-0.38	0.073	1.40	0.75
NCF4	neutrophil cytosolic factor 4, 40kDa	-0.38	0.078	0.21	0.20
PAICS	phosphoribosylaminoimidazole carboxylase, ...	-0.38	0.075	1.02	0.94
CEP164	centrosomal protein 164kDa	-0.37	0.078	0.91	0.61
CECR5	cat eye syndrome chromosome region, candidate 5	-0.37	0.098	2.40	1.02
KIAA0174	KIAA0174	-0.36	0.100	4.34	2.09
ATRX	alpha thalassemia/mental retardation syndrome X-linked	-0.36	0.090	6.84	1.85
DNAJC5	DnaJ (Hsp40) homolog, subfamily C, member 5	-0.36	0.097	15.11	3.70
FRMD4A	FERM domain containing 4A	-0.36	0.097	5.73	1.89
RAP1GAP	RAP1 GTPase activating protein	-0.36	0.092	14.17	3.32
CDK4	cyclin-dependent kinase 4	-0.36	0.590	1.35	1.23

Table 1.5: Top 25 genes with anti-correlated expression patterns between mouse (Mm) and human H0351.2001 (Hs) sorted by Spearman rank correlation (rho). Expr corresponds to the mean expression level across brain regions.

Our analysis also shows that *in situ* data from mouse can be meaningfully compared to microarray data from human. As Lee et al. (2008) pointed out, comparisons between ISH and microarray data are challenging due to technical differences such as probe sequence sensitivity and specificity, dynamic range and normalization and mapping of ISH data [23]. Despite these technical differences, we report gene expression pattern similarities as exemplified by the anti-correlation between neuron and oligodendrocyte enriched patterns. Our interpretation of the cell-type enriched pattern in human is similar to our previous interpretation in mouse [17]. A simple explanation is that neurons and glia vary in inverse proportions across brain regions in both human and mouse, which shows an anterior-posterior gradient (Figure 1.3). However, it is difficult to fully verify this because we currently have limited information on the details of the size and proportions of cell types within each brain region sampled.

The strength of the cell type marker enrichment suggests that many other genes, while not reported as cell type markers by Cahoy et al. (2008), are likely to be expressed in a cell type enriched manner. Genes in this category include ones we predict based on our readings to be expressed in neurons such as neural

Gene symbol	Gene description	rho	q-value	Mm Expr	Hs Expr
SLC6A12	solute carrier family 6 (neurotransmitter transporter, betaine/GABA), member 12	-3.13E-01	0.18	0.26	4.43
GPR85	G protein-coupled receptor 85	-2.66E-01	0.29	4.94	5.48
CACNG8	calcium channel, voltage-dependent, gamma subunit 8	-1.64E-01	0.60	3.07	5.21
COL6A1	collagen, type VI, alpha 1	-1.42E-01	0.68	2.47	6.42
KIAA1370	KIAA1370	-1.27E-01	0.72	1.71	5.56
CLCN2	chloride channel 2	-1.08E-01	0.77	1.83	4.09
MFGE8	milk fat globule-EGF factor 8 protein	-8.71E-02	0.82	0.75	3.64
PDYN	prodynorphin	-6.68E-02	0.87	2.31	3.75
PCDH20	protocadherin 20	-3.60E-02	0.94	2.45	4.95
LGALS1	lectin, galactoside-binding, soluble, 1	-8.12E-03	0.99	1.44	9.01
SNCG	synuclein, gamma (breast cancer-specific protein 1)	-2.37E-03	1.00	6.76	9.92
SLC6A7	solute carrier family 6 (neurotransmitter transporter, L-proline), member 7	-4.31E-04	1.00	4.49	3.75

Table 1.6: Negatively correlated genes that show discordant patterns in Zeng et al. Genes are sorted by increasing q-value. Mm Expr and Hs Expr correspond to the mean expression level across mouse and human brain regions respectively.

Gene symbol	Gene description	rho	q-value	Mm Expr	Hs Expr
KIAA0174	KIAA0174	-3.62E-01	0.10	4.34	5.17
ADK	adenosine kinase	-3.21E-01	0.15	5.47	7.89
P2RX7	purinergic receptor P2X, ligand-gated ion channel, 7	-2.90E-01	0.23	0.25	6.41
HSPA8	heat shock 70kDa protein 8	-2.03E-01	0.49	19.90	12.10
LEPROT	leptin receptor overlapping transcript	-9.69E-02	0.80	0.44	8.02
CBFB	core-binding factor, beta subunit	-8.12E-02	0.84	0.21	5.49
INPP1	inositol polyphosphate-1-phosphatase	-6.90E-02	0.87	4.61	6.09
TGFBR3	transforming growth factor, beta receptor III	-4.40E-02	0.92	2.38	5.70
COL4A5	collagen, type IV, alpha 5	-4.18E-02	0.92	0.29	5.37
RNF103	ring finger protein 103	-2.35E-02	0.96	2.59	7.92
UQCRC2	ubiquinol-cytochrome c reductase core protein II	-2.32E-02	0.96	11.48	7.10
PSEN1	presenilin 1	-1.71E-02	0.97	0.45	6.35
DYNC1I2	dynein, cytoplasmic 1, intermediate chain 2	-8.67E-03	0.99	12.95	9.65
CHD1L	chromodomain helicase DNA binding protein 1-like	-8.31E-04	1.00	0.21	6.39

Table 1.7: Negatively correlated genes that show discordant patterns in Miller et al. Genes are sorted by increasing q-value. Mm Expr and Hs Expr correspond to the mean expression level across mouse and human brain regions respectively

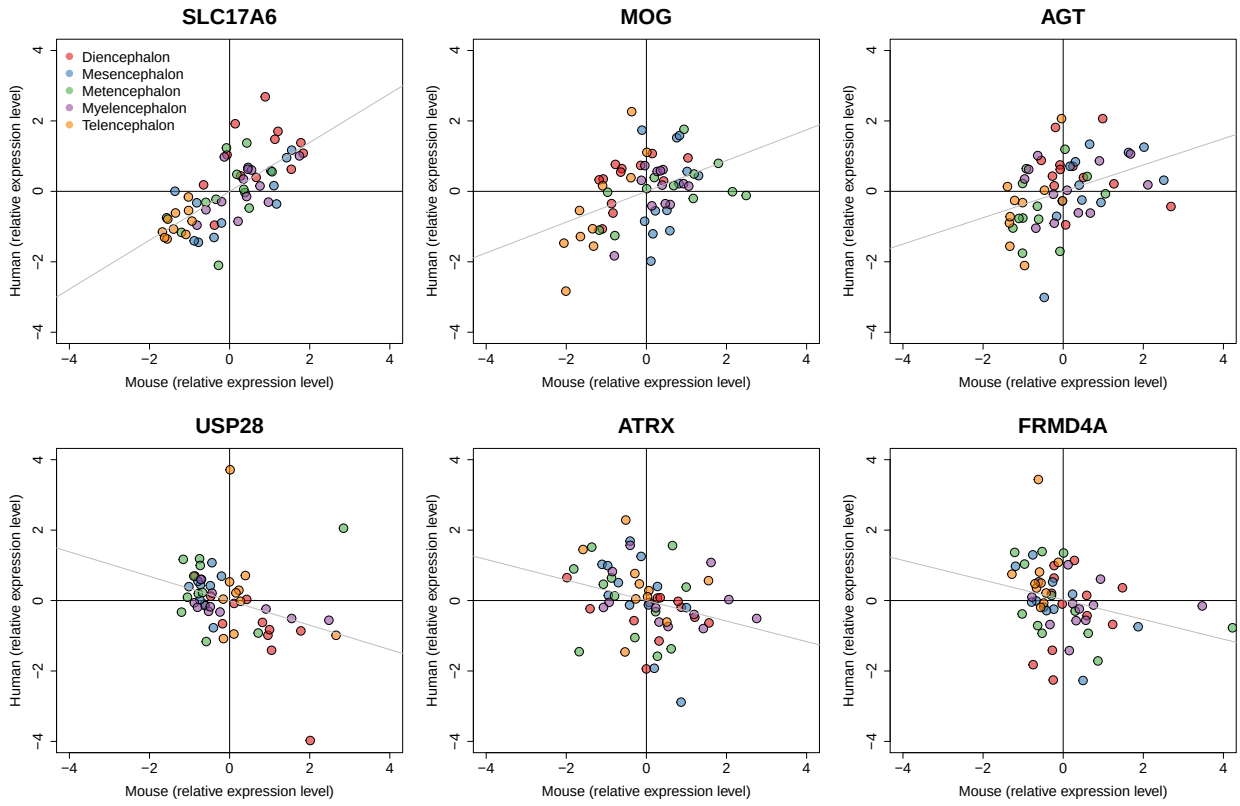


Figure 1.5: Examples of positively and negatively correlated gene expression patterns between mouse and human H0351.2001. Dots represent brain region samples colored by major brain divisions. Three genes with expression patterns that are positively correlated are shown at the top (*SLC17A6* rho = 0.73, *MOG* rho = 0.39, *AGT* rho = 0.44) while negatively correlated gene expression patterns are shown at the bottom (*USP28* rho = -0.44, *ATRX* rho = -0.36, *FRMD4A* rho = -0.36). All six genes have q-values < 0.3. Expression levels are scaled and centered at zero for visualization.

epidermal growth factor-like 2 (*NELL2*), reticulon 4 receptor (*RTN4R*), potassium channel, subfamily K, member 1 (*KCNK1*), and glutaminase (*GLS*) as well as ones we predict to be expressed in oligodendrocytes such as chloride intracellular channel 4 (*CLIC4*), crystallin, alpha B (*CRYAB*), prostaglandin D2 synthase 21kDa (*PTGDS*), quinoid dihydropteridine reductase (*QDPR*) and G protein-coupled receptor, family C, group 5, member B (*GPRC5B*). Using a literature review, we have confirmed some of these, suggesting their absence from the lists given by Cahoy et al. to be due to technical factors or the choice of cells used in their study. For example, ISH of the adult mouse and rat brains show *RTN4R* (reticulon 4 receptor or Nogo receptor) is strongly expressed within neurons of the neocortex, hippocampal formation, and granule cells of the cerebellum [21]. On the other hand, it is also apparent that what we term the "oligodendrocyte-enriched" and "neuron-enriched" patterns are not purely populated by genes specific for those cell types. For example in the H0351.2001 dataset, *TMEM163*, *CNTN1*, and *TMEM2* are Cahoy oligodendrocyte marker genes but are found close to neuronal markers in our PCA, while the converse is true for the neuronal markers *ST8SIA2*

and *GPR12*. This complexity presumably in part reflects sub-populations of neurons which have a different physical or regulatory relationship to glial cells than those which occur in the "neuron-enriched" pattern, or vice versa.

A second goal of our study was to identify similarities and differences in expression pattern between mouse and human brains. Our overall conclusion is that the similarities vastly outnumber the differences. We found that the similarities are most striking for genes which are known to be enriched in neurons and oligodendrocytes (Table 1.4 and Figure 1.5). In contrast, markers of astrocytes demonstrate more differences between mouse and human. In mouse, astrocyte markers were equally represented in both "neuron-enriched" and "oligodendrocyte-enriched" patterns [17]. In contrast, in the human data, astrocyte markers coordinately vary in expression levels considerably across regions. Astrocytes support the metabolically demanding tasks of neurons by recycling neurotransmitters and maintaining ion homeostasis in the brain [8]. The enrichment seen in humans could be caused by the increased complexity found only in human astrocytes [32] or by the higher astrocyte to neuron ratio observed with increasing brain complexity [29]. Aside from astrocyte markers, we found evidence for other genes showing discordant expression patterns. For example, the mouse *ATRX* (alpha thalassemia/mental retardation syndrome X-linked) expression pattern is negatively correlated with human (rho = -0.36, q = 0.09) (Figure 1.5 and Table 1.5). In adult mouse, this gene has a higher expression in the medulla compared to the amygdala, while the opposite is true in human. We caution that from the available data it is difficult to determine which of the differences we observe reflect true biological differences (e.g., different species isoforms), and which are due to differences between ISH and microarray. However, the partial overlap of our negative correlations with previous reports of mouse-human differences [28, 45] suggests that some the other differences we report are worthy of further study.

In summary, using PCA, we provide a candidate list of cell type markers which could be useful for targeting specific cell types or specific regional patterns of interest. In addition, we report correlations for the regional expression of genes between mouse and human which can be useful in the development of mouse disease models or in the study of the molecular evolution of the brain. Future studies that explore the different regulatory mechanisms of genes with discordant expression patterns might provide insights into the evolution of brain structure and function. Furthermore, future high-resolution large-scale studies that examine gene expression in developing mouse and human brain will uncover genes that are only active in early development and thus provide a better understanding of human brain evolution.

Chapter 2

Large-scale survey of tissue types and experimental conditions across datasets

2.1 Introduction

Genes are regulated in different ways. For example, housekeeping genes (e.g. *HSP90* [18]) are expressed at similar levels in nearly every cell under normal conditions. On the other hand, other genes are known to be expressed in only a few cell types. For example, the neurotransmitter GABA (gamma-aminobutyric acid) is found in both synaptic vesicles in GABAergic neurons in the brain and also in synaptic like microvesicles secreted from pancreatic beta-cells [7, 15]. Whether GABA is also expressed in other cell types remains unclear. Aside from genes differentially expressed (DE) in different cell types, genes are also DE under different experimental conditions (e.g. disease state, sampling time point or drug treatment).

Sample groups are often chosen with a specific biological question in mind. For example, control samples are compared to neurological disease samples with the goal of identifying disease-associated genes [36]. However, this targeted approach limits our understanding to comparisons within these two groups that were selected *a priori*. A subset of DE genes may also be DE in seemingly unrelated diseases. An understanding of how diseases are related is helpful for drug re-purposing where old drugs are applied to new diseases. It may also be useful to apply knowledge from a well-studied disease to a less-studied one provided that both diseases share a common biological pathway.

The idea of integrating microarray datasets is not new. Large-scale microarray studies such as Lukk et al. [27] (Array Expression ID: E-MTAB-62) have combined thousands of datasets to identify DE genes between biological groups such as cell type, tissue type, disease state and cell lines. Lukk et al. grouped samples by using sample annotations which were provided by the submitters of the dataset. They identified tissue-enriched genes by comparing the gene's expression in each tissue group against the global mean expression level.

Aside from using sample annotations, the gene expression content can also be used as well. Engreitz et al. [12] formulated this problem in a data-driven fashion. First, they created a library to search against by calculating the DE profile of each experiment. Next, the DE profile of the input query was compared against

the library of DE profiles based on a weighted Pearson correlation similarity measure.

To better explore gene function in human and identify which experimental conditions affect it, we compiled a list of publicly available human microarray datasets in an unsupervised manner. From these datasets, we created matrices that represent gene expression levels (EE) and DE levels. We compared these two matrices in two ways. First, we investigated if tissue-enriched genes cluster by tissue type. Second, we identified modules that are enriched for biological processes. The preliminary work and conclusions presented here will provide additional insights into future large-scale gene expression investigations.

2.2 Methods

2.2.1 Data overview and pre-processing

Both EE and DE matrices were obtained from the human microarray datasets deposited in Gemma [46]. We term each comparison between two sample groups in each dataset as a "result set (RS)", e.g. GSE12860 has two RS, control vs. treatment and rheumatoid arthritis fibroblast vs. normal fibroblast. Matrix rows correspond to genes annotated in the microarray platform. EE columns correspond to datasets while DE columns correspond to RS. Since most datasets have only one RS (91%), we refer to RS by their Gene Expression Omnibus (GEO) series ID in our results [3]. We measured gene expression levels at the resolution of each dataset. Expression levels in the EE matrix were obtained by calling the "dEDVRank" webservice in Gemma where genes were averaged across all samples in each dataset and normalized from zero (low expression) to one (high expression) across all genes in the dataset. Since we are also interested in gene expression changes within a dataset, the Gemma framework was used to calculate the p-values in the DE matrix. Briefly, a one-way analysis of variance (ANOVA) was performed for all the genes in each RS.

The following filters were applied to the human microarray studies. First, we selected RS that were derived from the GPL570 and GPL96 platforms which corresponds to the Affymetrix Human Genome U133 Plus 2.0 Array and U133A Array respectively. These platforms are the most common human microarray platforms in Gemma. Restricting our analysis to within these two similar platforms also reduces variability between different microarray platforms. Moreover, since these platforms have similar probe sets, there were fewer instances of probe sets with missing values across RS. Those RS with missing values in more than 10% of the total number of probe sets were excluded. Missing values could be attributed to the filtering process applied by submitters of the dataset. We also checked for missing values across rows. Those probe sets with missing values in more than 10% of the total RS were excluded. These probe sets mapped to RNA genes and pseudogenes (e.g. *MIR4680*, *TEN1-CDK3*, and *A2MP1*). Moreover, we chose RS with annotations that relate to disease state, treatment, and sampling time point, excluding organism part (e.g., different brain regions) from our analysis. Aside from missing values, we also removed those RS with too many (q-values less than 0.05 for more than 50% of the genes, e.g. GSE16385) or too few (no q-values less than 0.3, e.g. GSE12644) DE probe sets. We excluded RS (e.g. GSE13501 and GSE7753 (Spearman rho = 0.66, $P < 0.001$) that we deemed as outliers due to their high correlations with other RS across all probe sets (Spearman rho $> 75^{th}$ percentile). We believe that these highly correlated RS will always be clustered together regardless of which genes were selected and therefore are not as informative. Finally, the

p-value histogram distribution of DE genes are expected to be similar to a beta-uniform mixture distribution [37]. We developed a simple heuristic to evaluate this metric where we subtracted the gene counts from the thirteenth histogram bin from the gene counts of the first histogram bin. P-values were binned every 0.1. Those datasets that differ by more than 100 gene counts were excluded from our analysis. This eliminated 28 RS from the DE matrix (e.g. GSE16385 and GSE11839). Finally, we filtered the EE matrix by selecting those genes and datasets from the filtered DE matrix.

2.2.2 Tissue-enriched genes

The list of tissue-enriched genes were selected from Lukk et al. [27]. Briefly, Lukk et al. integrated microarray datasets to from a final expression matrix of ∼14,000 genes times ∼5,000 samples. They compared each tissue group to the global mean by performing a one-way ANOVA. From this list of genes, we selected those genes that were upregulated (t-statistic greater than the 75^{th} percentile) in the brain, muscle and hematopoietic system meta groups. We call these genes as "tissue-enriched" and annotated these genes as brain-enriched, muscle-enriched and hematopoietic system-enriched genes respectively. Moreover, the two genes (*PDE4DIP* and *ARHGAP19*) that were upregulated in more than one tissue type were excluded from our analysis. RS were grouped as brain, muscle and "other" by manually curating the sample tissue type annotation. We combined those RS with the same dataset by averaging p-values across tissue-enriched genes. Finally, the tissue-enriched DE and EE matrices have the same number of rows (844 genes) and columns (163 datasets).

2.2.3 Biclustering

The Iterative Signature Algorithm (ISA) (isa2 version 0.3.1-1 R package http://cran.r-project.org/web/packages/isa2/index.html) was used to identify modules in both the EE and DE matrices [5]. First, DE p-values were $-log_{10}$ transformed prior to biclustering. Modules were created by setting the random seed to 1, number of seeds to 100 and direction up for both rows and columns. We adjusted the parameters to meet the following criteria: first, the number of clusters must be small (< 50) and second, the size of the clusters must be reasonable (∼100 rows and ∼10 columns). DE modules were identified using a row threshold of 4 and a column threshold of 1. EE modules were identified using a row and column thresholds of 2.5 and 1.5 respectively. For both cases, removal of similar modules was performed by calling the "isa.unique" function with a correlation limit of 0.6.

2.2.4 GO enrichment

Gene Ontology (GO) enrichment analysis was performed using the topGO (Version 2.8.0) Bioconductor R package (http://www.bioconductor.org/packages/2.10/bioc/html/topGO.html). We used all the human gene annotations from the org.Hs.eg.db database in Bioconductor as our background gene list. For each module, we performed the Biological Process GO enrichment by selecting all the genes in the module as our input genes. The classic algorithm was run using the Fisher's Exact Test as the test statistic. The classic algorithm scores each GO group independently of its neighbouring GO groups and is also independent of the test statistic used [1]. The top (most significant) GO group was assigned to each module.

2.2.5 Statistical analysis

Agglomerative hierarchical clustering was performed using the Ward method with Euclidean distances [13]. In each iteration, the Ward method minimizes the variance within newly formed clusters. Correlations were calculated using Spearman rank correlations and the default two sided alternative hypothesis was used for calculating Spearman rho p-values.

2.3 Results

2.3.1 Experimental design and analysis

Our workflow is summarized in Figure 2.1. We performed strict quality checks and selected high quality human microarray datasets deposited in Gemma, a database and framework for analyzing expression profiling studies (see Methods) [46]. Initially, the DE matrix had 64,260 genes (64,594 probe sets) across 1,298 RS. Further data processing and quality checks reduced the matrix to 12,881 genes (12,892 probe sets) and 180 RS (see Methods). Genes with two probe sets included *CCR2, DDX39B, EIF2D, FXYD6-FXYD2, HSFX1, MICA, RAD21L1, RBMXL1, SBF1P1, TRMT1L* and *ZNF559-ZNF177*. We obtained the corresponding EE matrix by measuring the corresponding gene expression levels in each dataset. In our first analysis, we analysed the tissue-enriched subset of each matrix (844 genes) (see Methods). Next, using all 12,881 genes, we identified EE and DE modules by applying the ISA biclustering algorithm in an unsupervised manner. Each module was then annotated with a biological processes by applying GO enrichment analysis.

2.3.2 Tissue-enriched gene expression and differential expression

We asses the quality of the datasets by clustering gene expression and differential expression of tissue-enriched genes. We hypothesize that tissue-enriched genes are highly expressed in datasets that use the corresponding tissue as sample source. From the study of Lukk et al. [27], we found 844 tissue-enriched genes in our dataset (526 brain-enriched genes, 214 heamtopoeitic system-enriched genes and 104 muscle-enriched genes). In the EE matrix, there are 15 (9%) brain datatasets, 11 (7%) muscle datatasets and 137 (84%) "other" datatasets (Tables 2.1, 2.2 and 2.3 respectively). We found similar proportions of RS in the DE matrix (16 brain RS, 11 muscle RS and 153 "other" RS). Our results show that the EE matrix reflects experimental source tissue type more so than the DE matrix.

As shown in Figure 2.2a, muscle-enriched genes such as *MYL3*, *FXYD1*, and *DES* are highly expressed in datasets where muscles are used as the tissue source (e.g. GSE9103 and GSE1551; Table 2.2) with low expression in non-muscle datasets. Likewise, brain-enriched genes such as *S100B*, *NEFL*, and *GRIA2* are highly expressed in datasets where nervous tissues or cell lines are used (e.g. GSE1993 and GSE21858; Table 2.1) with low expression in non-nervous related datasets. This reflects the high-quality of our datasets such that the majority of tissue-enriched genes are highly expressed in datasets with similar tissue annotations. However, there are a few muscle-enriched genes *PDK2, PTP4A3, TMOD1*, brain-enriched genes *CALM3, PPP3CA, PPP3CB, CNP, B3GNT1, MAPK8IP3*, and hematopoietic system-enriched genes such as *RAB8A, DDX5, GNAI3* which are highly expressed in almost all datasets, except for a few cancer and

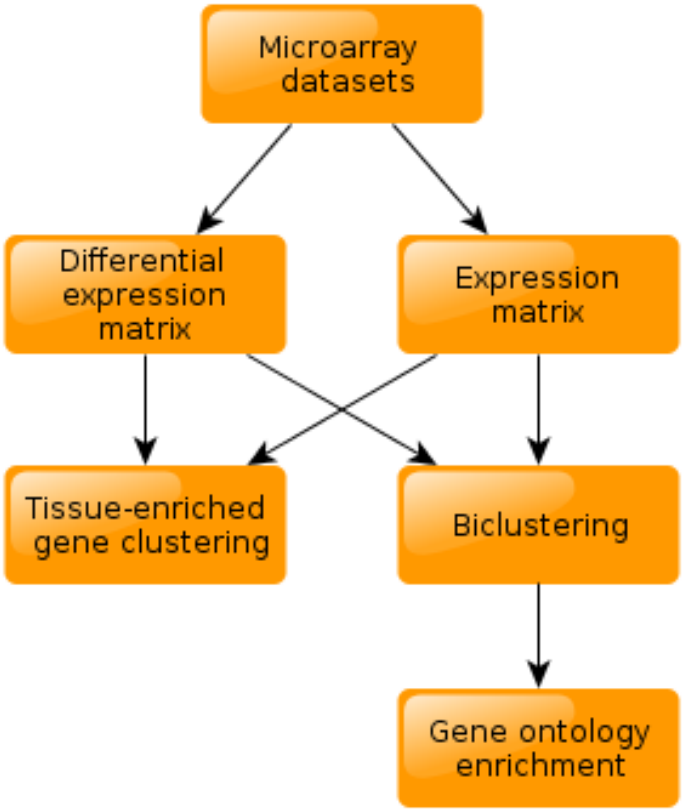


Figure 2.1: Experimental design and analysis

GEO.ID	Tissue source	Sample Size
GSE1993	Glioblastoma and Astrocytoma frozen primary human tissue	65
GSE21858	Frontal and temporal cortex	8
GSE5388	Human postmortem brain tissue	61
GSE11208	Peripheral ganglion	11
GSE5389	Human post-mortem brain tissue	21
GSE12460	Neuroblastic tumor	64
GSE24072	Fresh glioma tissue	32
GSE4773	SK-N-MC neuroblastoma cells	21
GSE1297	Hippocampal CA1 tissue	31
GSE19728	Normal brain tissue, glioblastoma, and astrocytoma	21
GSE17440	Frontal cortex	8
GSE7621	Postmortem human substantia nigra	25
GSE20168	Postmortem brain prefrontal cortex	29
GSE20141	Substantia Nigra pars compacta	18
GSE2732	Human brain neuronal SH-SY5Y cell lines	18

Table 2.1: Nervous tissue datasets. These datasets involve brain tissues or neuronal cell lines. Tissue sources were manually curated from GEO.

GEO.ID	Tissue source	Sample Size
GSE9103	quadriceps (Vastus Lateralis) muscle biopsy samples	40
GSE10161	cardiac biopsies	27
GSE1551	muscle biopsies	23
GSE24235	biceps brachii muscle biopsies	28
GSE1869	human heart	37
GSE3112	muscle biopsies	40
GSE11686	arm (Extensor and Flexor carpi radialis) muscles	16
GSE13070	quadriceps (Vastus Lateralis) muscle	364
GSE6798	quadriceps (Vastus Lateralis) muscle	29
GSE1145	human heart	107
GSE14901	quadriceps	72

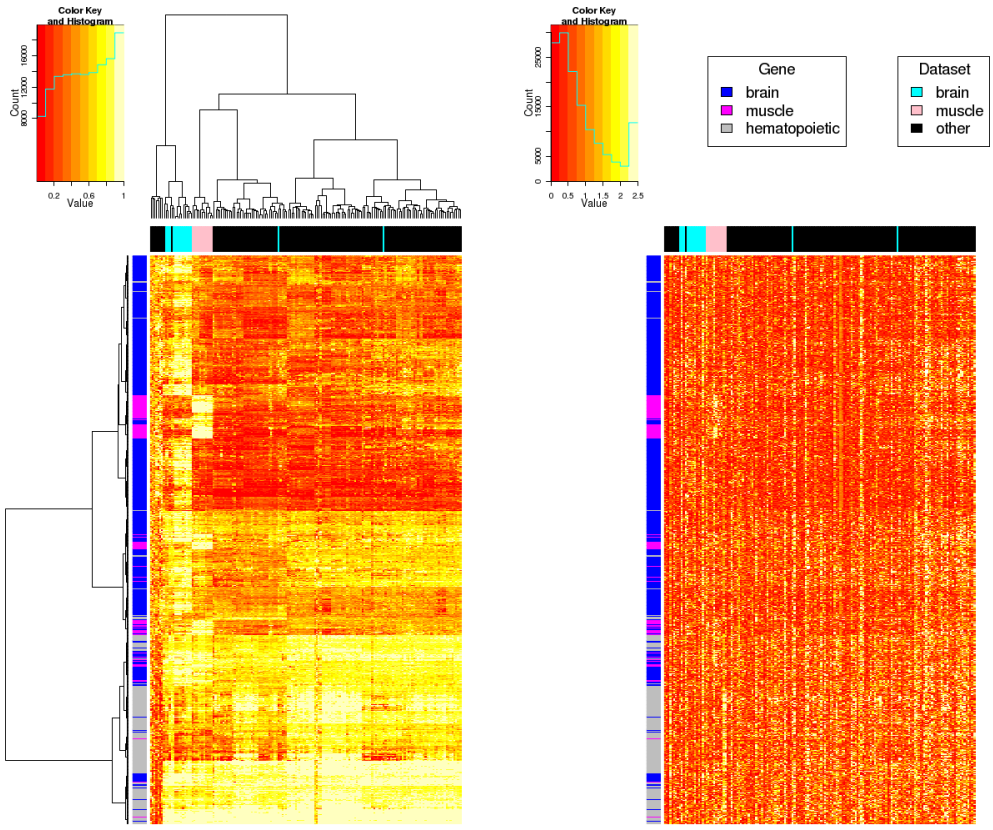
Table 2.2: Muscle tissue datasets. Tissue sources were manually curated from GEO.

GEO.ID	Name	RS
GSE16844	Integrated pathways for neutrophil recruitment and inflammation in leprosy	erythema nodosum leprosum VS Lepromatous Leprosy
GSE12860	Antirheumatic Drug Response in Human Chondrocytes: Potential Molecular Targets to Stimulate Cartilage Regeneration	Rheumatoid arthritis synovial fibroblasts vs. Normal donor synovial fibroblasts
GSE15132	Riboflavin depletion impairs cell proliferation in intestinal cells: Identification of mechanisms and consequences	Control VS Ribodeficient
GSE9971	CYP3A5 Gene Expression is Associated with Early Recurrence of Non-small Cell Lung Cancer	recurrent vs non-recurrent (control) non-small cell lung cancer
GSE15132	Riboflavin depletion impairs cell proliferation in intestinal cells: Identification of mechanisms and consequences	24 H VS 48 H VS 72 H
GSE26713	Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T-ALL	normal bone marrow control vs pediatric T-ALL

Table 2.3: Examples of datasets and RS from "other" tissues. These datasets were not classified as brain nor muscle.

inflammatory related datasets (e.g. GSE16844 and GSE12860; Table 2.4). These genes may be highly expressed in more than one tissue type under normal conditions. It may also be caused by mutations that upregulate gene expression.

In contrast to the EE matrix, it is not immediately obvious that tissue-enriched genes are differentially expressed in those RS that use the same tissue type (Figure 2.2b). At first glance, the EE and DE matrices do not seem to have anything in common. However, a closer inspection of the heatmap shows that the hematopoietic-system enriched cluster which corresponds to the bottom EE cluster, has more differentially expressed genes compared to genes in other clusters. On the other hand, most brain-enriched genes (top cluster) show little expression and are not DE in most datasets. Next, we calculated the Spearman rank



(a) EE matrix

(b) DE matrix

Figure 2.2: EE vs DE tissue-enriched matrices. (a) EE values are the relative average expression. EE values were binned from 0 (low) to 1 (high). EE rows and columns were clustered using the Ward method. (b) DE values were $-log_{10}$ transformed. DE values were binned from 0 (low) to 2.5 (high). DE rows and columns have the same order as that of the EE heatmap in (a).

GEO.ID	Name	Tissue source
GSE13849	Expression Signatures in Polyarticular JIA Show Heterogeneity and Offer a Molecular Classification of Disease Subsets	blood
GSE2405	NIH/NIAID Neutrophil Response to A. phagocytophilum	leukocytes
GSE20266	Salivary Transcriptomic and Proteomic Biomarkers for Breast Cancer Detection	saliva
GSE22377	mRNA expression data from human adenocarcinomas of the stomach	intestine
GSE11341	Lung selective gene responses to alveolar hypoxia	lung
GSE28796	Gene expression profiles of pretreatment biopsies from dose-dense-docetaxel-treated breast cancers	breast

Table 2.4: Cancer-related datasets that have low hematopoietic gene expression. Tissue sources were manually curated from GEO.

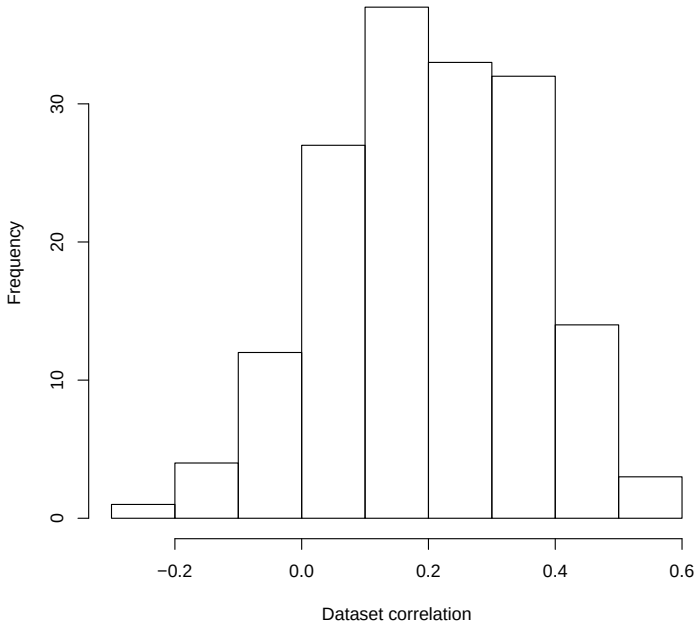


Figure 2.3: The distribution of Spearman rank correlations between EE and DE datasets.

correlation between matching EE and DE datasets across genes. Figure 2.3 shows the correlation distribution is positively skewed. We interpret this as more datasets show agreement in both EE and DE levels than not.

2.3.3 Modules enriched for biological processes

We hypothesize that genes and experimental datasets cluster together in modules because genes are either expressed in the same tissue type or share a common functional pathway. From the matrix of 12,881 genes, we are interested in finding modules that represent a common biological process. An EE module is a cluster of genes with high expression values across a subset of datasets while a DE module is a cluster of genes that are significantly differentially expressed across a subset of RS. A close inspection of these modules may provide novel insights toward genes or disease states.

Traditional clustering methods such as hierarchical clustering only allows one gene to exclusively belong to a single cluster. Moreover genes are clustered using all samples in the matrix while genes can be expressed or differentially expressed in only a subset of conditions. To address these issues, we applied the ISA biclustering algorithm [5]. Biclustering is a technique that overcomes these limitations by simultaneously clustering rows and columns. ISA has been used to identify regulatory modules (a sub-cluster of genes and samples) in the yeast transcriptional network [5]. The algorithm uses a set of random genes as input and iteratively selects genes and samples that are significantly different based on a threshold. The algorithm ends when the average expression across genes and samples have become very similar as indicated by a

GO.ID	Term	Significant	classicFisher	Module.ID	Rows	Cols
GO:0006936	muscle contraction	50	$<$ 1e-30	3	175	16
GO:0006955	immune response	53	$<$ 1e-30	4	148	19
GO:0019226	transmission of nerve impulse	36	4.7e-26	1	149	14
GO:0006968	cellular defense response	2	0.0016	2	39	14

Table 2.5: Top GO biological process in each module from the EE matrix. The table is sorted by increasing p-value from the Fisher's Exact Test statistic for gene over representation. The number of significant genes found in each group are shown. Rows represent the number of genes and Cols represent the number of datasets in each module.

high Pearson correlation. These set of genes and samples are now part of the same module. We applied the ISA algorithm to both EE and DE matrices separately. Finally, we applied GO enrichment analysis to each module for biological interpretation.

EE modules cluster by source tissue type

We identified four EE modules after applying ISA on the EE matrix (Table 2.5). On average, most modules have over 100 genes and over 10 datasets. Figure 2.4 displays the hierarchical clustering of EE modules based on overlapping genes. This clustering shows that there is greater overlap between brain and muscle related genes than hematopoietic system genes which is consistent with the clustering of biological groups in Supplementary Figure 4a of Lukk et al. [27].

We have chosen to take a closer look at Module 3, the muscle contraction module (GO:0006936) as an example. The Top 5 GO annotations for this module is available in Table 2.6. The expression levels of genes within this module is more homogeneous relative to randomly chosen genes outside this module (Figure 2.5). Datasets that belong to this module involve muscle or heart tissues (e.g., GSE1551 and GSE13070) with a few exceptions (Table 2.7). These exceptions include cancer-related datasets such as GSE2405 (from leukocytes), GSE22377 (from intestine), GSE11341 (from lung cells) and GSE28796 (from breast) (Table 2.4). In addition to the muscle-enriched genes (e.g., *MYH6*, *TTN* and *MYL2*), we also identified genes that are known to be highly expressed in the brain such as *CAMK2A*, *AQP4*, and *S100A1* and other genes that are not muscle-enriched (e.g., *HSPB2*, *PPP1R1A* and *EBF2*). This is perhaps due to the involvement of calcium signalling pathways in both muscle contraction and neuronal transmission [6].

DE modules have diverse biological processes

We hypothesize that differential expression can offer additional insights toward gene function in addition to tissue type specificity. To test this, we first applied the same biclustering parameters to the DE matrix (see Methods). This resulted in 33 DE modules, which on average have $\sim$387 genes and $\sim$3 RS. These modules have approximately three times the number of genes compared to the EE modules. This makes it difficult to meaningfully compare between EE and DE modules. To circumvent this, we applied stricter biclustering parameters to the DE matrix and found 37 modules with comparable module sizes. Table 2.8 shows the top GO annotations for each module. Our results show that there are more DE modules compared

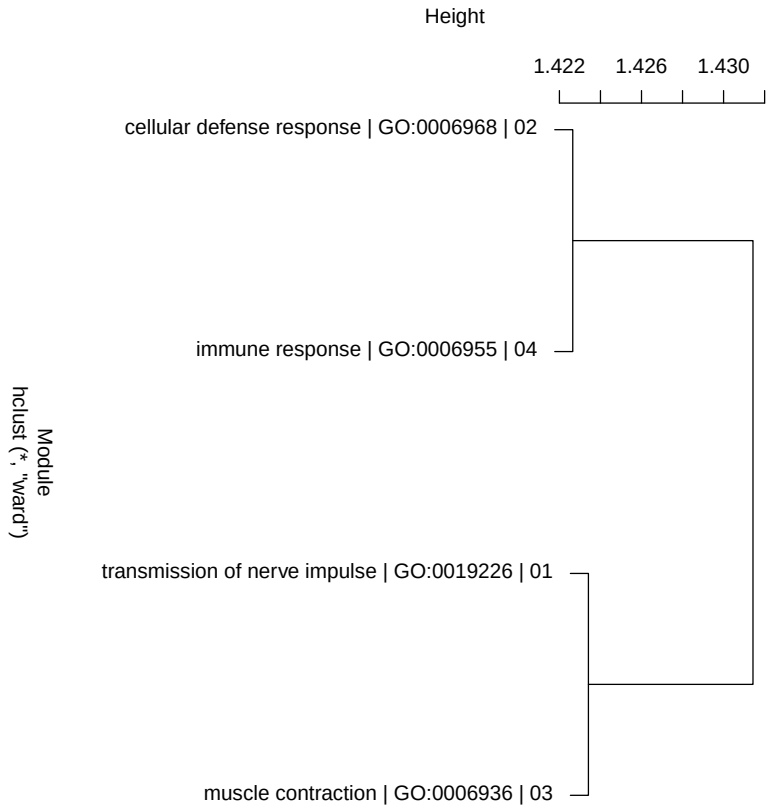


Figure 2.4: Clustering of GO-enriched EE modules. Tree leaves are labelled with the following format: GO Term | GO.ID | Module.ID.

GO.ID	Term	Significant	classicFisher	bicluster
GO:0006936	muscle contraction	50	< 1e-30	3
GO:0003012	muscle system process	51	< 1e-30	3
GO:0006941	striated muscle contraction	26	< 1e-30	3
GO:0061061	muscle structure development	38	1.3e-29	3
GO:0007517	muscle organ development	35	1.0e-28	3
GO:0003008	system process	65	3.1e-22	3
GO:0003009	skeletal muscle contraction	11	2.1e-18	3
GO:0050879	multicellular organismal movement	12	8.4e-18	3
GO:0050881	musculoskeletal movement	12	8.4e-18	3
GO:0014706	striated muscle tissue development	21	6.8e-17	3

Table 2.6: Top 5 GO annotations for the muscle contraction EE module

GEO.ID	Description	RS
GSE9103	Skeletal Muscle Transcript Profiles in Trained or Sedentary Young and Old Subjects	sedentary VS trained
GSE10161	Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass	aortic stenosis vs. healthy control
GSE1551	dermatomyositis	dermatomyositis VS healthy
GSE24235	Skeletal muscle gene expression in response to resistance exercise: sex specific regulation	resting vs 24 hrs post acute resistance exercise
GSE1869	Ischemic and Nonischemic CM and NF Hearts	non-ischemic vs ischemic cardiomyopathy
GSE13849	Expression Signatures in Polyarticular JIA Show Heterogeneity and Offer a Molecular Classification of Disease Subsets	Healthy (control group) vs. juvenile idiopathic arthritis
GSE3112	Plasma Cells in Muscle in Inclusion Body Myositis and Polymyositis	inclusion body myositis VS polymyositis VS normal control
GSE11686	Unique Transcriptional Profile in Wrist Muscles From Cerebral Palsy Patients	cerebral palsy VS healthy control
GSE13070	Human Insulin Resistance and Thiazolidinedione-Mediated Insulin Sensitization	preClamp vs postClamp vs noClamp
GSE6798	Reduced expression of mitochondrial oxidative metabolism genes in skeletal muscle of women with PCOS	control vs insulin-resistant polycystic ovary syndrome
*GSE2405	NIH/NIAID Neutrophil Response to A. phagocytophilum	control vs anaplasma phagocytophilum
GSE1145	changes in cardiac transcription profiles brought about by heart failure	ischemic, idiopathic dilated and normal hearts
GSE14901	Limb immobilization induces a coordinate down-regulation of mitochondrial and other metabolic pathways in men and women	pre or post-cast
*GSE22377	mRNA expression data from human adenocarcinomas of the stomach	diffuse adenocarcinoma vs intestinal adenocarcinomas
*GSE11341	Lung selective gene responses to alveolar hypoxia	Normoxia VS 3 H Hypoxia VS 24 H Hypoxia VS 48 H Hypoxia
*GSE28796	Gene expression profiles of pretreatment biopsies from dose-dense-docetaxel-treated breast cancers	pathological complete response (pCR) vs residual disease (NR)

Table 2.7: Muscle contraction EE module RS. All datasets use muscle or heart as source tissue except for datasets marked with *.

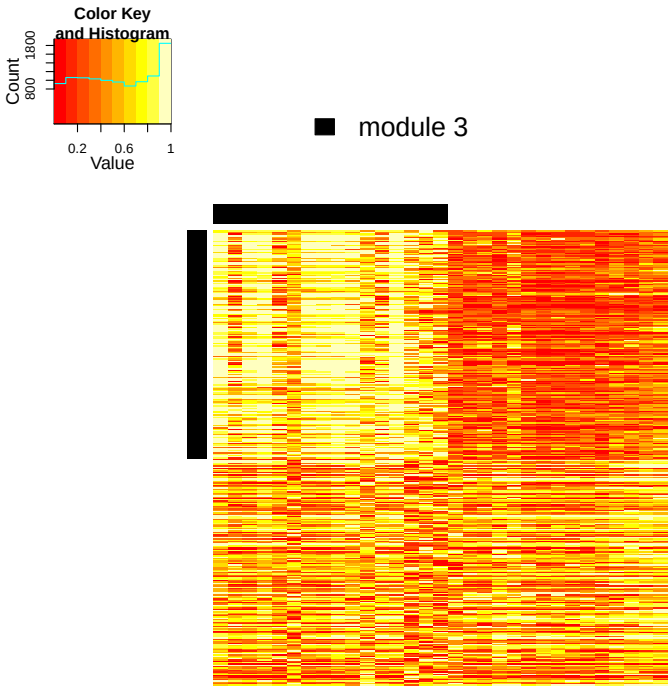


Figure 2.5: Muscle contraction EE module. Rows are genes and columns are datasets. Those genes and datasets that belong to the module are highlighted in black. Genes and datasets that do not belong to the module were randomly chosen.

to EE modules which reflects the increased variation found in experimental conditions relative to tissue types. For example, we find GSE9103 and GSE1551 in two separate DE modules (DE Module 18 and DE Module 1 respectively) even though both datasets are found in only one EE module (Module 3). Finally, similar to the EE modules, we clustered the DE modules based on gene overlap as shown in Figure 2.6. In comparison to the GO EE hierarchy, the GO DE hierarchy also includes a module enriched in "immune response genes" (GO:0006955). Moreover, the GO DE hierarchy also includes modules related to the "M phase" (GO:0000279) and "response to DNA damage stimulus" (GO:0006974) which suggests the involvement of a common set of genes with respect to the "immune response genes" enriched module. We observe a few outliers as well ("generation of precursor metabolites and energy" (GO:0006091) and "translational elongation" (GO:0006414)). These biological processes might involve a specialized set of genes that is different from those of other biological processes.

As a use case, we selected DE module 18 (GO:0006091, generation of precursor metabolites and energy). To visualize the quality of this module's clustering, Figure 2.7 shows the differential expression of all genes in this module compared a random set of genes of the same size. The top 5 GO annotations for this module is available in Table 2.9 which include closely related biological processes such as "oxidative phosphorylation" and "cellular respiration". Similar to the muscle contraction EE module before, this DE module also includes a few muscle-related RS such as GSE9103, GSE10161, GSE11686, and GSE14901

GO.ID	Term	Significant	Pval	Module	Rows	Cols
GO:0006414	translational elongation	49	< 1e-30	36	87	23
GO:0006091	generation of precursor metabolites and energy	54	< 1e-30	18	114	19
GO:0000279	M phase	49	< 1e-30	10	98	23
GO:0006955	immune response	51	< 1e-30	1	143	14
GO:0007156	homophilic cell adhesion	15	2.5e-19	13	56	16
GO:0007586	digestion	13	1.9e-17	11	62	4
GO:0034470	ncRNA processing	13	3.7e-16	33	41	19
GO:0006695	cholesterol biosynthetic process	8	1.8e-15	29	35	13
GO:0006955	immune response	29	4.7e-15	34	110	12
GO:0006334	nucleosome assembly	8	6.5e-10	2	79	3
GO:0048259	regulation of receptor-mediated endocytosis	6	1.3e-09	12	88	4
GO:0007186	G-protein coupled receptor protein signalng pathway	15	2.8e-09	3	78	12
GO:0006974	response to DNA damage stimulus	16	4.3e-09	19	112	2
GO:0016339	calcium-dependent cell-cell adhesion	5	1.9e-08	24	61	13
GO:0048869	cellular developmental process	31	3.1e-08	20	97	1
GO:0009611	response to wounding	15	2.9e-07	15	83	12
GO:0006915	apoptosis	28	6.8e-07	14	148	5
GO:0019219	regulation of nucleobase-containing compound metabolic process	22	1.6e-06	37	60	16
GO:0007154	cell communication	18	8.6e-06	9	68	15
GO:0050994	regulation of lipid catabolic process	3	2.0e-05	26	31	7
GO:0065008	regulation of biological quality	24	5.1e-05	31	101	2
GO:0042776	mitochondrial ATP synthesis coupled proton transport	3	7.5e-05	5	114	4
GO:0001539	ciliary or flagellar motility	3	7.7e-05	17	122	4
GO:0007267	cell-cell signaling	12	8.2e-05	25	71	4
GO:0006887	exocytosis	6	0.00014	30	96	3
GO:0051967	negative regulation of synaptic transmission, glutamatergic	2	0.00014	16	55	12
GO:0007215	glutamate signaling pathway	3	0.00014	32	77	6
GO:0009987	cellular process	83	0.00037	8	114	5
GO:2000027	regulation of organ morphogenesis	4	0.00039	35	93	4
GO:0010739	positive regulation of protein kinase A signaling cascade	2	0.00041	7	113	1
GO:0045449	regulation of transcription	21	0.00048	4	71	5
GO:0022617	extracellular matrix disassembly	2	0.0005	23	112	1
GO:0010467	gene expression	39	0.00056	22	122	2
GO:0006814	sodium ion transport	5	0.0012	28	114	5
GO:0021761	limbic system development	3	0.0013	6	82	2
GO:0007601	visual perception	6	0.0017	21	121	2
GO:0007283	spermatogenesis	4	0.0049	27	70	10

Table 2.8: Top GO biological process in each module from the DE matrix. The table is sorted by increasing p-value from the Fisher's Exact Test statistic for gene over representation. Significant column indicates the number of significant genes in the module that were found in the GO group.

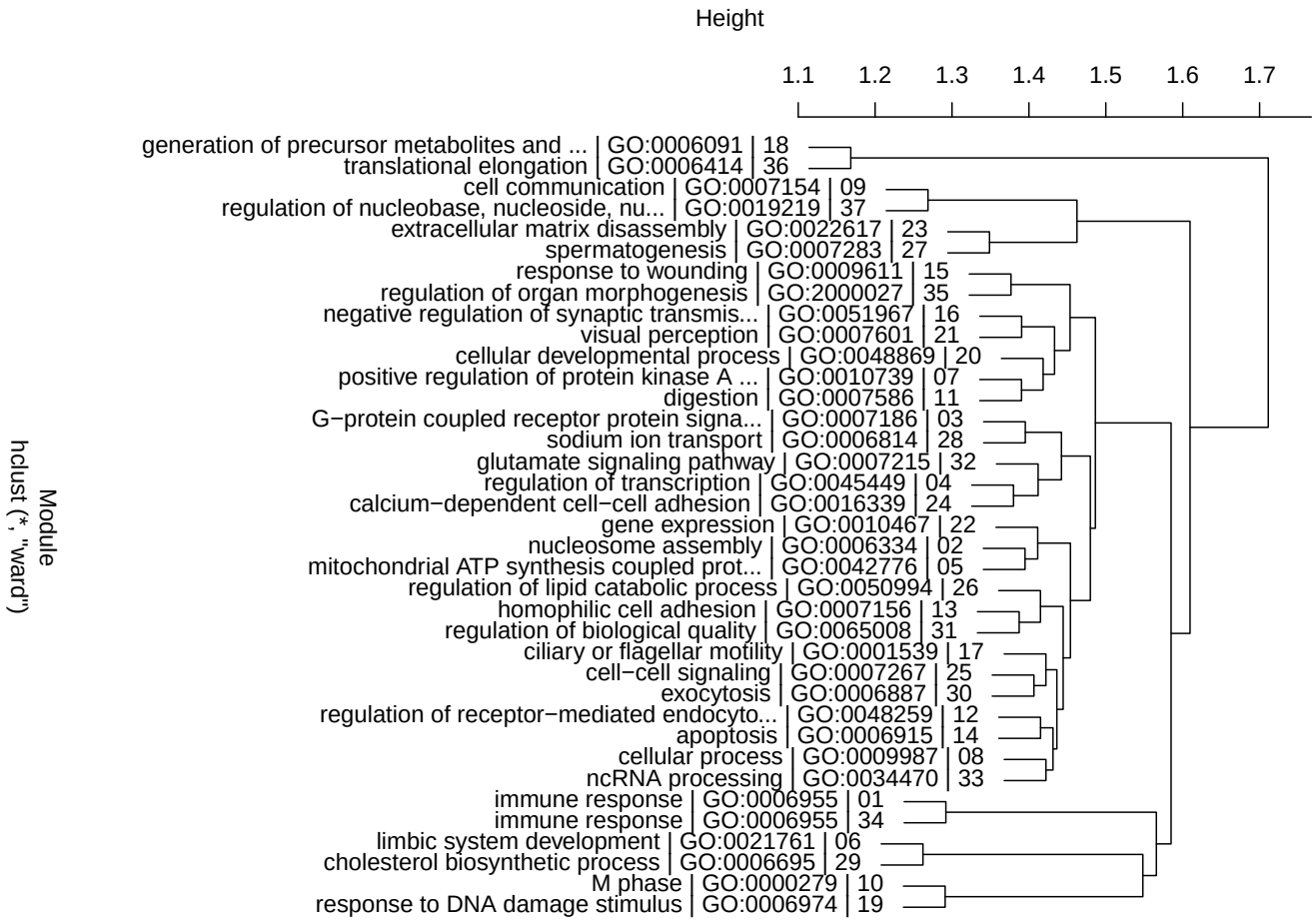


Figure 2.6: Clustering of GO-enriched DE modules. Tree leaves are labelled with the following format: GO Term | GO.ID | Module.ID

(Table 2.10) and muscle-enriched genes such as *ACTN2*, *CPT1B*, and *CKMT2*. We also find brain-related RS in this module such as GSE5388, GSE1297, and GSE2732 (Table 2.10). We suspect that some of these genes are involved in both metabolism and neuronal systems. To verify this, we compared our list of DE genes in Module 18 with those genes from another study that reported a list of proteins linked to metabolic abnormalities in the dorsolateral prefrontal cortex of bipolar disorder post-mortem brain tissues [36]. As expected, we found 5 out of 46 Pennington et al. genes in common (*ATP5B*, *ATP5C1*, *ATP5D*, *UQCRC1*, and *SUCLA2*; $P < 0.0001$, hypergeometric). Since these genes are also DE in non-brain related RS (Table 2.10), these genes are probably involved in different pathways as well.

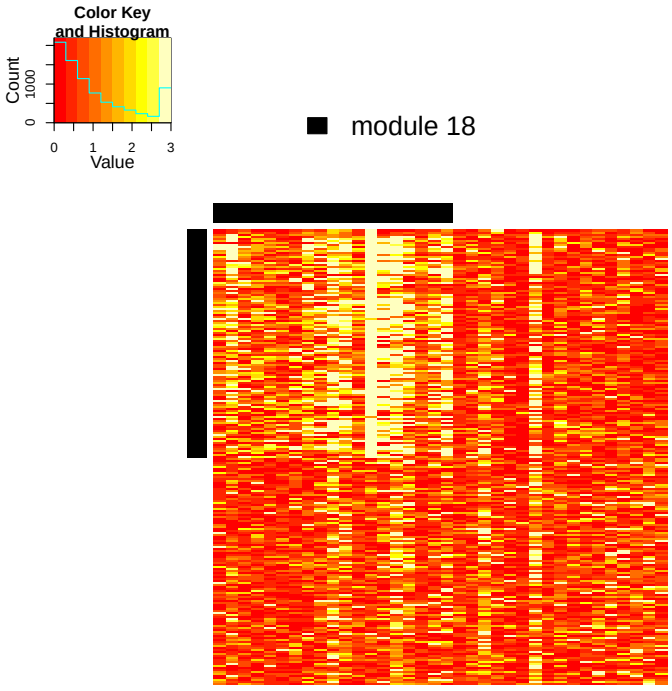


Figure 2.7: Generation of precursor metabolites DE module. Values correspond to $-log_{10}$(P-value). Rows are genes and columns are RS. Those genes and RS that belong to the module are highlighted in black. Genes and RS that do not belong to the module were randomly chosen.

GO.ID	Term	Significant	classicFisher	bicluster
GO:0006091	generation of precursor metabolites and ...	54	< 1e-30	18
GO:0006119	oxidative phosphorylation	30	< 1e-30	18
GO:0045333	cellular respiration	27	< 1e-30	18
GO:0022900	electron transport chain	26	< 1e-30	18
GO:0015980	energy derivation by oxidation of organi...	28	< 1e-30	18
GO:0022904	respiratory electron transport chain	20	1.20E-028	18
GO:0042773	ATP synthesis coupled electron transport	19	4.90E-028	18
GO:0042775	mitochondrial ATP synthesis coupled elec...	19	4.90E-028	18
GO:0055114	oxidation reduction	35	7.30E-023	18
GO:0006120	mitochondrial electron transport, NADH t...	15	1.20E-022	18

Table 2.9: Top 5 GO annotations for the generation of precursor metabolites DE module.

GEO.ID	Description	RS
GSE15132	Riboflavin depletion impairs cell proliferation in intestinal cells: Identification of mechanisms and consequences	Control VS Ribodeficient
GSE9103	Skeletal Muscle Transcript Profiles in Trained or Sedentary Young and Old Subjects	sedentary VS trained
GSE10161	Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass	aortic stenosis vs. healthy control
GSE5388	Adult postmortem brain tissue (dorsolateral prefrontal cortex) from subjects with bipolar disorder and healthy controls	Control group VS Bipolar Disorder
GSE6927	Gingival Epithelial Cell Transcriptional Responses to Commensal and Opportunistic Oral Microbial Species.	affected vs control group
GSE2443	Prostate cancer - comparison of androgen-dependent vs androgen-independent prostate cancer	
GSE12288	Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease	Coronary artery disease vs. healthy
GSE11686	Unique Transcriptional Profile in Wrist Muscles From Cerebral Palsy Patients	cerebral palsy VS healthy control
GSE9006	Gene expression in PBMCs from children with diabetes	sampling timepoint
GSE5900	Gene Expression of Bone Marrow Plasma Cells from Healthy Donors (N=22), MGUS (N=44), and Smoldering Myeloma (N=12)	Control (healthy) vs. smoldering myeloma
GSE25518	Testis developmental gene expression in cryptorchid boys at risk of azoospermia	cryptorchidism vs control
GSE29605	Gene expression data from chronic lymphocytic leukemia samples	control vs mutated IgVH
GSE14901	Limb immobilization induces a coordinate down-regulation of mitochondrial and other metabolic pathways in men and women	pre or post-cast
GSE13762	Comparative gene expression profile of 1,25-dihydroxyvitamin D3-treated human monocyte-derived dendritic cells	Vehicle Treated VS 1,25 dihydroxyvitamin D3
GSE30499	Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis	0 vs 1.5 vs 3 vs 4.5 hr
GSE21942	Expression data from peripheral blood mononuclear cells in multiple sclerosis patients and controls	Multiple sclerosis vs control group
GSE1297	Incipient Alzheimer's Disease: Microarray Correlation Analyses	Control vs. AD
GSE16581	Genomic landscape of meningiomas: gene expression	benign vs anaplastic
GSE2732	Global gene expression pattern of human brain neuronal (SH-SY5Y) cell lines exposed to sarin (GB).	Sarin Treated VS Control group

Table 2.10: Generation of precursor metabolites DE module RS

2.4 Discussion

Microarray experiments are often designed to identify genes that are involved in a particular biological phenomenon. For example, genes that are differentially expressed between blood samples from control and Parkinson's disease patients may play an important role in the disease. However, which of these DE genes are also differentially expressed in other neurological disease such as Alzheimer's disease is also of interest.

To this end, we have integrated hundreds of curated human microarray datasets and investigated expression and differential expression of genes across datasets and result sets. Our preliminary results show that as expected, tissue-enriched genes are highly expressed in those samples with the same tissue type. With respect to modules, there are fewer EE modules than DE modules. We attribute differences in the biological process of EE modules to tissue type variation. In contrast, there are more differences in experimental conditions and this is reflected in the biological process of each DE module.

We highlight possible extensions to our current work. First, most of the datasets in our study are related to cancer. As more datasets from studies such as drug-effects, neurological and developmental disorders become available, re-analysis of additional datasets may uncover new relationships between different biological pathways. Second, our current work can be applied to gene expression studies of model organisms such as mouse as well. Model organisms allow researchers to discover gene function through controlled genetic manipulations which are not possible in human. Third, the corresponding fold change of differentially expressed genes can be incorporated. Genes that change expression in the same direction could be co-regulated by the same transcription factors. Fourth, the interpretation of differentially expressed genes can be confounded by experimental artifacts. Examples of experimental artifacts include differences in reagents, equipment, or the time the experiment was conducted. Batch correction can be incorporated during data processing or experimental designs can be revisited during data analysis.

In our current work, we used the ISA biclustering algorithm for identifying modules in our data. While other methods could also be applied, we focused on designing a simple workflow for discovering biological knowledge from the data that is currently available.

In summary, we have provided a glimpse into gene expression and differential expression of many seemingly unrelated datasets and experimental conditions in humans. By combining information from other datasets in an unbiased manner, we can better interpret different biological mechanisms in a broader context.

Bibliography

[1] A. Alexa, J. Rahnenfhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, July 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl140. URL http://bioinformatics.oxfordjournals.org/content/22/13/1600.

[2] F. A. C. Azevedo, L. R. B. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. L. Ferretti, R. E. P. Leite, W. Jacob Filho, R. Lent, and S. Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology*, 513(5):532–541, Apr. 2009. ISSN 1096-9861. doi: 10.1002/cne.21974. URL http://www.ncbi.nlm.nih.gov/pubmed/19226510. PMID: 19226510.

[3] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. NCBI GEO: archive for functional genomics data sets10 years on. *Nucleic Acids Research*, 39(Database issue):D1005–D1010, Jan. 2011. ISSN 0305-1048. doi: 10.1093/nar/gkq1184. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013736/. PMID: 21097893 PMCID: PMC3013736.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1):pp. 289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

[5] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 67(3 Pt 1): 031902, Mar. 2003. ISSN 1539-3755. URL http://www.ncbi.nlm.nih.gov/pubmed/12689096. PMID: 12689096.

[6] M. J. Berridge, P. Lipp, and M. D. Bootman. The versatility and universality of calcium signalling. *Nature Reviews Molecular Cell Biology*, 1(1):11–21, Oct. 2000. ISSN 1471-0072. doi: 10.1038/35036035. URL http://www.nature.com/nrm/journal/v1/n1/full/nrm1000_011a.html.

[7] M. Braun, A. Wendt, B. Birnir, J. Broman, L. Eliasson, J. Galvanovskis, J. Gromada, H. Mulder, and P. Rorsman. Regulated exocytosis of GABA-containing synaptic-like microvesicles in pancreatic -cells. *The Journal of General Physiology*, 123(3):191–204, Mar. 2004. ISSN 0022-1295. doi: 10.1085/jgp.200308966. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2217446/. PMID: 14769845 PMCID: PMC2217446.

[8] M. Blanger, I. Allaman, and P. J. Magistretti. Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell Metabolism*, 14(6):724–738, Dec. 2011. ISSN 1932-7420. doi: 10.1016/j.cmet.2011.08.016. URL http://www.ncbi.nlm.nih.gov/pubmed/22152301. PMID: 22152301.

[9] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, W. J. Thompson, and B. A. Barres. A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function. *The Journal of Neuroscience*, 28(1):264 –278, Jan. 2008. doi: 10.1523/JNEUROSCI.4178-07.2008. URL http://www.jneurosci.org/content/28/1/264.abstract.

[10] S. Cohen and M. E. Greenberg. Communication between the synapse and the nucleus in neuronal development, plasticity, and disease. *Annual review of cell and developmental biology*, 24:183–209, 2008. ISSN 1081-0706. doi: 10.1146/annurev.cellbio.24.110707.175235. URL http://www.ncbi.nlm.nih.gov/pubmed/18616423. PMID: 18616423.

[11] J. Dennis, Glynn, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(5):P3, 2003. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/12734009. PMID: 12734009.

[12] J. Engreitz, A. Morgan, J. Dudley, R. Chen, R. Thathoo, R. Altman, and A. Butte. Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, 11(1):603, Dec. 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-603. URL http://www.biomedcentral.com/1471-2105/11/603/abstract.

[13] B. S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis, 5th Edition*. Wiley Series in Probability and Statistics, Jan. 2011. ISBN 9780470749913, 9780470977811. URL http://onlinelibrary.wiley.com/book/10.1002/9780470977811.

[14] J. Feher. *Quantitative Human Physiology: An Introduction*. Academic Press, 2012.

[15] I. K. Franklin and C. B. Wollheim. GABA in the endocrine pancreas its putative role as an islet cell paracrine-signalling molecule. *The Journal of General Physiology*, 123(3):185–190, Mar. 2004. ISSN 0022-1295, 1540-7748. doi: 10.1085/jgp.200409016. URL http://jgp.rupress.org/content/123/3/185.

[16] L. French and P. Pavlidis. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Computational Biology*, 7(1):e1001049, 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1001049. URL http://www.ncbi.nlm.nih.gov/pubmed/21253556. PMID: 21253556.

[17] L. French, P. P. C. Tan, and P. Pavlidis. Large-scale analysis of gene expression and connectivity in the rodent brain: Insights through data integration. *Front Neuroinform*, 5, 2011. doi: 10.3389/fninf.2011.00012. PMID: 21863139 PMCID: 3149147.

[18] S. Greer, R. Honeywell, M. Geletu, R. Arulanandam, and L. Raptis. Housekeeping genes; expression levels may change with density of cultured cells. *Journal of Immunological Methods*, 355(12):76–79, Apr. 2010. ISSN 0022-1759. doi: 10.1016/j.jim.2010.02.006. URL http://www.sciencedirect.com/science/article/pii/S0022175910000402.

[19] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. v. d. Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, R. A. Dalley, B. D. Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R.

Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399, Sept. 2012. ISSN 0028-0836. doi: 10.1038/nature11405. URL http://www.nature.com/nature/journal/v489/n7416/full/nature11405.html?WT.ec_id=NATURE-20120920.

[20] S. Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3:31, 2009. doi: 10.3389/neuro.09.031.2009. URL http://www.frontiersin.org/human_neuroscience/10.3389/neuro.09/031.2009/abstract.

[21] D. Hunt, M. R. J. Mason, G. Campbell, R. Coffin, and P. N. Anderson. Nogo receptor mRNA expression in intact and regenerating CNS neurons. *Molecular and cellular neurosciences*, 20(4): 537–552, Aug. 2002. ISSN 1044-7431. PMID: 12213438.

[22] H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, and N. Sestan. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, Oct. 2011. ISSN 0028-0836. doi: 10.1038/nature10523. URL http://dx.doi.org/10.1038/nature10523.

[23] C.-K. Lee, S. M. Sunkin, C. Kuan, C. L. Thompson, S. Pathak, L. Ng, C. Lau, S. Fischer, M. Mortrud, C. Slaughterbeck, A. Jones, E. Lein, and M. Hawrylycz. Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome biology*, 9(1):R23, 2008. ISSN 1465-6914. doi: 10.1186/gb-2008-9-1-r23. URL http://www.ncbi.nlm.nih.gov/pubmed/18234097. PMID: 18234097.

[24] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T.-M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H.-W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K.-R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, Jan. 2007. ISSN 1476-4687. doi: 10.1038/nature05453. URL http://www.ncbi.nlm.nih.gov/pubmed/17151600. PMID: 17151600.

[25] B.-Y. Liao and J. Zhang. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution*, 23(3):530–540, Mar. 2006. ISSN 0737-4038. doi: 10.1093/molbev/msj054. URL http://www.ncbi.nlm.nih.gov/pubmed/16280543. PMID: 16280543.

[26] D. J. Lockhart and C. Barlow. Expressing what's on your mind: DNA arrays and the brain. *Nat Rev Neurosci*, 2(1):63–68, Jan. 2001. ISSN 1471-003X. doi: 10.1038/35049070. URL http://dx.doi.org/10.1038/35049070. PMID: 11253360.

[27] M. Lukk, M. Kapushesky, J. Nikkil, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–324, 2010. ISSN 1087-0156. doi: 10.1038/nbt0410-322. URL http://www.nature.com/nbt/journal/v28/n4/abs/nbt0410-322.html.

[28] J. A. Miller, S. Horvath, and D. H. Geschwind. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 107(28):12698–12703, July 2010. ISSN 1091-6490. doi: 10.1073/pnas.0914257107. URL http://www.ncbi.nlm.nih.gov/pubmed/20616000. PMID: 20616000.

[29] M. Nedergaard, B. Ransom, and S. A. Goldman. New roles for astrocytes: redefining the functional architecture of the brain. *Trends in Neurosciences*, 26(10):523–530, Oct. 2003. ISSN 0166-2236. URL http://www.ncbi.nlm.nih.gov/pubmed/14522144. PMID: 14522144.

[30] L. Ng, S. Pathak, C. Kuan, C. Lau, H.-w. Dong, A. Sodt, C. Dang, B. Avants, P. Yushkevich, J. Gee, D. Haynor, E. Lein, A. Jones, and M. Hawrylycz. Neuroinformatics for genome-wide 3-d gene expression mapping in the mouse brain. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4(3): 382393, July 2007. ISSN 1545-5963. doi: 10.1109/tcbb.2007.1035. URL http://dx.doi.org/10.1109/tcbb.2007.1035.

[31] L. Ng, A. Bernard, C. Lau, C. C. Overly, H.-W. Dong, C. Kuan, S. Pathak, S. M. Sunkin, C. Dang, J. W. Bohland, H. Bokil, P. P. Mitra, L. Puelles, J. Hohmann, D. J. Anderson, E. S. Lein, A. R. Jones, and M. Hawrylycz. An anatomic gene expression atlas of the adult mouse brain. *Nature neuroscience*, 12(3):356–362, Mar. 2009. ISSN 1546-1726. doi: 10.1038/nn.2281. URL http://www.ncbi.nlm.nih.gov/pubmed/19219037. PMID: 19219037.

[32] N. A. Oberheim, T. Takano, X. Han, W. He, J. H. C. Lin, F. Wang, Q. Xu, J. D. Wyatt, W. Pilcher, J. G. Ojemann, B. R. Ransom, S. A. Goldman, and M. Nedergaard. Uniquely hominid features of adult human astrocytes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(10):3276–3287, Mar. 2009. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.4707-08.2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19279265. PMID: 19279265.

[33] M. C. Oldham, G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. H. Geschwind. Functional organization of the transcriptome in human brain. *Nature Neuroscience*, 11(11): 1271–1282, Oct. 2008. ISSN 1097-6256. doi: 10.1038/nn.2207. URL http://www.nature.com/neuro/journal/v11/n11/full/nn.2207.html.

[34] P. Pavlidis and W. S. Noble. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, 2(10):RESEARCH0042, 2001. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/11597334. PMID: 11597334.

[35] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[36] K. Pennington, C. L. Beasley, P. Dicker, A. Fagan, J. English, C. M. Pariante, R. Wait, M. J. Dunn, and D. R. Cotter. Prominent synaptic and metabolic abnormalities revealed by proteomic analysis of the dorsolateral prefrontal cortex in schizophrenia and bipolar disorder. *Molecular psychiatry*, 13(12): 1102–1117, Dec. 2008. ISSN 1476-5578. doi: 10.1038/sj.mp.4002098. PMID: 17938637.

[37] S. Pounds and S. W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, July 2003. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btg148. URL http://bioinformatics.oxfordjournals.org/content/19/10/1236.

[38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org. ISBN 3-900051-07-0.

[39] R. Sandberg, R. Yasuda, D. G. Pankratz, T. A. Carter, J. A. Del Rio, L. Wodicka, M. Mayford, D. J. Lockhart, and C. Barlow. Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20): 11038–11043, Sept. 2000. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/11005875. PMID: 11005875.

[40] E. Sibille, V. Arango, J. Joeyen-Waldorf, Y. Wang, S. Leman, A. Surget, C. Belzung, J. J. Mann, and D. A. Lewis. Large-scale estimates of cellular origins of mRNAs: enhancing the yield of transcriptome analyses. *Journal of Neuroscience Methods*, 167(2):198–206, Jan. 2008. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2007.08.009. URL http://www.ncbi.nlm.nih.gov/pubmed/17889939. PMID: 17889939.

[41] A. D. Strand, A. K. Aragaki, Z. C. Baquet, A. Hodges, P. Cunningham, P. Holmans, K. R. Jones, L. Jones, C. Kooperberg, and J. M. Olson. Conservation of regional gene expression in mouse and human brain. *PLoS Genet*, 3(4):e59, Apr. 2007. doi: 10.1371/journal.pgen.0030059. URL UR-http://dx.plos.org/10.1371/journal.pgen.0030059,http://dx.plos.org/10.1371/journal.pgen.0030059. PMID: 17447843.

[42] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236, Mar. 1963. ISSN 01621459. doi: 10.2307/2282967. URL http://www.jstor.org/discover/10.2307/2282967?uid=3739400&uid=2&uid=3737720&uid=4&sid=21101299103581.

[43] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(Database issue):D5–12, Jan. 2007. ISSN 1362-4962. doi: 10.1093/nar/gkl1031. PMID: 17170002.

[44] M. A. Zapala, I. Hovatta, J. A. Ellison, L. Wodicka, J. A. D. Rio, R. Tennant, W. Tynan, R. S. Broide, R. Helton, B. S. Stoveken, C. Winrow, D. J. Lockhart, J. F. Reilly, W. G. Young, F. E. Bloom, D. J. Lockhart, and C. Barlow. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29): 10357–10362, July 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0503357102. URL http://www.pnas.org/content/102/29/10357.

[45] H. Zeng, E. H. Shen, J. G. Hohmann, S. W. Oh, A. Bernard, J. J. Royall, K. J. Glattfelder, S. M. Sunkin, J. A. Morris, A. L. Guillozet-Bongaarts, K. A. Smith, A. J. Ebbert, B. Swanson, L. Kuan, D. T. Page, C. C. Overly, E. S. Lein, M. J. Hawrylycz, P. R. Hof, T. M. Hyde, J. E. Kleinman, and A. R. Jones. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, 149(2):483–496, Apr. 2012. ISSN 0092-8674. doi:

10.1016/j.cell.2012.02.052. URL
http://www.sciencedirect.com/science/article/pii/S0092867412003480.

[46] A. Zoubarev, K. M. Hamer, K. D. Keshav, E. L. McCarthy, J. R. C. Santos, T. V. Rossum, C. McDonald, A. Hall, X. Wan, R. Lim, J. Gillis, and P. Pavlidis. Gemma: A resource for the re-use, sharing and meta-analysis of expression profiling data. *Bioinformatics*, July 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts430. URL
http://bioinformatics.oxfordjournals.org/content/early/2012/07/10/bioinformatics.bts430.