Wide-scale Comparison of Transcriptome Data and the Role of MicroRNA in Major Depression and Suicide

by

Raymond Lim

B. Sc, University of British Columbia, 2008

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)

October 2011

Abstract

The first chapter of this thesis addresses a common problem in genomics experiments: interpreting a resulting "hit list" of interesting genes. We present work on an approach for summarizing and exploring "hit lists" that makes use of the large amount of gene expression data in public repositories such as the Gene Expression Omnibus. We compare the query list with datasets that we have analyzed for differential expression of genes. Studies that have similarities to the given hit list yield additional insights, help contextualize studies, and serve as a basis for future meta-analysis. A conceptually similar problem that we addressed is the classification or clustering of datasets based on patterns of differential expression. Both problems required a method for determining distances between datasets based on rankings of genes. We tested and benchmarked several methods using manually annotated datasets. The method that performed best according to our evaluation process is based on Kendall's Tau top-k distance. We investigated potential sources of confounds, finding that the largest challenge may be posed by the high prevalence of certain gene expression patterns. These highly prevalent patterns tended to dominate search results. Nonetheless, we demonstrated the effectiveness of this approach in a case study.

In the second chapter, we investigated the role of microRNAs in the context of major depression and suicicide. We profiled microRNA and messenger RNA levels in post-mortem prefrontal cortex and hippocampus brain tissue of depressed suicides, suicides, and controls. In the prefrontal cortex, we found miR-1202 to be down-regulated in suicides versus controls, and LCT (lactase enzyme) was up-regulated in suicides or depressed suicides compared to controls. The former result was independently confirmed using quantitative PCR. While further study is needed, our results have the potential to provide insight into molecular changes in the brains of depressed and suicidal individuals.

Preface

Chapter 2 presents work that was done in collaboration with Dr. Gustavo Turecki, Juan Pablo Lopez, and Bharatkumar Patel at the McGill Group for Suicide Studies. I was responsible for the bioinformatic analysis.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Leon French, Jesse Gillis, Artemis Lai, Willie Kwok, Suzanne Lane, Lydia Xu, and Tamryn Loo for their contributions to the work presented in chapter 1.

I offer my sincerest gratitude to my supervisor, Dr. Paul Pavlidis, for providing his wisdom and kind support throughout the thesis project.

For their collaboration with work detailed in chapter 2, I owe my thanks to Juan Pablo Lopez, Bharatkumar Patel, and Dr. Gustavo Turecki at the McGill Group for Suicide Studies (McGill University).

Funding was provided by the CIHR/MSFHR Bioinformatics Training Program.

Finally, I thank my parents to whom I am especially indebted for their support throughout my years of education.

To my parents

Chapter 1

Wide-scale Comparison of Transcriptome Data

1.1 Introduction

Public repositories of gene expression data provide a wealth of opportunities for re-using data to generate novel findings and hypotheses for follow-up analysis [14, 105, 118]. However, the repositories merely store the data and do not provide a straightforward means of conducting queries, e.g. finding studies where a gene of interest is correlated with an experimental factor. Gemma (www.chibi.ubc.ca/Gemma/), a framework for the meta-analysis of gene expression data, provides an interface for querying data to find experiments where a query gene is relevant. The natural progression is to allow querying of groups of genes, i.e. gene sets, and ranked continua of genes to allow comparisons among complete experiments. Retrieving similar experiments in this data-driven fashion would aid biologists in viewing their experiments within the context of previously published studies and provide the potential for novel insight. Studies with similar expression signatures may have conditions that modulate the same pathways. Comparing complete experiments would also offer a better characterization of the complete transcriptome landscapes, e.g. the set of gene expression states available.

Comparing data that has been independently collected presents a number of challenges. Differences in a single study may be subtle; sometimes only involving a small number of genes. At the same time, there may be experimental and biological confounding factors, which can make comparing studies difficult. A suitable metric for comparing studies needs to be both sensitive and robust. Current methods are based on correlation, gene overlap, or similarity in enriched gene sets, and may employ a threshold to select differentially expressed genes (DEGs). Thresholds are typically used, as genes that are not differentially expressed, usually did not pass the threshold and may have random ranks. A drawback of a threshold is that it is difficult to decide upon a universally acceptable one due to the diversity of statistical power within studies, resulting in relevant genes below the threshold. In contrast, threshold-free methods, such as Gene Set Enrichment Analaysis (GSEA) [109] and Rank-rank Hypergeometric Overlap (RRHO) [87], use the complete ranked list of genes rather than a threshold truncated set of genes. GSEA is gene set-based

approach relying on gene annotations. Gene set-based approaches are recommended in order to increase the overlap in signatures; however, many genes lack annotations and they can be biased to certain biological processes [55].

We approached the project from two perspectives: information retrieval and a wide-scale meta-analysis perspective. In the former perspective, the query is a gene expression profile, and the target database is a set of profiles from public repositories. We compare the query to all the profiles in our database to return a list of profiles ranked by relevance, which we call a "similarity profile". The latter perspective of wide-scale meta-analysis may offer new insight into the reproducibility and comparability of public microarray data, as well as offer a broad characterization of the transcriptome landscape. Previous such studies have offered insight into pathophysiological reproducibility, gene dynamics (when and how genes change expression), disease signatures, and gene-phenotype signatures [26, 73, 76]. We intend to link both perspectives; in developing a better understanding of the general characteristics of microarray data, we can refine the information retrieval framework.

Existing tools or resources for querying microarray data are Connectivity Map (cMap) [64], GEM-TREND [31], MARQ [114], MASTA [94], NextBio [62], HORMONOMETER [115], FARO [68], and AtCAST [100]. The cMap compared gene expression profiles of cells treated with small molecules, inferring similarities between drugs. MARQ is a tool for mining the Gene Expression Omnibus (GEO) [118] freely available at marq.dacya.ucm.es. They applied their tool towards identifying the common signature for cell wall remodeling in response to cell wall stress. GEM-TREND is another tool used for mining GEO, and is targeted towards network discovery, having implemented a network visualization interface. NextBio provides their service for mining publicly shared repositories to paying customers (www.nextbio.com). This service was used by Hoenerhoff *et al.* to compare human hepatocellular carcinoma (HCC) to mouse HCC, identifying the dysregulation of several mediators similarly altered across species [47]. The AtCAST tool explored relationships among *Arabidopsis thaliana* datasets by building a "module-based correlation network". cMap, MARQ, GEM-TREND, NextBio and AtCAST are all based on the GSEA method [109]. In contrast, while also search tools for *Arabidopsis thaliana* microarray datasets, MASTA and FARO use overlap in DEG lists, and HORMONOMETER is correlation-based. MASTA is focused on identifying potential chemical inhibitors/activators and genetic suppressors/enhancers. HORMONOMETER was used for evaluating transcriptome response similarities between hormones and external pH [63].

The first part of the project was an assessment of the general characteristics and trends of microarray data. Next, we performed pair-wise comparisons of all datasets using multiple metrics: a rank threshold approach, a gene set threshold approach, and a simple threshold based approach. The three method types were evaluated using a framework that utilized the metadata or annotations from dataset descriptors to give some idea of the relative performance (see Figure 1.1). This framework relied on real microarray data downloaded from public repositories, and as a result, we lacked a true gold standard. However, by using real biological data, we can give insight into challenges under real-world conditions and suggest possible avenues of future development.

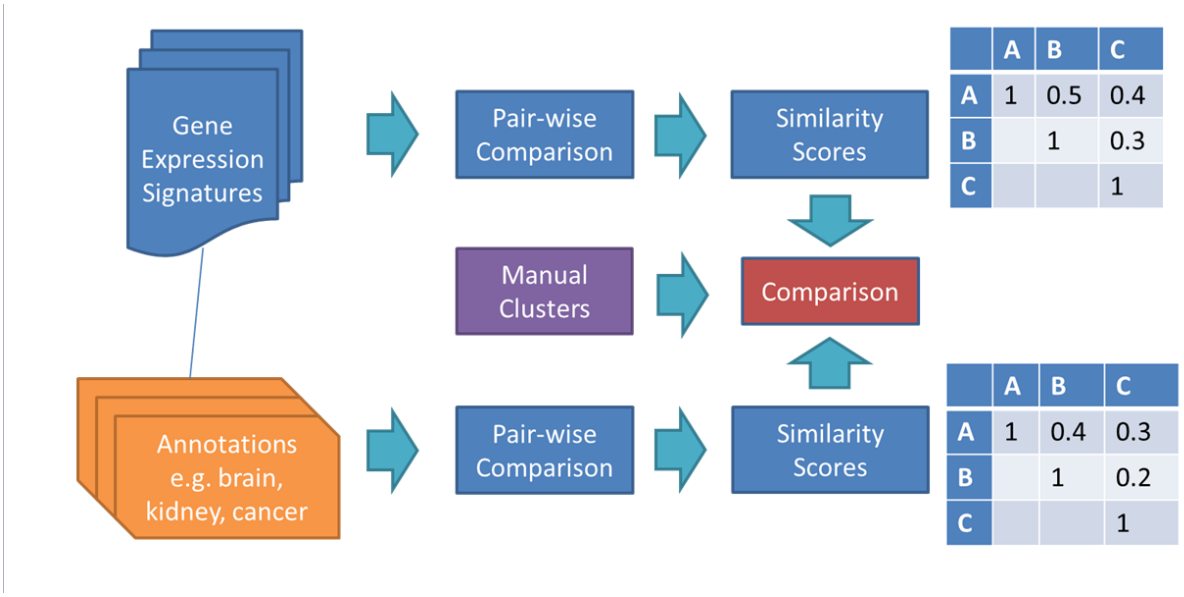


Figure 1.1: Evaluation pipeline for comparing gene expression signatures.

1.2 Methods

1.2.1 Data Pre-processing

We detail the procedure for the statistical analysis and probe mapping, i.e. the mapping of microarray probes to genes, though these procedures were Gemma-based analyses outside the scope of this project.

Statistical Analysis

Microarray data from public repositories such as the Gene Expression Omnibus (GEO) and ArrayExpress [14, 105, 118] were loaded into Gemma, a framework for the meta-analysis of gene expression data (www.chibi.ubc.ca/Gemma/). A subset of studies were automatically analyzed for differential gene expression for up to 3 factors; interactions were omitted if there was more than 2 factors. Thus, the following analysis types were supported, one-sample t-tests (rare), t-tests between two levels, one-way ANOVA, two-way ANOVA with or without interactions, and three-way ANOVA without interactions. T-test or ANOVA p-values were corrected for multiple testing using the method described in [108]. We refer to the analysis results for a single factor or interaction in an experiment as a "result set". Thus, a single dataset may have more than one result set, e.g. GSE9806, a time course study on the effects marinobufagenin in human dermal fibroblasts, has three associated result sets: sampling time, treatment, and the interaction between sampling time and treatment. This is of significance in the evaluation where we used dataset annotations so that result sets had to be summarized.

Probe Mapping

The Gemma framework does not use the annotations provided by the microarray manufacturers and instead uses an in-house protocol for mapping, employing the probe sequences to improve cross-platform consistency [72]. Briefly, probes were aligned to the reference genome using BLAT, and then the UCSC GoldenPath database was used to identify the gene at the aligned region [5]. We identified and removed non-specific alignments such that we only used probes which mapped to a single gene. When multiple probes mapped to a single gene, we took the median across the probes.

To compare mouse and human transcriptome profiles, we used the HomoloGene database to define orthologs [118].

Quality Control

In pilot studies, we found that datasets which have extreme amounts of differential expression resulted in low performance when comparing profiles, particularly with rank-based methods, as differential expression of a gene tended to be binary in such cases. Similarly, result sets which have no differential expression were uninformative. Thus, we filtered out result sets that have no q-values less than 0.3, or q-values less than 0.05 for more than 50% of the genes.

We assessed the impact of the data quality on the evaluation, examining outlier samples and batch effects. A sample was considered an outlier if it was not well correlated with any other sample (a correlation coefficient higher than 0.90 or the 85th percentile of all sample pair-wise correlation coefficients). Under this criterion, roughly 30% of the datasets in Gemma had an outlier. Batch effects were also investigated in a separate analysis using batch information available in Gemma. The batch factor was determined from sample processing dates, and the degree to which it affected the data was determined by its correlation with the principal components, with other experimental factors, and with the expression of individual probe sets.

1.2.2 Comparison Methods

Expression profiles were compared as binary DEG profiles, continuous rank profiles, or binary enriched Gene Ontology (GO) term profiles.

In the binary DEG profiles, a gene has a score of 1 if it was considered differentially expressed (q-value < 0.05), otherwise its score was 0 or NA if the gene is not tested on that platform. GO profiles were similar to binary profiles except they were at the GO gene set level. A GO term had a score of 1 if a Fisher's exact test (implemented in GOstat) determined that its members were over-represented among the DEGs (p-value < 0.05) [6]. Binary gene and GO profiles were compared using Jaccard distance $J_\delta(A,B)$, which is complementary to the Jaccard similarity coefficient $J(A,B)$:

$$J_\delta(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{1.1}$$

The Jaccard coefficient, when comparing two profiles, can be thought of as the fraction of concordant DEGs.

Rank profile comparisons used a rank transformation of the p-value scores. Rank profiles were compared using the top-k Kendall distance algorithm described in [29]. This method gave a measure of distance

between the top or high scoring genes of a data set profile. We focused on the top of the gene list, since these genes were less likely to be influenced by experimental variability. Specifically, we used the top 5% of genes in a dataset; we experimented with other threshold valuesfor optimal results according to our evaluation process.

As there was a concern that genes with low information content were resulting in uninformative similarity between expression profiles, we also compared rank profiles between datasets using the top-k Kendall distance after filtering out genes with low information content. Information content was defined as the $-log_2(P(g))$ where $P(g)$ was the fraction of all differential expression accounted for by gene g (q-value $<$ 0.05). The filter threshold was the 10th percentile of the information content distribution; this was set using our evaluation protocol.

1.2.3 Evaluation of Results

We assessed the distance methods using dataset annotations and manual classification of datasets. Since datasets may be associated with up to three result sets, we summarized the associated result sets by taking the minimum p-value for a gene. This merged information from all three result sets, such that a dataset expression profile may represent information from multiple experimental contrasts, e.g. age and treatment level.

Leon French designed an automated pipeline to annotate each dataset with concepts linked to classes in open biomedical ontologies [36]. Concept tags were then manually curated for improved accuracy[1]. We compared concept profiles between datasets using Jaccard distance.

Evaluation also included the use of a set of result sets manually annotated with a disease. We chose diseases that were sufficiently represented within our database, and for each such disease, we selected relevant and comparable result sets. For example, for Huntington's disease, we selected all result sets that contrasted Huntington's disease samples versus controls. Result sets classified under the same disease were compared to each other.

1.2.4 Annotation Enrichment

We analyzed the similarity profile of each dataset to determine enriched dataset annotations. This yielded some insight into the type of datasets driving the correlation. Term enrichment was calculated using partial area under the receiver operating characteristic curve (ROC AUC), limiting fall-out, i.e. setting a false positive rate (FPR) threshold to control the probability of non-relevant datasets in the similarity profile.

1.3 Results and Discussion

1.3.1 Data Overview

After filtering, we were left with 349 mouse and 224 human datasets. These datasets used a variety of different platforms, although the most common ones use Affymetrix technology (see Table 1.1).

[1]Manual curations were performed by Suzanne Lane, Lydia Xu, Tamryn Loo, Artemis Lai, and Willie Kwok

ID	Name	Count
GPL570	Affymetrix GeneChip Human Genome U133 Plus 2.0	79
GPL96	Affymetrix GeneChip Human Genome U133A	46
GPL91	Affymetrix GeneChip Human Genome U95A	22
	(Other)	67
		214

(a) Human

ID	Name	Count
GPL1261	Affymetrix GeneChip Mouse Genome 430 2.0	137
GPL81	Affymetrix GeneChip Mouse Genome U74A V2	70
GPL339	Affymetrix GeneChip Mouse Genome 430A 2.0	32
GPL260	Caltech 16K cDNA Mouse	29
	(Other)	81
		349

(b) Mouse

Table 1.1: Number of datasets per platform.

1.3.2 General Data Characteristics

Using our diverse set of experiments studying a variety of factors, we can give some insight into the general characteristics of gene expression.

Using several statistics concerning the expression dynamics of genes, we made a couple simple observations. Frequently expressed genes were more commonly differentially expressed, and more variably expressed genes were more frequently differentially expressed (see Figure 1.2). This made sense as genes have to be expressed in order to be differentially expressed, and we would expect genes which are more frequently differentially expressed to exhibit more dynamic range in expression levels (or vice versa).

Awareness of gene dynamics or the probability that a gene is identified as differentially expressed is important for tasks such as "gene prioritization", where limited resources impose constraints on the number of genes that can be included in follow-up analysis. An intuitive notion is that if a gene is frequently observed as differentially expressed across many different studies, it is likely not pertinent to any particular target disease or domain of interest. We hypothesized that these genes were involved in a wide range of different processes, i.e. that they were multifunctional. In fact, the probability that a gene was differentially expressed was also correlated with its multifunctionality [39] (Figure 1.3). For example, the least frequently DEGs were enriched for sensory perception GO groups, which were among the least multifunctional (see Supplementary Table 1.5).

We investigated the gene sets that were enriched among the frequently DEGs. Enriched biological process gene sets included "protein biosynthesis and degradation", "NF-κB regulation", and "cellular respiration" (see Supplementary Table 1.4). Infrequently DEGs had enriched associations with "sensory perception", "regulation of nucleotide metabolism", and "regulation of cAMP" (cyclic adenosine monophosphate) (see Supplementary Table 1.5). Similar findings were observed by Morgan *et al.* [76]. They attributed increased frequency with a higher variety of transcription factor regulatory site annotations in the molecular signatures database [109]. They also observed a conservation of gene dynamics across taxa using homolo-

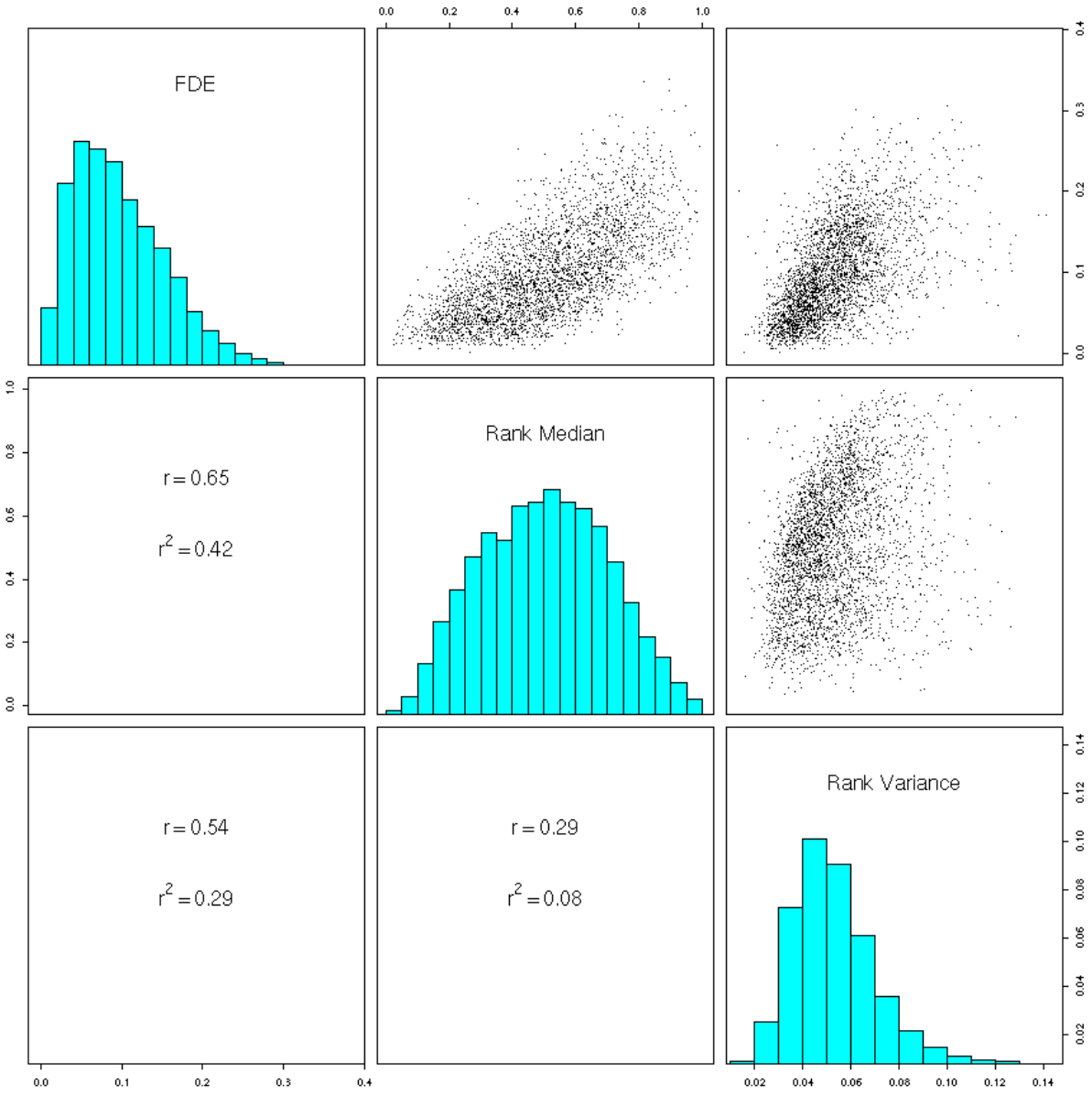


Figure 1.2: Global patterns of differential expression correlated with gene variability and expression levels. FDE: fraction of result sets where a gene is differentially expressed. Rank Median: median rank of a gene's expression value. Rank Variance: variance of a gene's expression rank.

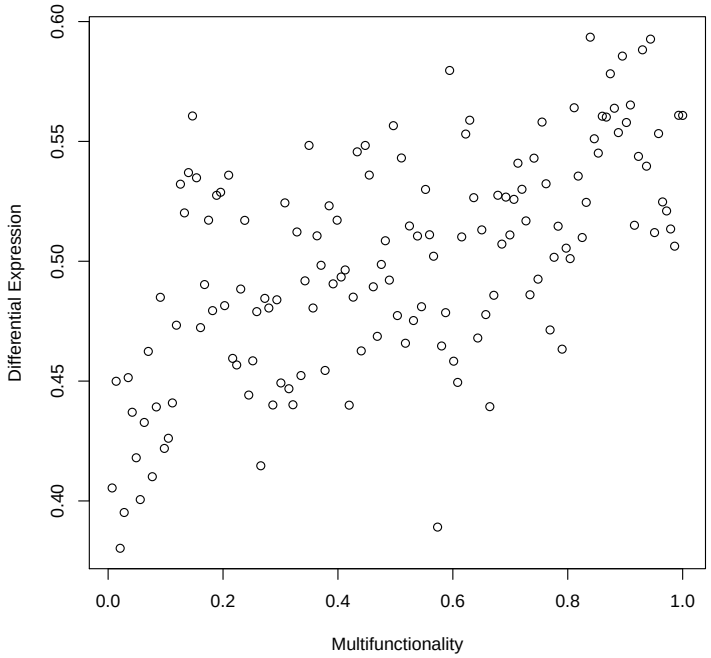


Figure 1.3: Multifunctional genes tended to be frequently differentially expressed (in mouse). Differential expression is the mean rank ratio of frequency of differential expression (centile bins). Multifunctionality is the multifunctionality rank ratio and derived from the number of Gene Ontology terms annotated to a gene. Credit to Jesse Gillis for providing the multifunctionality data.

		Human		
		FDE	Rank Median	Rank Variance
Mouse	FDE	0.37	0.37	0.12
	Rank Median	0.35	0.69	-0.23
	Rank Variance	0.09	-0.16	0.35

Table 1.2: Mouse and human gene dynamic signatures were correlated (Spearman's correlation).

gous genes. We observed that this was the case (see Table 1.2).

1.3.3 Query Use Case: Tauopathies

To demonstrate a use case scenario for dataset querying, we applied the top-*k* Kendall's correlation method with several queries to find similar datasets within our database. In this scenario, we confirmed known similarities in expression signatures, providing evidence of the robustness of the query signature. Additionally, we demonstrated how our results may be used as a starting point for meta-analysis.

The query experiment was a profiling of multiple tauopathies using post-mortem tissue from the medial temporal lobe (E-MEXP-2280) [17]. Four neurodegenerative diseases were studied in the query experiment: Alzheimer's disease, Pick's disease, progressive supranuclear palsy, and frontotemporal dementia. The

Disease	Count	top-k p-value	Overlap p-value	GO p-value
myopathy	14	$<$ **0.0001**	0.30	**0.001**
breast cancer	11	**0.0001**	0.53	**0.03**
arthritis	6	**0.0004**	0.22	**0.01**
Huntington's	5	**0.004**	0.49	0.33
lung disease	11	0.12	**0.05**	0.17
leukemia	10	0.12	0.54	**0.01**
encephalopathy	8	0.33	**0.001**	0.49

Table 1.3: The empirical p-value of average pair-wise expression profile distance in a disease classification.

most similar dataset was a study of the aging human brain (GSE1572), where they profiled the postmortem brain tissue of the prefrontal cortex in neuropathologically normal individuals ranging from 26 to 106 years of age [67]. A large challenge in studying tauopathies is discriminating between tauopathic-related and normal age-related changes, due to considerable clinical and molecular overlap. Another significant hit was a study on Alzheimer's disease (GSE1297), which profiled brain hippocampi from postmortem subjects with incipient, moderate, and severe Alzheimer's disease.

We performed a simple meta-analysis, comparing and contrasting the results of each study using gene set enrichment analysis (Figure 1.4). From our simple comparison, we can identify gene sets which may be specific to disease or age, e.g. "generation of precursor metabolites and energy" appeared to be more specific to the disease state.

1.3.4 Metric Evaluation

Comparing expression profile distances to annotation distances has the caveat that for a single comparison, we would not necessarily expect the annotation similarity to be similar to the expression similarity. For example, two experiments studying Huntington's disease in different mouse models, i.e. with similar annotations, may have quite different differential expression profiles due to the nature of the mouse models. We would also not expect expression profile distances to have a high correlation with annotation distances for distantly related datasets, as these longer distances are much noisier.

We evaluated three types of metrics for comparing expression datasets: a simple gene overlap method, a non-parametric method, and a gene set-based method. According to our comparison using the annotations (see Figure 1.6), the top-k Kendall's method has the highest correlation with the annotations, although it did not always perform better (see Figure 1.7). The gene overlap method underperformed likely due to its reliance on a fixed threshold for identifying genes as differentially expressed. As a result, datasets with a similar number of differentially expressed genes tended to cluster together (see Supplementary Figure 1.5).

We observed how closely expression profile distances corresponded to the manual classifications by calculating the average pair-wise expression profile distance between datasets with the same classification and its empirical p-value using Monte Carlo simulations (see Table 1.3). The dataset disease labels were used to to calculate average precision of a similarity profile (see Figure 1.7).

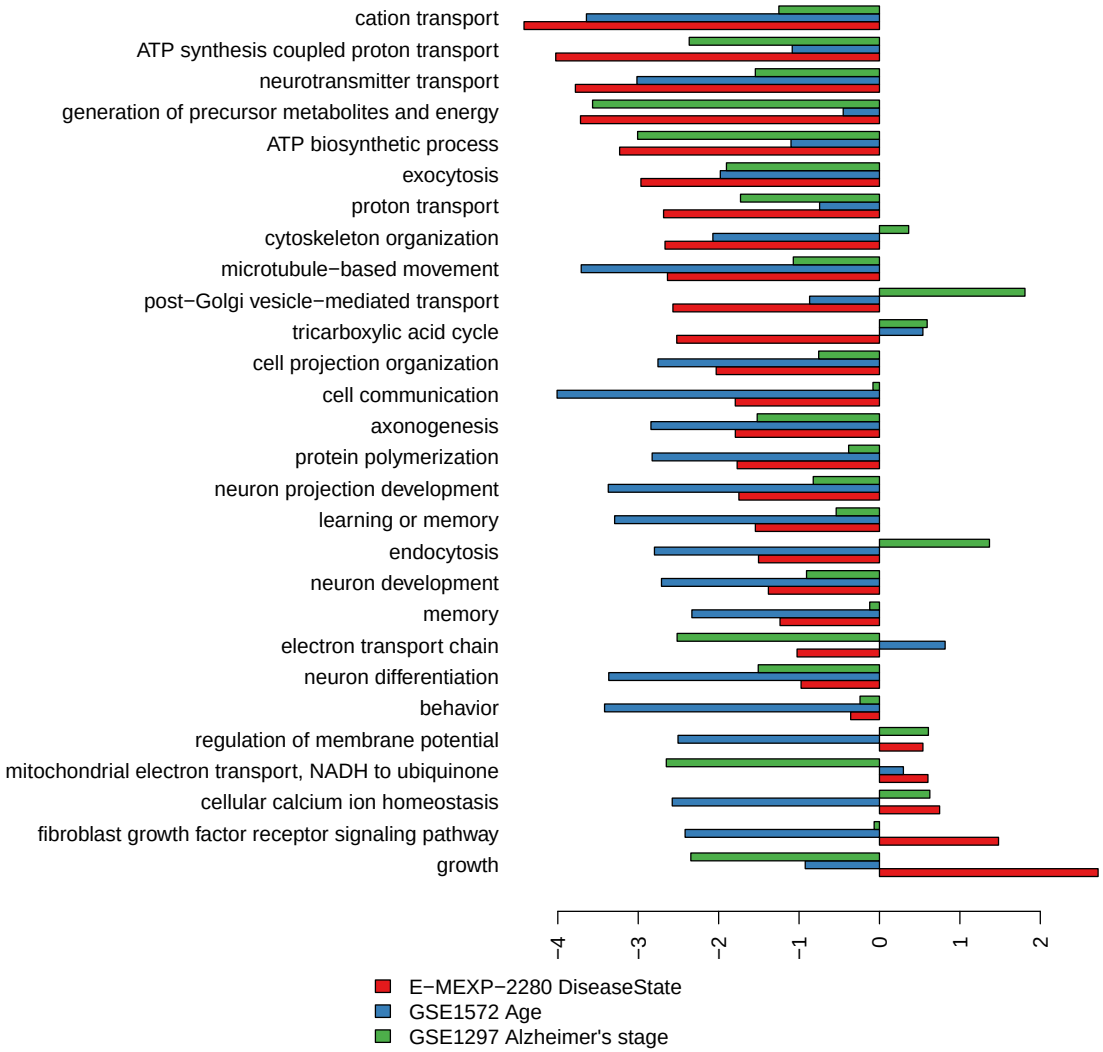


Figure 1.4: Barplot of enriched GO terms in result sets similar to the multiple tauopathies differential expression signature (E-MEXP-2280). Values are probit-transformed p-values.

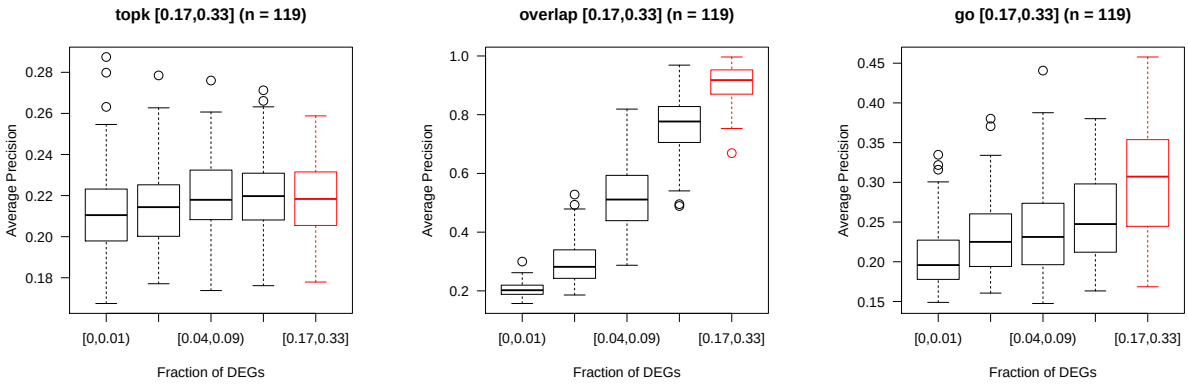


Figure 1.5: Certain methods (overlap and go) cluster datasets based on fraction of DEGs. We plot the distributions of average precision for recovering result sets according to the fraction of DEGs using similarity profiles of result sets with a high fraction of DEGs ([0.17,0.33]). (topk: top-k Kendall's distance; overlap: binary gene Jaccard's distance; go: enriched GO Jaccard's distance)

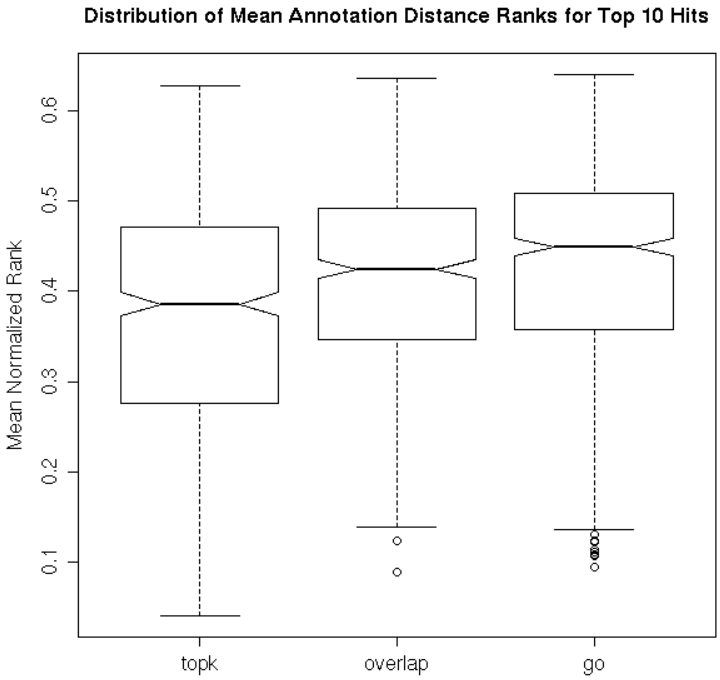


Figure 1.6: The distribution of mean rank ratios across the top 10 hits for a dataset. Each dataset was queried against all other datasets using one of the three methods (topk: top-k Kendall's distance, overlap: binary gene Jaccard's distance, go: enriched GO Jaccard's distance). We took the mean rank of annotation profile distances from the query to the top 10 hits. Lower ranks indicate higher concordance. Notches indicate roughly a 95% confidence interval for the median ($\pm 1.58 IQR/\sqrt{n}$).

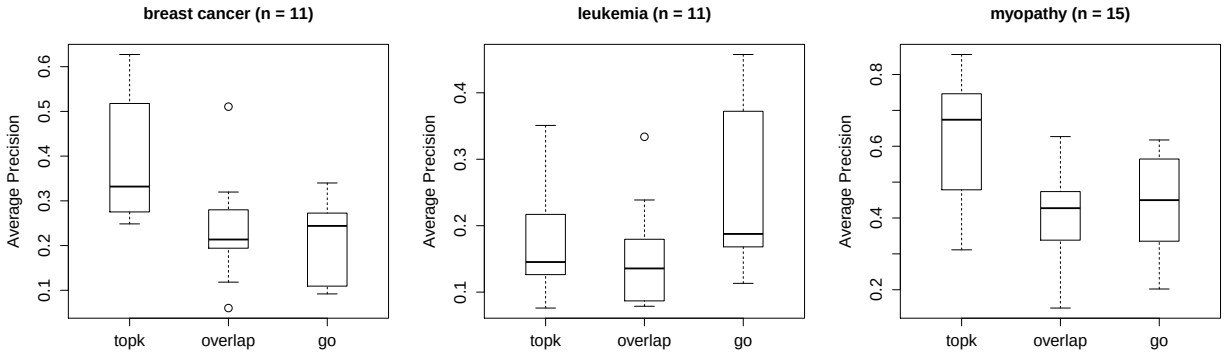


Figure 1.7: The distribution of average precisions of disease classified dataset similarity profiles. We took the average precision of a query for each result set classified to a disease against all other disease-classified result. (topk: top-k Kendall's distance, overlap: binary gene Jaccard's distance, go: enriched GO Jaccard's distance.)

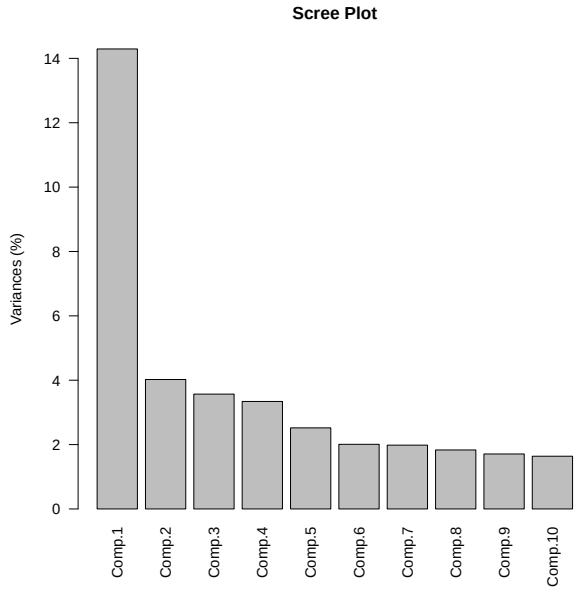
We have tested other methods including Spearman's rank correlation, weighted rank-based methods, and gene coexpression-based methods that are not shown here. Unweighted rank methods did not perform well due to the noisiness of genes that are not significantly differentially expressed. Weighted rank-based methods require a suitable weight function which we have yet to determine. Gene coexpression-based methods hold similar promise to Gene Ontology based methods in that they may mitigate bias due to the univariate statistics yet without some of the drawbacks of relying on the Gene Ontology, e.g. term coverage and overlap of terms.

1.3.5 Platform Effect

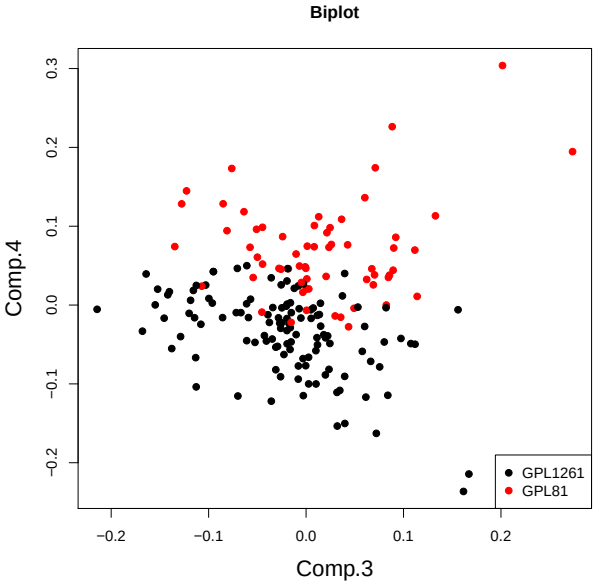
According to the clustering of datasets, platforms may play some role in determining dataset similarity (see Figure 1.10). We examined the platform effect more closely for two main Affymetrix platforms: Mouse Genome 430 2.0 (GPL1261, 137 experiments) and Mouse Genome U74 Version 2 (GPL81, 70 experiments). We performed principal component analysis (PCA) on the differential expression results (probit-transformed p-values), restricted to datasets using only these two platforms and their common genes. There was a clear separation of datasets according to the loadings of components 3 and 4 (Figure 1.8b); however, components 3 and 4 only accounted for a small percent of the variance (Figure 1.8a). The separation of platforms was possibly due to differing number of probes per gene, resulting in small statistical power differences between platforms. The component 4 scores correlated with the number of probes per gene (see Figure 1.9). Gene set-based methods seemed to mitigate this effect as we may expect from methods that summarize information across genes.

1.3.6 Dominant Differential Expression Patterns

Certain datasets are frequently returned near the top of search results. For example, GSE18597 was in the top 10 results for 72/633 query datasets using the top-k rank method. Thus, it was returned as similar to



(a) Scree Plot



(b) Biplot

Figure 1.8: Platform effects account for a significant fraction of the variance in differential expression. We performed principal component analysis on the probit-transformed differential expression p-values of common set of genes on two platforms (GPL81 and GPL1261). Biplot of components 3 and 4 coloured by platform. The major component (component 1) tended to describe the shape of the p-value distribution and the number of DEGs.

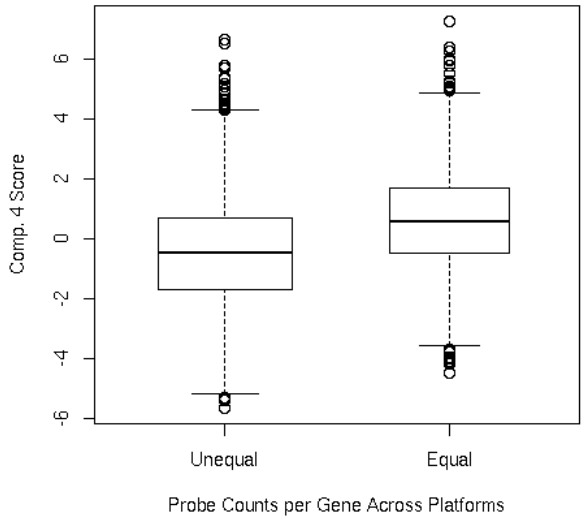


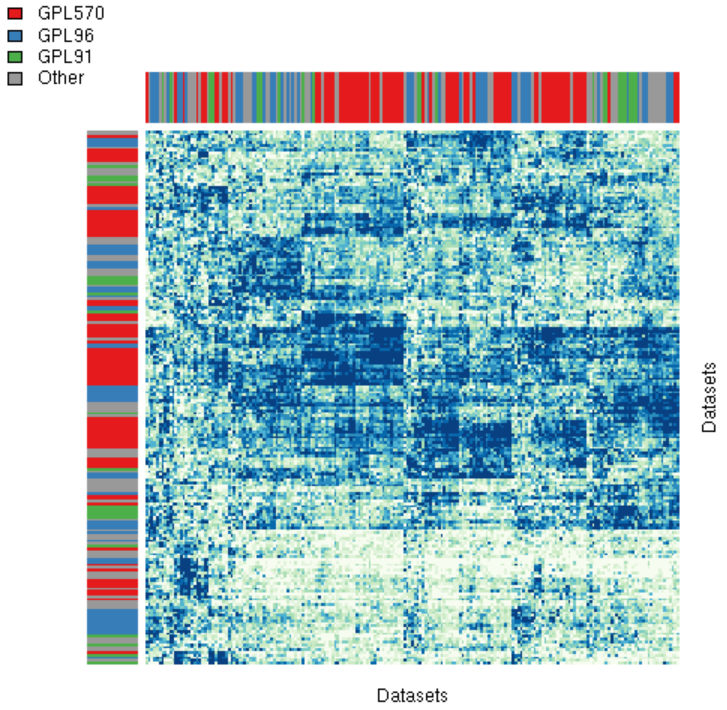
Figure 1.9: Component 4 scores were correlated with the number of probes per gene. We plotted the distribution of component 4 scores split by equal and unequal probe per gene counts across two mouse platforms (GPL1261 and GPL81).

datasets on a wide range of topics. We plot this "dominance", i.e. the number of times that a dataset was ranked in the top 10 for a query, in Figure 1.11. Again, using the top-k rank method, there were 42 datasets which come up in the top 10 results for at least 30 other datasets. While these dominant datasets were not very numerous, they account for around 23% of the top 10 results for all dataset queries, and more than 60% of dataset queries have a dominant dataset ranked in the top 5. A similar phenomenon can be observed using other methods, i.e. the dominant pattern is reasonably robust, showing generally high correlation between methods (Figure 1.11).

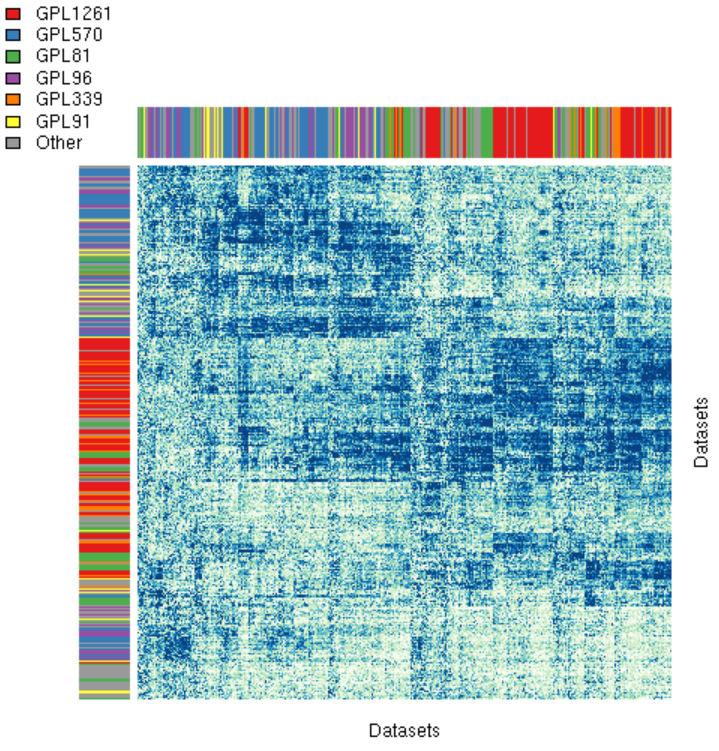
The dominance phenomenon can be partly attributed to some datasets having a higher degree of differential expression; datasets with little to no differential expression tended to correlate more with datasets that have common DEGs. However, even if we removed datasets with few DEGs, certain expression profiles remained highly enriched (in Figure 1.11, such datasets have been removed).

Dominant datasets can be characterized by their own gene rankings, which we call "dominant expression patterns". Dominant expression patterns likely reflected real biological properties of the data, which should be captured by the similarity metric. However, we were concerned that dominant patterns were masking the presence of more subtle similarities. Also, from an information retrieval perspective, it is not particularly useful to always return the same results, no matter what the query.

In order to examine the biological properties reflected by dominant expression patterns, we performed a GO term enrichment analysis. Enriched biological processes included "immune reponse", "cell proliferation", "cell death", and "protein biosynthesis". However, with this straightforward examination of enriched gene sets, it was not clear which gene sets are driving dominance, or how many dominant expression patterns exist. We thus selected several of the most dominant datasets, and for each one, we identified similar



(a) Human



(b) Mouse

Figure 1.10: Clustered heatmap of rank-transformed top-k Kendall similarities. Darker colours indicate higher rank of similarity or higher relative similarity. Side bars indicate array platform.

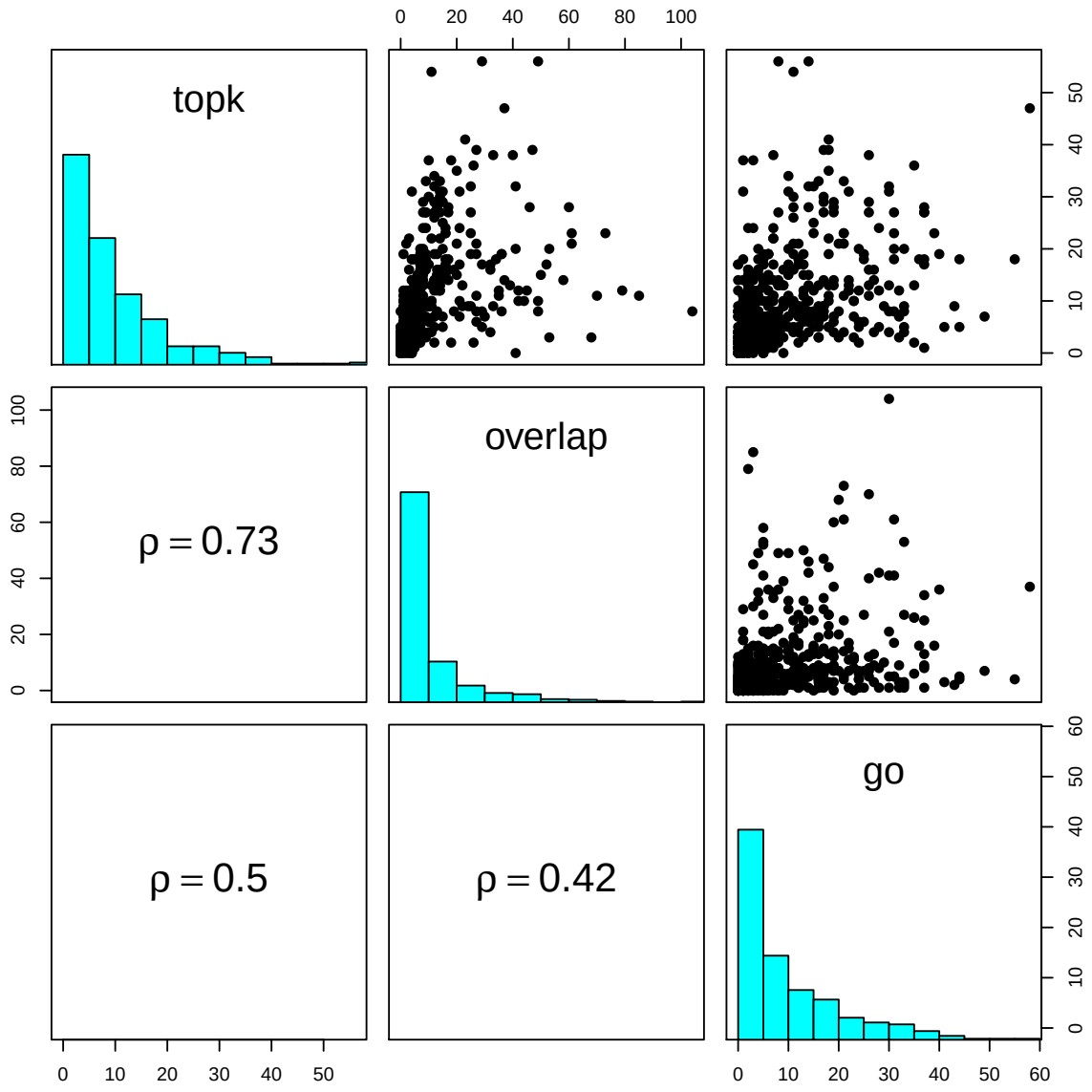


Figure 1.11: Dominance distributions and pair-wise scatter plots: dominance refers to the number of times that a result set appeared in the top 10 of a query. Spearman's rank correlation coefficient is shown. topk: top-*k* Kendall's Tau distance, overlap: binary gene Jaccard's distance, go: enriched GO Jaccard's distance.

datasets, i.e. datasets which had the dominant dataset in the top 10 results. We summarized across these similar datasets by taking average gene ranks to construct a "meta-signature". For example, as mentioned previously, GSE18597 was in the top 10 results for 72 query datasets, so the meta-signature for GSE18597 consisted of average gene ranks across the 72 query datasets. Compared to the individual signatures (Figure 1.12), the meta-signatures (Figure 1.13) exhibited many similarities, yet had reduced heterogeneity. Nonetheless, there is still evidence of the presence of multiple dominant expression patterns that may involve gene sets such as "gene expression", "signal transduction", and "cell differentiation" due to their heterogeneity in enrichment.

To examine the annotations associated with dominant expression patterns, we conducted an annotation enrichment analysis on all query results (described in Section 1.2.4). Common expression patterns were thus associated with annotations such as muscular dystrophy, and immune response (see Figure 1.16). Muscular dystrophy studies involved a large cell death and inflammation signal [89, 90, 112], which was consistent with our enriched GO term findings in dominant expression patterns (Figure 1.12 and 1.13).

Dominant expression profiles were also enriched for frequently DEGs (Figure 1.15), suggesting that search results may be driven by genes with low information content (which is correlated with frequency of differential expression). Filtering based on gene information content mitigated the problem of non-specific correlation with dominant signatures, most notably with the top-k Kendall's distance method (Figure 1.14). In addition, filtering low information content genes slightly decorrelated dataset enrichment for frequently DEGs with dominance (Figure 1.15). Removing genes or ontology terms with low information content neither degraded nor improved performance significantly (Figure 1.17b), but this may be due to the coarseness of our performance metrics.

An important question is whether the effect of dominant expression patterns is a function of our choice of metrics. That is, would different methods be less sensitive to these patterns. As mentioned in the Introduction, there are three fundamental types of similarity algorithms employed: correlation-based, overlap-based, and gene set-based, which we have implemented. While we have not implemented a threshold-free method such as RRHO [87] or GSEA [109], results from a scaled down analysis on a single platform using Spearman's correlation distance gave similar results. We have partially addressed the problem of dominant expression patterns by removing low information content genes. To further address the problem, we may need to employ more sophisticated filtering or weighting of genes that takes into account patterns of co-differential expression, i.e. groups of genes that behave similarly under different subsets of experimental conditions.

In contrast to dominant expression patterns, there were expression profiles that were uncommonly observed, i.e. signatures which did not correlate relatively highly with any other expression profiles. It is possible that this was due to gaps in our database, or that the dominant signatures were masking similarities. Some of these datasets did exhibit a low proportion of statistically significant differentially expressed genes or had non-uniform null p-value distributions, suggesting the presence of unmodeled sources of variation [65]. This suggested that part of the problem may be addressed by improving pre-processing and statistical analysis to help control for unmodeled variation. We discuss this further in the next section.

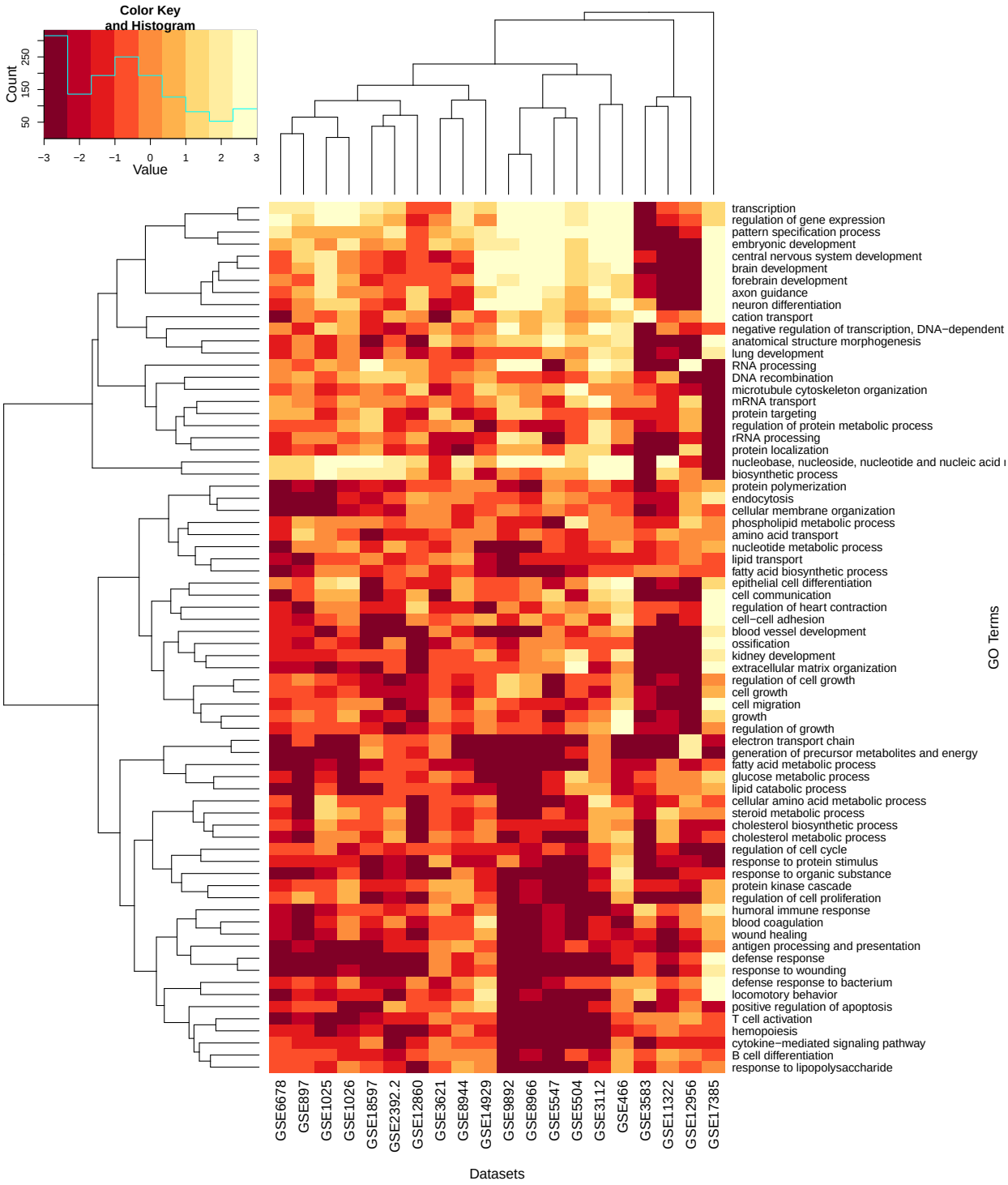


Figure 1.12: Enriched GO terms among top-20 dominant expression signatures (datasets which showed up most frequently in the top-10 hits of all top-k Kendall queries). Values are probit-transformed p-values restricted to the range [-3, 3].

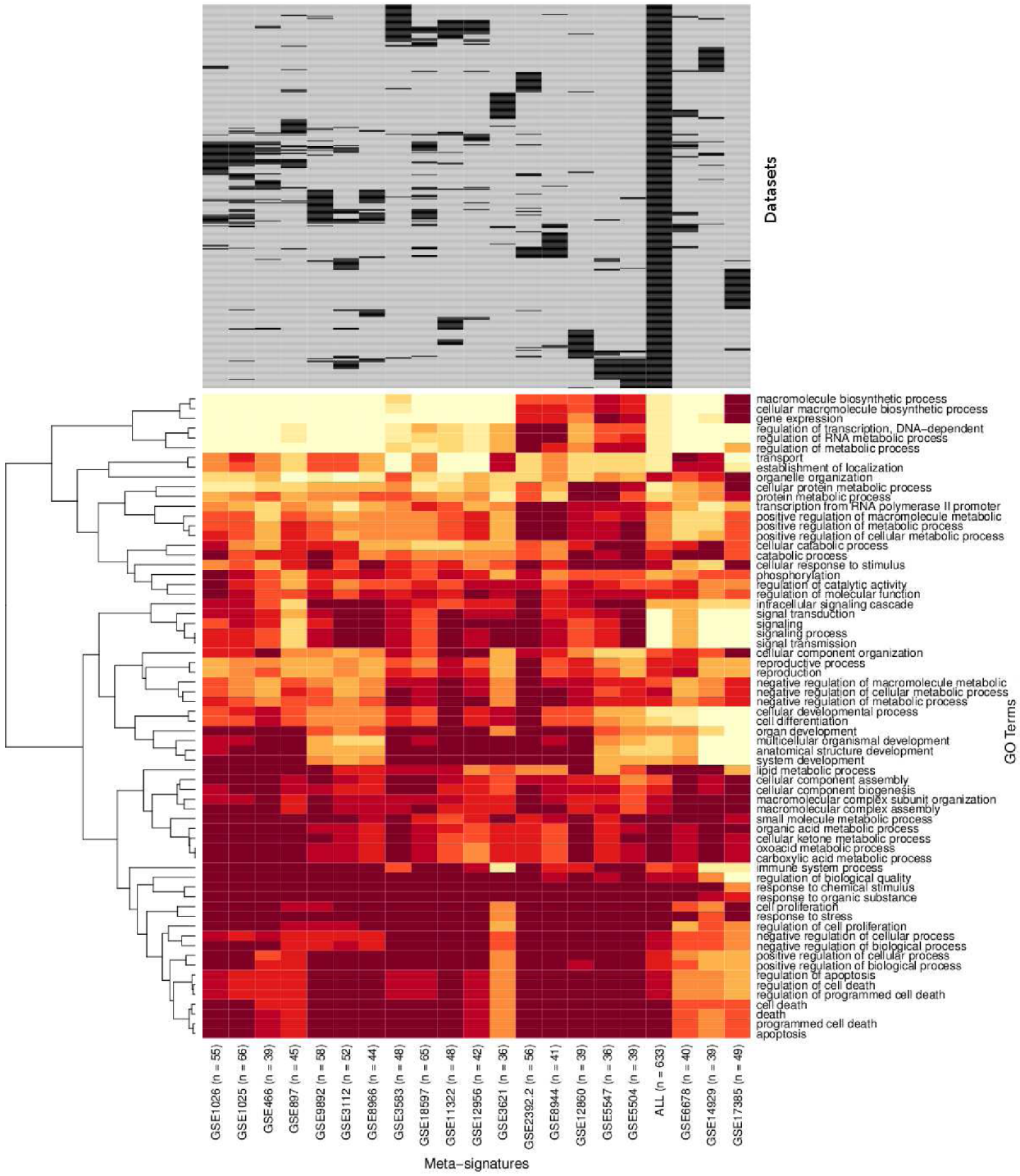


Figure 1.13: Heterogeneity in enriched GO terms of meta-signatures of dominant datasets. Darker colours indicate higher enrichment of the gene set. We constructed meta-signatures from summarizing expression profiles of query datasets which gave the dominant dataset as a top 10 result. An exception is the "ALL" meta-signature which summarizes across all datasets. The datasets included in each meta-signature are shown in the barcode blot above the heatmap (black: included, grey: excluded). Summarization of expression profiles consisted of averaging gene ranks.

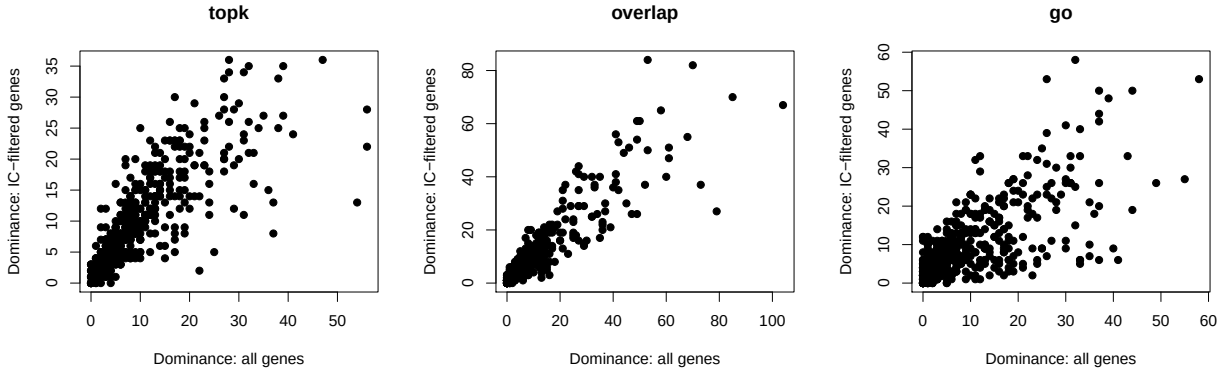


Figure 1.14: Filtering out low information content genes (IC-filtered genes) tended to reduce result set dominance especially using top-k Kendall's distance. Dominance refers to the number of times that a result set appeared in the top 10 of a query. Information content was defined as the $-log_2(P(g))$ where $P(g)$ was the fraction of all differential expression accounted for by gene g (q-value < 0.05). The filter threshold was the 10th percentile of the information content distribution, and was set using our evaluation protocol. topk: top-k Kendall's distance, overlap: binary gene Jaccard's distance, go: enriched GO Jaccard's distance.

1.3.7 Outliers and Batch Effects

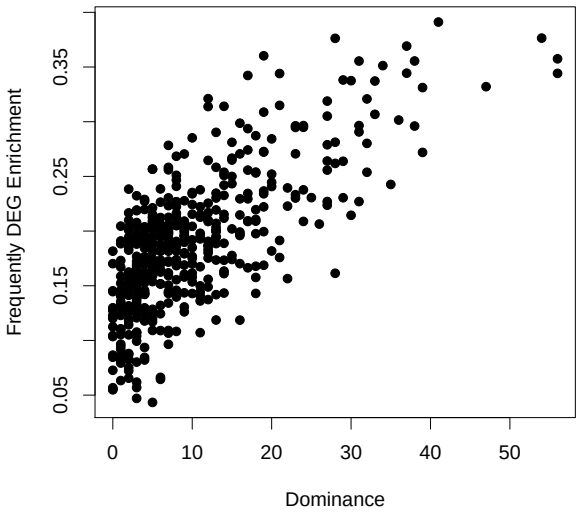
We investigated whether the presence of outliers may be adversely affecting the results. Datasets lacking outlier samples had expression signatures which were more similar on average (see Figure 1.18), which suggested that the presence of outliers may be indicative of lower data quality. Outliers appeared to have a small effect on the performance of methods as measured using annotations (see Figure 1.17a), which may be due to the coarseness of this performance metric.

Some of the studies included in our analysis had batch affected probes or experimental designs confounded with batch. Preliminary analysis did not indicate a large effect on dataset similarity.

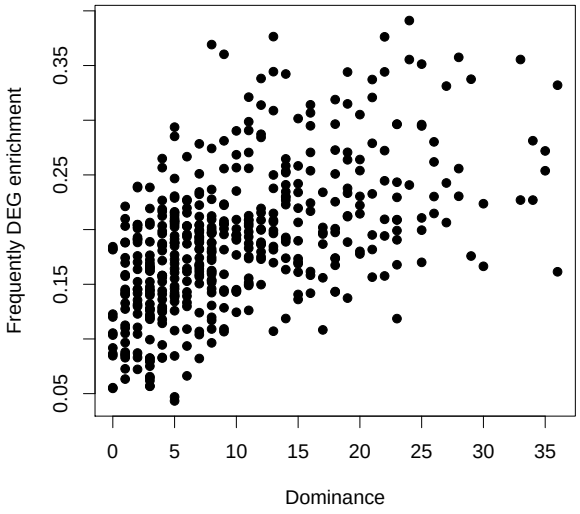
1.4 Concluding Remarks and Future Work

We have described a framework for the evaluation of methods for comparing gene lists from gene expression studies and given a use case example that details the potential worth of such a method. One of my main contributions was identifying "dominant expression patterns" dominating search results, which was not completely due to frequently differentially expressed genes. Future work will be concentrated in addressing this issue.

As mentioned in the Introduction, there exist a number of other algorithms and techniques for querying gene expression data, most of which are based on GSEA. We would like to implement them into our framework as they may offer increased levels of sensitivity or robustness. Different statistical analysis of the data may be helpful as well. For increased inter- and intra-laboratory data comparability, Reina-Pinto *et al.* suggested a rank-product analysis for determining DEGs [94]. Another idea is to incorporate ideas from the information retrieval literature. Term frequency-inverse document frequency weighting measures and its



(a) No Gene Filtering



(b) Low Information Content Genes Filtered

Figure 1.15: Expression pattern dominance is correlated with enrichment for frequently DEGs. Dominance was measured by the number of times that a dataset appears in the top 10 results for a query. In Figure 1.15b, genes with low information content are not considered when comparing datasets. Enrichment for frequently DEG genes was measured by Kendall's top-k similarity to genes scored by frequency of differential expression. Spearman's rho: (1.15a) -0.67, (1.15b) -0.55

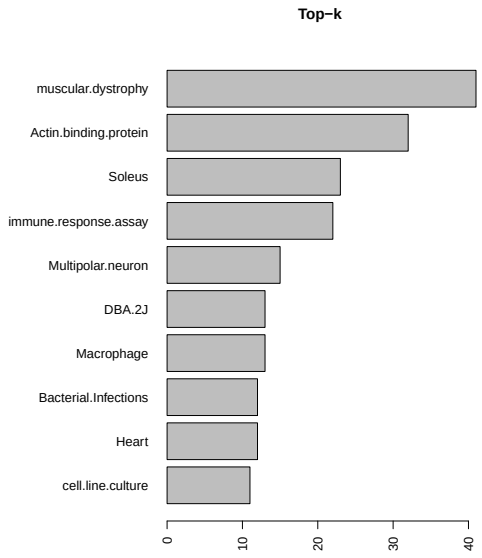
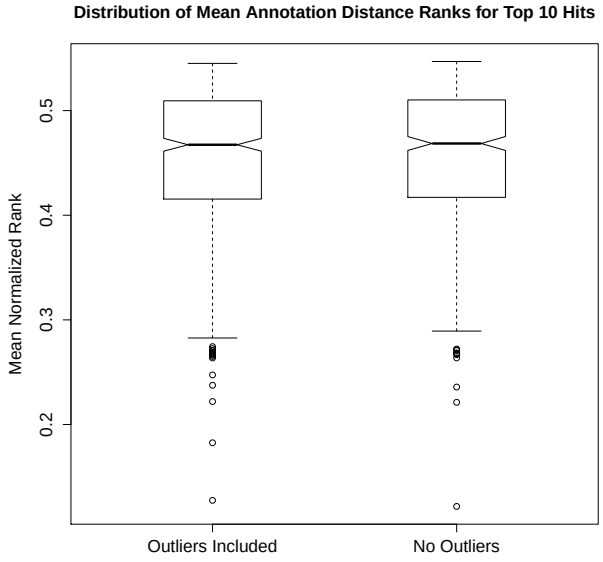


Figure 1.16: Barplot of top 10 enriched dataset annotations using top-k Kendall's tau similarity profiles (partial ROC AUC > 0.7 and p-value < 0.01, FDR < 0.5). Highly co-occuring annotations are not shown.

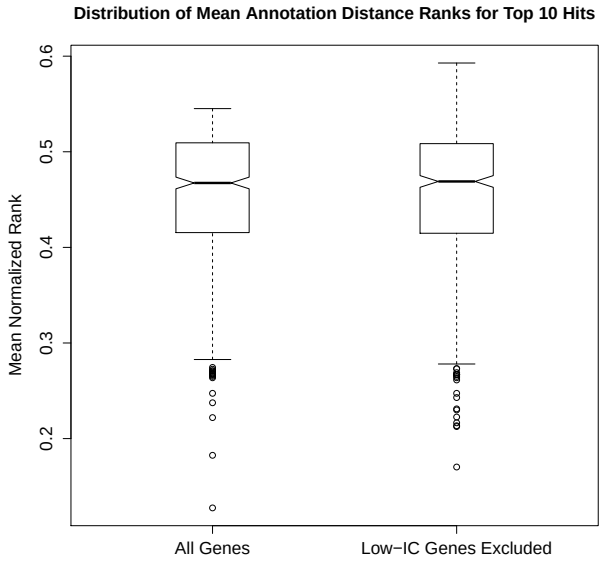
variants may be employed to moderate the presence of frequent terms (genes) [3, 98]. Similarly, document length normalization may be a useful method. Document length normalization would make retrieval of gene lists of different length, i.e. from different platforms or from studies with differing numbers of differentially expressed genes, roughly equally probable [106].

The current framework did not take into account directional change in expression, i.e. whether gene expression is up or down-regulated relative to the control. Making this distinction may allow for a finer level of resolution when comparing gene expression signatures.

In closing, the identification of similarities between datasets is an unsolved problem, with a key challenge being dealing with the impact of the distortive effect of non-specific patterns of differential expression.



(a) Outliers



(b) Gene Information Content

Figure 1.17: Outliers (1.17a) or low information content genes (1.17b) did not influence much the performance according to annotations. We plot the distribution of mean rank ratios across the top 10 hits for a dataset with outliers or low information content genes included and excluded. Low information content genes are those below the 10th percentile. Each dataset was queried against other datasets using top-k Kendall's tau. We took the mean rank of annotation profile distances from the query to the top 10 hits. Lower ranks indicate higher concordance. Notches indicate roughly a 95% confidence interval for the median ($\pm 1.58 IQR/\sqrt{n}$).

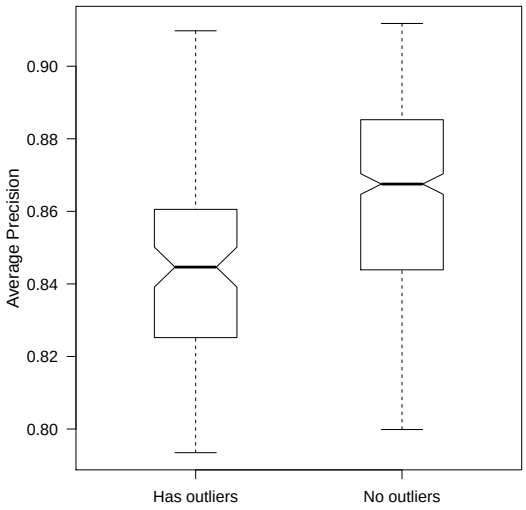


Figure 1.18: Removing outliers improved quality of data retrieval results. We plotted the average precision for recovering other datasets with outliers and datasets without outliers using similarity profiles of datasets without outliers. Notches indicate roughly a 95% confidence interval for the median ($\pm 1.58 IQR/\sqrt{n}$).

1.5 Supplementary Data

Name	ID	Num Genes	ROC AUC	P-value	Multifunctionality Bias
tRNA metabolic process	GO:0006399	93	0.69	1.395E-07	0.65
positive regulation of I-kappaB kinase/NF-kappaB cascade	GO:0043123	84	0.67	7.993E-05	0.67
tRNA processing	GO:0008033	58	0.70	6.649E-05	0.62
regulation of I-kappaB kinase/NF-kappaB cascade	GO:0043122	91	0.66	9.015E-05	0.67
electron transport chain	GO:0022900	80	0.66	1.464E-04	0.67
ribosome biogenesis	GO:0042254	88	0.65	4.556E-04	0.59
DNA-dependent DNA replication	GO:0006261	47	0.69	8.2E-04	0.77
peptide metabolic process	GO:0006518	41	0.70	9.883E-04	0.78
cell cycle checkpoint	GO:0000075	72	0.65	9.992E-04	0.79
centromere complex assembly	GO:0034508	5	0.95	1.804E-03	0.93
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	GO:0031145	44	0.69	1.93E-03	0.69
regulation of ubiquitin-protein ligase activity	GO:0051438	56	0.66	2.066E-03	0.7
positive regulation of ubiquitin-protein ligase activity	GO:0051443	50	0.67	1.954E-03	0.7
rRNA processing	GO:0006364	67	0.65	1.946E-03	0.61
regulation of ubiquitin-protein ligase activity during mitotic cell cycle	GO:0051439	49	0.67	2.027E-03	0.7
chromosome segregation	GO:0007059	66	0.65	2.411E-03	0.76
amino acid activation	GO:0043038	39	0.69	2.657E-03	0.69
rRNA metabolic process	GO:0016072	70	0.64	2.747E-03	0.62
cellular respiration	GO:0045333	66	0.65	2.751E-03	0.77
regulation of ligase activity	GO:0051340	59	0.65	2.631E-03	0.71

Table 1.4: Enriched GO terms among commonly differentially expressed genes scored using ROC. P-values are multiple test corrected using Benjamini-Hochberg. Multifunctionality bias is the degree to which that gene set contains multifunctional genes (ROC AUC).

Name	ID	Num Genes	ROC AUC	P-value	Multifunctionality Bias
sensory perception of chemical stimulus	GO:0007606	67	0.77246726	9.467E-13	0.58
G-protein signaling, coupled to cyclic nucleotide second messenger	GO:0007187	98	0.73009712	2.663E-12	0.77
regulation of cyclic nucleotide metabolic process	GO:0030799	97	0.71586494	1.011E-10	0.84
regulation of nucleotide biosynthetic process	GO:0030808	93	0.72027197	7.88E-11	0.84
sensory perception of smell	GO:0007608	43	0.79951866	1.539E-10	0.51
regulation of lyase activity	GO:0051339	81	0.72993986	2.068E-10	0.83
regulation of nucleotide metabolic process	GO:0006140	99	0.70742789	2.112E-10	0.84
G-protein signaling, coupled to cAMP nucleotide second messenger	GO:0007188	69	0.73984301	1.926E-10	0.81
regulation of cyclase activity	GO:0031279	80	0.72833432	2.714E-10	0.82
regulation of cAMP biosynthetic process	GO:0030817	85	0.72126153	2.614E-10	0.83
regulation of cAMP metabolic process	GO:0030814	88	0.71341308	6.07E-10	0.83
regulation of adenylate cyclase activity	GO:0045761	78	0.72520848	7.493E-10	0.82
cAMP-mediated signaling	GO:0019933	77	0.70890241	2.61E-08	0.8
positive regulation of lyase activity	GO:0051349	47	0.75274024	4.818E-08	0.83
positive regulation of cyclase activity	GO:0031281	46	0.75047609	1.004E-07	0.83
positive regulation of adenylate cyclase activity	GO:0045762	45	0.74795326	2.147E-07	0.83
activation of adenylate cyclase activity	GO:0007190	44	0.74264085	6.933E-07	0.82
digestion	GO:0007586	66	0.68965455	4.639E-06	0.76
feeding behavior	GO:0007631	56	0.70234012	6.598E-06	0.87
regulation of heart contraction	GO:0008016	72	0.68116185	7.646E-06	0.87

Table 1.5: Enriched GO terms among uncommonly differentially expressed genes scored using ROC. P-values are multiple test corrected using Benjamini-Hochberg. Multifunctionality bias is the degree to which that gene set contains multifunctional genes (ROC AUC).

Chapter 2

Role of MicroRNA in Major Depression and Suicide

2.1 Introduction

Major depressive disorder (MDD), commonly referred to as major depression, is a possibly recurrent mood disorder with symptoms that include low mood, loss of interest in pleasure, insomnia or hypersomnia, fatigue or loss of energy, and suicidal ideation [97]. Major depression affects a large proportion of the population with one-year prevalence estimates ranging between 6.4% and 10.1% [53, 54, 77, 117]. Consequently, major depression imposes a large economic burden on society; for instance, the direct and indirect costs were estimated at about $83 billion for the year 2000 in the United States [40].

Suicide is another major public health problem that is often associated with MDD; suicide rates among mood disorder sufferers are more than 20-fold higher than the general population [88]. There is mounting evidence that individuals who commit suicide have a genetic predisposition [56, 69].

Studies of MDD and suicide have suggested several molecular causes. Although it is possible that the genetic factors for predisposition for MDD and suicide are independent [16, 56, 107], there is evidence that they may be linked. For example, spermidine/spermine N^1-acetyltransferase (SAT1) was observed to have decreased expression in suicide completers [35, 42, 58, 101, 102], and a genetic variant that predicts SAT1 expression was associated with depressed suicide completers compared to depressed non-suicides [34]. Dysregulation of several neurotransmitter systems in different brain regions have been suggested: the serotonergic [78, 81, 95, 113], dopimanergic [12, 85, 119], noradrenergic [11, 41, 59, 84, 96], glutamate [20, 23, 32, 33, 48, 52, 79] and GABAergic [18, 20, 57, 61, 74, 91, 99] systems. Other molecular systems that have been implicated include the cylic adenosine monophosphate response element binding (CREB) protein signaling pathway [27], the immune system [8, 38], fibroblast growth factor [28], ATP biosynthesis [57], and cell proliferation [110].

Neuroimaging and histopathological studies of major depression and suicide have revealed abnormalities in the prefrontal cortex (PFC) [25, 92], and hippocampus [15, 44, 104]. The PFC plays a role in a diverse range of executive processes, including impulse control, working memory, attention, and judgement. The

hippocampus plays a role in the formation of new memories as well as cognitive maps. Additionally, the PFC and hippocampus have been implicated in other neurological disorders such as schizophrenia [46, 116], and bipolar disorder [37, 93].

MicroRNAs (miRNAs) are small single-stranded non-coding RNA molecules (about 19-24 nucleotides) that function primarily to down-regulate gene expression [19, 70]. They act by complementary binding to the 3' untranslated region of a transcript, resulting in translational repression or transcript degradation [66]. There is also evidence that miRNAs may serve to induce transcription [86].

MiRNAs have been implicated in a number of diseases, including some types of cancer, heart disease [49, 50]. Moreover, miRNAs are abundantly expressed in the brain [60], and there is evidence that dysregulation of miRNA function may cause neurodegeneration [45]. In Alzheimer's disease subjects, miR-9, miR-125b and miR-146a were found to be up-regulated in the temporal lobe neocortex [103]. Rare genetic variants of SLITRK1, associated with Tourette's syndrome patients, have a frame shift mutation in its binding site for miR-189, altering their interaction [1]. Spinal muscular atrophy is associated with the Survival of Motor Neuron (SMN) complex, and it was shown that two components of the SMN complex associate with miRNAs to form ribonucleoprotein complexes [24]. Altered miRNA expression was observed in autism post-mortem cerebellar cortex tissue [2], and in the prefrontal cortex of schizophrenia subjects [83].

The objective of this project is to find evidence of differential miRNA regulation in depressed suicides compared to controls. We profiled major depressive and suicide subjects at the mRNA and miRNA level in the hippocampus and prefrontal cortex of postmortem brain tissue.

The first part of this project entailed independent statistical analysis of miRNA and mRNA data. We then attempted to detect correlations between mRNA and miRNA expression levels. Any promising candidates may be experimentally validated.

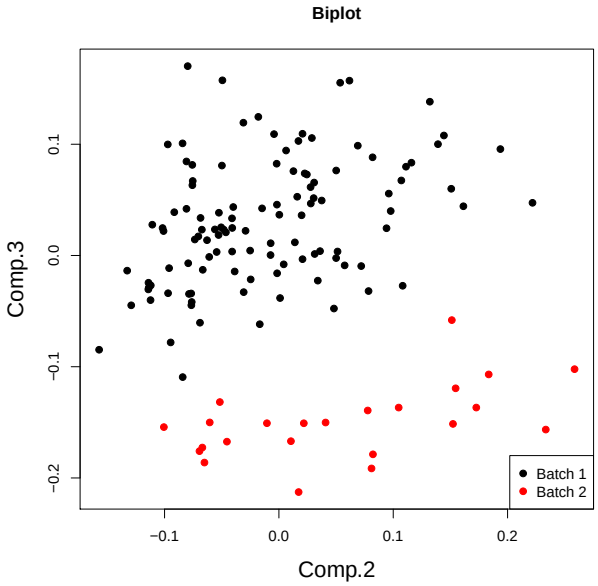
2.2 Methods

2.2.1 Data Overview and Pre-processing

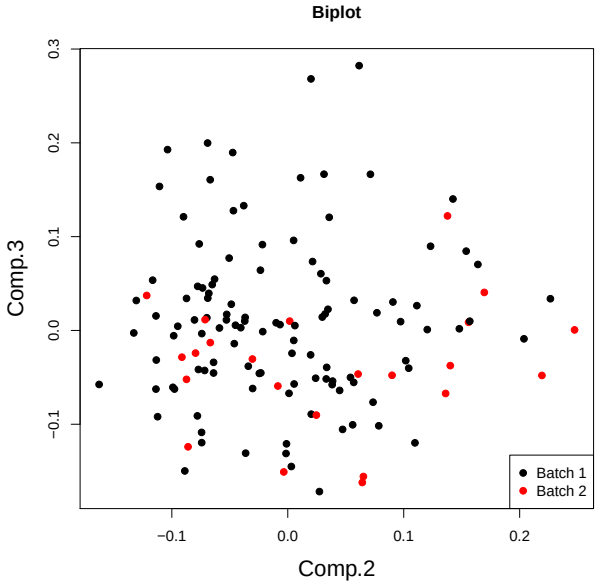
Brain tissue for this project was available through the Québec Suicide Brain Bank, which is a facility managed by Dr. Gustavo Turecki and Dr. Naguib Mechawar of the McGill Group for Sucide Studies (MGSS; Douglas Mental Health University Institute, Mcgill University, Montreal, Québec, Canada). Dr. Gustavo Turecki, our collaborator at the MGSS, oversaw the collection of miRNA and mRNA data using Human Agilent miRNA microarrays and Human Affymetrix U133 Plus 2.0 arrays. They profiled two brain regions: Brodmann area 44 (BA44) and the hippocampus. However, not all brain samples run on the miRNA microarrays were run on the U133 Plus 2.0 arrays and vice versa.

In initial quality control, we found many samples yielded low-quality mRNA or miRNA profiles. Some of these samples were excluded from the analysis, or re-run in new batches. In an attempt to remove the batch effect, we subtracted out for each gene the median difference in gene expression between samples run on both batches (see Figure 2.1).

We used the AgiMicroRna package in Bioconductor to read the miRNA data. The robust multiarray



(a) Pre-batch effect correction



(b) Post-batch effect correction

Figure 2.1: The batch effect was reduced in component 3 loadings of the principal component analysis after batch effect correction. To remove the batch effect, for each gene, we subtracted the median difference in gene expression between samples run on both batches.

	Region	SMDD	S	C
miRNA	BA44	20	13	10
	Hippo	33	17	19
mRNA	BA44	15	9	12
	Hippo	11	8	10
mRNA/miRNA	BA44	8	5	5
	Hippo	3	2	6

Table 2.1: Breakdown of sample counts per brain region and diagnosis. S: suicide, SMDD: major depressive suicides, C: control.

average algorithm, developed for Affymetrix arrays, was used to summarize the data. We normalized the data using using the quantile method. We removed genes that were flagged as absent, leaving us with 447 genes (from 939). Not all samples run on the miRNA arrays were run on the mRNA arrays and vice versa (see Table 2.1).

We processed the mRNA data using Affymetrix's MAS 5.0 expression algorithm, before applying quantile normalization. Genes with low or undetectable expression level (below the 30th percentile) within a diagnosis group were filtered out to minimize spurious hits in the combined analysis.

2.2.2 miRNA Microarray Data: Statistical Analysis

We applied standard linear regression techniques in conjunction with surrogate variable analysis (SVA) [65]. SVA attempts to capture the heterogeneity involved in a gene expression study by incorporating so called "surrogate variables" into the model. We fitted models using the limma Bioconductor package for linear regression [51]. To select a model, we fitted a number of different linear models that had been augmented with surrogate variables to each gene, and scored each model fit using Akaike information criterion (AIC) [13]. AIC measures the goodness of fit while penalizing for greater number of terms in the model. We then chose the model that had the highest number of best AIC scores.

We split the analysis between the two brain regions, such that two models were employed. Aside from the diagnosis and surrogate variables, the BA44 model included effects for pH and batch, and the hippocampus model included pH and PMI. 4 surrogate variables were found significant for BA44, and 6 for hippocampus. Significance in this case means that the surrogate variables captured more expression heterogeneity than expected by chance.

2.2.3 mRNA Microarray Data: Statistical Analysis

We performed a similar statistical analysis of the mRNA microarray data; that is, linear regression in conjunction with SVA. Again, we analyzed BA44 and hippocampus separately, and we fitted a number of different linear models augmented with surrogate variables to each gene, scoring and selecting models using AIC. The selected models included the diagnosis effect, and 3 and 7 surrogate variables (as determined by SVA) for the BA44 and hippocampus models, respectively. Multiple test correction was performed using the Benjamani-Hochberg method [7].

	ID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0007602	0.0011	55.20	0.05	2	7	phototransduction
2	GO:0043627	0.0051	9.55	0.35	3	47	response to estrogen stimulus
3	GO:0031032	0.0054	21.21	0.11	2	15	actomyosin structure organization
4	GO:0048008	0.0054	21.21	0.11	2	15	platelet-derived growth factor receptor signaling pathway
5	GO:0009888	0.0055	3.63	2.18	7	291	tissue development
6	GO:0009991	0.0058	6.11	0.73	4	97	response to extracellular stimulus
7	GO:0009749	0.0078	17.22	0.13	2	18	response to glucose stimulus
8	GO:0034284	0.0078	17.22	0.13	2	18	response to monosaccharide stimulus
9	GO:0007584	0.0083	7.92	0.42	3	56	response to nutrient
10	GO:0007167	0.0085	3.72	1.80	6	240	enzyme linked receptor protein signaling pathway
11	GO:0009582	0.0086	16.21	0.14	2	19	detection of abiotic stimulus
12	GO:0001707	0.0096	15.30	0.15	2	20	mesoderm formation

Table 2.2: BA44 suicide vs. control genes (p-value < 0.01): over-represented biological process GO terms

We tested for over-represented Gene Ontology (GO) terms among the list of differentially expressed genes from each model fit, using the R package GOstats [30], a hypergeometric-based test that uses the relationships among GO terms to decorrelate the results. For each gene, we used the probe with the highest expression variance across samples, and we selected genes for the hypergeometric test with a test p-value less than 0.01.

2.2.4 Combined miRNA-mRNA Analysis

We used Spearman's rank correlation to assess correlation between all possible mRNA transcript probes and miRNA genes from the same samples. We then performed multiple test correction using the method described in [108].

2.3 Results and Discussion

The analysis of the mRNA data only returned one statistically significant hit after multiple test correction: LCT (lactase). In BA44, LCT was up-regulated significantly in suicides vs. controls (q-value: 0.039) as well as major depressive suicides (q-value: 0.058). We list enriched GO terms in Tables 2.2, 2.3, 2.4, and 2.5.

Statistical analysis of the miRNA data yielded one significant (q-value < 0.05) result after multiple test correction: hsa-miR-1202. Hsa-miR-1202 was down-regulated in suicides versus controls in BA44 (q-value: 0.0061, see Figure 2.3). Our collaborators at the MGSS performed several follow-up qRT-PCR experiments on hsa-miR-1202 among other highly ranked miRNAs using the same samples. The results of the qRT-PCR

	ID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0008633	0.0059	20.74	0.12	2	12	activation of pro-apoptotic gene products
2	GO:0045778	0.0059	20.74	0.12	2	12	positive regulation of ossification
3	GO:0007243	0.007	3.12	2.84	8	289	protein kinase cascade
4	GO:0045669	0.008	17.28	0.14	2	14	positive regulation of osteoblast differentiation
5	GO:0031032	0.0092	15.95	0.15	2	15	actomyosin structure organization

Table 2.3: BA44 depressed suicide vs. control genes (p-value < 0.01): over-represented biological process GO terms

	ID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0008380	0.0036	5.57	1.06	5	135	RNA splicing
2	GO:0016071	0.0077	4.61	1.26	5	161	mRNA metabolic process

Table 2.4: Hippocampus suicide vs. control genes (p-value < 0.01): over-represented biological process GO terms

validated the microarray finding (Figure 2.4).

We hypothesized that miRNA and mRNA expression may be correlated. However, we did not find any statistically significant correlated or anti-correlated genes after multiple test correction (0.05 FDR level). We also hypothesized that correlation in general may be enriched among predicted miRNA-mRNA target pairs. Nunez-Iglesias *et al.* and Tsang *et al.* observed an increase in positively-correlated target pairs within the brain [80, 111], but we were unable to replicate this finding in our data, possibly due to small sample size and data quality.

Based on the correlation analysis, we identified several candidates for regulation by hsa-miR-1202 in Table 2.6. The expression of the candidate genes were correlated with hsa-miR-1202 expression and were decreased or increased in suicides versus control samples. Two interesting targets were ATXN7 and KIAA0319, for which both have support for hsa-miR-1202 binding sites according to the miRanda algorithm [9]. A polyglutamine expansion in ATXN7 is the cause of spinocerebellar ataxia type 7 (SCA7), a neu-

	ID	P-value	Odds Ratio	Exp Count	Count	Size	Term
1	GO:0045785	0.0018	39.63	0.07	2	14	positive regulation of cell adhesion
2	GO:0007422	0.0024	33.95	0.07	2	16	peripheral nervous system development
3	GO:0022610	0.0026	6.47	1.01	5	217	biological adhesion
4	GO:0007601	0.0088	16.33	0.14	2	31	visual perception

Table 2.5: Hippocampus depressed suicide vs. control genes (p-value < 0.01): over-represented biological process GO terms

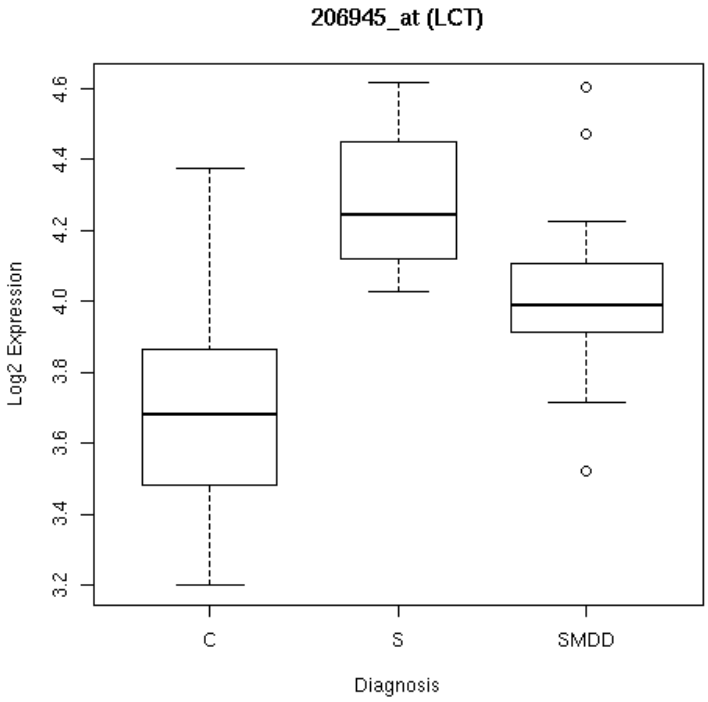


Figure 2.2: Boxplot of lactase (LCT) log2 expression under each diagnosis level in BA44. C: control ($n = 11$). S: suicides ($n = 8$). SMDD: major depressive suicides ($n = 15$). S vs. C q-value is 0.039. SMDD vs. C q-value is 0.058.

rodegenerative disorder [22, 75] that is characterized by macular degeneration, dysphagia, and dysarthria [4]. Spinocerebellar ataxia patients also tend to exhibit depressive symptoms [71]. KIAA0319 encodes a transmembrane protein that when defective may cause susceptibility to dyslexia type 2 [21, 43, 82].

2.4 Concluding Remarks and Future Direction

We encountered many challenges regarding the quality of the data, and much effort was spent in quality control, normalization, and increasing statistical power. Further efforts in this area may be focused on different normalization methods and more sophisticated batch correction algorithms such as using regression modeling to identify differences between batches for removal [10].

In conclusion, we are following up on a few promising candidates experimentally. Further miRNA genes may be tested for differential expression using qRT-PCR. Certain anti-correlated miRNA mRNA-target pairs may also be tested experimentally, particularly those anti-correlated with hsa-miR-1202, in order to validate interaction.

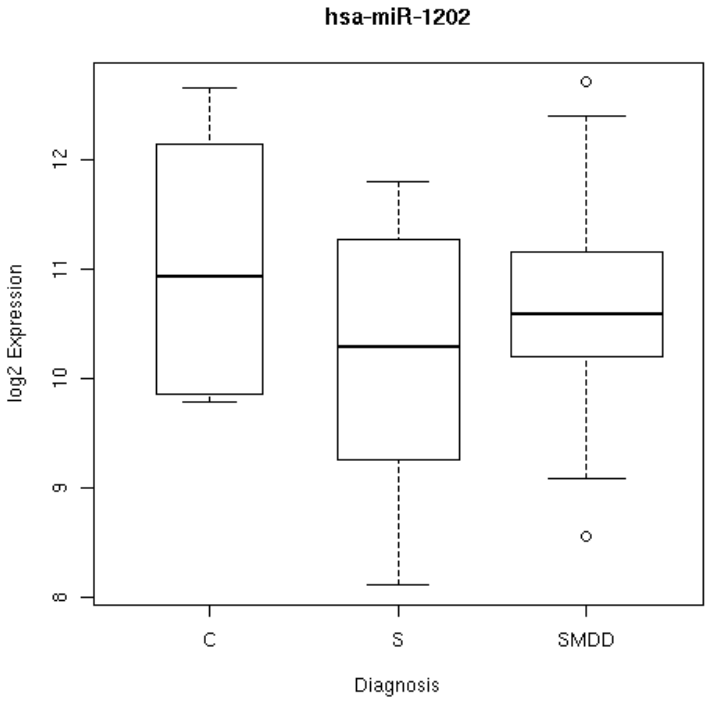


Figure 2.3: Boxplot of hsa-miR-1202 log2 expression under each diagnosis level in BA44. C: control ($n = 10$). S: suicides ($n = 13$). SMDD: depressed suicides ($n = 20$). S vs. C q-value is 0.0061.

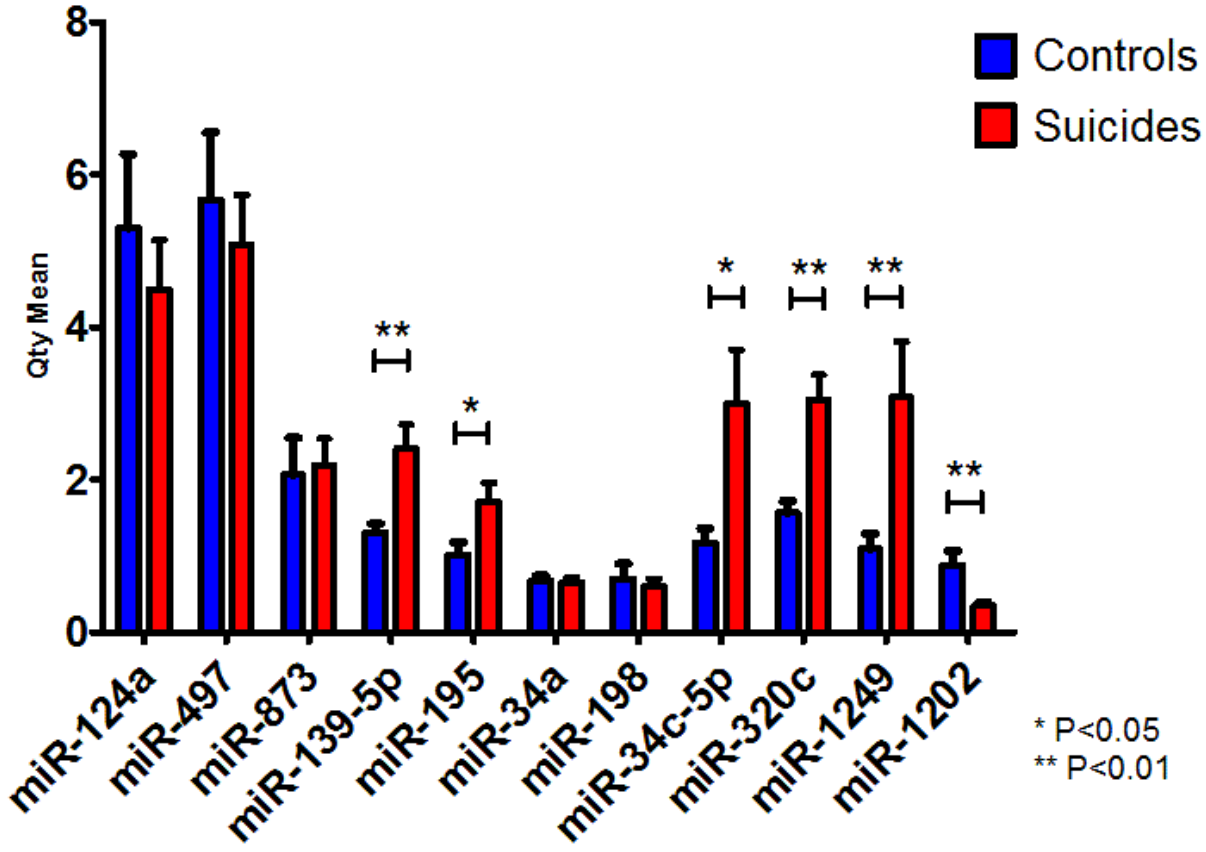


Figure 2.4: qRT-PCR validation of candidate miRNAs in BA44. This validation was performed by our collaborators at the McGill Group for Suicide Studies. Error bars indicate standard error and p-values are from one-sided Mann-whitney U-tests.

Probe	Gene	S vs. C p-value	S vs. C $\log_2$-FC	Cor Coef	Cor p-value
229336_at	**ST3GAL2**	0.020	0.384	0.670	0.003
218733_at	MSL2	0.005	-0.343	-0.628	0.006
226214_at	**GDE1**	0.032	0.279	0.626	0.007
223330_s_at	**SUGT1**	0.015	-0.367	-0.614	0.008
234073_at	**SDK2**	0.001	-0.390	-0.583	0.013
209696_at	FBP1	0.036	-0.306	-0.567	0.016
240451_at	**HIRA**	0.004	-0.352	-0.548	0.020
225288_at	KIAA1870	0.005	0.659	0.529	0.026
206209_s_at	CA4	0.038	0.269	0.521	0.028
224715_at	WDR34	0.018	-0.278	0.513	0.031
213730_x_at	**TCF3**	0.013	0.273	0.513	0.031
235731_at	AIPL1	0.012	-0.473	-0.509	0.033
228415_at	AP1S2	0.027	-0.398	-0.507	0.034
221928_at	**ACACB**	0.027	-0.625	-0.496	0.038
236748_at	**RASGEF1C**	0.016	0.396	0.494	0.039
229153_at	SLC7A6OS	0.020	-0.324	-0.490	0.041
243259_at	**ATXN7**	0.0003	-0.438	-0.482	0.045
244320_at	**NHLRC2**	0.009	-0.610	-0.480	0.046
231960_at	**BRWD1**	0.0001	0.559	0.478	0.047
230533_at	ZMYND8	0.001	-0.299	-0.478	0.047
206017_at	**KIAA0319**	0.049	0.471	0.478	0.047
244239_at	**ANKH**	0.041	-0.489	-0.474	0.049
208621_s_at	**EZR**	0.021	0.467	-0.474	0.049

Table 2.6: Correlated putative targets of hsa-miR-1202. Hsa-miR-1202, which is down-regulated in suicides vs. controls, is correlated with these mRNA probes. Bolded gene symbols indicate that there is support from miRNA target prediction algorithms (miRanda) for the pairing.

Bibliography

[1] J. F. Abelson, K. Y. Kwan, B. J. O'Roak, D. Y. Baek, A. A. Stillman, T. M. Morgan, C. A. Mathews, D. L. Pauls, M.-R. Rašin, M. Gunel, N. R. Davis, A. G. Ercan-Sencicek, D. H. Guez, J. A. Spertus, J. F. Leckman, L. S. Dure, R. Kurlan, H. S. Singer, D. L. Gilbert, A. Farhi, A. Louvi, R. P. Lifton, N. Šestan, and M. W. State. Sequence Variants in SLITRK1 Are Associated with Tourette's Syndrome. *Science*, 310(5746):317 –320, oct 2005. doi: 10.1126/science.1116502.

[2] K. Abu-Elneel, T. Liu, F. S. Gazzaniga, Y. Nishimura, D. P. Wall, D. H. Geschwind, K. Lao, and K. S. Kosik. Heterogeneous dysregulation of microRNAs across the autism spectrum. *Neurogenetics*, 9(3):153–161, jun 2008. ISSN 1364-6745. doi: 10.1007/s10048-008-0133-5.

[3] A. Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1):45–65, jan 2003. ISSN 0306-4573. doi: 10.1016/S0306-4573(02)00021-3.

[4] T. S. Aleman, A. V. Cideciyan, N. J. Volpe, G. Stevanin, A. Brice, and S. G. Jacobson. Spinocerebellar Ataxia Type 7 (SCA7) Shows a Cone-Rod Dystrophy Phenotype. *Experimental Eye Research*, 74(6):737–745, jun 2002. ISSN 0014-4835. doi: 06/exer.2002.1169.

[5] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic acids research*, 33(18):5914, 2005. doi: 10.1093/nar/gki890.

[6] T. Beiß barth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464 –1465, 2004. doi: 10.1093/bioinformatics/bth088.

[7] 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):289–300, jan 1995. ISSN 0035-9246.

[8] J. Bergquist, L. Träskman-Bendz, M. B. Lindström, and R. Ekman. Suicide-attempters having immunoglobulin G with affinity for dopamine in cerebrospinal fluid. *European Neuropsychopharmacology*, 12(2):153–158, apr 2002. ISSN 0924-977X. doi: 16/S0924-977X(02)00002-0.

[9] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(Database):D149–D153, dec 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm995.

[10] S. Bhattacharya and T. Mariani. Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression modeling. *Nucleic Acids Research*, 33(18):e157, 2005. doi: 10.1093/nar/gni159.

[11] H. R. Bourne, W. E. J. Bunney, R. W. Colburn, J. M. Davis, J. N. Davis, D. M. Shaw, and A. J. Coppen. Noradrenaline, 5-hydroxytryptamine, and 5-hydroxyindoleacetic acid in hindbrains of suicidal patients. *Lancet*, 2(7572):805–808, oct 1968. ISSN 0140-6736.

[12] C. Bowden, A. E. Theodorou, S. C. Cheetham, S. Lowther, C. L. E. Katona, M. Rufus Crompton, and R. W. Horton. Dopamine D1 and D2 receptor binding sites in brain samples from depressed suicides and controls. *Brain Research*, 752(1-2):227–233, mar 1997. ISSN 0006-8993. doi: 16/S0006-8993(96)01460-6.

[13] H. Bozdogan. Model selection and akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, sep 1987. ISSN 0033-3123. doi: 10.1007/BF02294361.

[14] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S.-A. Sansone. ArrayExpressa public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68 –71, 2003. doi: 10.1093/nar/gkg091.

[15] J. D. Bremner, M. Narayan, E. R. Anderson, L. H. Staib, H. L. Miller, and D. S. Charney. Hippocampal Volume Reduction in Major Depression. *Am J Psychiatry*, 157(1):115–118, jan 2000. doi: 10.1176/appi.ajp.157.1.115.

[16] D. A. Brent, J. Bridge, B. A. Johnson, and J. Connolly. Suicidal Behavior Runs in Families: A Controlled Family Study of Adolescent Suicide Victims. *Arch Gen Psychiatry*, 53(12):1145–1152, dec 1996. doi: 10.1001/archpsyc.1996.01830120085015.

[17] I. F. Bronner, Z. Bochdanovits, P. Rizzu, W. Kamphorst, R. Ravid, J. C. van Swieten, and P. Heutink. Comprehensive mRNA expression profiling distinguishes tauopathies and identifies shared molecular pathways. *PloS One*, 4(8):e6826, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006826.

[18] S. C. Cheetham, M. R. Crompton, C. L. E. Katona, S. J. Parker, and R. W. Horton. Brain GABAA/benzodiazepine binding sites and glutamic acid decarboxylase activity in depressed suicide victims. *Brain Research*, 460(1):114–123, sep 1988. ISSN 0006-8993. doi: 16/0006-8993(88)91211-5.

[19] C.-Y. A. Chen and A.-B. Shyu. AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences*, 20(11):465–470, nov 1995. ISSN 0968-0004. doi: 10.1016/S0968-0004(00)89102-1.

[20] P. V. Choudary, M. Molnar, S. J. Evans, H. Tomita, J. Li, M. P. Vawter, R. M. Myers, W. E. Bunney, H. Akil, S. J. Watson, and E. G. Jones. Altered cortical glutamatergic and GABAergic signal transmission with glial involvement in depression. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15653 –15658, oct 2005. doi: 10.1073/pnas.0507901102.

[21] N. Cope, D. Harold, G. Hill, V. Moskvina, J. Stevenson, P. Holmans, M. J. Owen, M. C. O'Donovan, and J. Williams. Strong Evidence That KIAA0319 on Chromosome 6p Is a Susceptibility Gene for Developmental Dyslexia. *The American Journal of Human Genetics*, 76(4):581–591, apr 2005. ISSN 0002-9297. doi: 86/429131.

[22] G. David, N. Abbas, G. Stevanin, A. Dürr, G. Yvert, G. Cancel, C. Weber, G. Imbert, F. Saudou, E. Antoniou, H. Drabkin, R. Gemmill, P. Giunti, A. Benomar, N. Wood, M. Ruberg, Y. Agid, J. L.

Mandel, and A. Brice. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nature Genetics*, 17(1):65–70, sep 1997. ISSN 1061-4036. doi: 10.1038/ng0997-65.

[23] N. DiazGranados, L. Ibrahim, N. Brutsche, R. Ameli, I. D. Henter, D. A. Luckenbaugh, R. Machado-Vieira, and C. A. Zarate. Rapid Resolution of Suicidal Ideation after a Single Infusion of an NMDA Antagonist in Patients with Treatment-Resistant Major Depressive Disorder. *The Journal of clinical psychiatry*, 71(12):1605–1611, dec 2010. ISSN 0160-6689. doi: 10.4088/JCP.09m05327blu.

[24] J. DOSTIE, Z. MOURELATOS, M. YANG, A. SHARMA, and G. DREYFUSS. Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, 9(2):180 –186, feb 2003. doi: 10.1261/rna.2141503.

[25] W. Drevets. Neuroimaging abnormalities in the amygdala in mood disorders. *Annals of the New York Academy of Sciences*, 985(1):420–444, apr 2003. ISSN 1749-6632. doi: 10.1111/j.1749-6632.2003.tb07098.x.

[26] J. Dudley, R. Tibshirani, T. Deshpande, and A. Butte. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5(1):307–307, 2009. doi: 10.1038/msb.2009.66.

[27] Y. Dwivedi, J. S. Rao, H. S. Rizavi, J. Kotowski, R. R. Conley, R. C. Roberts, C. A. Tamminga, and G. N. Pandey. Abnormal Expression and Functional Characteristics of Cyclic Adenosine Monophosphate Response Element Binding Protein in Postmortem Brain of Suicide Subjects. *Arch Gen Psychiatry*, 60(3):273–282, mar 2003. doi: 10.1001/archpsyc.60.3.273.

[28] S. J. Evans, P. V. Choudary, C. R. Neal, J. Z. Li, M. P. Vawter, H. Tomita, J. F. Lopez, R. C. Thompson, F. Meng, J. D. Stead, D. M. Walsh, R. M. Myers, W. E. Bunney, S. J. Watson, E. G. Jones, and H. Akil. Dysregulation of the fibroblast growth factor system in major depression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15506 –15511, oct 2004. doi: 10.1073/pnas.0406788101.

[29] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, page 36, Baltimore, Maryland, 2003. Society for Industrial and Applied Mathematics. ISBN 0898715385.

[30] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257 –258, jan 2007. doi: 10.1093/bioinformatics/btl567.

[31] C. Feng, M. Araki, R. Kunimoto, A. Tamon, H. Makiguchi, S. Niijima, G. Tsujimoto, and Y. Okuno. GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genomics*, 10:411, 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-411.

[32] A. M. Feyissa, A. Chandran, C. A. Stockmeier, and B. Karolewicz. Reduced levels of NR2A and NR2B subunits of NMDA receptor and PSD-95 in the prefrontal cortex in major depression. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 33(1):70–75, feb 2009. ISSN 0278-5846. doi: 10.1016/j.pnpbp.2008.10.005.

[33] A. M. Feyissa, W. L. Woolverton, J. J. Miguel-Hidalgo, Z. Wang, P. B. Kyle, G. Hasler, C. A. Stockmeier, A. H. Iyo, and B. Karolewicz. Elevated level of metabotropic glutamate receptor 2/3 in the prefrontal cortex in major depression. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 34(2):279–283, mar 2010. ISSN 1878-4216. doi: 10.1016/j.pnpbp.2009.11.018.

[34] L. M. Fiori and G. Turecki. Association of the SAT1 in/del polymorphism with suicide completion. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153B(3):825–829, apr 2010. ISSN 1552-485X. doi: 10.1002/ajmg.b.31040.

[35] L. M. Fiori, N. Mechawar, and G. Turecki. Identification and Characterization of Spermidine/Spermine N1-Acetyltransferase Promoter Variants in Suicide Completers. *Biological Psychiatry*, 66(5):460–467, sep 2009. ISSN 0006-3223. doi: 16/j.biopsych.2009.04.002.

[36] L. French, S. Lane, T. Law, L. Xu, and P. Pavlidis. Application and evaluation of automated semantic annotation of gene expression experiments. *Bioinformatics*, 25(12):1543 –1549, jun 2009. doi: 10.1093/bioinformatics/btp259.

[37] B. N. Frey, A. C. Andreazza, F. G. Nery, M. R. Martins, J. Quevedo, J. C. Soares, and F. Kapczinski. The role of hippocampus in the pathophysiology of bipolar disorder. *Behavioural Pharmacology*, 18 (5-6):419–430, sep 2007. ISSN 0955-8810. doi: 10.1097/FBP.0b013e3282df3cde.

[38] V. Gabbay, R. Klein, L. Guttman, J. Babb, C. Alonso, M. Nishawala, Y. Katz, M. Gaite, and C. Gonzalez. A preliminary study of cytokines in suicidal and nonsuicidal adolescents with major depression. *Journal of child and adolescent psychopharmacology*, 19(4):423–430, aug 2009. ISSN 1044-5463. doi: 10.1089/cap.2008.0140.

[39] J. Gillis and P. Pavlidis. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. *PLoS ONE*, 6(2):e17258, feb 2011. doi: 10.1371/journal.pone.0017258.

[40] P. E. Greenberg, R. C. Kessler, H. G. Birnbaum, S. A. Leong, S. W. Lowe, P. A. Berglund, and P. K. Corey-Lisle. The economic burden of depression in the United States: How did it change between 1990 and 2000?. *Journal of Clinical Psychiatry*, 2003.

[41] F. Grossman and W. Z. Potter. Catecholamines in depression: a cumulative study of urinary norepinephrine and its major metabolites in unipolar and bipolar depressed patients versus healthy volunteers at the NIMH. *Psychiatry Research*, 87(1):21–27, jul 1999. ISSN 0165-1781. doi: 16/S0165-1781(99)00055-4.

[42] M. Guipponi, S. Deutsch, K. Kohler, N. Perroud, F. Le Gal, M. Vessaz, T. Laforge, B. Petit, F. Jollant, S. Guillaume, P. Baud, P. Courtet, R. La Harpe, and A. Malafosse. Genetic and epigenetic analysis of SSAT gene dysregulation in suicidal behavior. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 150B(6):799–807, sep 2009. ISSN 1552-485X. doi: 10.1002/ajmg.b.30901.

[43] D. Harold, S. Paracchini, T. Scerri, M. Dennis, N. Cope, G. Hill, V. Moskvina, J. Walter, A. J. Richardson, M. J. Owen, J. F. Stein, E. D. Green, M. C. O'Donovan, J. Williams, and A. P. Monaco. Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. *Mol Psychiatry*, 11(12):1085–1091, oct 2006. ISSN 1359-4184.

[44] R. S. Hastings, R. V. Parsey, M. A. Oquendo, V. Arango, and J. J. Mann. Volumetric Analysis of the Prefrontal Cortex, Amygdala, and Hippocampus in Major Depression. *Neuropsychopharmacology*, 29(5):952–959, mar 2004. ISSN 0893-133X.

[45] S. S. Hébert and B. De Strooper. Alterations of the microRNA network cause neurodegenerative disease. *Trends in Neurosciences*, 32(4):199–206, apr 2009. ISSN 0166-2236. doi: 16/j.tins.2008.12.003.

[46] S. Heckers. Neuroimaging studies of the hippocampus in schizophrenia. *Hippocampus*, 11(5): 520–528, oct 2001. ISSN 1098-1063. doi: 10.1002/hipo.1068.

[47] M. J. Hoenerhoff, A. R. Pandiri, S. A. Lahousse, H.-H. Hong, T.-V. Ton, T. Masinde, S. S. Auerbach, K. Gerrish, P. R. Bushel, K. R. Shockley, S. D. Peddada, and R. C. Sills. Global Gene Expression Profiling of Spontaneous Hepatocellular Carcinoma in B6C3F1 Mice: Similarities in the Molecular Landscape with Human Liver Cancer. *Toxicologic Pathology*, may 2011. doi: 10.1177/0192623311407213.

[48] S. Holemans, F. De Paermentier, R. W. Horton, M. R. Crompton, C. L. Katona, and J.-M. Maloteaux. NMDA glutamatergic receptors, labelled with [3H]MK-801, in brain samples from drug-free depressed suicides. *Brain Research*, 616(1-2):138–143, jul 1993. ISSN 0006-8993. doi: 16/0006-8993(93)90202-X.

[49] S. Ikeda, S. W. Kong, J. Lu, E. Bisping, H. Zhang, P. D. Allen, T. R. Golub, B. Pieske, and W. T. Pu. Altered microRNA expression in human heart disease. *Physiological Genomics*, 31(3):367 –373, nov 2007. doi: 10.1152/physiolgenomics.00144.2007.

[50] M. V. Iorio, M. Ferracin, C.-G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Ménard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini, and C. M. Croce. MicroRNA Gene Expression Deregulation in Human Breast Cancer. *Cancer Research*, 65(16):7065 –7070, 2005. doi: 10.1158/0008-5472.CAN-05-1783.

[51] S. G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. doi: 10.2202/1544-6115.1027.

[52] B. Karolewicz, K. Szebeni, T. Gilmore, D. Maciag, C. A. Stockmeier, and G. A. Ordway. Elevated levels of NR2A and PSD-95 in the lateral amygdala in depression. *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)*, 12(2):143–153, mar 2009. ISSN 1461-1457. doi: 10.1017/S1461145708008985.

[53] R. C. Kessler, P. Berglund, O. Demler, R. Jin, and E. E. Walters. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, 62(6):593–602, 2005.

[54] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, 62(6):617–628, 2005.

[55] P. Khatri. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti565.

[56] C. D. Kim, M. Seguin, N. Therrien, G. Riopel, N. Chawky, A. D. Lesage, and G. Turecki. Familial Aggregation of Suicidal Behavior: A Family Study of Male Suicide Completers From the General Population. *Am J Psychiatry*, 162(5):1017–1019, may 2005. doi: 10.1176/appi.ajp.162.5.1017.

[57] T. A. Klempan, A. Sequeira, L. Canetti, A. Lalovic, C. Ernst, J. Ffrench-Mullen, and G. Turecki. Altered expression of genes involved in ATP biosynthesis and GABAergic neurotransmission in the

ventral prefrontal cortex of suicides with and without major depression. *Mol Psychiatry*, 14(2): 175–189, oct 2007. ISSN 1359-4184.

[58] T. A. Klempan, A. Sequeira, L. Canetti, A. Lalovic, C. Ernst, J. Ffrench-Mullen, and G. Turecki. Altered expression of genes involved in ATP biosynthesis and GABAergic neurotransmission in the ventral prefrontal cortex of suicides with and without major depression. *Molecular Psychiatry*, 14 (2):175–189, feb 2009. ISSN 1476-5578. doi: 10.1038/sj.mp.4002110.

[59] V. Klimek, C. Stockmeier, J. Overholser, H. Y. Meltzer, S. Kalka, G. Dilley, and G. A. Ordway. Reduced Levels of Norepinephrine Transporters in the Locus Coeruleus in Major Depression. *The Journal of Neuroscience*, 17(21):8451 –8458, nov 1997.

[60] A. M. KRICHEVSKY, K. S. KING, C. P. DONAHUE, K. KHRAPKO, and K. S. KOSIK. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10): 1274 –1281, oct 2003. doi: 10.1261/rna.5980303.

[61] J. H. Krystal, G. Sanacora, H. Blumberg, A. Anand, D. S. Charney, G. Marek, C. N. Epperson, A. Goddard, and G. F. Mason. Glutamate and GABA systems as targets for novel antidepressant and mood-stabilizing treatments. *Molecular Psychiatry*, 7 Suppl 1:S71–80, 2002. ISSN 1359-4184. doi: 10.1038/sj.mp.4001021.

[62] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-Based Meta-Analysis of Global Collections of High-Throughput Public Data. *PLoS ONE*, 5(9):e13066, 2010. doi: 10.1371/journal.pone.0013066.

[63] I. LAGER, O. ANDRÉASSON, T. L. DUNBAR, E. ANDREASSON, M. A. ESCOBAR, and A. G. RASMUSSON. Changes in external pH rapidly alter plant gene expression and modulate auxin and elicitor responses. *Plant, Cell & Environment*, 33(9):1513–1528, sep 2010. ISSN 1365-3040. doi: 10.1111/j.1365-3040.2010.02161.x.

[64] J. Lamb, E. Crawford, D. Peck, J. Modell, I. Blat, M. Wrobel, J. Lerner, J. Brunet, A. Subramanian, K. Ross, and Others. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929, sep 2006. ISSN 1095-9203. doi: 10.1126/science.1132939.

[65] J. T. Leek and J. D. Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*, 3(9):e161, 2007. doi: 10.1371/journal.pgen.0030161.

[66] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, feb 2005. ISSN 0028-0836. doi: 10.1038/nature03315.

[67] T. Lu, Y. Pan, S.-Y. Kao, C. Li, I. Kohane, J. Chan, and B. A. Yankner. Gene regulation and DNA damage in the ageing human brain. *Nature*, 429(6994):883–891, June 2004. ISSN 1476-4687. doi: 10.1038/nature02661.

[68] M. Manijak and H. Nielsen. FARO server: Meta-analysis of gene expression by matching gene expression signatures to a compendium of public gene expression data. *BMC Research Notes*, 4(1): 181, 2011. ISSN 1756-0500.

[69] J. J. Mann, D. A. Brent, and V. Arango. The neurobiology and genetics of suicide and attempted suicide: a focus on the serotonergic system. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 24(5):467–477, may 2001. ISSN 0893-133X. doi: 10.1016/S0893-133X(00)00228-1.

[70] J. S. Mattick. Small regulatory RNAs in mammals. *Human Molecular Genetics*, 14(suppl_1): R121–R132, apr 2005. ISSN 0964-6906. doi: 10.1093/hmg/ddi101.

[71] A. M. McMurtray, D. G. Clark, M. K. Flood, S. Perlman, and M. F. Mendez. Depressive and memory symptoms as presenting features of spinocerebellar ataxia. *J Neuropsychiatry Clin Neurosci*, 18(3):420–422, aug 2006. doi: 10.1176/appi.neuropsych.18.3.420.

[72] B. Mecham, G. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Wetmore, T. Mariani, I. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32(9):e74, may 2004. ISSN 0305-1048. doi: 10.1093/nar/gnh071.

[73] M. Mehan, J. Nunez-Iglesias, C. Dai, M. Waterman, and X. Zhou. An integrative modular approach to systematically predict gene-phenotype associations. *BMC Bioinformatics*, 11(Suppl 1):S62, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-S1-S62.

[74] Z. Merali, L. Du, P. Hrdina, M. Palkovits, G. Faludi, M. O. Poulter, and H. Anisman. Dysregulation in the Suicide Brain: mRNA Expression of Corticotropin-Releasing Hormone Receptors and GABAA Receptor Subunits in Frontal Cortical Brain Region. *The Journal of Neuroscience*, 24(6): 1478 –1485, feb 2004. doi: 10.1523/JNEUROSCI.4734-03.2004.

[75] A. Michalík, J. Del-Favero, C. Mauger, A. Löfgren, and C. Van Broeckhoven. Genomic organisation of the spinocerebellar ataxia type 7 (SCA7) gene responsible for autosomal dominant cerebellar ataxia with retinal degeneration. *Human Genetics*, 105(5):410–417, nov 1999. ISSN 0340-6717.

[76] A. A. Morgan, J. T. Dudley, T. Deshpande, and A. J. Butte. Dynamism in gene expression across multiple studies. *Physiol. Genomics*, 40(3):128–140, feb 2010. doi: 10.1152/physiolgenomics.90403.2008.

[77] W. E. Narrow, D. S. Rae, L. N. Robins, and D. A. Regier. Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. *Archives of General Psychiatry*, 59(2):115–123, feb 2002. ISSN 0003-990X.

[78] C. B. Nemeroff. Recent advances in the neurobiology of depression. *Psychopharmacology Bulletin*, 36 Suppl 2:6–23, 2002. ISSN 0048-5764.

[79] G. Nowak, G. A. Ordway, and I. A. Paul. Alterations in the N-methyl-D-aspartate (NMDA) receptor complex in the frontal cortex of suicide victims. *Brain Research*, 675(1-2):157–164, mar 1995. ISSN 0006-8993.

[80] J. Nunez-Iglesias, C.-C. Liu, T. E. Morgan, C. E. Finch, and X. J. Zhou. Joint Genome-Wide Profiling of miRNA and mRNA Expression in Alzheimer's Disease Cortex Reveals Altered miRNA Regulation. *PLoS ONE*, 5(2):e8898, feb 2010. doi: 10.1371/journal.pone.0008898.

[81] G. N. Pandey, Y. Dwivedi, H. S. Rizavi, X. Ren, S. C. Pandey, C. Pesold, R. C. Roberts, R. R. Conley, and C. A. Tamminga. Higher Expression of Serotonin 5-HT2A Receptors in the Postmortem Brains of Teenage Suicide Victims. *Am J Psychiatry*, 159(3):419–429, mar 2002. doi: 10.1176/appi.ajp.159.3.419.

[82] S. Paracchini, A. Thomas, S. Castro, C. Lai, M. Paramasivam, Y. Wang, B. J. Keating, J. M. Taylor, D. F. Hacking, T. Scerri, C. Francks, A. J. Richardson, R. Wade-Martins, J. F. Stein, J. C. Knight, A. J. Copp, J. LoTurco, and A. P. Monaco. The chromosome 6p22 haplotype associated with dyslexia reduces the expression of KIAA0319, a novel gene involved in neuronal migration. *Human Molecular Genetics*, 15(10):1659 –1666, may 2006. doi: 10.1093/hmg/ddl089.

[83] D. O. Perkins, C. D. Jeffries, L. F. Jarskog, J. M. Thomson, K. Woods, M. A. Newman, J. S. Parker, J. Jin, and S. M. Hammond. microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biology*, 8(2):R27, 2007. ISSN 14656906. doi: 10.1186/gb-2007-8-2-r27.

[84] W. Pitchot, M. Ansseau, A. Gonzalez Moreno, J. Wauthy, M. Hansenne, and R. von Frenckell. Relationship between alpha 2-adrenergic function and suicidal behavior in depressed patients. *Psychiatry Research*, 52(2):115–123, may 1994. ISSN 0165-1781.

[85] W. Pitchot, J. Reggers, E. Pinto, M. Hansenne, S. Fuchs, S. Pirard, and M. Ansseau. Reduced dopaminergic activity in depressed suicides. *Psychoneuroendocrinology*, 26(3):331–335, apr 2001. ISSN 0306-4530.

[86] R. F. Place, L.-C. Li, D. Pookot, E. J. Noonan, and R. Dahiya. MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences*, 105(5):1608 –1613, feb 2008. doi: 10.1073/pnas.0707594105.

[87] S. B. Plaisier, R. Taschereau, J. A. Wong, and T. G. Graeber. Rankrank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, 38(17):e169, 2010. doi: 10.1093/nar/gkq636.

[88] M. Pompili, Z. Rihmer, M. Innamorati, D. Lester, P. Girardi, and R. Tatarelli. Assessment and treatment of suicide risk in bipolar disorders. *Expert Review of Neurotherapeutics*, 9(1):109–136, jan 2009. ISSN 1473-7175. doi: 10.1586/14737175.9.1.109.

[89] J. Porter. Persistent over-expression of specific CC class chemokines correlates with macrophage and T-cell recruitment in mdx skeletal muscle. *Neuromuscular Disorders*, 13(3):223–235, mar 2003. ISSN 09608966. doi: 10.1016/s0960-8966(02)00242-0.

[90] J. D. Porter, A. P. Merriam, P. Leahy, B. Gong, J. Feuerman, G. Cheng, and S. Khanna. Temporal gene expression profiling of dystrophin-deficient (mdx) mouse diaphragm identifies conserved and muscle group-specific mechanisms in the pathogenesis of muscular dystrophy. *Human Molecular Genetics*, 13(3):257–269, feb 2004. ISSN 0964-6906. doi: 10.1093/hmg/ddh033.

[91] M. O. Poulter, L. Du, I. C. G. Weaver, M. Palkovits, G. Faludi, Z. Merali, M. Szyf, and H. Anisman. GABAA receptor promoter hypermethylation in suicide brain: implications for the involvement of epigenetic processes. *Biological Psychiatry*, 64(8):645–652, oct 2008. ISSN 1873-2402. doi: 10.1016/j.biopsych.2008.05.028.

[92] G. Rajkowska, J. J. Miguel-Hidalgo, J. Wei, G. Dilley, S. D. Pittman, H. Y. Meltzer, J. C. Overholser, B. L. Roth, and C. A. Stockmeier. Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. *Biological Psychiatry*, 45(9):1085–1098, may 1999. ISSN 0006-3223. doi: 16/S0006-3223(99)00041-4.

[93] G. Rajkowska, A. Halaris, and L. D. Selemon. Reductions in neuronal and glial density characterize the dorsolateral prefrontal cortex in bipolar disorder. *Biological Psychiatry*, 49(9):741752, 2001.

[94] J. J. Reina?Pinto, D. Voisin, R. Teodor, and A. Yephremov. Probing differentially expressed genes against a microarray database for in silico suppressor/enhancer and inhibitor/activator screens. *The Plant Journal*, 61(1):166–175, jan 2010. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2009.04043.x.

[95] P. Rosel, B. Arranz, M. Urretavizcaya, M. Oros, L. San, and M. A. Navarro. Altered 5-HT2A and 5-HT4 Postsynaptic Receptors and Their Intracellular Signalling Systems IP¡sub¿3¡/sub¿ and cAMP in Brains from Depressed Violent Suicide Victims. *Neuropsychobiology*, 49(4):189–195, 2004. ISSN 1423-0224. doi: 10.1159/000077365.

[96] A. Roy, D. Pickar, J. De Jong, F. Karoum, and M. Linnoila. Suicidal behavior in depression: relationship to noradrenergic function. *Biological Psychiatry*, 25(3):341–350, feb 1989. ISSN 0006-3223.

[97] B. J. Sadock, H. I. Kaplan, and V. A. Sadock. *Kaplan & Sadock's synopsis of psychiatry: behavioral sciences/clinical psychiatry*. Lippincott Williams & Wilkins, 2007. ISBN 9780781773270.

[98] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0.

[99] G. Sanacora, R. Gueorguieva, C. N. Epperson, Y.-T. Wu, M. Appel, D. L. Rothman, J. H. Krystal, and G. F. Mason. Subtype-specific alterations of gamma-aminobutyric acid and glutamate in patients with major depression. *Archives of General Psychiatry*, 61(7):705–713, jul 2004. ISSN 0003-990X. doi: 10.1001/archpsyc.61.7.705.

[100] E. Sasaki, C. Takahashi, T. Asami, and Y. Shimada. AtCAST, a Tool for Exploring Gene Expression Similarities among DNA Microarray Experiments Using Networks. *Plant and Cell Physiology*, 52 (1):169 –180, jan 2011. doi: 10.1093/pcp/pcq185.

[101] A. Sequeira, F. G. Gwadry, J. M. H. Ffrench-Mullen, L. Canetti, Y. Gingras, R. A. Casero, G. Rouleau, C. Benkelfat, and G. Turecki. Implication of SSAT by Gene Expression and Genetic Variation in Suicide and Major Depression. *Arch Gen Psychiatry*, 63(1):35–48, jan 2006. doi: 10.1001/archpsyc.63.1.35.

[102] A. Sequeira, T. Klempan, L. Canetti, J. Ffrench-Mullen, C. Benkelfat, G. A. Rouleau, and G. Turecki. Patterns of gene expression in the limbic system of suicides with and without major depression. *Mol Psychiatry*, 12(7):640–655, mar 2007. ISSN 1359-4184.

[103] P. Sethi and W. J. Lukiw. Micro-RNA abundance and stability in human brain: Specific alterations in Alzheimer's disease temporal lobe neocortex. *Neuroscience Letters*, 459(2):100–104, aug 2009. ISSN 0304-3940. doi: 10.1016/j.neulet.2009.04.052.

[104] Y. I. Sheline, P. W. Wang, M. H. Gado, J. G. Csernansky, and M. W. Vannier. Hippocampal atrophy in recurrent major depression. *Proceedings of the National Academy of Sciences*, 93(9):3908 –3913, apr 1996.

[105] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, and J. M. Cherry. The Stanford Microarray Database. *Nucleic Acids Research*, 29(1):152 –155, 2001. doi: 10.1093/nar/29.1.152.

[106] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing & Management*, 32(5):619–633, sep 1996. ISSN 0306-4573. doi: 10.1016/0306-4573(96)00008-8.

[107] D. J. STATHAM, A. C. HEATH, P. A. F. MADDEN, K. K. BUCHOLZ, L. BIERUT, S. H. DINWIDDIE, W. S. SLUTSKE, M. P. DUNNE, and N. G. MARTIN. Suicidal Behaviour: An Epidemiological and Genetic Study. *Psychological Medicine*, 28(04):839–855, 1998. doi: null.

[108] J. D. Storey. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics*, 31(6):2013–2035, dec 2003. ISSN 00905364.

[109] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, oct 2005. doi: 10.1073/pnas.0506580102.

[110] M. Tochigi, K. Iwamoto, M. Bundo, T. Sasaki, N. Kato, and T. Kato. Gene expression profiling of major depression and suicide in the prefrontal cortex of postmortem brains. *Neuroscience Research*, 60(2):184–191, feb 2008. ISSN 0168-0102. doi: 16/j.neures.2007.10.010.

[111] J. Tsang, J. Zhu, and A. van Oudenaarden. MicroRNA-mediated Feedback and Feedforward Loops are Recurrent Network Motifs in Mammals. *Molecular cell*, 26(5):753–767, jun 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.05.018.

[112] B. S. Tseng, P. Zhao, J. S. Pattison, S. E. Gordon, J. A. Granchelli, R. W. Madsen, L. C. Folk, E. P. Hoffman, and F. W. Booth. Regenerated mdx mouse skeletal muscle shows differential mRNA expression. *Journal of Applied Physiology*, 93(2):537 –545, 2002. doi: 10.1152/japplphysiol.00202.2002.

[113] G. Turecki, R. Brière, K. Dewar, T. Antonetti, A. Lesage, M. Seguin, N. Chawky, C. Vanier, M. Alda, R. Joober, and Others. Prediction of level of serotonin 2A receptor binding by serotonin receptor 2A genetic variation in postmortem brain samples from subjects who did or did not commit suicide. *American Journal of Psychiatry*, 156(9):1456, sep 1999.

[114] M. Vazquez, R. Nogales-Cadenas, J. Arroyo, P. Botias, R. Garcia, J. M. Carazo, F. Tirado, A. Pascual-Montano, and P. Carmona-Saez. MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Research*, 38(Web Server): W228–W232, may 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq476.

[115] D. Volodarsky, N. Leviatan, A. Otcheretianski, and R. Fluhr. HORMONOMETER: A Tool for Discerning Transcript Signatures of Hormone Action in the Arabidopsis Transcriptome. *Plant Physiology*, 150(4):1796 –1805, 2009. doi: 10.1104/pp.109.138289.

[116] D. R. Weinberger, K. F. Berman, and R. F. Zec. Physiologic dysfunction of dorsolateral prefrontal cortex in schizophrenia: I. regional cerebral blood flow evidence. *Archives of General Psychiatry*, 1986.

[117] M. M. Weissman, R. C. Bland, G. J. Canino, C. Faravelli, S. Greenwald, H.-G. Hwu, P. R. Joyce, E. G. Karam, C.-K. Lee, J. Lellouch, J.-P. Lépine, S. C. Newman, M. Rubio-Stipec, J. E. Wells, P. J. Wickramaratne, H.-U. Wittchen, and E.-K. Yeh. Cross-National Epidemiology of Major Depression and Bipolar Disorder. *JAMA: The Journal of the American Medical Association*, 276(4):293 –299, jul 1996. doi: 10.1001/jama.1996.03540040037030.

[118] 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–D12, jan 2007. ISSN 0305-1048. doi: 10.1093/nar/gkl1031.

[119] G. Zalsman, A. Frisch, R. Lewis, E. Michaelovsky, H. Hermesh, L. Sher, E. Nahshoni, L. Wolovik, S. Tyano, A. Apter, R. Weizman, and A. Weizman. DRD4 receptor gene exon III polymorphism in inpatient suicidal adolescents. *Journal of Neural Transmission*, 111(12):1593–1603, jun 2004. ISSN 0300-9564. doi: 10.1007/s00702-004-0182-3.