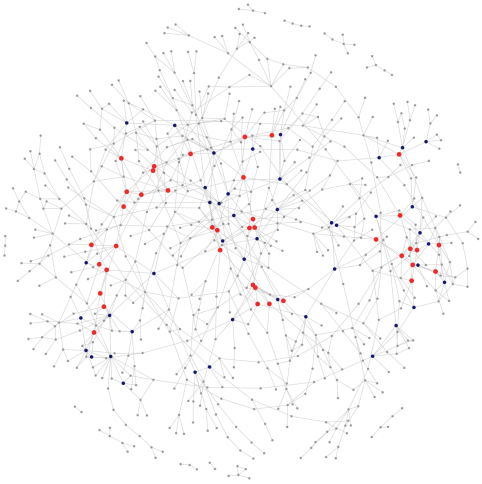


Identifying Schizophrenia Risk Genes and Sub-networks Using DAWN Framework

Author

Julian Q. Zhou[1]

Advisor

Kathryn Roeder, Ph.D

Dietrich College Honors Thesis

Carnegie Mellon University

May 2015

[1]Officially Quan Zhou. Correspondence: julian.q.zhou@gmail.com.

Abstract

Human geneticists in the post-genomics era are blessed with unprecedentedly powerful genomic technologies such as next-generation sequencing to uncover the mysteries of complex human diseases. On the other hand, nevertheless, new practical and analytical challenges that arise with the technological revolutions abound. Working in the context of schizophrenia, a neuropsychiatric disease with a strong genetic basis, we take advantage of genomic datasets produced by modern genomic technologies, as well as novel statistical methods developed in response to the analytical challenges. Specifically, we apply a new meta-analysis framework – *Detecting Association With Network* (DAWN) – to high-dimensional gene expression datasets in an attempt to identify potential risk genes and sub-networks for schizophrenia. We also address a practical measurement issue that arises with the transition between different genomic technologies. By proposing a procedure that transforms datasets measured using two different technologies to achieve comparable measurements, we combine both data sources, thereby increasing sample size. Using DAWN, we identify a set of 39 primary risk genes and 44 secondary risk genes. We conclude by visualizing the risk gene network and 6 sub-networks surrounding the primary risk genes.

Keywords: mapping, genetic association scores, correlation-wise odd pairs, transformation, partial neighborhood selection, parameter tuning, co-expression network, hidden Markov random field, Bayesian false discovery rate control, risk genes, sub-networks

This page intentionally left blank

Acknowledgments

First and foremost, I am grateful to my advisor, Dr. Kathryn Roeder, who has not only endowed me with her exceptional research mentorship, but also shared with me invaluable advice for my graduate school applications and academic career in general. I feel most fortunate to have been able to work with her.

Standing at the end of this year-long journey, I would like to thank Dr. Bernie Devlin, for providing access to the CommonMind data and sharing his insightful feedback; Dr. Li Liu, for sharing DAWN's source code; Dr. A. Ercument Cicek, for helping me with running DAWN; and Cong Lu, for offering guidance at various stages of my thesis. I would also like to thank other members of Dr. Roeder's Bioinformatics & Statistical Genetics Group (BiGGS) and of Dr. Devlin's Computational Genetics Lab for helping me with my project in one way or another. I have been fond of the Tuesday lab meetings.

The preliminary phase of this project was supported by Summer Undergraduate Research Fellowship (SURF) 2014 through the Undergraduate Research Office at Carnegie Mellon University. I would like to express my gratitude towards the Office of the Provost and the Walter P. Ketterer Undergraduate Research Fund for sponsoring my fellowship.

I would also like to thank Dr. Bill Eddy, director of Summer Undergraduate Research Experience (SURE) in the Department of Statistics, for offering a life-changing opportunity through the program's 2013 iteration that introduced me to bioinformatics research with Dr. Roeder.

Last but not least, I am grateful to the faculty and staff in the Department of Statistics for a superb education in statistics and a marvelous undergraduate experience overall.

This page intentionally left blank

To Mom and Dad,

without whom

I would not be where I am.

献给赋予我
一切的父母。

This page intentionally left blank

Contents

1 Introduction **1**

1.1 Summary of DAWN Framework 2

2 Derivation of Schizophrenia-Specific Genetic Scores **3**

2.1 Mapping of SNPs to Genes 4

2.2 Hyper-mapped Genes 7

2.3 Derivation of Genetic Association Scores 8

3 Gene Expression Data, Regression, and Transformation **11**

3.1 Data Cleaning and Quality Control 11

3.2 Removal of Age Effect through Regression 13

3.3 Correlation-wise Odd Pairs 15

3.4 'Fixing' of COPs via Transformation 17

3.5 Comparison of COPs before and after Transformation 20

4 DAWN Analysis **25**

4.1 Co-expression Network 25

4.2 Hidden Markov Random Field Model 26

4.3 Schizophrenia Risk Genes and Sub-networks 28

5 Discussion **35**

5.1 Reflections 35

5.2 Future Directions 37

6 Conclusion **39**

References **41**

7 Supplemental Information **45**

7.1 List of Hyper-mapped Genes 45

7.2 Annotation of Genes Using *Ensembl* 49

7.3 Regression Diagnostics 51

7.4 List of COP Hubs before Transformation 66

7.5 List of COP Genes for 'Fixing' via Transformation 66

7.6 List of COP Hubs after Transformation 68

7.7 List of Primary and Secondary Risk Genes . . . 69
7.8 Complete DAWN Network . . . 71
7.9 Code . . . 72

List of Tables

2.1.1 Significant eQTL Associations at Various q-value Cut-offs . . . 5
3.1.1 Periods of Human Brain Development . . . 11
7.1.1 Genes Mapped to Over 500 SNPs . . . 45
7.4.1 Genes Involved in Most COPs on Average across Thresholds before Transformation . . . 66
7.5.1 Genes to be 'Fixed' in Microarray Data via Transformation . . . 66
7.6.1 Genes Involved in Most COPs on Average across Thresholds after Transformation . . . 68
7.7.1 Potential Primary and Secondary Risk Genes in Schizophrenia Gene Co-expression Network . 69
7.9.1 Code by Section . . . 72

List of Figures

2.0.1 Mapping of SNPs to Genes and Derivation of Genetic Association Scores of Genes . . . 3
2.1.1 Distribution of q-values of All eQTL Associations . . . 4
2.1.2 Distributions of Numbers of Significant eQTL Associations per Gene at Various q-value Cut-offs 5
2.1.3 Distribution of q-values of eQTL Associations Involving SNPs in PGC Data at Cut-off of 0.05 6
2.2.1 Distribution of Numbers of SNPs Mapped to a Gene . . . 7
2.3.1 Distribution of z-scores of SNPs Mapped to 500 Randomly Selected Genes . . . 8
2.3.2 Density Estimates of z-scores of Genes and z-scores of SNPs . . . 9
2.3.3 Number of SNPs Mapped to a Gene vs. z-score of a Gene . . . 10
3.2.1 Regression Diagnostics for *BTN3A2* . . . 14
3.2.2 Distributions of Adjusted R^2 from Linear Regressions Against Period for All Genes . . . 15
3.3.1 COPs, COP Genes, and COP Hubs across Thresholds before Transformation . . . 16
3.4.1 Distributions of Expression Values after Nonparanormal Transformation . . . 18
3.4.2 Number of COPs That Each COP Gene is Involved in at $t_{COP} = 1.0$ . . . 19
3.5.1 COPs, COP Genes, and COP Hubs across Thresholds after Transformation . . . 20
3.5.2 Numerical Comparison of COPs before and after Transformation . . . 21
3.5.3 Visual Comparison of COPs before and after Transformation . . . 23
4.1.1 Parameter Tuning for λ . . . 26

4.3.1 Distributions of FPPs and Numbers of Global Neighbors of Genes in Co-expression Network . 29
4.3.2 Fraction of Risk Gene Neighbors vs. Genetics-based p-value of Primary and Secondary Risk Genes . . . 31
4.3.3 Network Between Primary and Secondary Risk Genes . . . 32
4.3.4 Network Between Primary Risk Genes and First-degree Neighbors . . . 34
7.2.1 Using *BioMart* – Step 1: Choose a Dataset . . . 49
7.2.2 Using *BioMart* – Step 2: Upload a List of Ensembl IDs as Filter . . . 49
7.2.3 Using *BioMart* – Step 3: Select Desired Attributes . . . 50
7.2.4 Using *BioMart* – Step 4: Export Annotations as a .csv File . . . 50
7.3.1 Regression Diagnostics for Randomly Selected Genes . . . 52
7.8.1 Complete DAWN Network . . . 71

List of Acronyms

ASD autism spectrum disorder

COP correlation-wise odd pair

DAWN Detecting Association With Network

eCDF empirical cumulative distribution function

eQTL expression quantitative trait loci

FDR false discovery rate

FPP FDR-controlled posterior probability

HMRF hidden Markov random field

MHC major histocompatibility complex

PCW post-conceptual week

PGC Psychiatric Genomics Consortium

PNS partial neighborhood selection

RIN RNA integrity number

SF-R^2 scale-free topology model R^2

SNP single nucleotide polymorphism

This page intentionally left blank

1 Introduction

HUMAN genetics researchers in the post-genomics era are blessed with unprecedentedly powerful genomic technologies such as next-generation sequencing to uncover the mysteries of complex human diseases. For instance, geneticists' understanding of ***autisim spectrum disorder* (ASD)**, a neurodevelopmental disorder with a heritable and complex genetic basis, has been growing rapidly thanks to advancements in sequencing technology [1, 2]. Through whole-exome sequencing, De Rubeis *et al.* recently identified 22 autosomal genes implicated for ASD that involve in pathways for synaptic formation, transcriptional regulation, and chromatin-remodelling, in addition to 107 genes that are strongly enriched [2]. On the other hand, nevertheless, geneticists face many new challenges, both practical and analytical, that arise with the technological revolutions. For example, due to incomplete penetrance and modest effects of risk genes for complex diseases, genome-wide association studies oftentimes require large sample sizes to overcome limitations of reduced analytical power [3]. However, it remains largely difficult for research groups to substantially increase their sample sizes as the operational cost to recruit human subjects and collect high-quality genetic data stays high, even though the cost of sequencing itself has dropped considerably in recent years. Analytically, the ability to measure the expression of thousands of genes simultaneously can be harnessed only if accompanied by statistical techniques tailored for high-dimensional settings where the genes by far outnumber the samples [4]. Moreover, as genes responsible for complex diseases are now widely believed to function in networks instead of acting in an isolated fashion [5], more sophisticated network-based analytical schemes are therefore necessary to make meaningful inferences on risk gene networks.

In this project, we take advantage of genomic datasets produced by modern genomic technologies, as well as novel statistical methods developed in response to the aforementioned analytical challenges. We do so in the context of schizophrenia, another neuropsychiatric disease for which evidence of a strong genetic basis has been shown [6]. Specifically, we apply a new meta-analysis framework – ***Detecting Association With Network* (DAWN)** – to high-dimensional gene expression datasets in an attempt to identify potential risk genes and sub-networks for schizophrenia. In the course of our analysis, we also address a measurement issue that arises with the transition between different genomic technologies. We propose a procedure that transforms datasets measured using different technologies to achieve comparable measurements, thereby making it possible to combine both data sources and increase sample size.

1.1 Summary of DAWN Framework

Developed by Liu *et al.*, DAWN uses a network-assisted approach to estimate the probability of each gene in the gene co-expression network being a risk gene [7]. In their paper presenting the debut version of DAWN, Liu *et al.* show that DAWN 'is effective in predicting ASD genes and sub-networks' and that it 'successfully predicts known ASD genes' [7]. While devised originally in the context of ASD studies, the framework can be applied to any generic complex disorder or disease with a strong genetic basis. We use in this project an updated version of DAWN which has not yet been published by the completion of this project [8]. The framework is based on the assumption that 'genes expressed at the same developmental period and brain region, and with highly correlated co-expression, are functionally interrelated and more likely to affect risk' [8]. It estimates risk gene probabilities through modeling of two types of data: gene co-expression in specific brain regions and periods of development, and disease-specific genetic association scores [8]. A brief summary of the new DAWN framework is outlined as follows:

(i) Obtain disease-specific p-value of each gene. This genetic association score serves as marginal evidence of a gene being a risk gene [8].

(ii) Estimate the gene co-expression network based on measurements of gene expression levels in specific tissues and periods of development. This step uses a partial neighborhood selection algorithm as described in Liu *et al.* to produce a network estimate [8, 9, 10].

(iii) Incorporate the disease-specific genetic association scores from (i) and the co-expression network from (ii) into a hidden Markov random field model, and estimate its parameters via an iterative algorithm also described in Liu *et al.* [8].

(iv) Based on the model from (iii), obtain posterior probability of a gene being a risk gene, while applying Bayesian false discovery rate control [8, 11].

(v) Risk genes are selected based on a chosen cut-off for their risk probabilities. Their sub-networks, if any, can be visualized [8].

This thesis is structured in an order that largely matches with the one outlined for the DAWN framework.

2 Derivation of Schizophrenia-Specific Genetic Scores

DAWN starts with ***genetic association scores*** of genes. These can be presented as p-values or z-scores. In our writing, we tend to use genetic association scores and p-values interchangeably, with the understanding that *higher* scores correspond with *smaller* p-values and *larger* z-scores. These scores are disease-specific and are considered marginal evidence on the likelihood of the genes being risk genes for the disease in question. The extent of usefulness of these scores, which take into no account of interactions amongst the genes, is regarded marginal because of the general consensus amongst modern geneticists that genes behind complex diseases rarely function in an isolated fashion [5]. Nonetheless, they serve as an appropriate starting point.

For schizophrenia, fortunately, the association scores of 9,444,230 ***single nucleotide polymorphisms* (SNPs)** have recently become available as part of a landmark study conducted by the ***Schizophrenia Working Group of the Psychiatric Genomics Consortium*** (hereafter referred to as **PGC**). In this study, '128 independent associations spanning 108 conservatively defined loci' were found to be significantly associated with schizophrenia [12]. Unfortunately, on the other hand, these scores belong to SNPs, which are sequence variations at single nucleotide positions in the genome. In order to build on the PGC results and meet DAWN's input requirement, we need to derive association scores of the genes from those of the SNPs. The process to achieve this is illustrated by Figure 2.0.1. With reference to Figure 2.0.1, we first examine the associations between the SNPs and a given gene, represented by $U_{1..N}$, and determine if a SNP is mapped to the gene. Next, we obtain association scores of the SNPs, $V_{1..N}$, from the PGC data. We then derive the association score of the gene for schizophrenia, z, based on the association scores of the SNPs mapped to the gene.

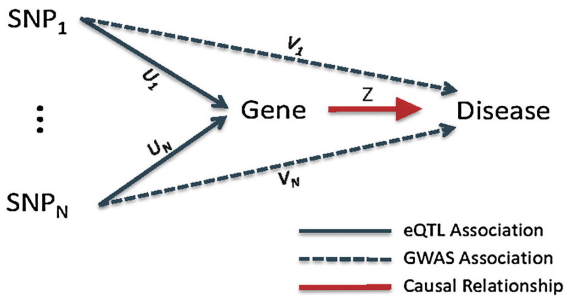


Figure 2.0.1: *Mapping of SNPs to Genes and Derivation of Genetic Association Scores of Genes*[2]. After mapping SNPs to a gene based on eQTL associations ($U_1...U_N$), we derive genetic association score (z) of the gene based on GWAS association scores ($V_1..V_N$) of the SNPs mapped to the gene.

[2]Reprinted from He *et al.* [13], Copyright (2013), with permission from Elsevier.

2.1 Mapping of SNPs to Genes

Mapping of a SNP to a gene is tissue-specific. That is, the same SNP may or may not be mapped to a gene depending on the tissue in which their association is being considered. We perform SNP-to-gene mapping in postmortem human brain tissue. Brain tissue is used, consistent with other genetics studies on ASD and schizophrenia [1, 2, 12], as the brain is the central organ of the nervous system. Specifically, we use ***expression quantitative trait loci* (eQTL)** data from the CommonMind Consortium. Note that we have had internal access to and used the eQTL data without SVA elements[3] for the Caucasian control subjects before Release 1 of CommonMind Consortium Data [14], and that the two versions may or may not differ.

The CommonMind data contain ***q-values*** of 133,159 trans-eQTL associations and 8,875,306 cis-eQTL associations, all corrected for multiple testing via ***false discovery rate* (FDR)** control. In loose terms, the q-value of an eQTL association between a SNP and a gene quantifies the significance of their association. If an association is statistically significant at a given cut-off, the SNP can be considered ***mapped*** to the gene. As visualized in Figure 2.1.1, the distribution of these q-values appears skewed severely to the left.

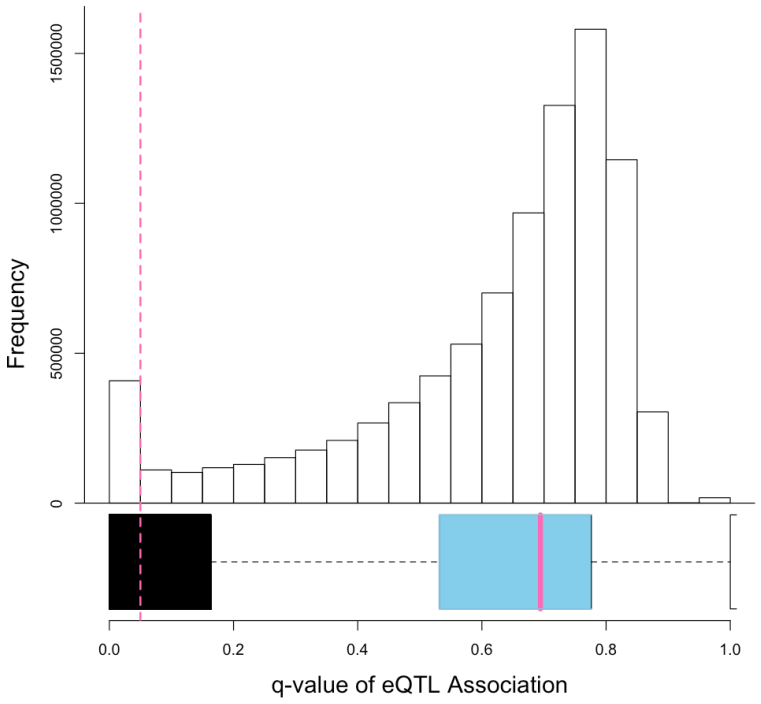


Figure 2.1.1: *Distribution of q-values of All Trans- and Cis-eQTL Associations Combined.* The *histbox* plot, produced using the *sfsmisc* package [15] in `R` [16], visualizes the distribution using a histogram and a horizontal boxplot. The left end of the boxplot appears as a black block due to many lines being drawn consecutively, each representing an outlier. The pink dotted line indicates a potential q-value cut-off at 0.05.

[3] A family of non-autonomous retroelements within the primate lineage.

In choosing a cut-off for q-values of the eQTL associations, we experiment with a range of possible cut-offs from 0.00001 to 0.1. We then compare the number of eQTL associations that are significant at different cut-offs, in addition to the numbers of unique genes (in terms of Ensembl IDs) and unique SNPs in those associations. The comparison results are presented in Table 2.1.1 and visualized in Figure 2.1.2. We pick 0.05 to be the cut-off largely out of consideration for having a sufficiently large yet manageable number of genes and SNPs to work with.

Table 2.1.1: *Significant eQTL Associations at Various q-value Cut-offs.* To aid in picking a q-value cut-off, we tally the numbers of significant eQTL associations, and the numbers of unique genes and unique SNPs involved in those eQTL associations at various cut-offs. we choose a cut-off of 0.05 as it corresponds to a sufficiently large yet manageable number of genes and SNPs.

Cut-off	0.00001	0.00005	0.0001	0.0005	0.001	0.005	0.01	0.025	0.05	0.1
# eQTLs	119658	141872	150787	178697	193246	245382	278723	337544	408054	518756
# Genes	843	988	1044	1286	1417	2013	2499	3657	5217	7718
# SNPs	86743	101954	108617	131469	143014	181663	203155	244423	294243	373267

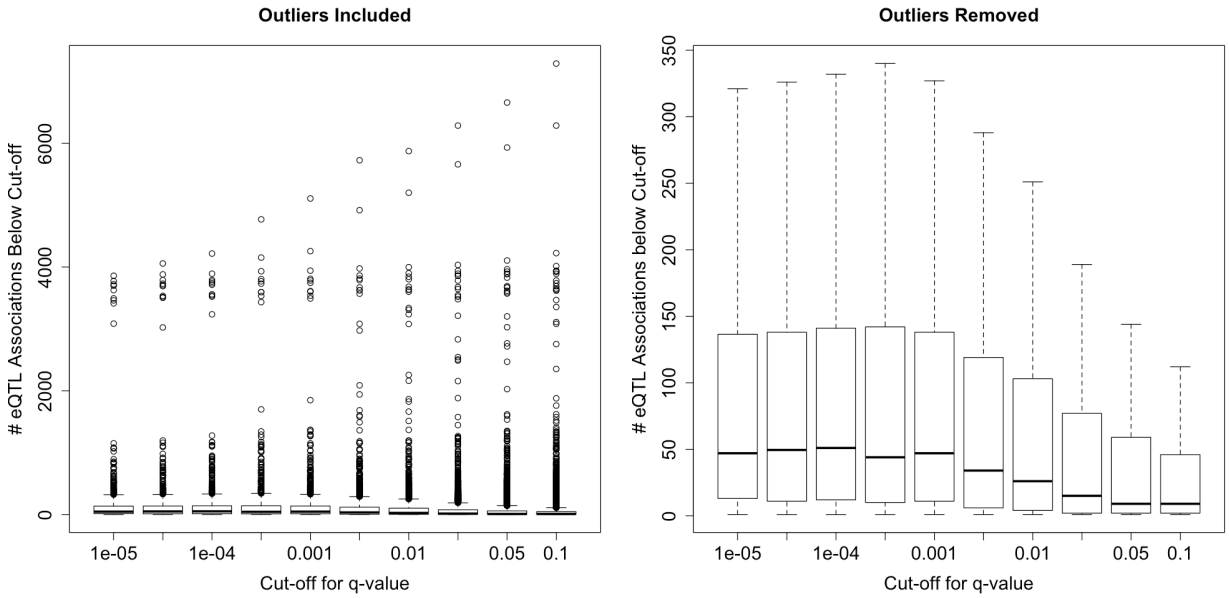


Figure 2.1.2: *Distributions of Numbers of Significant eQTL Associations per Gene at Various q-value Cut-offs.* At a given q-value cut-off, the distribution of number of significant eQTL associations per gene is visualized with a boxplot with and without outliers. While the distributions do not appear too different without the outliers, outliers in distributions at higher (i.e. more relaxed) cut-offs appear to be greater in both number and magnitude.

Imposing 0.05 as the cut-off for q-values of eQTL associations, we have 408,054 significant eQTL associations remaining, involving 5217 unique genes and 294,243 unique SNPs (Table 2.1.1). We then further screen these eQTL associations by keeping only those involving SNPs present in the PGC data. This is necessary because when deriving the genetic association score of a gene, without schizophrenia-specific association score of a SNP from the PGC data, we will not be able to make use of that SNP even if it has been mapped to the gene. Out of 294,243 unique SNPs, 261,189 are present in the PGC data. As a small number of unique genes are mapped solely to SNPs that are not in the PGC data, and are as a result excluded altogether with those SNPs, we are left with 357,834 significant eQTL associations, mapping 5049 unique genes with 261,189 unique SNPs. The distribution of q-values of the remaining eQTL associations, which appears to be skewed severely to the right, is shown in Figure 2.1.3.

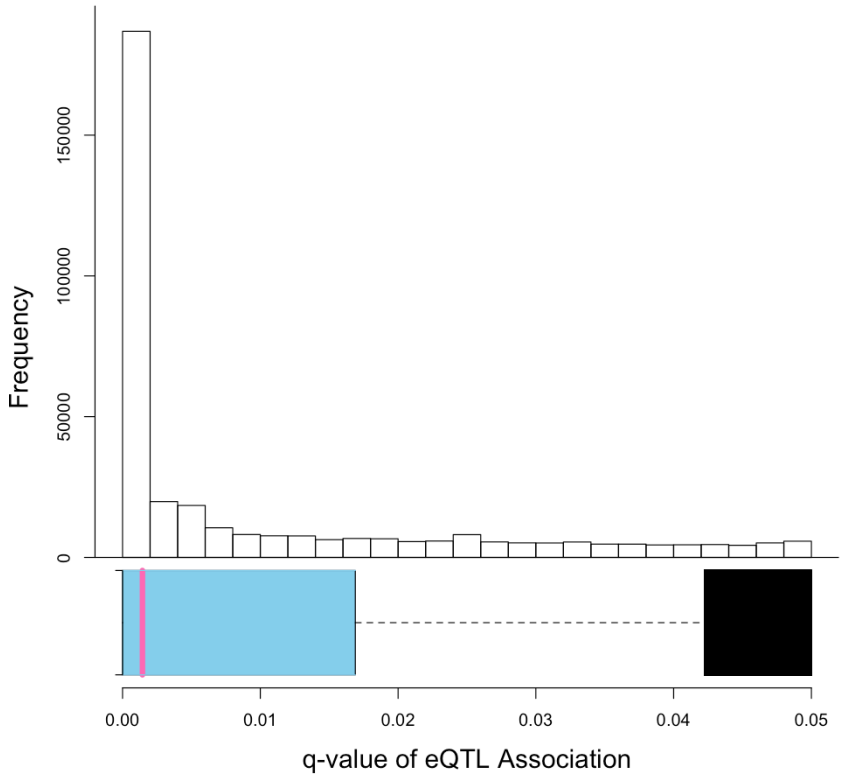
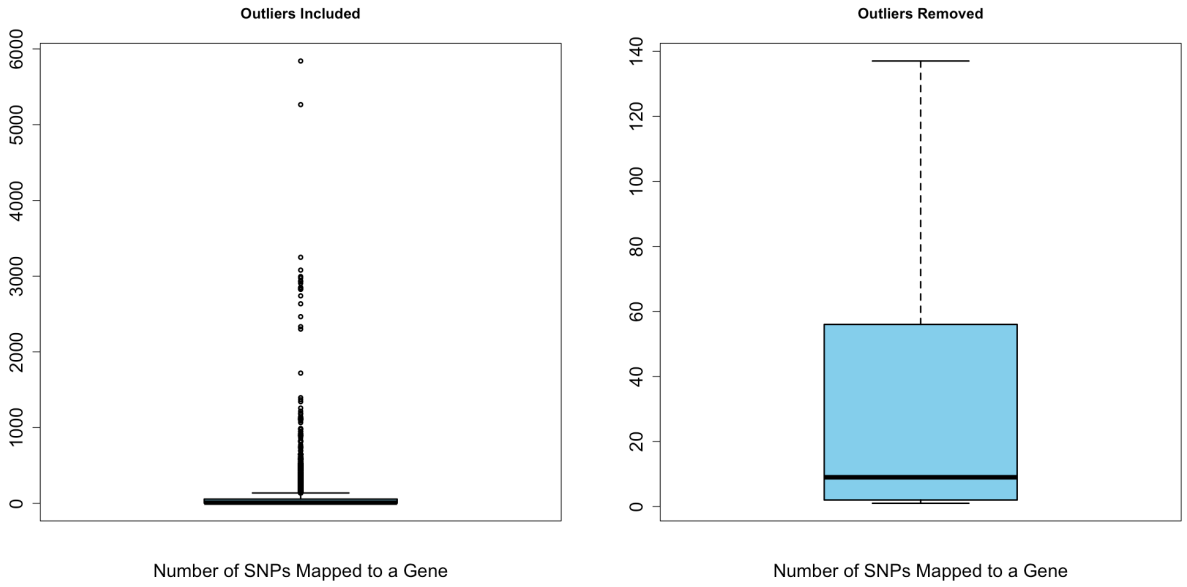


Figure 2.1.3: *Distribution of q-values of eQTL Associations Involving SNPs in PGC Data at a Cut-off of* 0.05. At a q-value cut-off of 0.05 and keeping only SNPs that are present in the PGC data, we have 357,834 significant eQTL associations that map 5049 unique genes with 261,189 unique SNPs. The distribution of q-values below the cut-off is visualized with a *histbox* plot, produced using the *sfsmisc* package [15] in `R` [16]. Skewed severely to the right, it has a large number of outliers with q-values close to 0.05, the lines for which when drawn consecutively appear as a black block in the horizontal boxplot.

2.2 Hyper-mapped Genes

In addition to q-values of our final SNP-to-gene mappings (Figure 2.1.3), we are also curious about the number of SNPs a gene is mapped to based on those q-values. This distribution is shown with and without outliers in Figures 2.2.1A and 2.2.1B respectively. Focusing on Figure 2.2.1A, the numbers of SNPs that some genes are mapped to are remarkably large – several genes are mapped to thousands of SNPs. Using 500 as a threshold for the number of SNPs mapped, we denote genes with more than 500 SNP mappings ***hyper-mapped genes***. This threshold is chosen so as to have a sensible number of hyper-mapped genes to look at. Details, such as the names, descriptions, and numbers of SNP mappings, of all 102 hyper-mapped genes are presented in Table 7.1.1 in Supplemental Information 7.1.



(A) With Outliers

(B) Without Outliers

Figure 2.2.1: *Distribution of Numbers of SNPs Mapped to a Gene.* After mapping, this distribution is visualized in boxplots with and without outliers. Genes with over 500 SNP mappings are considered *hyper-mapped genes*.

Upon examining the list, it appears that many of the hyper-mapped genes either are pseudogenes or encode less 'interesting' proteins with regards to schizophrenia, such as zinc finger proteins. Nevertheless, a few of them – *HLA-DQA1*, *HLA-DRB1*, and *HLA-C* – encode ***major histocompatibility complexes* (MHCs)**. MHCs are immune-related protein molecules that form part of epitopes and are therefore pivotal in antigen presentation. As there has been evidence linking MHCs as risk factors to schizophrenia [17, 18], we will keep in mind the MHC-encoding hyper-mapped genes as we derive their schizophrenia-specific genetic association

scores and as we review our final selection of risk genes for schizophrenia.

2.3 Derivation of Genetic Association Scores

With SNPs mapped to genes, we are ready to derive genetic association scores of the genes, based on schizophrenia-specific genetic association scores of the SNPs mapped to them. Recall that the genetic association scores of the SNPs for schizophrenia are quantified as p-values in the PGC data [12]. We convert these p-values into upper-tailed z-scores to avoid having to work with extremely small numbers. Figure 2.3.1 shows the distribution of z-scores of the SNPs mapped to 500 randomly selected genes. To derive the schizophrenia-specific, genetics-based p-value of a *gene*, we take the minimum of the p-values of all the *SNPs* mapped to that gene. With reference to Figure 2.3.1, in which each column of z-scores belong to SNPs mapped to a unique gene, this is equivalent to adopting the largest z-score in a column as the z-score of the gene to which that column corresponds.

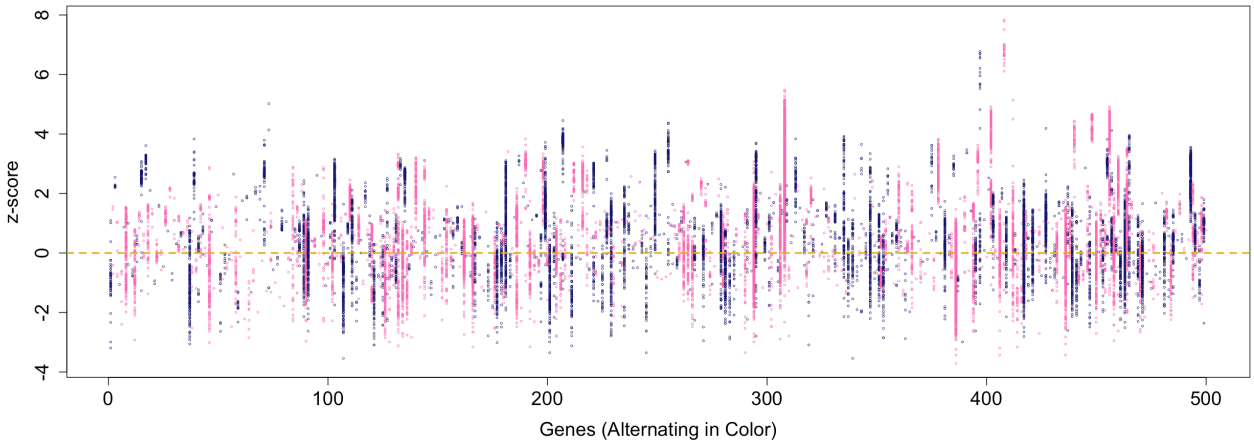


Figure 2.3.1: *Distribution of z-scores of SNPs Mapped to 500 Randomly Selected Genes.* Each column represents z-scores of SNPs mapped to a unique gene. Columns alternate in color for visualization. The yellow dotted line indicates where $z = 0$ for reference. To derive the genetic association score of a gene, we adopt the maximum z-score in its corresponding column.

While simple and straightforward, taking the minimum p-value – or equivalently, the maximum z-score – from those of the SNPs could introduce systematic bias. As illustrated in Figure 2.3.2, the distribution of z-scores of the genes appears to have shifted to the right, compared to that of the SNPs. This is unsurprising considering that the z-scores of the genes are derived by always taking the maxima of those of the SNPs, and that the maxima by definition lie towards the right end of the x-axis. We discuss alternative derivation methods in Section 5.

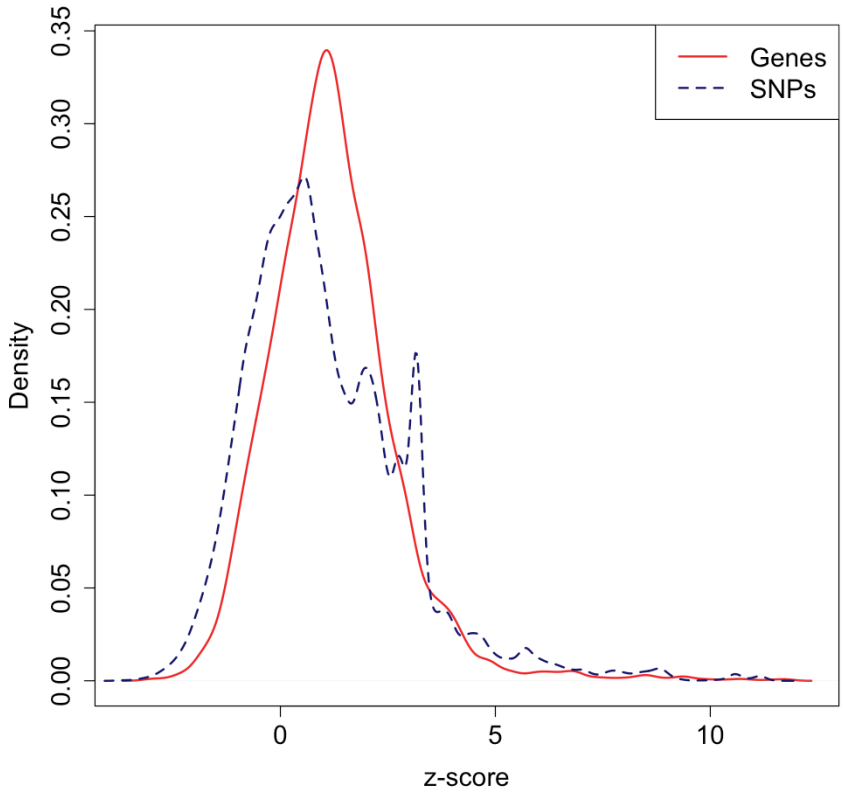
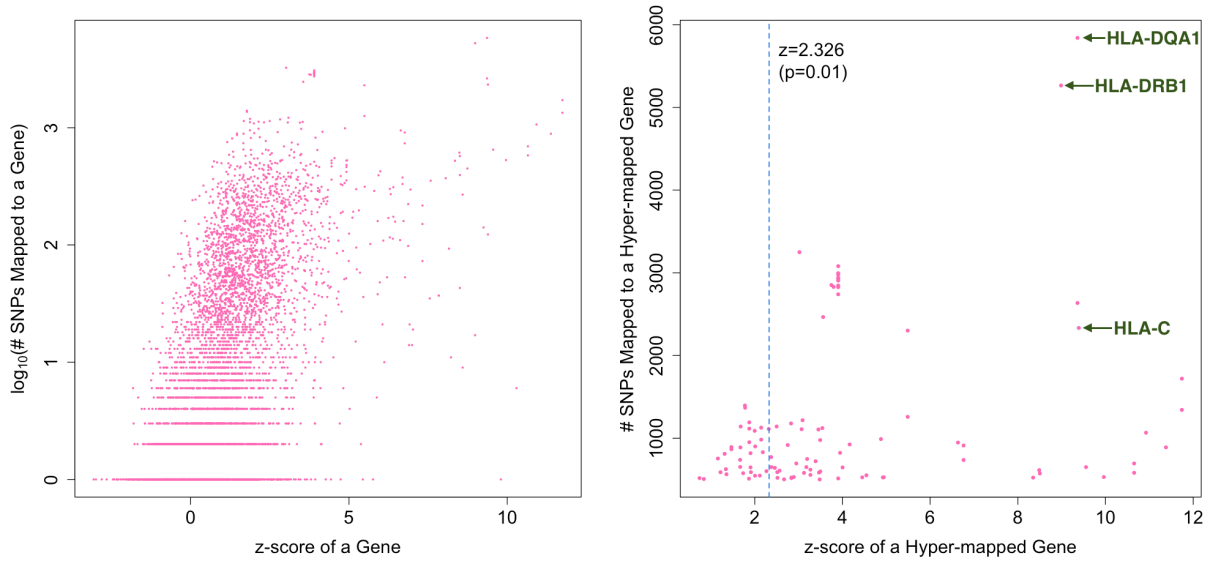


Figure 2.3.2: *Density Estimates of z-scores of Genes and z-scores of SNPs.* The z-score of a gene is derived by taking the maximum of the z-scores of the SNPs mapped to the gene. A shift to the right is observed in the density estimate of z-scores of the genes, suggesting bias introduced by always taking the maximum.

Now that we have finished mapping SNPs to genes and deriving p-values of genes, let us examine the relationship, if any, between the number of SNP mappings per gene and schizophrenia-specific z-score of the gene. While Figure 2.3.3A suggests that no particularly interesting pattern exists between the two variables when the number of SNP mappings is relatively small ($\leq log_{10}(500) \approx 2.7$), it becomes clear in Figure 2.3.3B that the hyper-mapped genes – those with over 500 SNP mappings per gene – tend to have larger z-scores. More specifically, many of their z-scores, including those of the 3 MHC-encoding genes, appear highly significant at a cut-off of $p = 0.01$, or equivalently, $z = 2.33$. We will continue monitoring these genes as we proceed with downstream DAWN analysis in Section 4.



(A) All 5049 Genes

(B) 102 Hyper-mapped Genes with MHC-encoding Genes Highlighted

Figure 2.3.3: *Number of SNPs Mapped to a Gene vs. z-score of a Gene.* The hyper-mapped genes, including the 3 highlighted MHC-encoding genes, appear to have larger z-scores.

In addition to assigning p-values and z-scores to the genes, we also annotate them with their commonly used names and descriptions of their functions, based on their Ensembl IDs. This is performed using *Ensembl* and its *BioMart* toolbox [19] with procedures documented in Supplemental Information 7.2.

3 Gene Expression Data, Regression, and Transformation

To estimate the gene co-expression network, we use publicly available gene expression datasets from the ***BrainSpan*** atlas [20]. These datasets measure ***developmental transcriptomes*** from brain tissues using both microarray[4] and RNA-seq[5] technologies.

3.1 Data Cleaning and Quality Control

We begin with 17,604 measurements of gene expression levels from 492 samples in the microarray dataset, and 52,376 measurements from 524 samples in the RNA-seq dataset. Each sample represents a brain region of an individual measured at a certain period of human brain development. The periods are specified and described in Table 3.1.1.

Table 3.1.1: *Periods of Human Brain Development*[6]. Periods are as defined by Kang *et al.* [21]. Ages are measured in post-conceptual weeks (PCW), post-natal months (M), and post-natal years (Y). Days are computed using the formula $7 * \#PCW$, $7 * 38 + 30 * \#M$, and $7 * 38 + 365 * \#Y$ for ages measured in PCW, M, and Y respectively. Later, we further restrict our samples to be between ages 8PCW and 12M.

Period	Description	Age	Days
1	Embryonic	4–8 PCW	28–56
2	Early fetal	8–10 PCW	56–70
3	Early fetal	10–13 PCW	70–91
4	Early mid-fetal	13–16 PCW	91–112
5	Early mid-fetal	16–19 PCW	112–133
6	Early mid-fetal	19–24 PCW	133–168
7	Late fetal	24–38 PCW	168–266
8	Neonatal & early infancy	0–6 M	266–446
9	Late infancy	6–12 M	446–626
10	Early childhood	1–6 Y	631–2456

As part of data cleaning, we first filter lowly-expressed genes, defined as those with gene expression values smaller than 1 in more than half of the samples. As a result, 15,760 measurements remain in the RNA-seq dataset, while none is excluded from the microarray dataset. Next, we combine multiple reads, if any, for the

[4] *'Exon microarray summarized to genes'* from `http://www.brainspan.org/static/download.html`.
[5] *'RNA-Seq Gencode v10 summarized to genes'* from `http://www.brainspan.org/static/download.html`.
[6] Adapted by permission from Macmillan Publishers Ltd: Nature [21], Copyright (2011).

same gene in the same sample by taking the average of those reads. No multiple reads exist in the RNA-seq dataset. This leaves us with 16,768 unique genes in the microarray dataset and 15,760 unique genes in the RNA-seq dataset. Note that we identify genes by their Ensembl IDs rather than Associated Names or Entrez IDs, consistent with the fact that the CommonMind dataset also identifies genes by Ensembl IDs. Another advantage of this is that some genes have different Associated Names and/or Entrez IDs in the microarray and the RNA-seq datasets, even though their Ensembl IDs are the same. Additionally, as the BrainSpan data are unique in that both microarray and RNA-seq measurements are available, we take advantage of this fact by using both the microarray and the RNA-seq datasets. A drawback of this, however, is that in order to have two 'symmetrical' datasets – one measured in microarray and the other measured in RNA-seq – we have to keep only common genes and common samples, and in doing so exclude some samples and additional unique genes. At this point, we have expression values of 10,969 genes from 433 samples, each measured using both microarray and RNA-seq.

We further screen the samples for quality. Specifically, we use the same quality control criteria adopted by Parikshak *et al.*, as their study uses the same BrainSpan datasets as we do [1]. These criteria are outlined as follows:

(a) Aged between 8 ***post-conceptual weeks*** **(PCWs)** and 12 post-natal months;

(b) Taken from one of the following brain regions[7]: DFC, VFC, MFC, OFC, M1C, S1C, A1C, IPC, STC, ITC, and V1C; and

(c) With an ***RNA integrity number*** **(RIN)** of at least 9 for RNA-seq measurements.

Imposing these criteria, we have 139 common samples remaining in the microarray and the RNA-seq datasets. Moreover, of 10,969 genes, we only keep 2971 for which genetic association scores derived in Section 2.3 are available. Last but not least, we perform a log-transformation on the RNA-seq expression values using the formulae $log_2(v + 1)$, where v is an expression value measured in RNA-seq. To summarize, we have microarray and RNA-seq measurements of expression levels of 2971 genes from 139 samples as our finalized gene expression data.

[7]See Kang *et al.* [21] for descriptions of the brain regions.

3.2 Removal of Age Effect through Regression

We decide to look into any age or gender effect on gene expression levels after failing to obtain a relatively scale-free co-expression network during preliminary analysis using the gene expression data obtained at the end of Section 3.1. For each gene, using its gene expression values as response, we attempt to fit linear regression models with different explanatory variables, ranging from age in terms of period as a single continuous variable, age in terms of days also as a single continuous variable, gender as a single categorical variable, to age in terms of period and gender as two explanatory variables, and age in terms of days and gender as two variables.

As it would be impractical to examine regression diagnostics for all 2971 genes, we select a small subset to look at. Specifically, in effort to be more representative, we run diagnostics on the aforementioned regression models for 5 genes with the largest genetic association scores (i.e. smallest p-values) derived in Section 2.3, 5 randomly selected genes with p-values smaller than 0.01, and 5 randomly selected genes with p-values equal to or greater than 0.01. Due to space limitation, we only show here in Figure 3.2.1 diagnostic plots for *BTN3A2*, the gene with the largest genetic association score for schizophrenia. Diagnostic plots for the rest of the semi-randomly selected genes are presented in Supplemental Information 7.3.

Based on the regression diagnostics, we determine that there is a reasonably linear relationship in most of the genes examined between age in terms of period and expression values, as well as between age in terms of days and expression values. We prefer using age in terms of period over age in terms of days as the explanatory variable, as the residuals vs. fitted values plots and the residuals vs. X values plots appear more evenly spread out and hence more pattern-less in the case of the former. Gender, on the other hand, appears to have little effect on the expression values of many of the genes examined. We therefore decide to remove only the age effect on the expression values of each gene through fitting a linear regression model of its expression values against the ages in period of its samples, and extracting the residuals for use as the new expression values with age effect removed. We show in Figure 3.2.2 the distributions of the adjusted R^2 of the regression models for all the genes, which appear to be skewed to the right in the cases of both microarray and RNA-seq measurements.

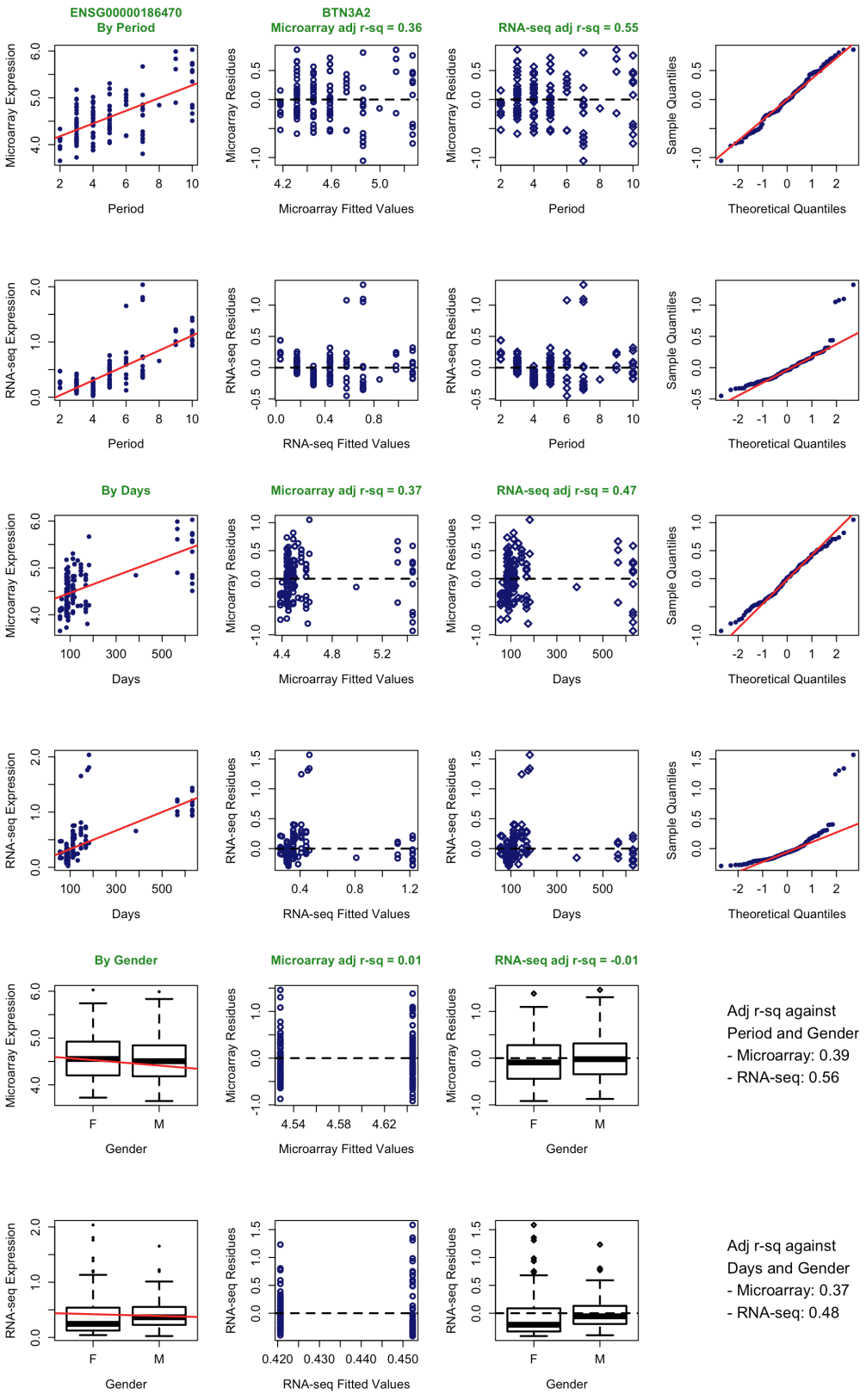
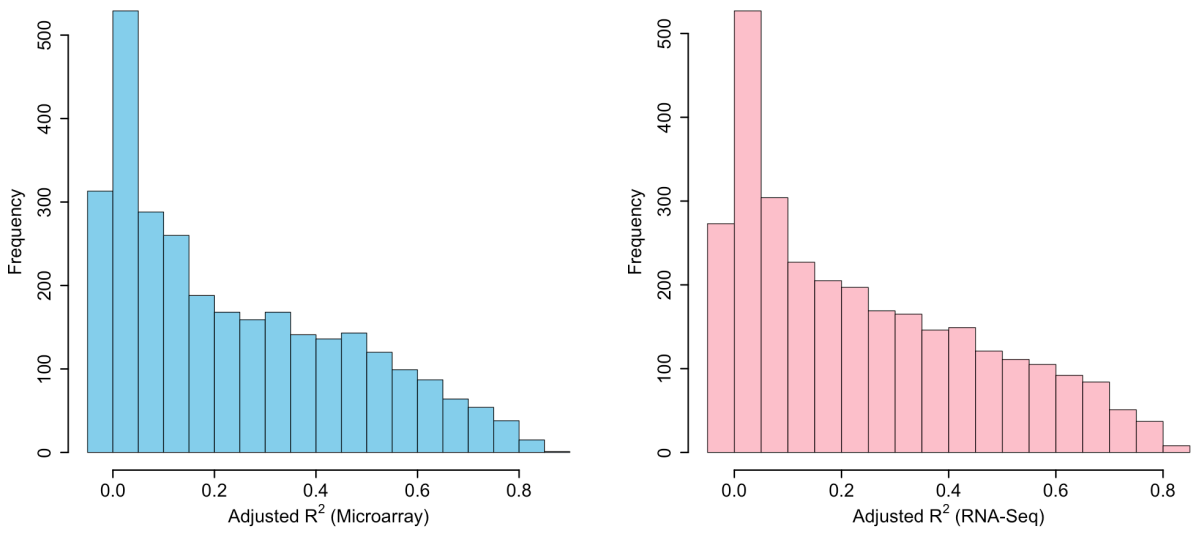


Figure 3.2.1: *Regression Diagnostics for BTN3A2.* This gene has the largest genetic association score for schizophrenia. Using expression values as response, diagnostics are shown for models using period, days, and gender as explanatory variable respectively. Adjusted R^2 of models using two explanatory variables are also shown.



(A) Using Microarray Expression Data

(B) Using RNA-Seq Expression Data

Figure 3.2.2: *Distributions of Adjusted R^2 from Linear Regressions Against Period for All Genes.* For each gene, we remove age effect on gene expression by fitting a regression model of its expression values against the ages in period of its samples, and extracting the residuals for use as new expression values.

3.3 Correlation-wise Odd Pairs

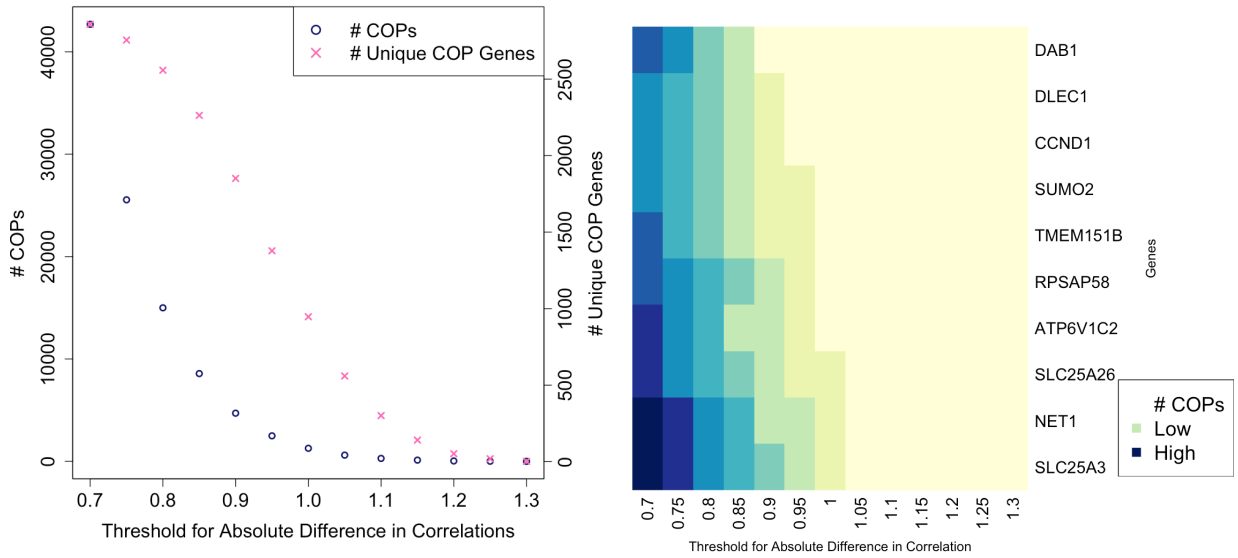
Through preliminary analysis, we also become aware of the existence of pairs of genes whose correlations in the microarray dataset differ considerably from those in the RNA-seq dataset. That is, let $\mathbf{r_{micro}}$ and $\mathbf{r_{RNA}}$ be the correlation coefficients of expression values of gene A and gene B measured using microarray and RNA-seq respectively; we find pairs of genes such as gene A and gene B for which the absolute difference in their correlation coefficients exceeds a non-trivial threshold $\mathbf{t_{COP}}$:

$$|r_{micro} - r_{RNA}| \geq t_{COP}. \tag{3.3.1}$$

While we by no means expect the correlation coefficients of two genes based on their microarray and RNA-seq measurements to match exactly, the extent of differences we discover is surprising, especially considering that we are looking at measurements of the same genes from the same samples with the only difference being the measurement technology. For instance, it would be odd to observe that gene A and gene B are positively correlated with an r_{micro} of 0.75 in the microarray dataset, whereas that the same pair of genes are negatively correlated with an r_{RNA} of -0.68 in the RNA-seq dataset. We therefore call the pairs of genes that exhibit such unexpected behavior ***Correlation-wise Odd Pairs*** **(COPs)**. It follows that any gene

involved in such a pair is called a ***COP gene***. Furthermore, an 'active' COP gene – one that is involved in a large number of COPs – is called a ***COP hub***.

We perform a global search across all genes and samples for COPs at varying thresholds for the absolute difference in r_{micro} and r_{RNA}. The numbers of COPs and COP genes detected at different t_{COP}'s are shown in Figure 3.3.1A. Unsurprisingly, the numbers drop as t_{COP} increases, indicating that fewer pairs of genes have very large absolute differences in microarray and RNA-seq correlations. Once we know the identity of the COP genes and thus the number of COPs a gene is involved in at various t_{COP}'s, we select as COP hubs those genes involved in the largest number of COPs on average across thresholds. In doing so, we identify 10 COP hubs, each involved in over 100 COPs on average. Figure 3.3.1B visualizes the number of COPs each COP hub is involved in at various thresholds. Details of the COP hubs are presented in Table 7.4.1 in Supplemental Information 7.4.



(A) Number of COPs and COP Genes vs. Threshold

(B) COP Hubs at Various Thresholds

Figure 3.3.1: *COPs, COP Genes, and COP Hubs across Thresholds before Transformation.* We search for COPs at various thresholds, and count the numbers of COPs and COP genes to aid in picking a threshold. We also count the number of COPs that each gene is involved in at various thresholds, and consider those involved in the largest number of COPs on average to be COP hubs.

Identification of COPs is important at a time where RNA-seq technology is gaining widespread popularity, yet at the same time microarray technology still offers much potential to be exploited. On the one hand, RNA-seq technology has been considered by many as an improvement over certain limitations of microarray technology, such as the latter's limited coverage tied to probes available [22]. Empirically, several known

ASD-associated genes such as *CDH8* had been excluded in the past when applying DAWN on microarray-only ASD data, because of poor measurement of these genes using microarray technology. On the other hand, however, it would be wasteful not to harness the rich biological information embedded in the abundantly available microarray datasets because of poor measurement of a small percentage of genes. In particular, suppose that expression of these genes have also been measured using RNA-Seq and with high quality, it could be useful to 'correct' for the poor microarray measurements of these genes based on their more reliable RNA-seq measurements.

3.4 'Fixing' of COPs via Transformation

Given the 'symmetric' nature of our microarray and RNA-seq datasets, we find ourselves in a perfect position to experiment with the idea of 'correcting' for poor microarray measurements of genes based on their high-quality RNA-seq measurements. Where the microarray correlation of two genes differs from the RNA-seq correlation for more than a given threshold (i.e. the two genes belong to a COP at a given t_{COP}), we conjecture that such difference is largely due to at least one of the COP genes being poorly measured on the microarray platform. We consider it reasonable to make this assumption for two reasons. First, we have imposed rather stringent quality control on the RNA-seq measurements by adopting an RIN threshold of at least 9 in Section 3.1. It would therefore be much less likely that the genes in our final dataset are poorly measured on the RNA-seq platform. Second, as previously discussed, RNA-seq technology is generally regarded as an improvement over microarray technology with less measurement bias [22]. To 'fix' COPs by 'correcting' for the microarray measurements of COP genes, we propose the following procedures:

1. ***Gaussianize*** high-quality RNA-seq measurements via ***nonparanormal transformation***. This is implemented using the *huge* package [23] in `R` [16]. Figure 3.4.1A visualizes the distribution of Gaussianized RNA-seq expression values of the genes in our final dataset.

2. Compute the ***empirical cumulative distribution function*** **(eCDF)** of Gaussianized RNA-seq data obtained in Step 1.

3. For a given set of COP genes, obtain the corresponding percentiles of the genes in Gaussianized RNA-seq data, based on the eCDF computed in Step 2.

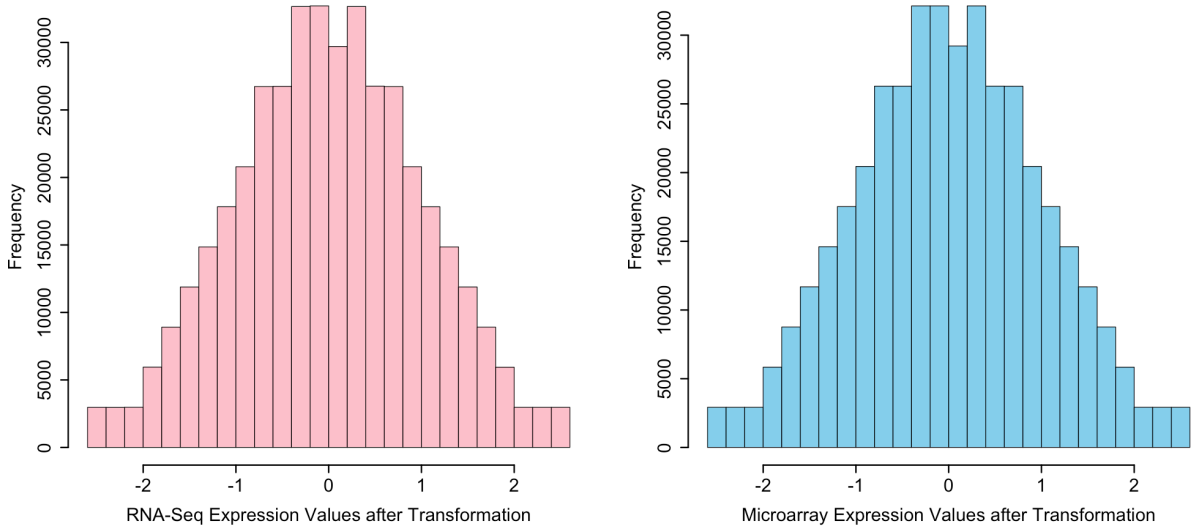
4. Remove measurements of the given genes from the microarray data, and Gaussianize the remaining microarray data via nonparanormal transformation. Figure 3.4.1B visualizes the distribution of Gaussianized microarray expression values of the remaining genes. Notice the similarity between the

distribution of Gaussianized RNA-seq data (Figure 3.4.1A) and that of Gaussianized microarray data (Figure 3.4.1B).

5. Using the percentiles obtained in Step 3 and based on Gaussianized microarray data obtained in Step 4, estimate Gaussianized microarray measurements for the given genes.

6. Add Gaussianzied microarray measurements for the given genes estimated in Step 5 back to Gaussianized microarray data obtained in Step 4.

7. Combine Gaussianized RNA-seq data obtained in Step 1 and Gaussianized microarray data containing 'fixed' estimates for the given set of COP genes obtained in Step5.



(A) RNA-Seq

(B) Microarray

Figure 3.4.1: *Distributions of Expression Values after Nonparanormal Transformation.* During transformation, we first Gaussianize high-quality RNA-seq measurements using the *huge* package [23] in `R` [16]. After removing microarray measurements of a given set of COP genes, we Gaussianize the remaining microarray data. Transformed microarray estimates for the genes removed are obtained based on percentiles of their transformed RNA-seq measurements and eCDF of Gaussianized RNA-seq data.

To summarize, we first perform a nonparanormal transformation on the RNA-seq measurements to get a Gaussianized distribution. We then estimate the percentiles of a given set of COP genes in the Gaussianized RNA-seq distribution. After removing their microarray measurements, we perform a nonparanormal transformation on the remaining microarray data. Next we estimate Gaussianized microarray measurements for the COP genes using the percentiles and the Gaussianized microarray distribution obtained earlier. Finally,

we combine Gaussianized RNA-seq data and Gaussianized micorarray data. An advantage of performing the above-mentioned series of transformation is that the transformed measurements in the end product – a combined Gaussianized dataset – are all on the same scale, as opposed to separate scales for the original RNA-seq and microarray measurements. This allows us to increase the sample size by incorporating two separate sources of data into a single analysis.

In practice, in order to determine a set of COP genes to be 'fixed', we first pick a threshold, t_{COP}, at which we capture COPs. Upon re-examining Figures 3.3.1A, we decide to choose $t_{COP} = 1.0$. This threshold is not as stringent as $t_{COP} = 1.3$, beyond which no more COPs exist. At the same time, it is not as relaxed as $t_{COP} = 0.7$, thus avoiding the need to attempt to 'fix' too many (≥ 1000) COP genes. In fact, upon examining the number of COPs each COP gene is involved in at $t_{COP} = 1.0$, we show in Figure 3.4.2 that there is great unevenness amongst the COP genes in the number of COPs they are each involved in. At $t_{COP} = 1.0$, majority of the COP genes are each involved in no more than 5 COPs. Only a small number – 50 – of the COP genes are involved in 6 or more COPs. We therefore choose to attempt 'fixing' on these 50 COP genes only, the details of which are presented in Table 7.5.1 in Supplemental Information 7.5. Note that this list of COP genes includes all the across-threshold COP hubs identified in Figure 3.3.1B.

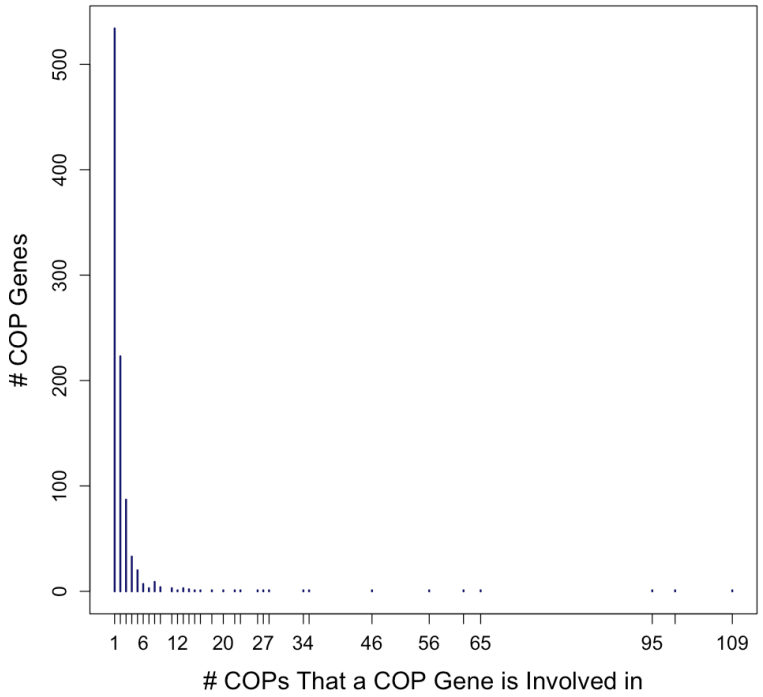
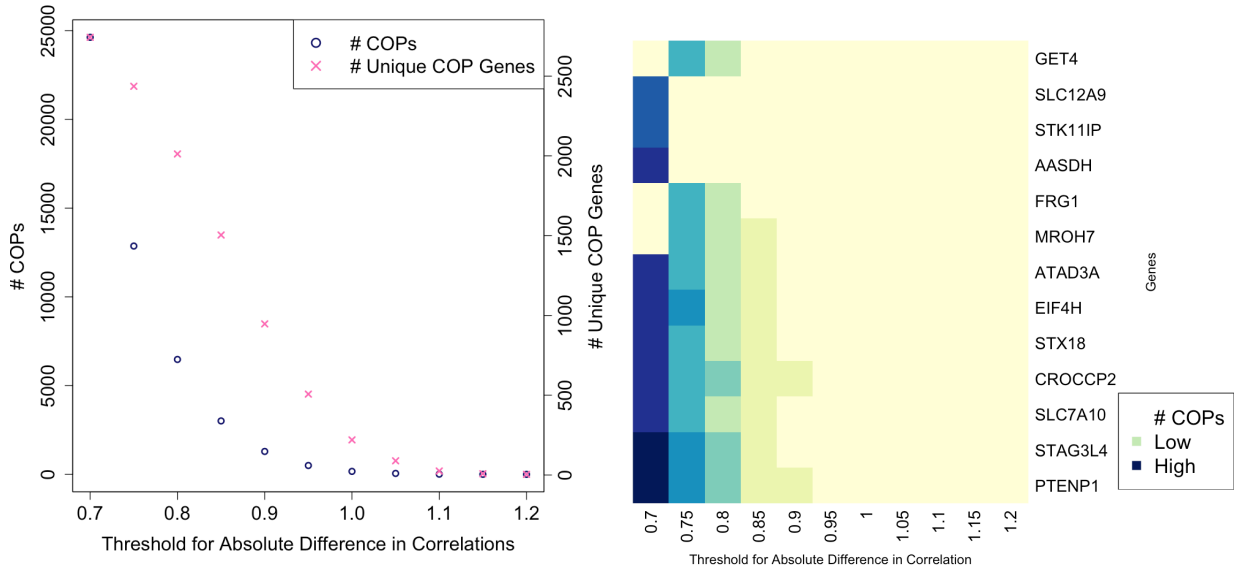


Figure 3.4.2: *Number of COPs That Each COP Gene is Involved in at* $t_{COP} = 1.0$. Majority of the COP genes are each involved in no more than 5 COPs, whereas only a small number (50) are involved in 6 or more COPs. The latter 50 are chosen to be 'fixed'.

3.5 Comparison of COPs before and after Transformation

We try out the idea of 'fixing' COPs via transformation on a set of 50 COP genes. Following transformation, we again perform searches across all genes and transformed samples for COPs at each t_{COP}, similar to that performed before transformation in Section 3.3. The numbers of COPs and COP genes detected – or rather, persisted – at different t_{COP}'s after transformation are shown in Figure 3.5.1A. Again, the numbers drop as t_{COP} increases, indicating that fewer pairs of genes have very large absolute differences in their transformed microarray and transformed RNA-seq correlations. Compared to Figure 3.3.1A, both the number of COPs and the number of unique COP genes at a given threshold appear to be lower after transformation than before.



(A) Number of COPs and COP Genes vs. Threshold

(B) COP Hubs at Various Thresholds

Figure 3.5.1: *COPs, COP Genes, and COP Hubs across Thresholds after Transformation.* Similar to the pre-transformation search, we search again for COPs at various thresholds, and count the numbers of COPs and COP genes at each threshold. The numbers at a given threshold are lower after transformation than before. We also count the number of COPs that each gene is involved in at various thresholds, and consider those involved in the largest number of COPs on average to be COP hubs. Post-transformation COP hubs have completely different identities and appear less hub-like compared to pre-transformation hubs.

COP hubs, defined as COP genes involved in the largest number of COPs on average across thresholds, are shown in Figure 3.5.1B, together with the number of COPs each hub is involved in at various thresholds. Compared to the 10 pre-transformation COP hubs previously identified in Figure 3.3.1B, the 13 post-transformation COP hubs in Figure 3.5.1B are completely new. This suggests that most, if not all, of the

pre-transformation COP hubs are no longer hubs across different thresholds, or in other words are 'fixed', by the transformation. As for the new COP hubs that emerge after transformation, their average numbers of COPs involved in range from 16 to 51, much smaller compared to a range between 101 and 202 before transformation. Details of the post-transformation COP hubs are presented in Table 7.6.1 in Supplemental Information 7.6.

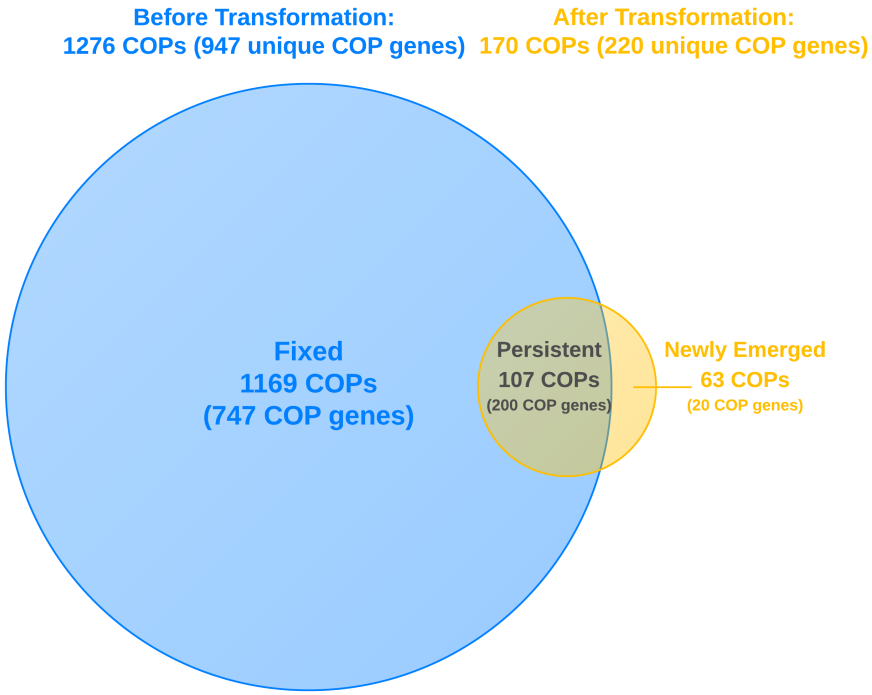
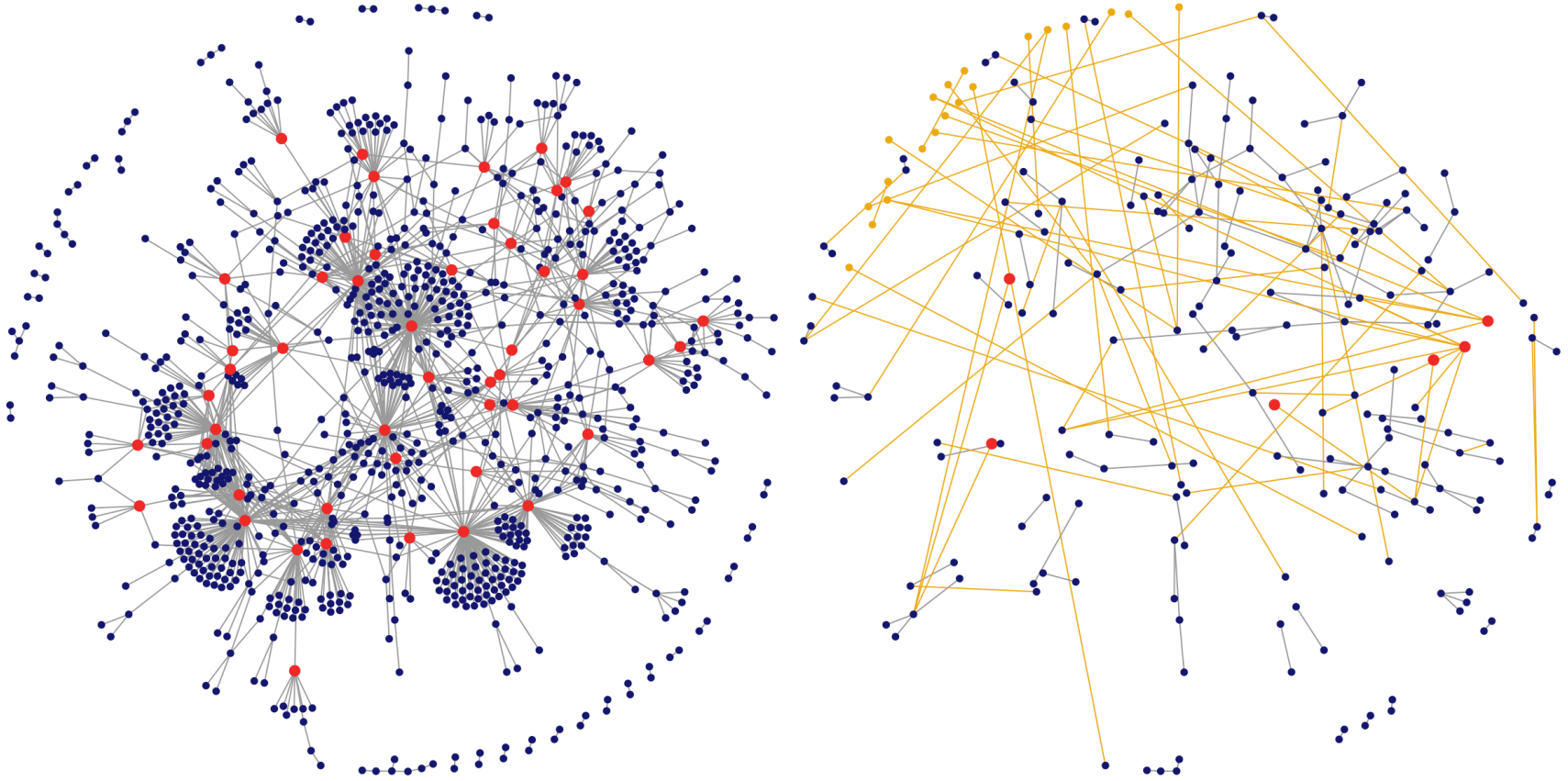


Figure 3.5.2: *Numerical Comparison of COPs before and after Transformation.* Of all the pre-transformation COPs, majority (91.6%) disappear after transformation, including those involving the 10 pre-transformation COP hubs. These are considered 'fixed'. A small fraction (8.4%) of COPs remain after transformation and are considered 'persistent'. Of all the post-transformation COPs, 37.1% are 'newly emerged' after transformation, possibly due to intrinsic stochasticity in the measurements.

Furthermore, we conduct a zoomed-in comparison of the COP networks at $t_{COP} = 1.0$ before and after transformation. Figure 3.5.2 provides a numerical summary of this comparison, and Figure 3.5.3 provides a COP-to-COP visual comparison. Prior to transformation, with reference to Figure 3.5.2, we capture 1276 COPs involving 947 unique COP genes at $t_{COP} = 1.0$. A complete network of these 1276 COPs is visualized in Figure 3.5.3A, with the 50 COP genes picked to be 'fixed' in Section 3.4 highlighted in red. It becomes clear in Figure 3.5.3A that these 50 genes indeed appear to be hub-like as they are involved in at least 6 and as many as 109 COPs. Following transformation, with reference to Figure 3.5.2, we identify 170 COPs involving 220 unique COP genes at $t_{COP} = 1.0$, this time based on the transformed data obtained at the end of Section 3.4. That is, as shown in Figure 3.5.3B, majority (91.6%) of the

COPs that exist in Figure 3.5.3A disappear. The COPs that disappear after transformation include all of those involving the 10 across-threshold COP hubs identified in Figure 3.3.1B. In fact, out of the 50 COP genes picked to be 'fixed' – a list that includes the 10 COP hubs, only 6 remain involved in COPs after transformation. Still highlighted in red in Figure 3.5.3B, they are: *SORBS2*, *BRAF*, *TAOK1*, *CEP192*, *EIF5A*, and *NOTCH3*. The numbers of COPs that these 'persistent' COP genes remain involved in are also considerably smaller compared to those before transformation. On the other hand, with reference to Figure 3.5.2, a small number of COPs (8.4%) remain after transformation. Additionally, 63 new COPs (Figure 3.5.2) emerge after transformation and are highlighted in yellow in Figure 3.5.3B, representing 37.1% of the post-transformation COPs. Notwithstanding, none of the persisting old COP genes or the new COP genes appear nearly as hub-like as the ones highlighted in red in Figure 3.5.3A. It is likely that inherent stochasticity in the original measurements gives rise to the newly emerged COPs. We therefore consider 'fixing' via transformation a success.

We are now ready for downstream DAWN analysis.



(A) Before Transformation

(B) After Transformation

Figure 3.5.3: *Visual Comparison of COPs before and after Transformation.* All COPs before and after transformation at $t_{COP} = 1.0$ are shown in a node-matched fashion. Pre-transformation COP genes picked for 'fixing' in Section 3.4 are colored in red. Of the 50 of them, only 6 'persist' after transformation. Other pre-transformation COP genes are colored in blue. After transformation, COPs and COP genes that 'persist' are colored as before. Disappearance of pre-transformation COPs or COP genes indicates successful 'fixing'. COPs and COP genes that newly emerge after transformation are colored in yellow.

This page intentionally left blank

4 DAWN Analysis

As we prepare to run the main DAWN algorithm, let us review our progress into the DAWN framework outlined in Section 1.1. We have completed Step (i) in Section 2.3, in which we derive genetic association scores of the genes. In Section 3, we clean and process our gene expression data, and get a combined dataset containing transformed microarray and RNA-seq measurements of 2971 genes from 139 samples. This has prepared us for Step (ii), to which we now proceed.

4.1 Co-expression Network

In Step (ii) of the DAWN framework, we construct a gene co-expression network based on genetic association scores of the genes and their correlation amongst each other, using a ***partial neighborhood selection* (PNS)** algorithm [8]. For this purpose, the PNS algorithm requires a threshold for the genetic association scores, often referred to interchangeably as the p-values; and a threshold for the correlation of gene expression values. Based on empirical experience and canonical practice in human genetics literature, we adopt 0.1 and 0.7 for the p-value and the correlation thresholds respectively.

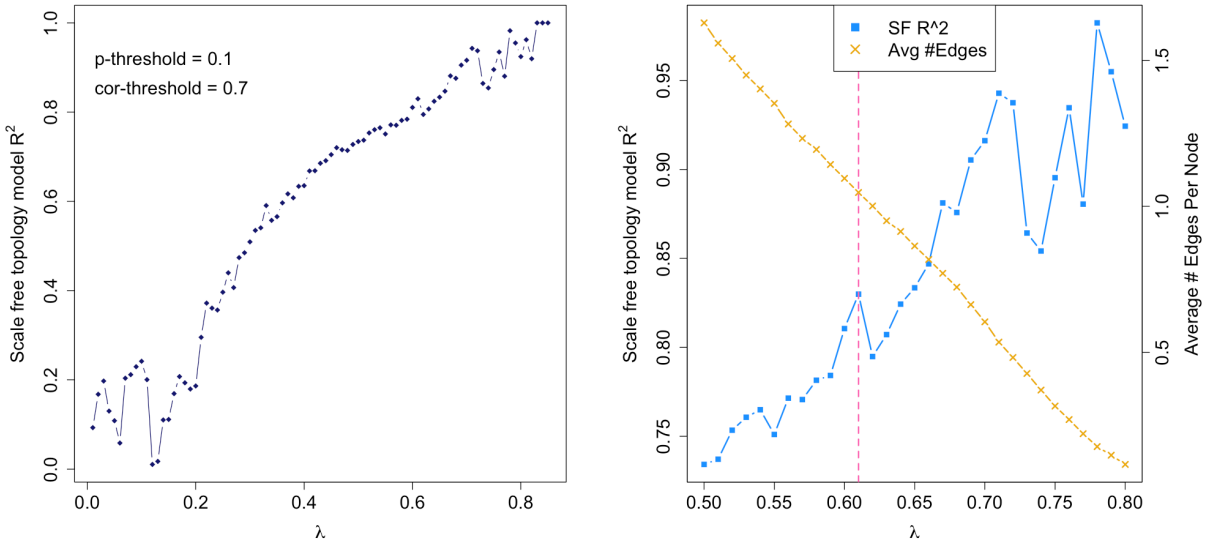
In addition, the PNS algorithm requires a ***regularization parameter***, λ, for the sparse lasso regression [9] that it uses to estimate the network. The DAWN framework supports the choice of a λ between 0 and 1 that achieves a reasonable tradeoff between a high degree of scale-freeness and moderate sparsity of the resultant network [8]. The degree of scale-freeness of a given network can be measured using a scale-free criterion proposed by Zhang and Horvath based on the observation that biological networks tend to be scale-free – that is, $p(k)$, the probability that a node connects to k other nodes, decays as a power law $p(k) \sim k^{-\gamma}$ ($\gamma > 1$) [10]. This criterion, ***scale-free topology model R^2* (SF-R^2)**, is defined as [10]

$$SF\text{-}R^2 = \Big)corr\Big]log\ p(k)\Big(,\ log\ k\Big(\ \sum^{2} \tag{4.1.1}$$

Ranging between 0 and 1, an SF-R^2 of 1 indicates that the network follows the power law perfectly [8]. The sparsity of the network, on the other hand, can be measured, albeit crudely, by the average number of edges per node. One could expect a positive correlation between λ and SF-R^2, and a negative correlation between λ and the average number of edges per node. As DAWN is a novel framework for which a canonical choice of λ is unavailable, we perform ***parameter tuning*** for λ while fixing the other two thresholds.

For λ from 0.01 to 0.85 with an increment of 0.01, we run the PNS algorithm and compute the SF-R^2's

of the corresponding networks. The results are shown in Figure 4.1.1A. Ideally, we would prefer an $SF\text{-}R^2$ around 0.9. In this case, with reference to Figure 4.1.1A, that would lead us to pick a λ close to 1, which might result in an overly sparse network. Based on the results in Figure 4.1.1A, we zoom into a smaller range of λ that gives relatively large $SF\text{-}R^2$'s and measure the sparsity of the corresponding networks. As suspected, with reference to Figure 4.1.1B, a λ greater than 0.7 tends to result in a network with an $SF\text{-}R^2$ around 0.9 but fewer than 1 edge per gene on average. We therefore make a compromise between the degree of scale-freeness of the network and its sparsity by choosing a λ of 0.61, which results in an estimated gene co-expression network with 950 genes, 995 edges, an average of 1.05 edges per gene, and an $SF\text{-}R^2$ of 0.83.



(A) SF-R^2 vs. λ

(B) SF-R^2 and Average Number of Edges vs. λ over a Narrower Range

Figure 4.1.1: *Parameter Tuning for* λ. When estimating the gene co-expression network, we aim to reach a reasonable trade-off between a high degree of scale-freeness (an SF-R^2 of around 0.9) and moderate sparsity (an average number of edges per gene of around 1.5). Picking a λ of 0.61 (pink dotted line), we obtain a gene co-expression network estimate with 950 genes, 995 edges, an average of 1.05 edges per gene, and an $SF\text{-}R^2$ of 0.83.

4.2 Hidden Markov Random Field Model

In Steps (iii) of the DAWN framework, we use a ***hidden Markov random field*** **(HMRF)** model to search for risk genes, based on p-values of the genes and the estimated gene co-expression network. The philosophy behind this approach stems from the observation that while very few genes have p-values that are significant

at the genome-wide level, some genes with small p-values appear clustered in the co-expression network [8]. It is considered 'highly unlikely to happen by chance' that a gene with a small schizophrenia-specific p-value has many risk gene neighbors [8].

The HMRF incorporates information embedded in the p-values by converting them to normal z-scores (Z_i's), and assuming that the z-scores follow a Gaussian mixture distribution, where the mixture membership of Z_i is determined by its hidden state I_i [8]. A true risk gene has a hidden state of 1, whereas a non-risk gene has a 0. The framework further assumes that Z_i with $I_i = 0$ is normally distributed with mean 0 and variance σ_0^2, that Z_j with $I_j = 1$ is approximately normally distributed with mean μ and variance σ_1^2, and that Z_i and Z_j are conditionally independent given their hidden states I_i and I_j [8]. Expressing the model as

$$Z_i \sim P(I_i = 0)\ N(0, \sigma_0^2) + P(I_i = 1)\ N(\mu, \sigma_1^2), \tag{4.2.1}$$

where σ_0^2, μ, and σ_1^2 remain to be estimated, Liu *et al.* show that this 'dependence structure reduces to the dependence of hidden states' [8]. The latter can be modeled using an Ising model with probability mass function

$$P(\mathbf{I} = \eta) \propto exp\Big(b^t\ \eta + c\ \eta^t\ \Omega\ \eta\Big(\text{ for all } \eta \in \{0,1\}^n, \tag{4.2.2}$$

where b and c are parameters to be estimated; Ω is the binary adjacency matrix of the co-expression network; and n is the number of genes in the network [8].

The iterative algorithm used to estimate the parameters requires us to know the hidden states of some genes in order to initialize. These states are also known as ***seed states***. With the true hidden states unknown to us, we make some educated choices. Of the 950 genes in the network, we assign a ***fixed hidden state*** of 1 to 10 of them whose p-values are in the lowest 1%. The rationale is that the marginal evidence presented by their extremely small p-values is strong enough for us to assume that their true hidden states are 1. These genes are: *RAI1*, *PCCB*, *ATAT1*, *MAPK7*, *GATAD2A*, *SRR*, *SEPT10*, *BRD2*, *DDAH2*, and *LY6G5B*. Additionally, we assign a seed state of 1 to the rest of the 14 genes whose p-values are in the lowest 2.5%. These genes are: *SPA17*, *LRCH4*, *WDR55*, *SCRN3*, *CISD2*, *INA*, *GPD1L*, *FAM167A*, *PDE9A*, *CPT1C*, *CXXC5*, *FAM221A*, *HSPA1A*, and *HSPA1B*. The hidden states of seed genes could change from iteration to iteration, whereas fixed hidden states remain unchanged. It should be noted, however, that whether a gene assigned with a fixed hidden state of 1 gets selected as a risk gene subsequently in Step (iv) depends on its posterior probability after FDR correction of being a risk gene in relation to that of the other genes.

Initializing the iterative algorithm with the above-mentioned seed states, we obtain after 17 iterations an

estimate of 1.64 for μ, and an estimate of 0.967 for both σ_0^2 and σ_1^2 (equal variance is further assumed by the algorithm for $Z_i|I_i = 0$ and $Z_j|I_j = 1$). In addition, we obtain estimates for b and c as -4.92 and 3.92 respectively. The fact that $c > 0$ suggests that genes with estimated hidden states of 1 tend to form clusters, a characteristic that we consider favorably.

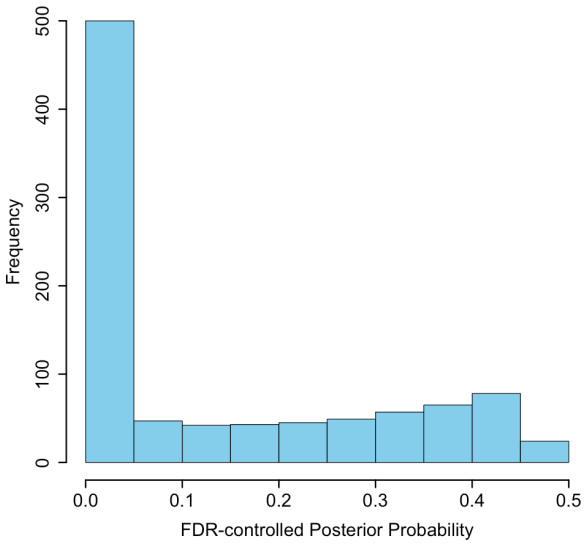
In Step (iv), in addition to estimating the parameters of the HMRF model, the algorithm also applies Gibbs sampling to estimate $pp_i = P(I_i = 0|\mathbf{Z})$, the posterior probability that the true hidden state of a gene is 0 given the z-score distribution of all the genes [8]. Lastly, we apply Bayesian FDR correction [11] to the posterior probabilities. To do so, we sort pp_i's in ascending order into $pp_{(i)}$'s, and compute the ***FDR-controlled posterior probability*** **(FPP)** that the true hidden state of the k^{th} sorted gene is 0 by [8]

$$FPP_k = \prod_{i=1}^{k} \frac{pp_{(i)}}{k}. \tag{4.2.3}$$

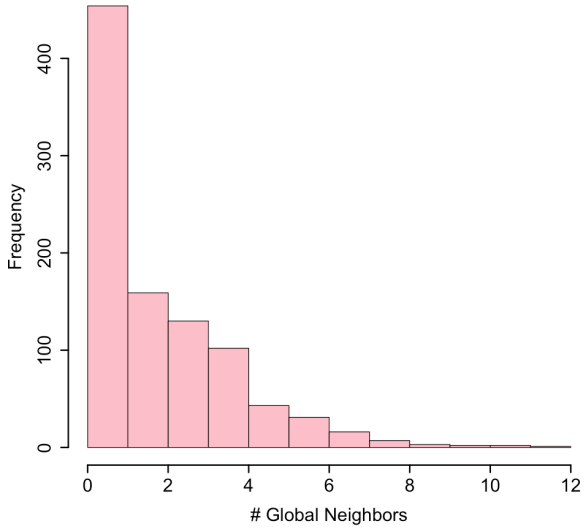
Important to note is that the FPP of a gene is the probability of that gene *not* being a risk gene for schizophrenia. Hence, the *smaller* the FPP of a gene is, the *more* likely that it is a risk gene. FPPs need to be formulated this way in order to accommodate the requirements of FDR correction.

4.3 Schizophrenia Risk Genes and Sub-networks

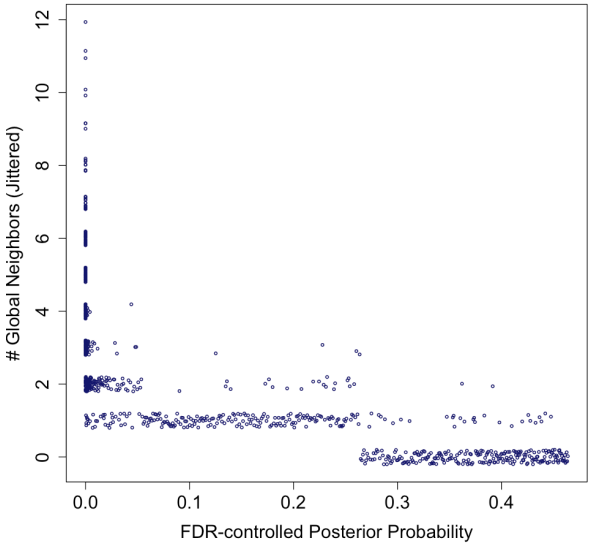
In the last step of the DAWN framework, we choose a cut-off for FPPs and select risk genes based on their posterior probabilities of being a risk gene (recall that smaller FPP means greater probability of being a risk gene). We show in Figure 4.3.1A the distribution of FPPs of the genes in our network. Because the vast majority of our genes appear to have rather small FPPs, as indicated by the distribution's severe skewedness to the right in Figure 4.3.1A, we anticipate a much smaller cut-off compared to the canonical choice of 0.1 adopted in past applications of the framework. We also show in Figure 4.3.1B the distribution of numbers of neighbors of each gene. As we are counting all the edges connected to a gene, as opposed to counting only edges from certain genes, we denote this the number of 'global neighbors'. With reference to Figure 4.3.1B, this is again a distribution that skews severely to the right, with a small number of genes having as many as 12 global neighbors and the majority of genes having fewer than 2 global neighbors. In addition, we examine the bivariate relationship between the FPP of a gene and its number of global neighbors. This is shown in Figure 4.3.1C. It becomes immediately clear that these two variables correlate negatively. Their correlation coefficient is -0.71.



(A) Distribution of FPPs



(B) Distribution of Numbers of Global Neighbors



(C) Number of Global Neighbors vs. FPP

Figure 4.3.1: *Distributions of FPPs and Numbers of Global Neighbors of Genes in Co-expression Network.* FPP estimates the probability that a gene is *not* a risk gene. Genes with *smaller* FPPs are more *likely* to be risk genes. The number of global neighbors of a gene counts all of its edges, as opposed to counting only edges from certain genes. These two variables correlate negatively with a correlation coefficient of -0.71.

Given the severely right-skewed distribution of FPPs (Figure 4.3.1A), what would be a reasonable cut-off? Instead of choosing an arbitrary number, for instance, 0.01; we consider genes whose FPPs are in the lowest

10% as ***small-FPP genes***. As we have 950 genes in our co-expression network, there are 95 small-FPP genes to begin with. Amongst them, 8 are assigned fixed hidden states of 1 for the HMRF model in Section 4.2, and 3 more serve as seed genes with an initial hidden state of 1 (recall that 10 genes are assigned fixed hidden states of 1 and 14 genes are chosen to be seed genes with an initial hidden state of 1). We further measure the inter-connectedness of these small-FPP genes by counting their numbers of global neighbors that are also small-FPP genes. From these 95 small-FPP genes, we select those that fulfill either one of the following criteria as ***primary risk genes*** for schizophrenia:

- The gene is assigned a fixed hidden state of 1 in Section 4.2. That is, it has a convincingly small genetics-based p-value as well as a small FPP according to the DAWN algorithm.
- The gene is well-connected to other small-FPP genes. That is, it has a small FPP as well as a large fraction of its global neighbors being also small-FPP genes. Specifically, we consider being 'well-connected' to other small-FPP genes as being in the 75^{th} percentile or higher in terms of the fraction of global neighbors that are also small-FPP genes.

Using these criteria, we identify 39 primary risk genes from the pool of small-FPP genes. Isolated small-FPP genes – small-FPP genes that neither are risk genes nor have any small-FPP neighbor – are removed, as it is our belief that risk genes function together as networks rather than alone. After excluding 12 isolated small-FPP genes, the remaining 44 small-FPP genes are then classified as ***secondary risk genes*** for schizophrenia. To summarize, we obtain a final set of 39 primary risk genes in addition to 44 secondary risk genes.

We examine the primary and secondary risk genes more closely in Figure 4.3.2, which shows the bivariate relationship between the genetics-based p-values of these genes and the fractions of their global neighbors that are risk genes, in addition to being size-coded by the absolute number of risk gene neighbors. Numerical values of FPPs are not shown since these genes all have small FPPs below a given cut-off. With reference to Figure 4.3.2, majority of the risk genes lie to the left of the pink dotted line indicating a genetics-based p-value of 0.1. Equivalently, approximately 80% of the risk genes selected by DAWN have genetics-based p-values smaller than the p-value threshold adopted for estimating the gene co-expression network in Section 4.1. We consider this favorably as it is not impossible for the HMRF model to favor assigning hidden states of 1 to genes with large genetics-based p-values, in which case DAWN results would be at odds with the marginal evidence represented by the p-values, signaling potential failure(s) during DAWN analysis. With regards to the fraction of global neighbors that are also risk genes, it is no surprise that all but 3 of the primary

risk genes have a larger fraction of risk gene neighbors than their secondary counterparts – majority of the primary risk genes are selected for being well-connected to other small-FPP genes. The 3 primary risk genes with smaller fractions of risk gene neighbors are selected based on the alternative criterion requiring them to have been previously assigned a fixed hidden state of 1. More details, such as their names, descriptions, FPPs, numbers of risk gene neighbors, etc., on the risk genes are presented in Table 7.7.1 in Supplemental Information 7.7.

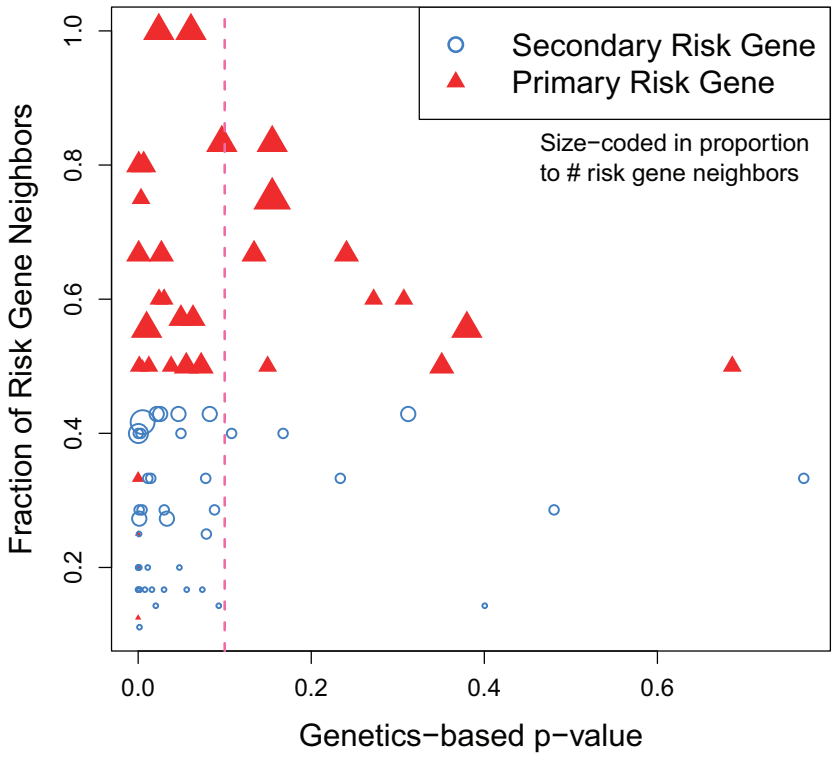


Figure 4.3.2: *Fraction of Risk Gene Neighbors vs. Genetics-based p-value of Primary and Secondary Risk Genes.* From a pool of small-FPP genes whose FPPs are in the lowest 10%, we select 39 genes that are either assigned fixed hidden states of 1 in Section 4.2 or well-connected to other small-FPP genes as primary risk genes. Isolated small-FPP genes are removed and the remaining 44 genes become secondary risk genes. Majority (80%) of the risk genes have p-values smaller than 0.1, the threshold used for estimating the gene co-expression network in Section 4.1, and thus lie to the left of the pink dotted line. Genes with a larger number of risk gene neighbors are plotted with bigger symbols.

We visualize the network amongst the primary and secondary risk genes themselves in Figure 4.3.3 using *igraph* [24]. With reference to Figure 4.3.3, primary risk genes (colored in red) appear by definition to be more well-connected in general than the secondary risk genes (colored in blue). Within the primary risk

genes, there appear to be two sub-types based on the type of genes they connect to. One sub-type appears to be more hub-like with respect to secondary risk genes, acting as a common co-expressed neighbor for several members of the latter. Examples include *CKAP2*, *ATAT1*, and *CRMP1*. Another sub-type, examples of which include *MNT* and *MAPK7*, appears to be mostly inter-connected with other primary risk genes as opposed to secondary risk genes. Of course, there are also some primary risk genes such as *PEX19* that connect to both primary and secondary risk genes and thus do not appear to fall into either sub-type in a clear-cut fashion.

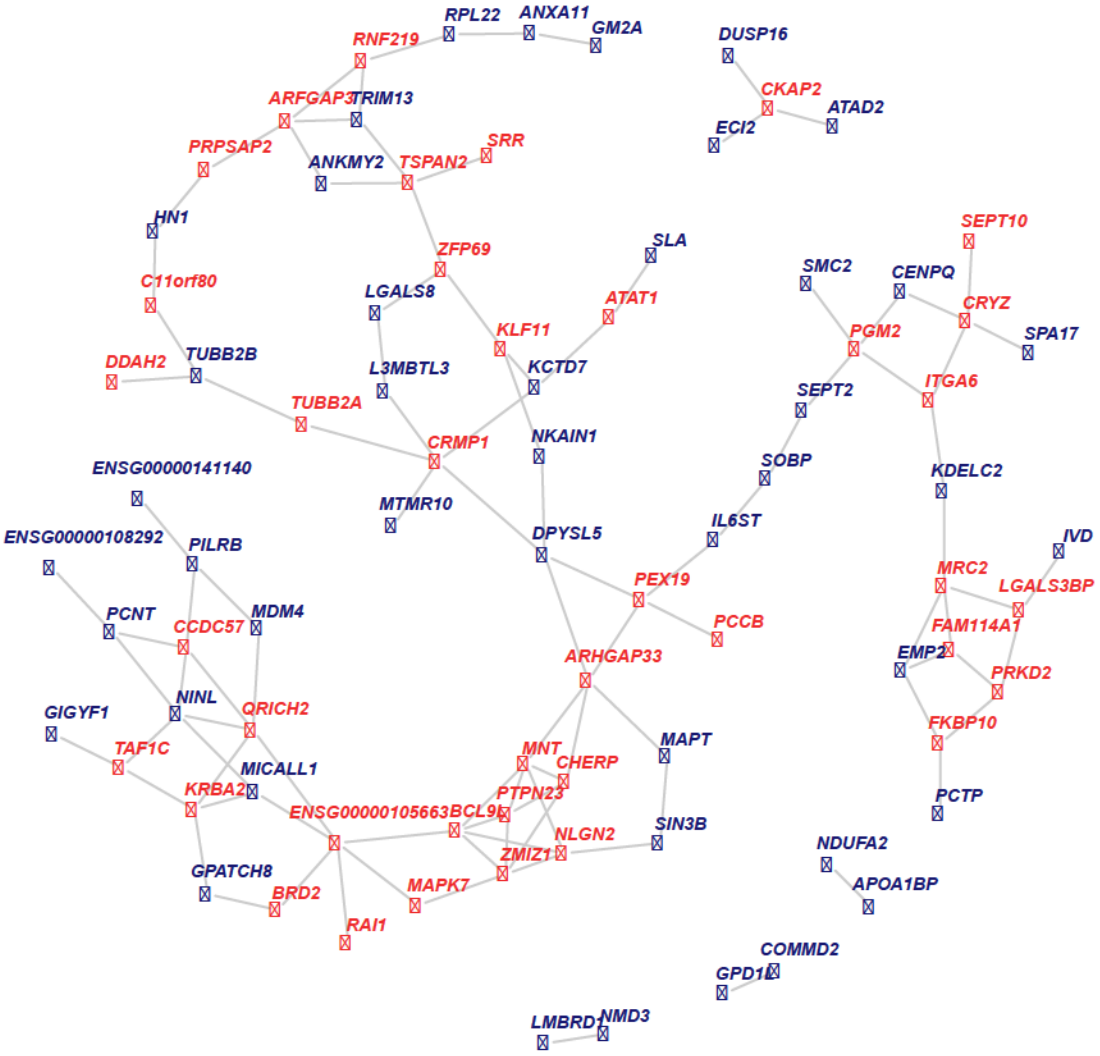


Figure 4.3.3: *Network Between Primary and Secondary Risk Genes.* Primary and secondary risk genes are colored in red and blue respectively. The former appears more well-connected than the latter. Some primary risk genes, such as *CKAP2* and *ATAT1*, appear to be more hub-like with respect to secondary risk genes. Others like *MNT* and *MAPK7* appear to be more inter-connected with primary risk genes themselves. (Genes for which no Associated Names are available in *Ensembl* are denoted by their Ensembl IDs.)

We also visualize the network amongst primary risk genes and their first-degree neighbors in Figure 4.3.4. In addition to 39 primary risk genes (colored in red), this network consists of secondary risk genes (colored in blue) and non-risk genes (colored in gray) that make up a total of 116 first-degree neighbors of the primary risk genes. Some secondary risk genes, such as *SOBP*, *GM2A*, and *NMD3* from Figure 4.3.3, are excluded as they do not connect to and therefore do not form part of a sub-network with any primary risk gene.

While examining Figure 4.3.4, several genes with known association with neuropsychiatric and/or neurological disorders immediately capture our attention. Amongst them is *MAPT*, a secondary risk gene that encodes microtubule-associated protein tau. Aggregation of tau proteins encoded by *MAPT* has long been recognized as a feature of tauopathy, a class of neurodegenerative diseases that includes Alzheimer's disease [25]. Recently, *MAPT* expression has been shown to also reduce adult neurogenesis, another characteristic of tauopathy [25]. Directly connected to *MAPT* in our network and of interest is *ARHGAP33*. Also known as *NOMA-GAP*, *ARHGAP33* has recently been shown to regulate synapse development and social behaviors that are often altered in neuropsychiatric developmental disorders such as ASD and schizophrenia in a mouse model [26]. Connecting to *ARHGAP33* via *MNT* is *PTPN23*, which has been identified as a novel candidate gene for neurological disorders in a recent whole-exome sequencing study [27]. In addition, *NLGN2*, which is an immediate neighbor of *PTPN23* and which encodes neuroligin-2, a protein vital for synaptogenesis and synaptic maturation, has been linked directly to schizophrenia [28]. Evidence suggests that rare mutations of *NLGN2* result in defects in GABAergic synapse formation, which may be an important trigger for schizophrenia [28].

Furthermore, upon visual inspection, with reference to Figure 4.3.4, there appears to be ***sub-networks*** formed around 6 subsets of primary risk genes. Each highlighted in a different background color in Figure 4.3.4, some sub-networks contain only one or two primary risk genes, while others are made up of as many as 15 primary risk genes.

Last but not least, see Figure 7.8.1 in Supplemental Information 7.8 for a visualization of the complete gene co-expression network and the positions of the risk genes in this network as predicted by DAWN.

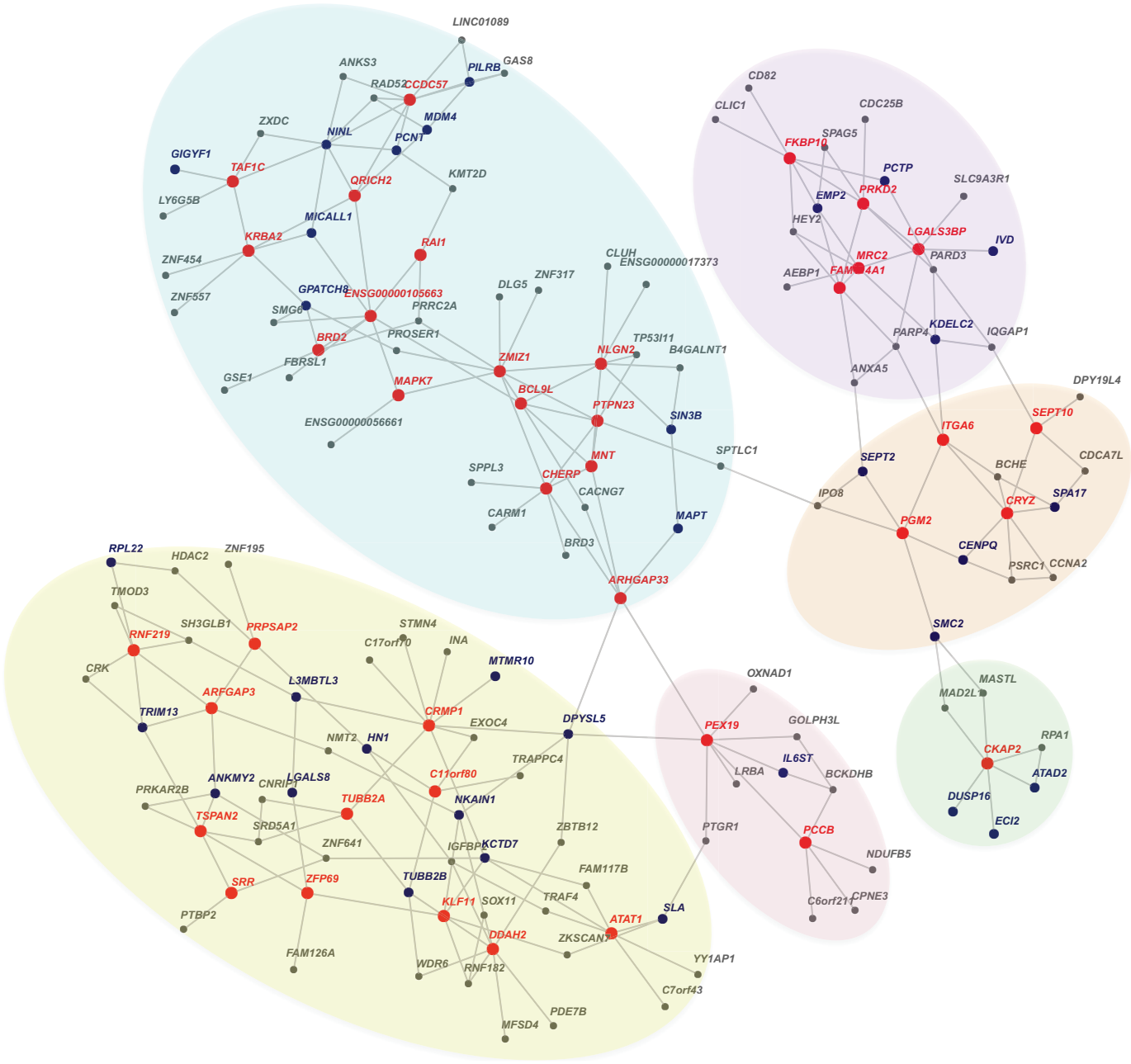


Figure 4.3.4: *Network Between Primary Risk Genes and First-degree Neighbors.* 39 primary risk genes, 33 secondary risk genes, and 83 non-risk genes are colored in red, blue, and gray respectively. Some individual genes that have been linked to neuropsychiatric and/or neurological disorders and that might thus be of interest include *MAPT*, *ARHGAP33*, *PTPN23*, and *NLGN2*. Based on visual examination of the network structure, 6 sub-networks formed around small subsets of primary risk genes are identified and highlighted in colors. (Genes for which no Associated Names are available in *Ensembl* are denoted by their Ensembl IDs.)

5 Discussion

THERE can be as much art to data analysis as there is science. From mapping, data transformation, to running through different steps of DAWN, multiple decisions regarding various cut-offs, thresholds, and parameters are made. While based largely on the objective facts presented by the data, many of these decisions also involve to varying degrees a subjective component. As a result, there is not a single correct answer to our question of interest, but instead different alternatives with their own pros and cons. Here, we reflect issues related to our particular approach, many of which are left as open-ended questions. We also discuss possible alternatives and future directions.

5.1 Reflections

During mapping of SNPs to genes in Section 2.1, given a SNP and a gene, we consider the q-value of their eQTL association. In the CommonMind data [14], all of the trans-eQTL associations are unique; and there is no overlap between trans-eQTL associations and cis-eQTL associations. There are, however, duplicate cis-eQTL associations. In other words, there are cases where a SNP and a gene have more than one cis-eQTL q-value available in the data. Fortunately, this is of little concern in our case after examining the distribution of q-values of the duplicate cis-eQTL associations. As these associations have a minimum q-value of 0.1992, which is much larger than our adopted q-value cut-off of 0.05, the fact that there are duplicates does not matter. However, what if the duplicates have q-values below our cut-off and as a result we do have to take them into account? Do we take the minimum, maximum, or average of the q-values of duplicate eQTL associations between a SNP and a gene? Do we consider a SNP mapped to a gene if the q-value of one of their eQTL associations is below the cut-off, while that of a duplicate association is above? More fundamentally, why are there different q-values of cis-eQTL associations between the same SNP and the same gene in the first place? We find these questions worth pondering even though they do not directly impact our particular analysis.

At the end of Section 2.2, we note that we will be looking out for *HLA-DQA1*, *HLA-DRB1*, and *HLA-C* – the MHC-encoding hyper-mapped genes with large z-scores – as we select risk genes. This has proved difficult as our final gene expression datasets obtained at the end of Section 3.1 do not contain *HLA-DQA1* and *HLA-C*. *HLA-DQA1* is absent in both the original microarray and the original RNA-seq datasets from BrainSpan [20]. While *HLA-C* is present in the original microarray dataset, it is missing from the original RNA-seq dataset. As our data cleaning protocol keeps only genes that are common to both the microarray

and the RNA-seq datasets, it has effectively excluded *HLA-C* from our analysis. As for *HLA-DRB1*, while it is commonly included in the microarray and the RNA-seq datasets, it is not chosen by DAWN to be part of the co-expression network estimated in Section 4.1. It therefore does not stand a chance to be selected as a risk gene for schizophrenia in our analysis. Amongst the risk genes that do get selected, with reference to Table 7.7.1 in Supplemental Information 7.7, only *SMC2* is considered hyper-mapped. Despite being mapped with 652 SNPs, *SMC2* as a secondary risk gene is connected to only 1 risk gene in the DAWN network.

The fact that some of the genes of potential interest, such as *HLA-DQA1* and *HLA-C*, are excluded from the final dataset prompts us to re-evaluate our data processing procedures. In particular, we have proposed to transform datasets measured using two different technologies – microarray and RNA-seq – to achieve comparable measurements. While doing so increases the sample size, it also requires that genes be present in both original datasets. As not every gene is measured or has measurements that pass quality control in both microarray and RNA-seq, some genes inevitably get excluded from the combined post-transformation dataset. In Section 3.1, for example, there are 16,768 and 15,760 unique genes in the original microarray and RNA-seq data respectively; but only 10,969 of them are common. The pertinent question to consider is then, which should we value more, a larger sample which confers more statistical power but which contains fewer genes, or a smaller sample which confers less power but which may contain more genes of potential interest?

In addition, we use regression in Section 3.2 to remove age effect from the gene expression measurements and use the residual values as the new measurements of levels of gene expression. The residual values, however, center around 0 and can be either positive or negative. Similarly, the post-transformation measurements in the combined dataset obtained at the end of Section 3.4 also have both positive and negative values centered around 0. While the negative values do not affect DAWN directly, they could make interpretation of the level of gene expression of a risk gene difficult. For instance, what does it mean for a risk gene to have a negative expression value after removal of age effect? What does it mean for a gene to have a post-transformation expression value of approximately 0?

In Section 3.5, after examining the remaining and newly emerged COPs after transformation, we decide that 'fixing' via transformation is a success. We base our judgment largely on the fact the post-transformation COP genes in Figure 3.5.3B do not appear nearly as hub-like as those in Figure 3.5.3A. For the purpose of drafting procedures for implementing transformation, however, it might be useful to also consider the possible scenario in which many pre-transformation COP hubs remain after applying transformation once.

In that case, what should we do with the persistent COPs and COP hubs? Should we apply transformation for a second time? Should we keep transforming the data until only a few COPs remain and no COP hubs exist? Would doing that help at all? In addition, in the event that persistent COP hubs exist even after transformation, would they appear different from the other genes in the DAWN network? In our set of risk genes, only *TUBB2B* is involved in any COP at all after transformation. Similarly unremarkable is that it is only involved in a single COP at $t_{COP} = 1.0$, and is hence certainly not a COP hub.

5.2 Future Directions

The method that we use in Section 2.3 to derive the genetic association scores of the genes may be too simplistic. By taking the minimum of the p-values for schizophrenia of the SNPs mapped to a gene as the genetic association score of that gene, we rely on the assumption that the degree of association between a gene and a disease correlates with that between the disease and the SNP to which the gene has the strongest association. There is nevertheless little evidence showing that this is always true. In the future, more sophisticated statistical methods may be adopted to derive these genetic association scores. Some suggested methods to try include He *et al.* 's *Sherlock* [13] and Conneely and Boehnke's *P Value Adjusted for Correlated Tests (P_{ACT})* [29].

In addition, while examining the distribution of genetic association scores derived for the genes in Figure 2.3.2 in Section 2.3, we note that bias favoring larger z-scores might be introduced by always taking the maximum z-score of the SNPs. More specifically, we assume in the HMRF model in Step (iii) of DAWN that the z-scores of genes with hidden states of 0 follow a normal distribution with mean 0 (Equation 4.2.1 in Section 4.2); whereas the mean of our derived z-score distribution shown in Figure 2.3.2 has clearly shifted to the right of 0. One way to adjust for the biased shift in the future would be to adjust the mean of the normal distribution assumed for z-scores of genes with $I_i = 0$ in Equation 4.2.1 from 0 to match that of the actual distribution of z-scores derived for the genes.

Last but not least, there is room for improvement in both detecting and characterizing sub-networks amongst the primary risk genes. Currently, the sub-networks in Figure 4.3.4 in Section 4.3 are identified by visual examination for obvious clusters formed around primary risk genes. A more robust way to detect sub-networks in the future would be to apply a community detection algorithm. Lancichinetti and Fortunato's comparative analysis on a wide spectrum of community detection algorithms may provide a good place to start [30]. After detection of sub-networks, bioinformatics databases such as *KEGG* [31, 32] may be consulted to characterize the metabolic functions of individual sub-networks. It might also be useful to

consider the functions of genes with known association with neuropsychiatric and/or neurological disorders, such as *MAPT* and *NLGN2* in Figure 4.3.4, in relation to the overall functions of the sub-networks; and vice versa.

6 Conclusion

In this project, we apply the DAWN framework in an attempt to identify schizophrenia risk genes and sub-networks. We first derive schizophrenia-specific genetic association scores of the genes. This is achieved through mapping thousands of SNPs to several thousand genes based on the CommonMind data [14], and taking advantage of association scores of the SNPs for schizophrenia from the PGC data [12]. In this process, we identify MHC-encoding hyper-mapped genes with association scores of potential interest. However, they do not become included in our analysis or selected as risk genes due to reasons discussed in Section 5.1.

Next, we prepare our gene expression data measured using microarray and RNA-seq from BrainSpan [20]. After removing age effect on gene expression levels using regression, we identify pairs of genes whose gene expression correlations appear drastically different in microarray and in RNA-seq. Defining them as COPs, we propose a series of transformation procedures that not only 'fix' the majority of the COPs but also render the microarray and the RNA-seq measurements comparable to each other, thereby increasing sample size.

We then perform parameter tuning for the PNS algorithm, seeking a reasonable trade-off between SF-R^2 and sparsity of the resultant co-expression network estimate. Based on the genetic association scores derived for the genes, and the gene co-expression network, we apply a HMRF model on our gene expression data combining transformed microarray and RNA-seq measurements. Applying Bayesian FDR control, we obtain FPPs of the genes in the co-expression network. A small subset of genes is selected as primary and secondary risk genes. Sub-networks consisting of the primary risk genes are identified and visualized.

Last but not least, we discuss future improvements in Section 5.2 and provide directions for using our code in Supplemental Information 7.9.

This page intentionally left blank

References

[1] N. N. Parikshak *et al.*, "Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism", *Cell*, vol. 155, no. 5, pp. 1008–21, 2013. DOI: 10.1016/j.cell.2013.10.031.

[2] S. De Rubeis *et al.*, "Synaptic, transcriptional and chromatin genes disrupted in autism", *Nature*, vol. 515, no. 7526, pp. 209–15, 2014. DOI: 10.1038/nature13772.

[3] B. E. Stranger, E. A. Stahl, and T. Raj, "Progress and promise of genome-wide association studies for human complex trait genetics", *Genetics*, vol. 187, no. 2, pp. 367–83, 2011. DOI: 10.1534/genetics.110.120907.

[4] J. Fan and R. Li, "Statistical challenges with high dimensionality: feature selection in knowledge discovery", in *Proceedings of the International Congress of Mathematicians*, M. Sanz-Solé *et al.*, Eds. Eur. Math. Soc., Zürich, 2006, vol. III, ch. 13, 595–622.

[5] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease", *Nat Rev Genet*, vol. 12, no. 1, pp. 56–68, 2011. DOI: 10.1038/nrg2918.

[6] J. Flint and M. Munafò, "Schizophrenia: genesis of a complex disease", *Nature*, vol. 511, no. 7510, pp. 412–3, 2014. DOI: 10.1038/nature13645.

[7] L. Liu *et al.*, "Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics", *Mol Autism*, vol. 5, no. 1, p. 22, 2014. DOI: 10.1186/2040-2392-5-22.

[8] L. Liu, J. Lei, and K. Roeder, "Network assisted analysis to reveal the genetic basis of autism", *Ann Appl Stat (Submitted)*, 2015.

[9] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso", *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006. DOI: 10.1214/009053606000000281.

[10] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis", *Stat Appl Genet Mol Biol*, vol. 4, Article17, 2005. DOI: 10.2202/1544-6115.1128.

[11] P. Müller, G. Parmigiani, and K. Rice, "Fdr and bayesian multiple comparisons rules", *Bayesian Statistics 8*, 2006.

[12] Schizophrenia Working Group of the Psychiatric Genomics Consortium, "Biological insights from 108 schizophrenia-associated genetic loci", *Nature*, vol. 511, no. 7510, pp. 421–7, 2014. DOI: 10.1038/nature13595.

[13] X. He *et al.*, "Sherlock: detecting gene-disease associations by matching patterns of expression qtl and gwas", *Am J Hum Genet*, vol. 92, no. 5, pp. 667–80, 2013. DOI: 10.1016/j.ajhg.2013.03.022.

[14] CommonMind Consortium, *Release 1 of commonmind consortium data*, Internet, Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH)., 2015. [Online]. Available: `http://commonmind.org/WP/data-generation/`.

[15] M. Maechler et al., *Sfsmisc: utilities from seminar fuer statistik eth zurich*, R package version 1.0-27, 2015. [Online]. Available: `http://CRAN.R-project.org/package=sfsmisc`.

[16] R Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: `http://www.R-project.org/`.

[17] Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2, "Genome-wide association study implicates hla-c*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia", *Biol Psychiatry*, vol. 72, no. 8, pp. 620–8, 2012. DOI: `10.1016/j.biopsych.2012.05.035`.

[18] A. K. McAllister, "Major histocompatibility complex i in brain development and schizophrenia", *Biol Psychiatry*, vol. 75, no. 4, pp. 262–8, 2014. DOI: `10.1016/j.biopsych.2013.10.003`.

[19] P. Flicek *et al.*, "Ensembl 2014", *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D749–55, 2014. DOI: `10.1093/nar/gkt1196`.

[20] BrainSpan, *Brainspan: atlas of the developing human brain*, Internet, Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01., 2011. [Online]. Available: `http://developinghumanbrain.org`.

[21] H. J. Kang *et al.*, "Spatio-temporal transcriptome of the human brain", *Nature*, vol. 478, no. 7370, pp. 483–9, 2011. DOI: `10.1038/nature10523`.

[22] J. C. Marioni *et al.*, "Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays", *Genome Res*, vol. 18, no. 9, pp. 1509–17, 2008. DOI: 10.1101/gr.079558.108.

[23] T. Zhao *et al.*, *Huge: high-dimensional undirected graph estimation*, R package version 1.2.6, 2014. [Online]. Available: http://CRAN.R-project.org/package=huge.

[24] G. Csardi and T. Nepusz, "The igraph software package for complex network research", *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: http://igraph.org.

[25] Y. Komuro *et al.*, "Human tau expression reduces adult neurogenesis in a mouse model of tauopathy", *Neurobiol Aging*, 2015. DOI: 10.1016/j.neurobiolaging.2015.03.002.

[26] S Schuster *et al.*, "Noma-gap/arhgap33 regulates synapse development and autistic-like behavior in the mouse", *Mol Psychiatry*, 2015. DOI: 10.1038/mp.2015.42.

[27] A. M. Alazami *et al.*, "Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families", *Cell Rep*, vol. 10, no. 2, pp. 148–61, 2015. DOI: 10.1016/j.celrep.2014.12.015.

[28] C. Sun *et al.*, "Identification and functional characterization of rare mutations of the neuroligin-2 gene (nlgn2) associated with schizophrenia", *Hum Mol Genet*, vol. 20, no. 15, pp. 3042–51, 2011. DOI: 10.1093/hmg/ddr208.

[29] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests", *Am J Hum Genet*, vol. 81, no. 6, pp. 1158–68, 2007. DOI: 10.1086/522036.

[30] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis", *Phys. Rev. E*, vol. 80, p. 056 117, 5 2009. DOI: 10.1103/PhysRevE.80.056117.

[31] M Kanehisa and S Goto, "Kegg: Kyoto encyclopedia of genes and genomes", *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.

[32] M. Kanehisa *et al.*, "Data, information, knowledge and principle: Back to metabolism in kegg", *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D199–205, 2014. DOI: 10.1093/nar/gkt1076.

This page intentionally left blank

7 Supplemental Information

7.1 List of Hyper-mapped Genes

Table 7.1.1: Genes Mapped to Over 500 SNPs

	Ensembl ID	Associated Name	Description	# SNPs	p-value	z-score
1	ENSG00000196735	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	5841	0.00000	9.36165
2	ENSG00000196126	HLA-DRB1	major histocompatibility complex, class II, DR beta 1	5265	0.00000	8.98600
3	ENSG00000205035	RP11-707M1.1	Unknown	3249	0.00127	3.01877
4	ENSG00000214425	LRRC37A4P	leucine rich repeat containing 37, member A4, pseudogene	3080	0.00005	3.90312
5	ENSG00000214401	KANSL1-AS1	KANSL1 antisense RNA 1	2995	0.00005	3.90312
6	ENSG00000120071	KANSL1	KAT8 regulatory NSL complex subunit 1	2976	0.00005	3.90312
7	ENSG00000238083	LRRC37A2	leucine rich repeat containing 37, member A2	2938	0.00005	3.90312
8	ENSG00000176681	LRRC37A	leucine rich repeat containing 37A	2928	0.00005	3.90312
9	ENSG00000185829	ARL17A	ADP-ribosylation factor-like 17A	2906	0.00005	3.90312
10	ENSG00000228696	ARL17B	ADP-ribosylation factor-like 17B	2852	0.00009	3.74779
11	ENSG00000266918	RP11-798G7.8	Unknown	2846	0.00005	3.90312
12	ENSG00000204650	CRHR1-IT1	CRHR1 intronic transcript 1 (non-protein coding)	2829	0.00007	3.79801
13	ENSG00000267198	RP11-798G7.6	Unknown	2826	0.00005	3.90312
14	ENSG00000232300	FAM215B	family with sequence similarity 215, member B (non-protein coding)	2740	0.00005	3.90312
15	ENSG00000244731	C4A	complement component 4A (Rodgers blood group)	2635	0.00000	9.36165
16	ENSG00000265218	RP11-927P21.1	Unknown	2465	0.00019	3.55877
17	ENSG00000204525	HLA-C	major histocompatibility complex, class I, C	2333	0.00000	9.39095
18	ENSG00000173295	FAM86B3P	family with sequence similarity 86, member B3, pseudogene	2301	0.00000	5.48787
19	ENSG00000216901	AL022393.7	Unknown	1720	0.00000	11.74245
20	ENSG00000226686	LINC01535	long intergenic non-protein coding RNA 1535	1395	0.03794	1.77511
21	ENSG00000163116	STPG2	sperm-tail PG-rich repeat containing 2	1368	0.03757	1.77961
22	ENSG00000187987	ZSCAN23	zinc finger and SCAN domain containing 23	1342	0.00000	11.74245
23	ENSG00000227888	FAM66A	family with sequence similarity 66, member A	1258	0.00000	5.48787
24	ENSG00000106610	STAG3L4	stromal antigen 3-like 4 (pseudogene)	1217	0.00100	3.09026
25	ENSG00000234585	CCT6P3	chaperonin containing TCP1, subunit 6 (zeta) pseudogene 3	1191	0.03042	1.87466
26	ENSG00000175170	FAM182B	family with sequence similarity 182, member B	1178	0.00235	2.82691

27	ENSG00000197465	GYPE	glycophorin E (MNS blood group)	1141	0.00623	2.49867
28	ENSG00000256274	TAS2R64P	taste receptor, type 2, member 64, pseudogene	1141	0.04662	1.67855
29	ENSG00000245958	RP11-33B1.1	Unknown	1127	0.01594	2.14591
30	ENSG00000246448	RP13-578N3.3	Unknown	1122	0.00020	3.54141
31	ENSG00000213462	ERV3-1	endogenous retrovirus group 3, member 1	1115	0.03042	1.87466
32	ENSG00000248828	RP11-673E1.4	Unknown	1111	0.01022	2.31817
33	ENSG00000171084	FAM86JP	family with sequence similarity 86, member J, pseudogene	1108	0.00108	3.06869
34	ENSG00000108883	EFTUD2	elongation factor Tu GTP binding domain containing 2	1105	0.00028	3.45125
35	ENSG00000261770	CTC-459F4.1	Unknown	1089	0.02284	1.99834
36	ENSG00000176998	HCG4	HLA complex group 4 (non-protein coding)	1066	0.00000	10.92442
37	ENSG00000170571	EMB	embigin	989	0.00000	4.87834
38	ENSG00000249244	RP11-548H18.2	Unknown	982	0.01594	2.14591
39	ENSG00000263142	LRRC37A17P	leucine rich repeat containing 37, member A17, pseudogene	977	0.00024	3.49145
40	ENSG00000164669	INTS4P1	integrator complex subunit 4 pseudogene 1	951	0.03042	1.87466
41	ENSG00000206344	HCG27	HLA complex group 27 (non-protein coding)	947	0.00000	6.63578
42	ENSG00000162753	SLC9C2	solute carrier family 9, member C2 (putative)	924	0.00002	4.16394
43	ENSG00000197134	ZNF257	zinc finger protein 257	917	0.00297	2.75163
44	ENSG00000172346	CSDC2	cold shock domain containing C2, RNA binding	911	0.00000	6.76290
45	ENSG00000162782	TDRD5	tudor domain containing 5	900	0.02224	2.00954
46	ENSG00000198039	ZNF273	zinc finger protein 273	892	0.07124	1.46662
47	ENSG00000226314	ZNF192P1	zinc finger protein 192 pseudogene 1	889	0.00000	11.37983
48	ENSG00000237636	ANKRD26P3	ankyrin repeat domain 26 pseudogene 3	889	0.04727	1.67192
49	ENSG00000182722	SEPHS1P1	selenophosphate synthetase 1 pseudogene 1	866	0.07124	1.46662
50	ENSG00000248044	ENSG00000248044	Unknown	830	0.01449	2.18376
51	ENSG00000172687	ZNF738	zinc finger protein 738	823	0.00004	3.94578
52	ENSG00000215146	RP11-313J2.1	Unknown	818	0.02854	1.90270
53	ENSG00000168038	ULK4	unc-51 like kinase 4	810	0.09477	1.31194
54	ENSG00000171806	METTL18	methyltransferase like 18	771	0.00881	2.37342
55	ENSG00000259905	PWRN1	Prader-Willi region non-protein coding RNA 1	754	0.12330	1.15865
56	ENSG00000182362	YBEY	ybeY metallopeptidase (putative)	748	0.00070	3.19416
57	ENSG00000221947	XKR9	XK, Kell blood group complex subunit-related family, member 9	738	0.04793	1.66526
58	ENSG00000100413	POLR3H	polymerase (RNA) III (DNA directed) polypeptide H (22.9kD)	736	0.00000	6.76290
59	ENSG00000108384	RAD51C	RAD51 paralog C	721	0.00035	3.38794
60	ENSG00000156253	RWDD2B	RWD domain containing 2B	694	0.00160	2.94765

61	ENSG00000186470	BTN3A2	butyrophilin, subfamily 3, member A2	694	0.00000	10.65362
62	ENSG00000136824	SMC2	structural maintenance of chromosomes 2	652	0.04768	1.66778
63	ENSG00000113593	PPWD1	peptidylprolyl isomerase domain and WD repeat containing 1	650	0.00872	2.37738
64	ENSG00000226752	PSMD5-AS1	PSMD5 antisense RNA 1 (head to head)	650	0.00074	3.17682
65	ENSG00000235109	ZSCAN31	zinc finger and SCAN domain containing 31	650	0.00000	9.55620
66	ENSG00000182632	CCNYL2	cyclin Y-like 2, pseudogene	647	0.02854	1.90270
67	ENSG00000255556	RP11-351I21.6	Unknown	646	0.00003	4.00279
68	ENSG00000214776	RP11-726G1.1	Unknown	641	0.00700	2.45747
69	ENSG00000185904	LINC00839	long intergenic non-protein coding RNA 839	625	0.09001	1.34069
70	ENSG00000176390	CRLF3	cytokine receptor-like factor 3	619	0.00054	3.26907
71	ENSG00000197279	ZNF165	zinc finger protein 165	613	0.00000	8.48790
72	ENSG00000215190	LINC00680	long intergenic non-protein coding RNA 680	607	0.00515	2.56532
73	ENSG00000165055	METTL2B	methyltransferase like 2B	601	0.01171	2.26651
74	ENSG00000163576	EFHB	EF-hand domain family, member B	597	0.00024	3.49470
75	ENSG00000137513	NARS2	asparaginyl-tRNA synthetase 2, mitochondrial (putative)	589	0.00598	2.51326
76	ENSG00000138829	FBN2	fibrillin 2	589	0.02792	1.91228
77	ENSG00000160321	ZNF208	zinc finger protein 208	588	0.11170	1.21754
78	ENSG00000214198	RP11-642P15.1	Unknown	584	0.00026	3.46512
79	ENSG00000111801	BTN3A3	butyrophilin, subfamily 3, member A3	581	0.00000	10.65362
80	ENSG00000198496	NBR2	neighbor of BRCA1 gene 2 (non-protein coding)	580	0.00090	3.12250
81	ENSG00000228716	DHFR	dihydrofolate reductase	576	0.04137	1.73500
82	ENSG00000214435	AS3MT	arsenite methyltransferase	573	0.00000	8.50303
83	ENSG00000117481	NSUN4	NOP2/Sun domain family, member 4	564	0.08775	1.35474
84	ENSG00000266490	CTD-2349P21.9	Unknown	556	0.00054	3.26907
85	ENSG00000154319	FAM167A	family with sequence similarity 167, member A	551	0.00000	4.54644
86	ENSG00000125804	FAM182A	family with sequence similarity 182, member A	547	0.01722	2.11488
87	ENSG00000152117	AC093838.4	Unknown	546	0.02243	2.00596
88	ENSG00000219392	RP1-265C24.5	Unknown	530	0.00000	9.96241
89	ENSG00000261556	SMG1P7	SMG1 pseudogene 7	530	0.00193	2.88987
90	ENSG00000173930	SLCO4C1	solute carrier organic anion transporter family, member 4C1	527	0.00000	4.94237
91	ENSG00000164037	SLC9B1	solute carrier family 9, subfamily B (NHA1, cation proton antiporter 1), member 1	526	0.00000	4.45148
92	ENSG00000250120	PCDHA10	protocadherin alpha 10	525	0.00000	4.92228
93	ENSG00000108592	FTSJ3	FtsJ homolog 3 (E. coli)	524	0.00227	2.83770
94	ENSG00000204267	TAP2	transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)	523	0.00000	8.34973

95	ENSG00000013573	DDX11	DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11	520	0.00591	2.51759
96	ENSG00000146530	VWDE	von Willebrand factor D and EGF domains	517	0.00230	2.83406
97	ENSG00000198874	TYW1	tRNA-yW synthesizing protein 1 homolog (S. cerevisiae)	517	0.22990	0.73918
98	ENSG00000180185	FAHD1	fumarylacetoacetate hydrolase domain containing 1	513	0.00005	3.90338
99	ENSG00000086991	NOX4	NADPH oxidase 4	510	0.03092	1.86744
100	ENSG00000159712	ANKRD18CP	ankyrin repeat domain 18C, pseudogene	506	0.20100	0.83805
101	ENSG00000168803	ADAL	adenosine deaminase-like	502	0.00381	2.66819
102	ENSG00000176927	EFCAB5	EF-hand calcium binding domain 5	501	0.00025	3.48022

7.2 Annotation of Genes Using *Ensembl*

Annotation of genes with their common names, descriptions, and possibly other information can be achieved using the *BioMart* toolbox of the databse *Ensembl*[8] [19]. The procedures are shown in Figures 7.2.1 through 7.2.4.

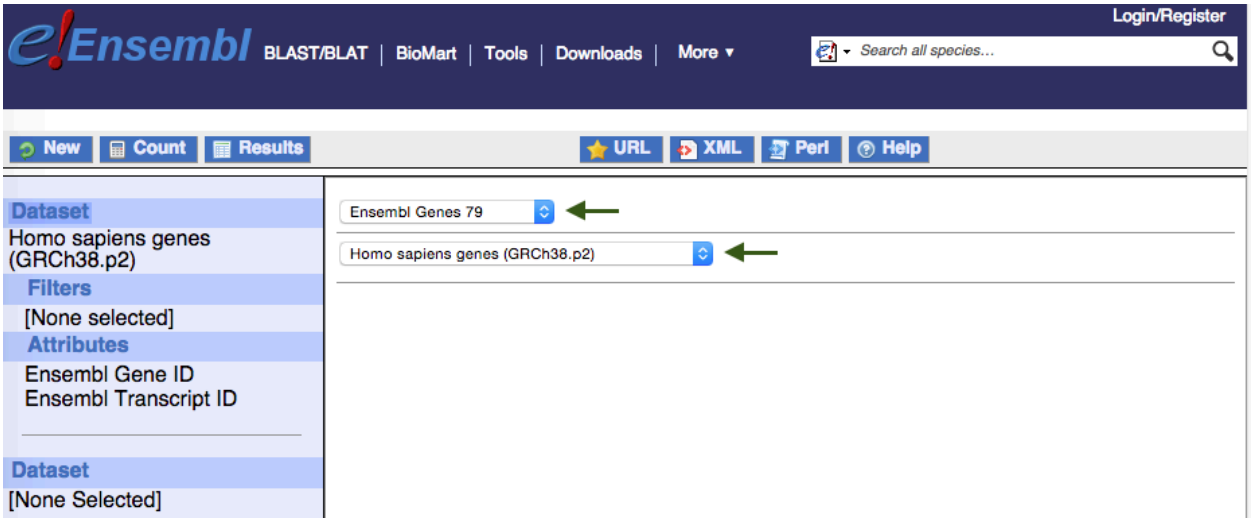


Figure 7.2.1: *Using BioMart – Step 1: Choose a Dataset.* We select the latest release of *Ensembl* (Ensembl Genes 79) and restrict the genes in the database to those of human.

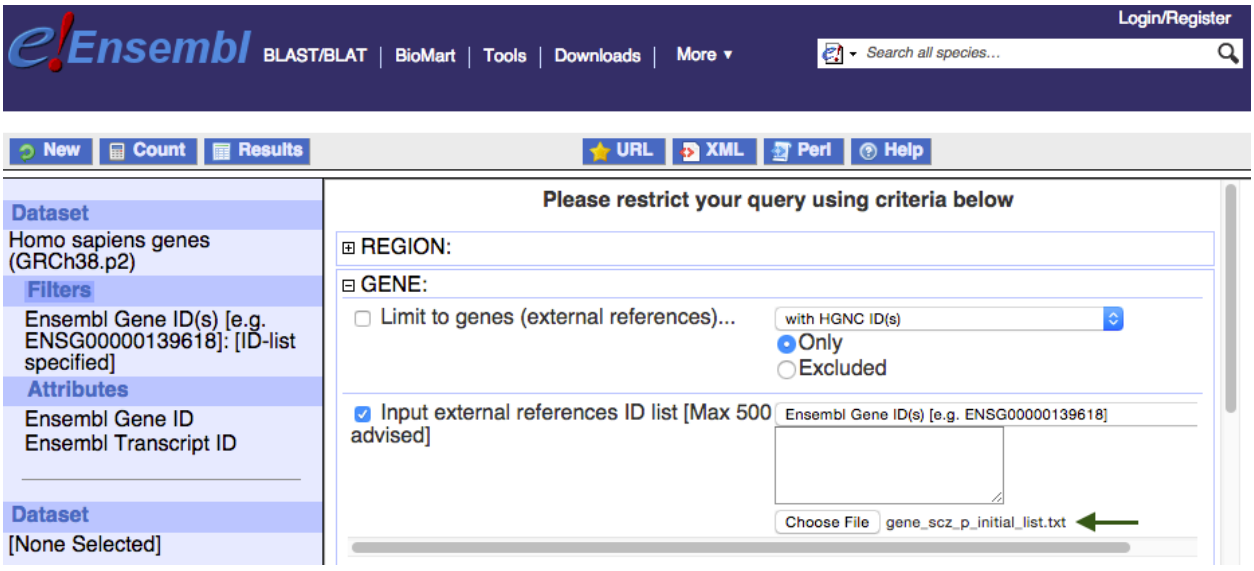


Figure 7.2.2: *Using BioMart – Step 2: Upload a List of Ensembl IDs as Filter.* We supply to *Ensembl* a text file containing a list of Ensembl IDs only. Each row of the file should be an Ensembl ID. No other characters or symbols should be included. This file may be created using the `R` function `write.table` with arguments `row.names=F, col.names=F, quote=F`.

[8] `http://www.ensembl.org/biomart/martview/`

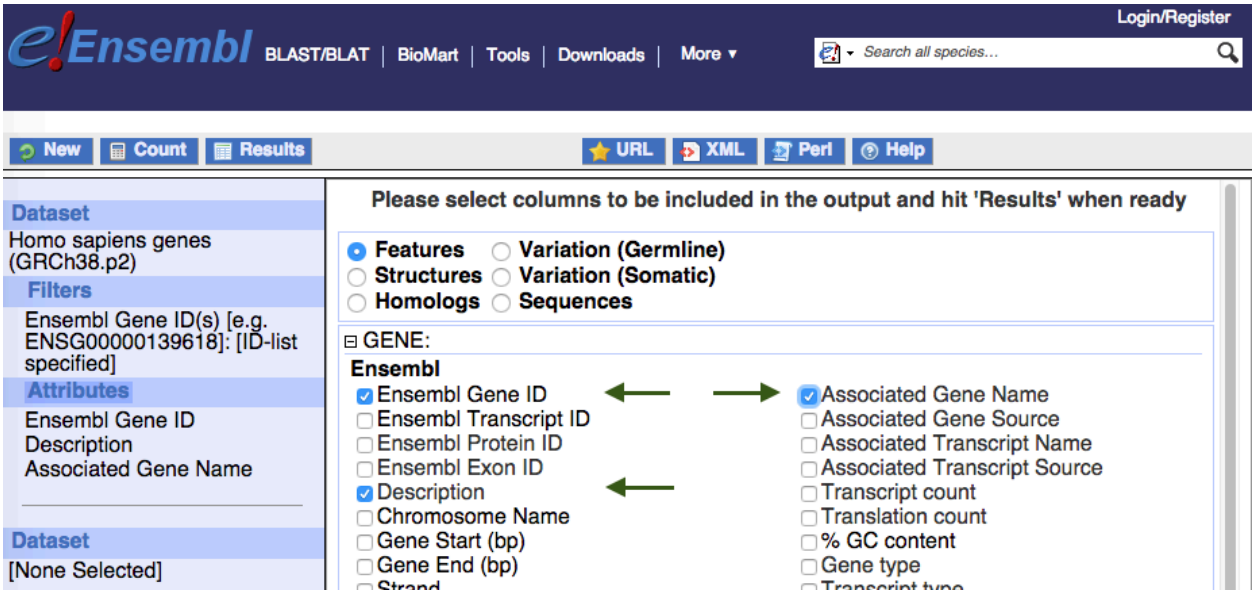


Figure 7.2.3: *Using BioMart – Step 3: Select Desired Attributes.* For our purpose, we only need the original Ensembl IDs (Ensembl Gene IDs), common names (Associated Gene Names), and descriptions of the gene functions (Descriptions) in the annotated file. Check more options if necessary.

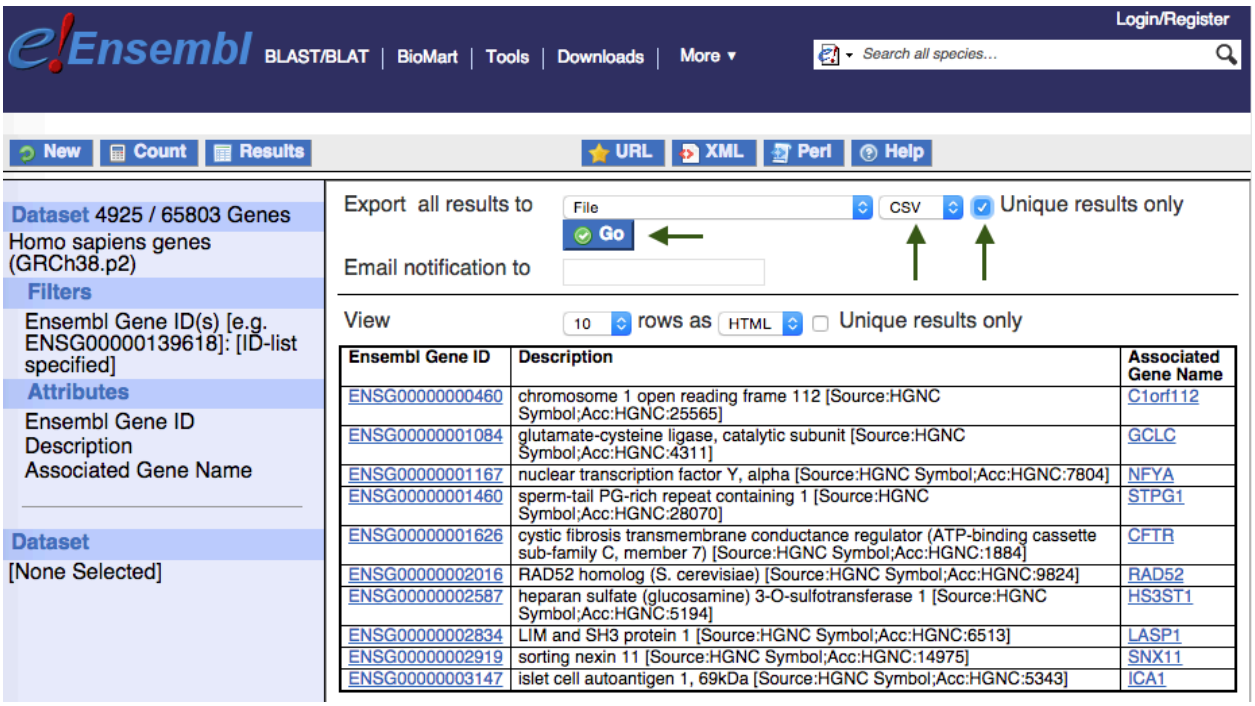


Figure 7.2.4: *Using BioMart – Step 4: Export Annotations as a* `csv` *File.* We check the box for 'Unique results only' to avoid duplicate entries in the annotation file. The resultant `csv` file may be read using the `R` function `read.csv` with arguments `header=T, stringsAsFactors=F`.

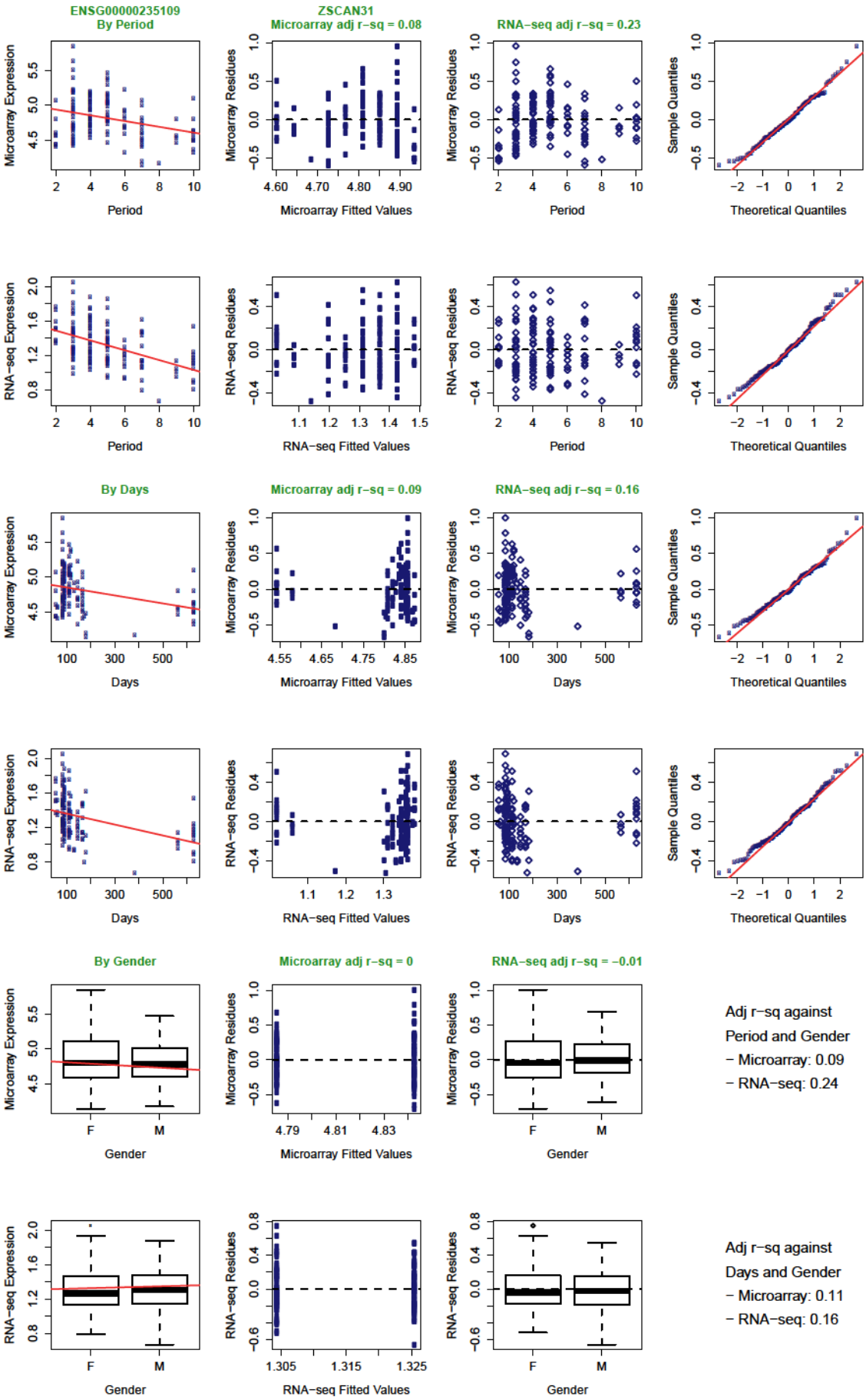
Note that not every Ensembl ID has an entry in *Ensembl*. After reading in the annotation file, Ensembl

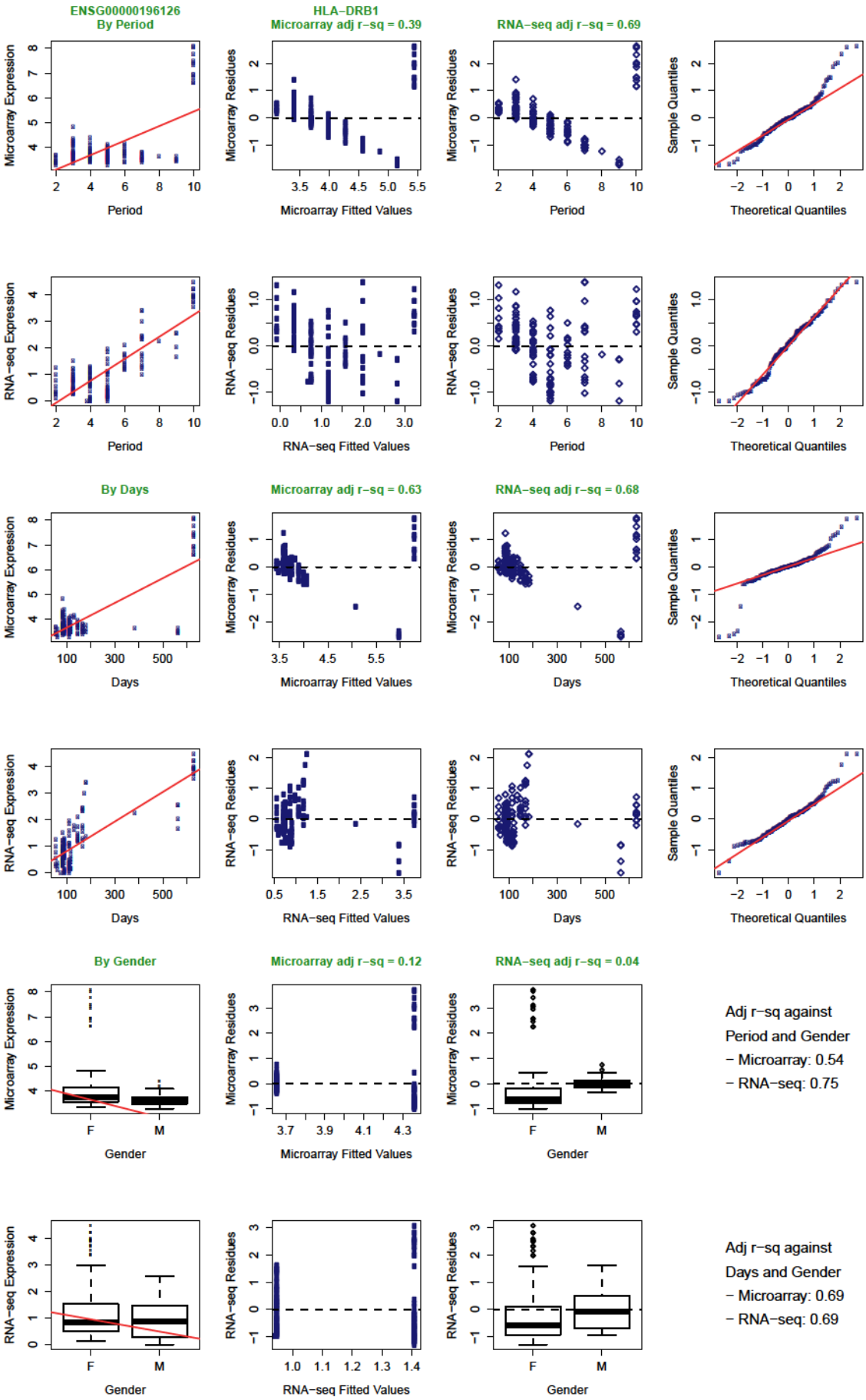
IDs for which no annotation is available can be annotated as such using relevant code in `mapping.R` (See Supplemental Information 7.9).

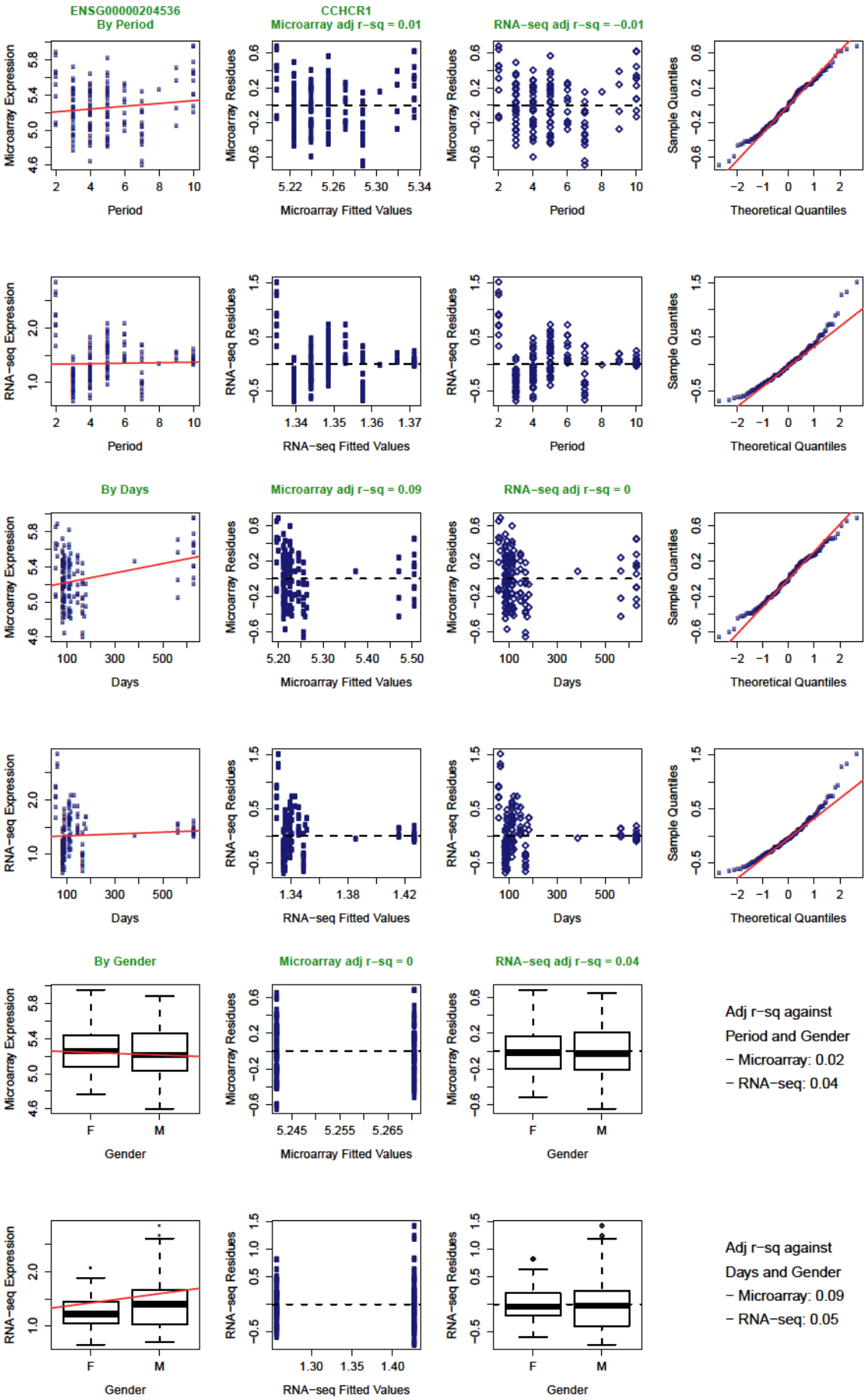
7.3 Regression Diagnostics

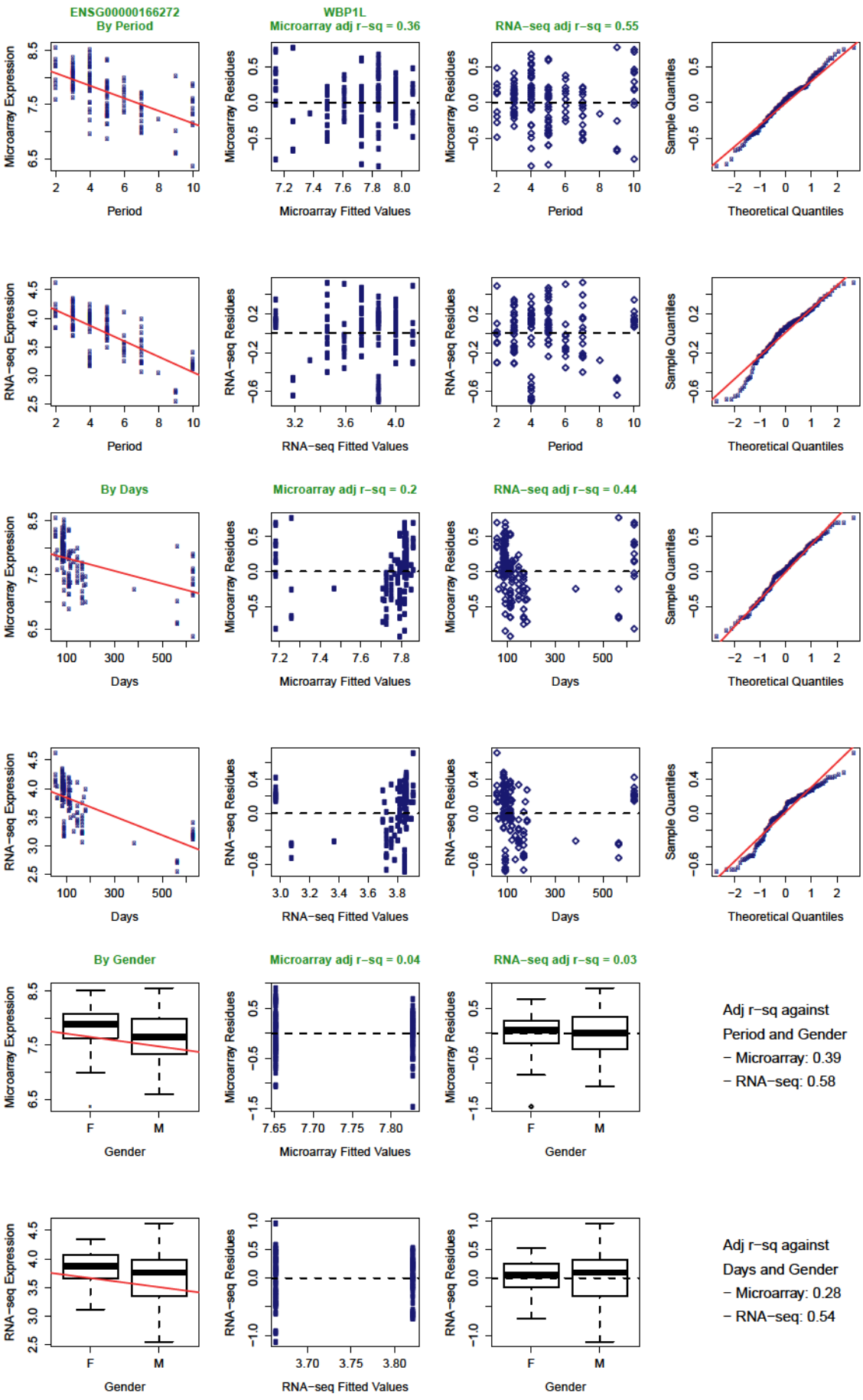
In addition to diagnostic plots (Figure 3.2.1) for *BTN3A2*, the gene with the largest genetic association score (i.e. smallest p-value) for schizophrenia, we present diagnostic plots for another 14 semi-randomly selected genes. They are:

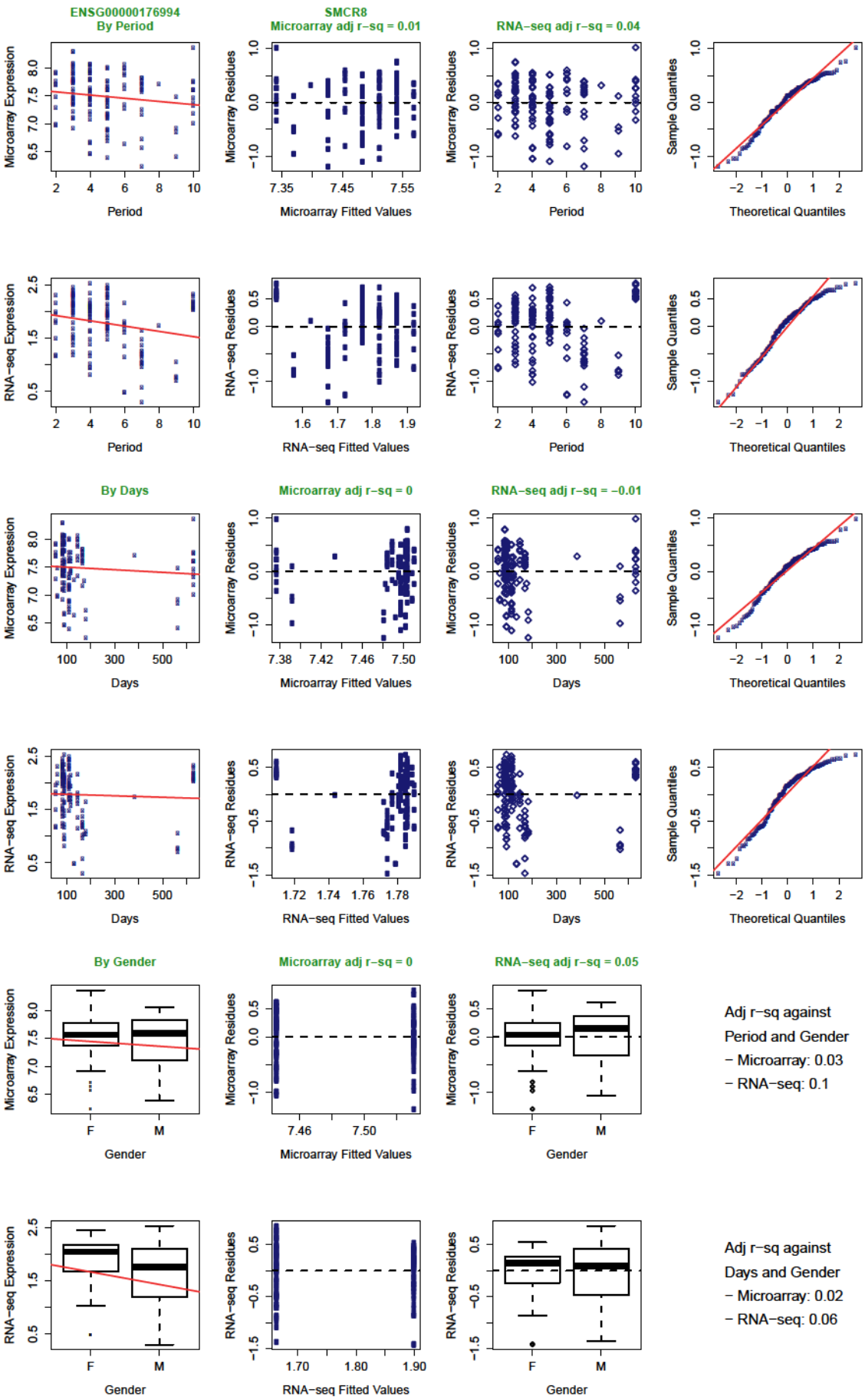
- 4 genes with the smallest p-values other than *BTN3A2*: *ZSCAN31*, *HLA-DRB1*, *CCHCR1*, and *WBP1L*;
- 5 randomly selected genes with p-values smaller than 0.01: *SMCR8*, *TRIM65*, *SOBP*, *JTB*, and *C15orf57*; and
- 5 randomly selected genes with p-values equal to or greater than 0.01: *ADAM19*, *VAT1*, *SLA*, *KLF13*, and *SPAG16*.

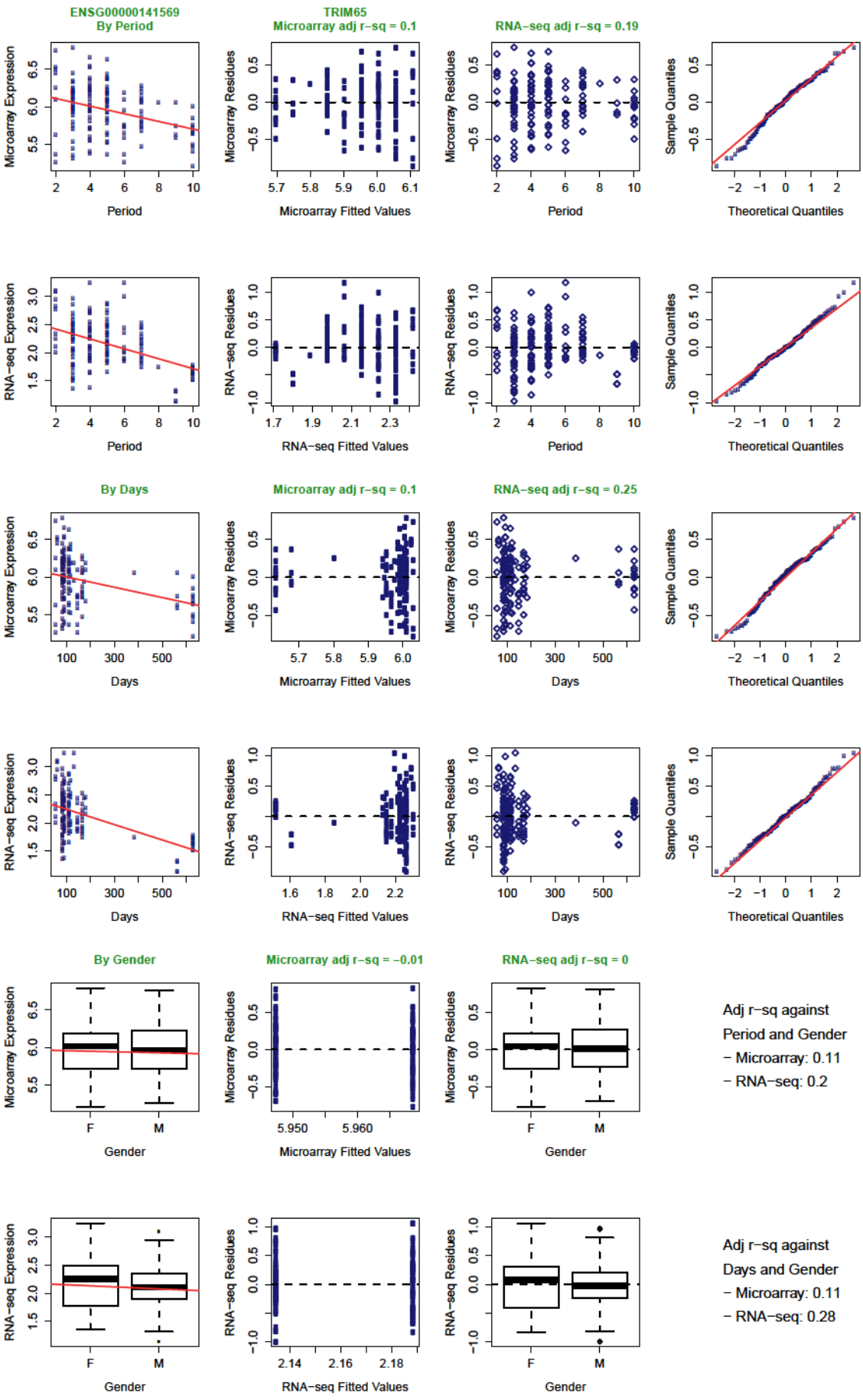

ENSG00000235109
By Period
ZSCAN31
Microarray adj r−sq = 0.08
RNA−seq adj r−sq = 0.23
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.09
RNA−seq adj r−sq = 0.16
Days
By Gender
Microarray adj r−sq = 0
RNA−seq adj r−sq = −0.01
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.09
− RNA−seq: 0.24
Adj r−sq against
Days and Gender
− Microarray: 0.11
− RNA−seq: 0.16

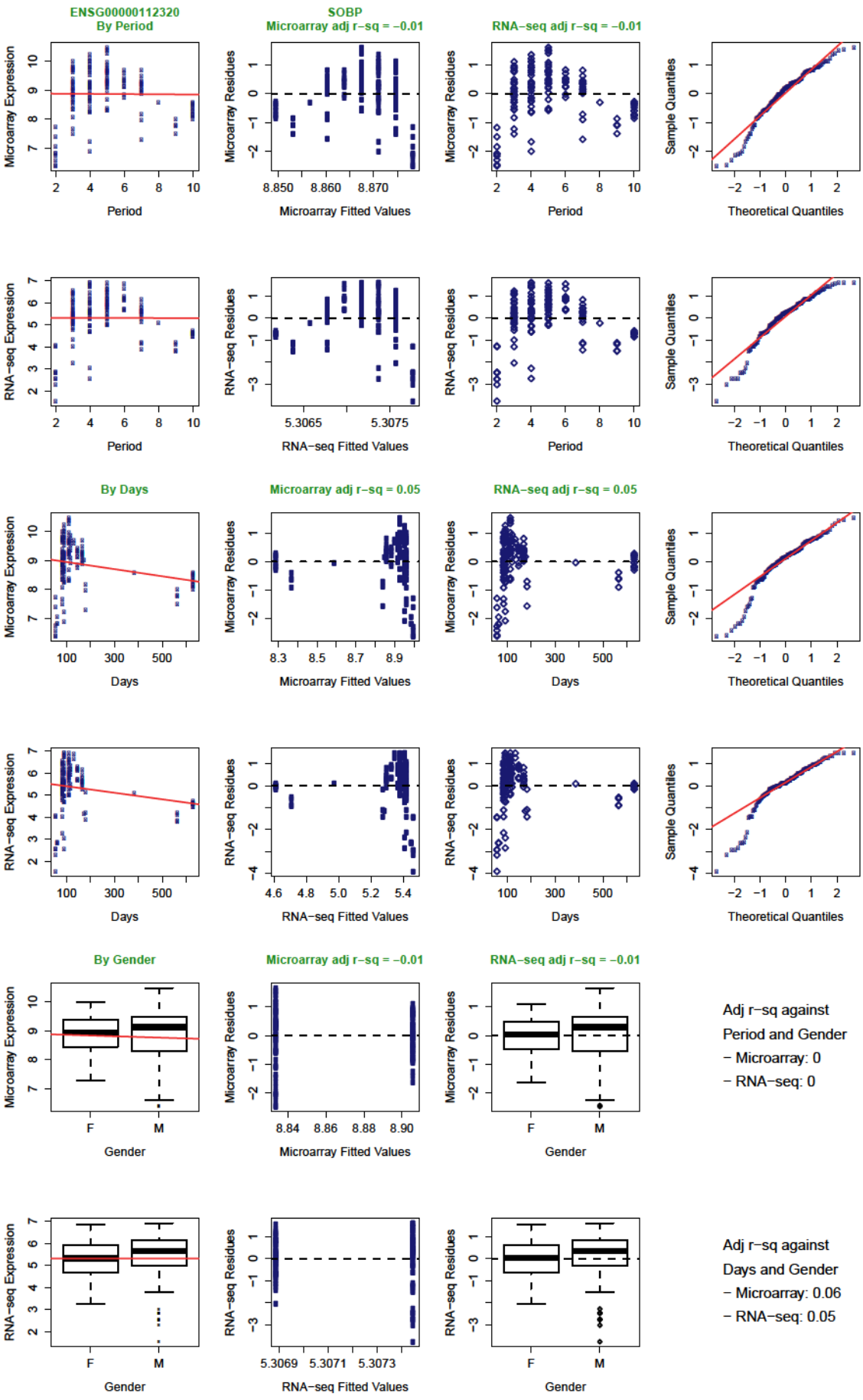

ENSG00000196126
By Period
HLA−DRB1
Microarray adj r−sq = 0.39
RNA−seq adj r−sq = 0.69
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.63
RNA−seq adj r−sq = 0.68
Days
By Gender
Microarray adj r−sq = 0.12
RNA−seq adj r−sq = 0.04
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.54
− RNA−seq: 0.75
Adj r−sq against
Days and Gender
− Microarray: 0.69
− RNA−seq: 0.69

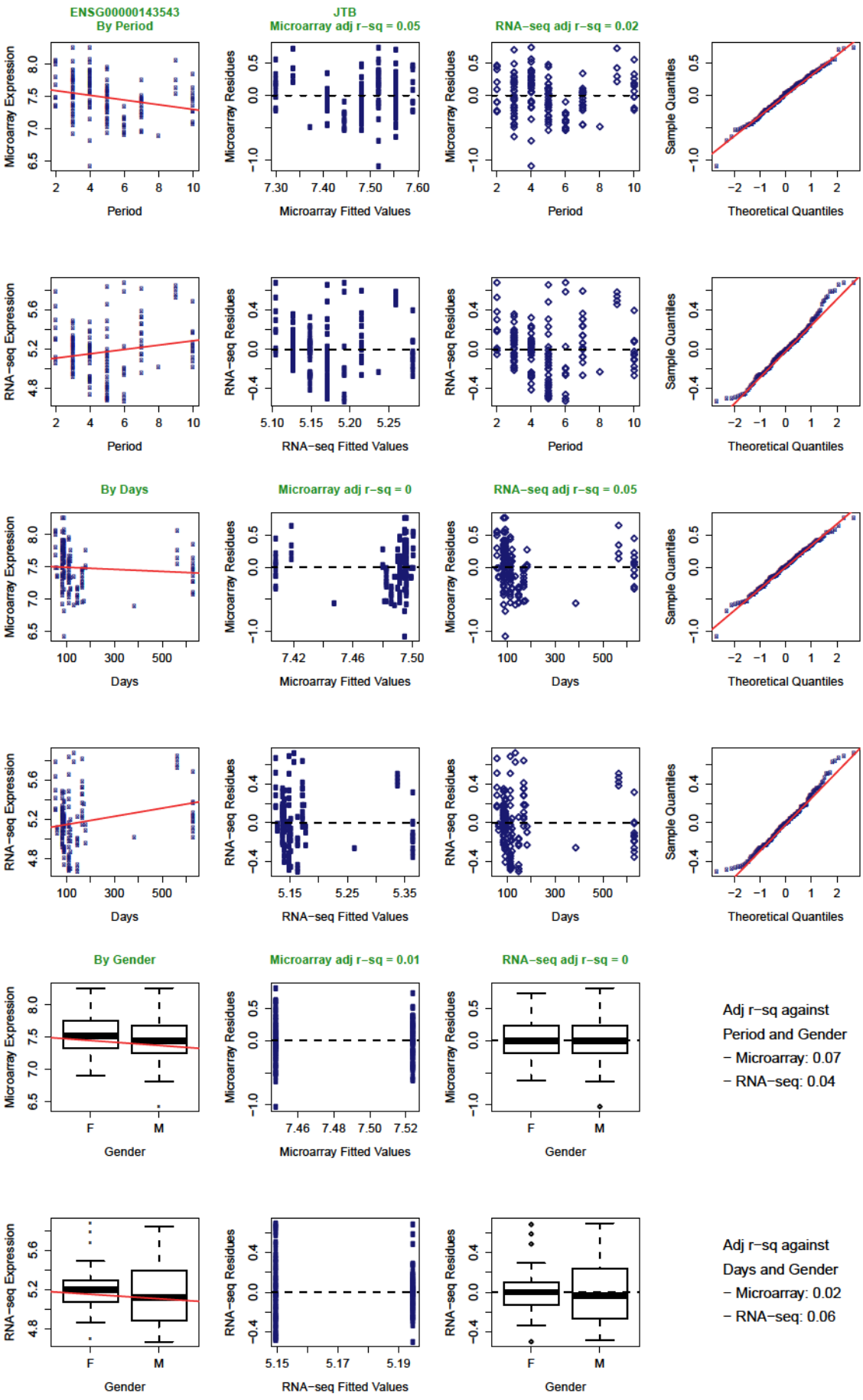

ENSG00000204536
By Period
CCHCR1
Microarray adj r−sq = 0.01
RNA−seq adj r−sq = −0.01
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.09
RNA−seq adj r−sq = 0
Days
By Gender
Microarray adj r−sq = 0
RNA−seq adj r−sq = 0.04
Gender
Adj r−sq against
Period and Gender
− Microarray: 0.02
− RNA−seq: 0.04
Adj r−sq against
Days and Gender
− Microarray: 0.09
− RNA−seq: 0.05

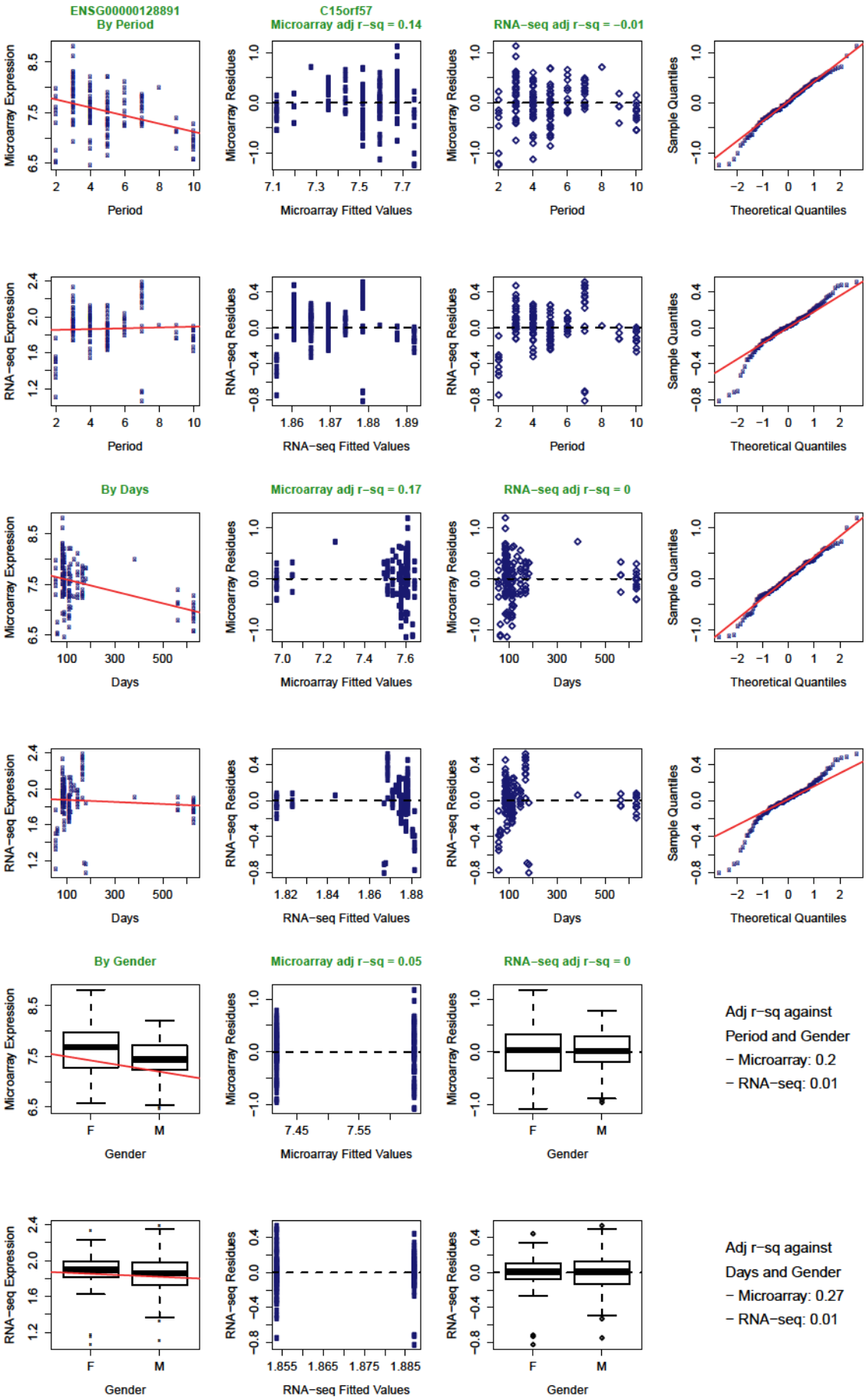

ENSG00000166272
By Period
WBP1L
Microarray adj r−sq = 0.36
RNA−seq adj r−sq = 0.55
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.2
RNA−seq adj r−sq = 0.44
Days
By Gender
Microarray adj r−sq = 0.04
RNA−seq adj r−sq = 0.03
Gender
Adj r−sq against
Period and Gender
− Microarray: 0.39
− RNA−seq: 0.58
Adj r−sq against
Days and Gender
− Microarray: 0.28
− RNA−seq: 0.54

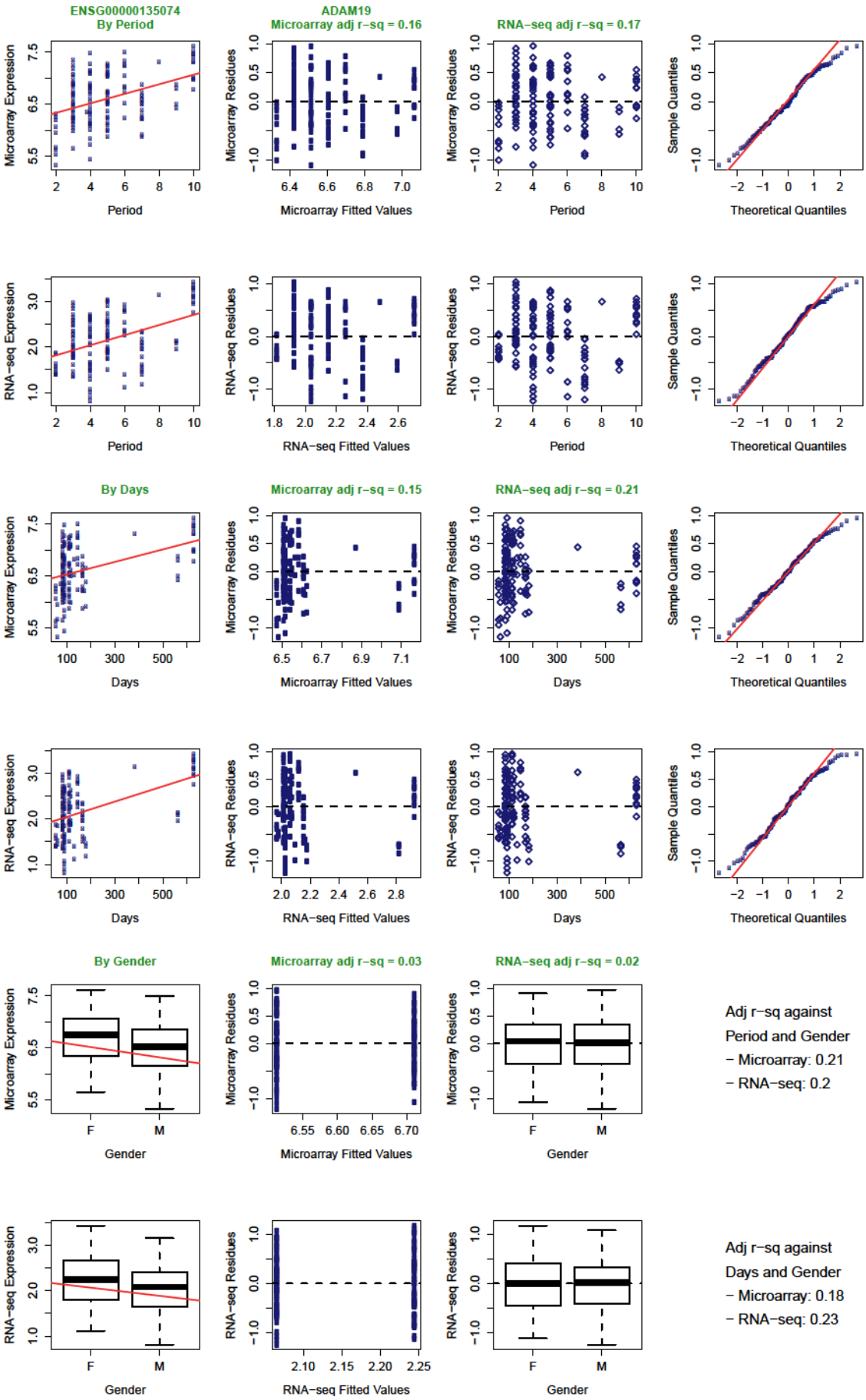

ENSG00000176994
By Period
SMCR8
Microarray adj r−sq = 0.01
RNA−seq adj r−sq = 0.04
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0
RNA−seq adj r−sq = −0.01
Days
By Gender
RNA−seq adj r−sq = 0.05
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.03
− RNA−seq: 0.1
Adj r−sq against
Days and Gender
− Microarray: 0.02
− RNA−seq: 0.06

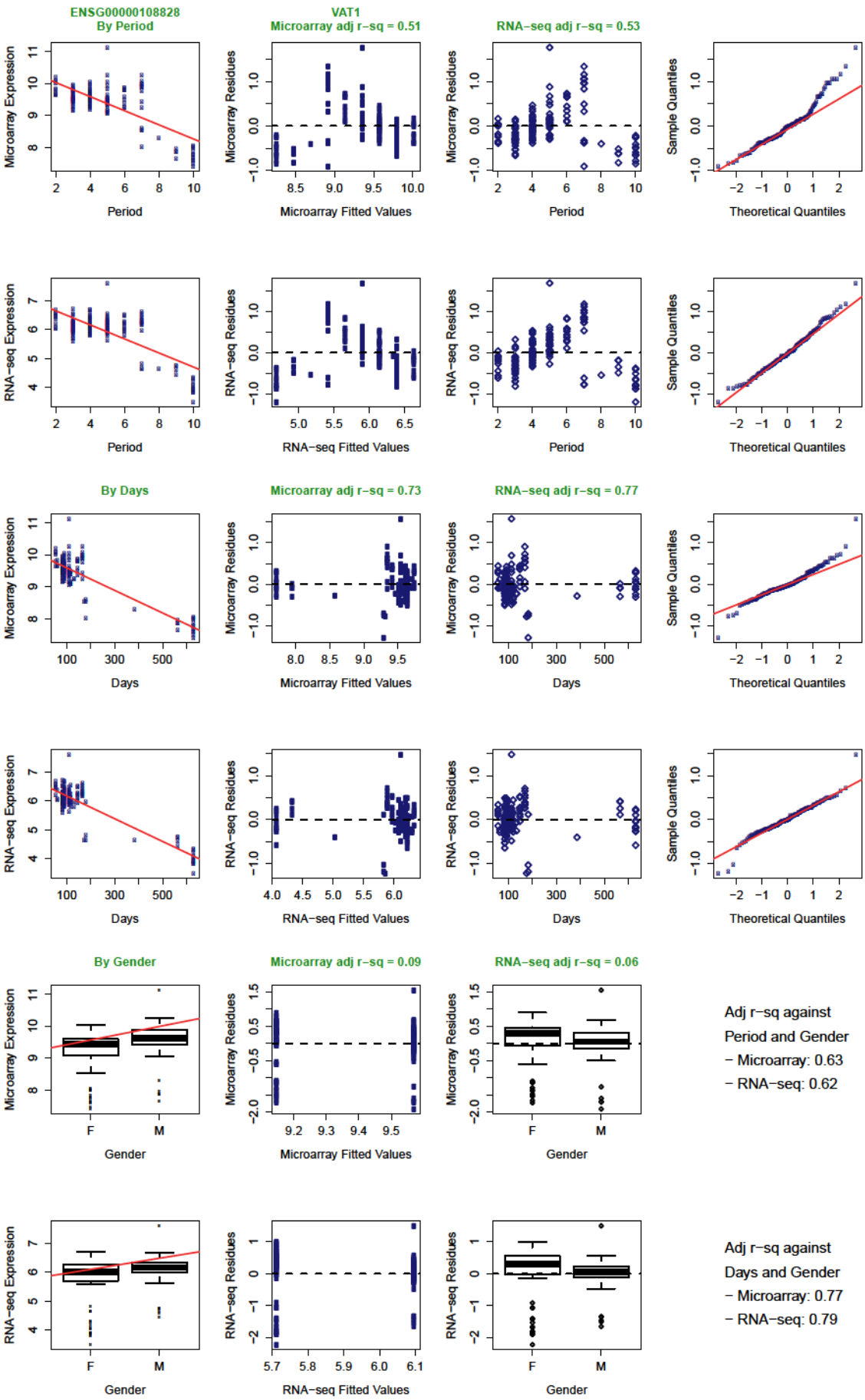

ENSG00000141569
By Period
Microarray Expression
Period
TRIM65
Microarray adj r−sq = 0.1
Microarray Residues
Microarray Fitted Values
RNA−seq adj r−sq = 0.19
Microarray Residues
Period
Sample Quantiles
Theoretical Quantiles
RNA−seq Expression
Period
RNA−seq Residues
RNA−seq Fitted Values
RNA−seq Residues
Period
Sample Quantiles
Theoretical Quantiles
By Days
Microarray Expression
Days
Microarray adj r−sq = 0.1
Microarray Residues
Microarray Fitted Values
RNA−seq adj r−sq = 0.25
Microarray Residues
Days
Sample Quantiles
Theoretical Quantiles
RNA−seq Expression
Days
RNA−seq Residues
RNA−seq Fitted Values
RNA−seq Residues
Days
Sample Quantiles
Theoretical Quantiles
By Gender
Microarray Expression
Gender
Microarray adj r−sq = −0.01
Microarray Residues
Microarray Fitted Values
RNA−seq adj r−sq = 0
Microarray Residues
Gender
Adj r−sq against
Period and Gender
− Microarray: 0.11
− RNA−seq: 0.2
RNA−seq Expression
Gender
RNA−seq Residues
RNA−seq Fitted Values
RNA−seq Residues
Gender
Adj r−sq against
Days and Gender
− Microarray: 0.11
− RNA−seq: 0.28

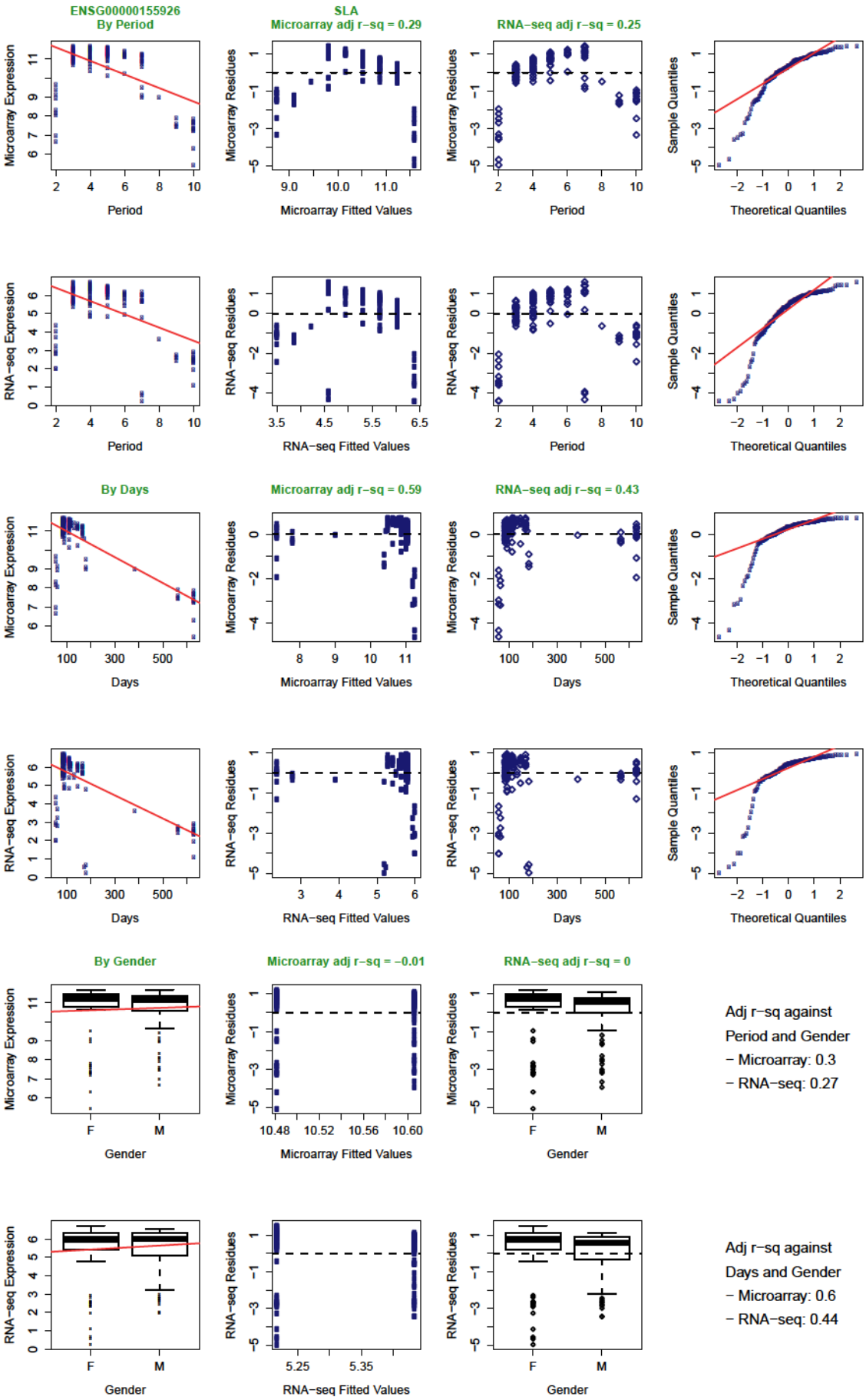

ENSG00000112320
By Period
SOBP
Microarray adj r−sq = −0.01
RNA−seq adj r−sq = −0.01
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.05
RNA−seq adj r−sq = 0.05
Days
By Gender
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0
− RNA−seq: 0
Adj r−sq against
Days and Gender
− Microarray: 0.06
− RNA−seq: 0.05

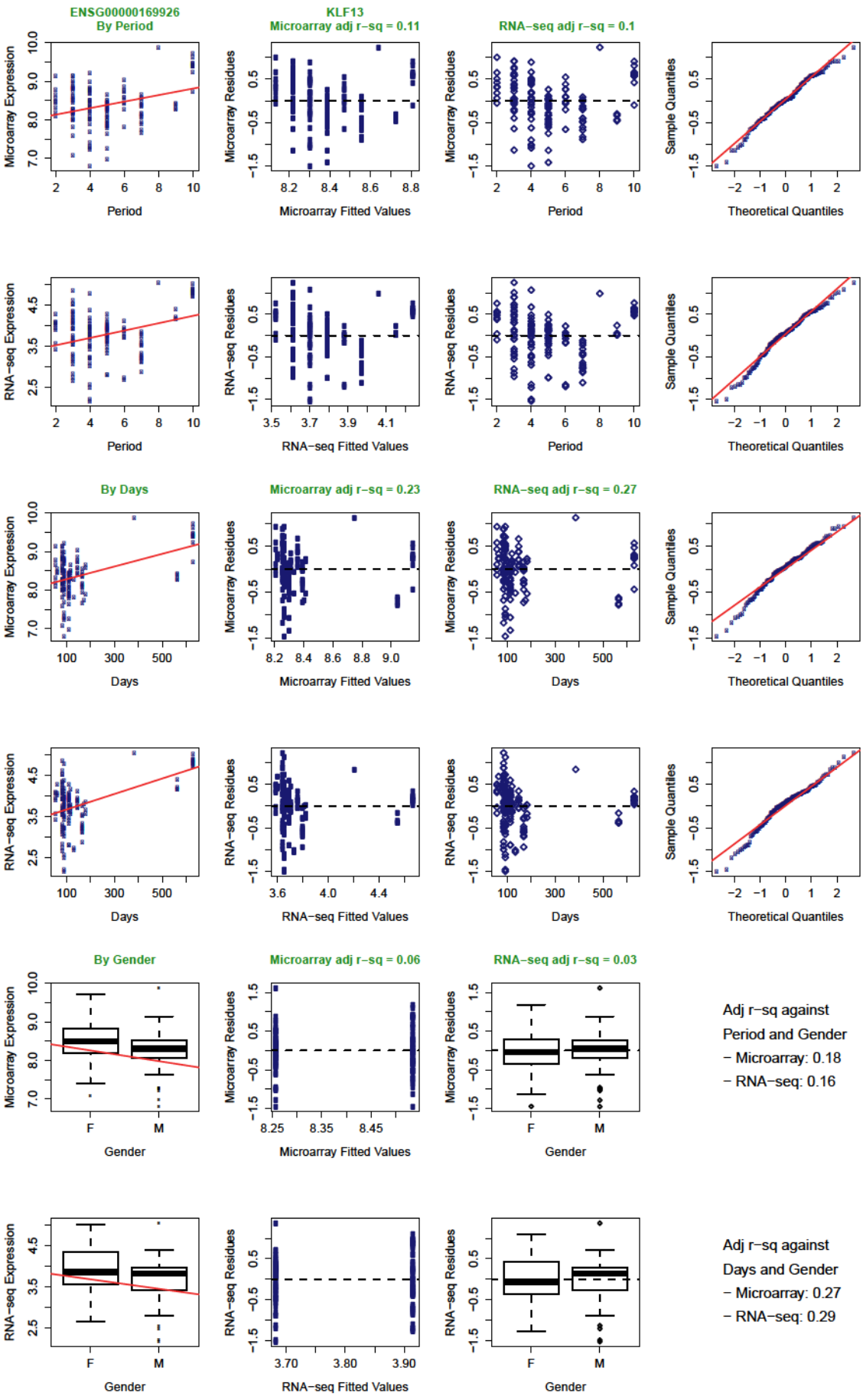

ENSG00000143543
By Period
JTB
Microarray adj r−sq = 0.05
RNA−seq adj r−sq = 0.02
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0
RNA−seq adj r−sq = 0.05
Days
By Gender
Microarray adj r−sq = 0.01
RNA−seq adj r−sq = 0
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.07
− RNA−seq: 0.04
Adj r−sq against
Days and Gender
− Microarray: 0.02
− RNA−seq: 0.06

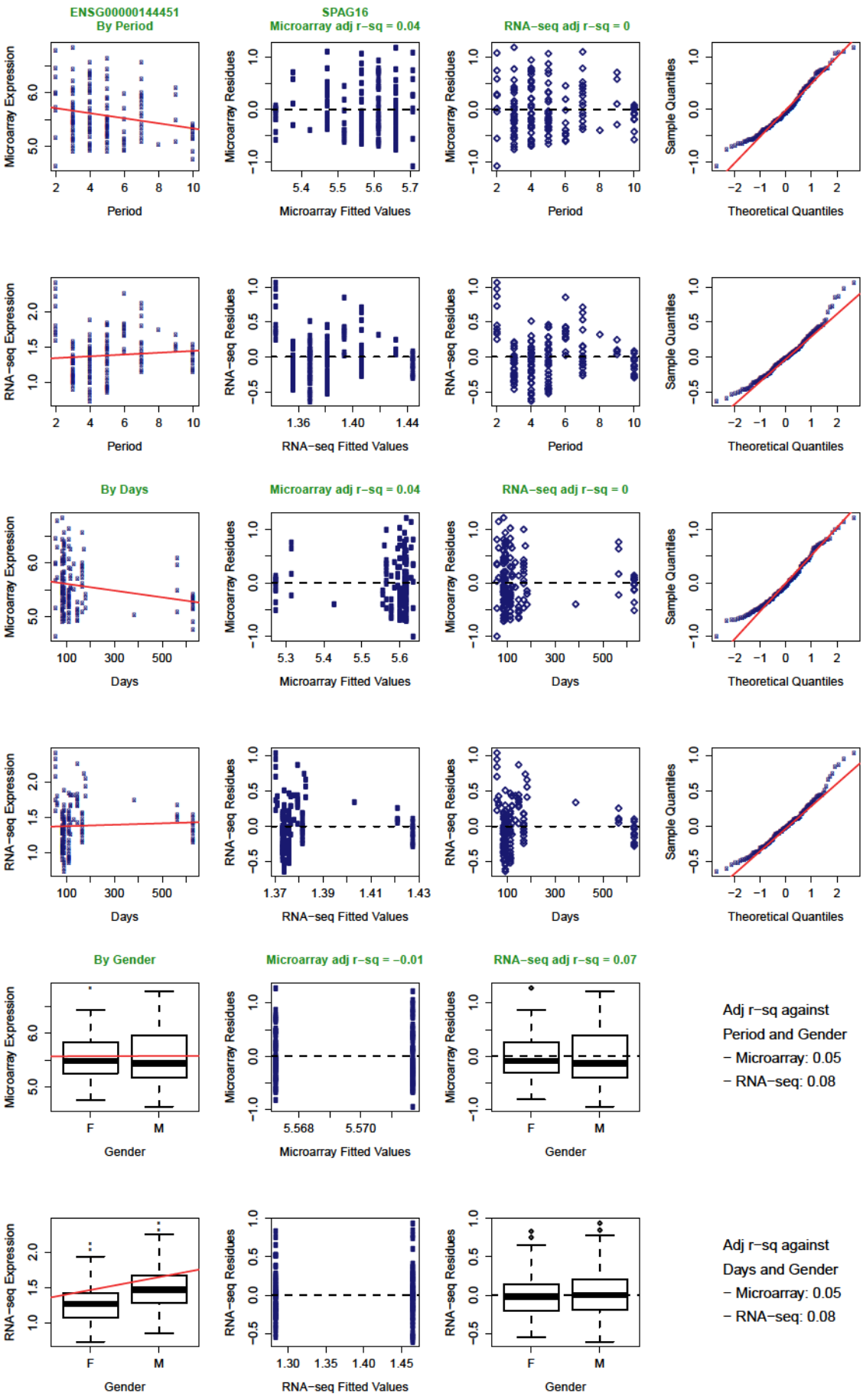

ENSG00000128891
By Period
C15orf57
Microarray adj r−sq = 0.14
RNA−seq adj r−sq = −0.01
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.17
RNA−seq adj r−sq = 0
Days
By Gender
Microarray adj r−sq = 0.05
RNA−seq adj r−sq = 0
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.2
− RNA−seq: 0.01
Adj r−sq against
Days and Gender
− Microarray: 0.27
− RNA−seq: 0.01


ENSG00000135074
By Period
ADAM19
Microarray adj r−sq = 0.16
RNA−seq adj r−sq = 0.17
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.15
RNA−seq adj r−sq = 0.21
Days
By Gender
Microarray adj r−sq = 0.03
RNA−seq adj r−sq = 0.02
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.21
− RNA−seq: 0.2
Adj r−sq against
Days and Gender
− Microarray: 0.18
− RNA−seq: 0.23


ENSG00000108828
VAT1
By Period
Microarray adj r−sq = 0.51
RNA−seq adj r−sq = 0.53
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.73
RNA−seq adj r−sq = 0.77
Days
By Gender
Microarray adj r−sq = 0.09
RNA−seq adj r−sq = 0.06
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.63
− RNA−seq: 0.62
Adj r−sq against
Days and Gender
− Microarray: 0.77
− RNA−seq: 0.79


ENSG00000155926
By Period
SLA
Microarray adj r−sq = 0.29
RNA−seq adj r−sq = 0.25
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.59
RNA−seq adj r−sq = 0.43
Days
By Gender
Microarray adj r−sq = −0.01
RNA−seq adj r−sq = 0
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.3
− RNA−seq: 0.27
Adj r−sq against
Days and Gender
− Microarray: 0.6
− RNA−seq: 0.44


ENSG00000169926
By Period
KLF13
Microarray adj r−sq = 0.11
RNA−seq adj r−sq = 0.1
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Microarray adj r−sq = 0.23
RNA−seq adj r−sq = 0.27
Days
By Gender
Microarray adj r−sq = 0.06
RNA−seq adj r−sq = 0.03
Gender
Adj r−sq against
Period and Gender
− Microarray: 0.18
− RNA−seq: 0.16
Adj r−sq against
Days and Gender
− Microarray: 0.27
− RNA−seq: 0.29


ENSG00000144451
By Period
SPAG16
Microarray adj r−sq = 0.04
RNA−seq adj r−sq = 0
Microarray Expression
Microarray Residues
Sample Quantiles
Period
Microarray Fitted Values
Theoretical Quantiles
RNA−seq Expression
RNA−seq Residues
RNA−seq Fitted Values
By Days
Days
By Gender
Microarray adj r−sq = −0.01
RNA−seq adj r−sq = 0.07
Gender
F
M
Adj r−sq against
Period and Gender
− Microarray: 0.05
− RNA−seq: 0.08
Adj r−sq against
Days and Gender
− Microarray: 0.05
− RNA−seq: 0.08

7.4 List of COP Hubs before Transformation

Table 7.4.1: Genes Involved in Most COPs on Average across Thresholds before Transformation

	Ensembl ID	Associated Name	Description	p-value	Avg #COPs
1	ENSG00000075415	SLC25A3	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3	0.10630	202
2	ENSG00000173848	NET1	neuroepithelial cell transforming 1	0.87420	188
3	ENSG00000144741	SLC25A26	solute carrier family 25 (S-adenosylmethionine carrier), member 26	0.02026	164
4	ENSG00000143882	ATP6V1C2	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C2	0.18060	147
5	ENSG00000205246	RPSAP58	ribosomal protein SA pseudogene 58	0.25290	141
6	ENSG00000178233	TMEM151B	transmembrane protein 151B	0.00009	120
7	ENSG00000188612	SUMO2	small ubiquitin-like modifier 2	0.10030	120
8	ENSG00000110092	CCND1	cyclin D1	0.20620	107
9	ENSG00000008226	DLEC1	deleted in lung and esophageal cancer 1	0.04607	106
10	ENSG00000173406	DAB1	Dab, reelin signal transducer, homolog 1 (Drosophila)	0.67540	101

7.5 List of COP Genes for 'Fixing' via Transformation

Table 7.5.1: Genes to be 'Fixed' in Microarray Data via Transformation

	Ensembl ID	Associated Name	# COPs	Fraction	p-value	Description
1	ENSG00000075415	SLC25A3	109	0.037	0.10630	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3
2	ENSG00000144741	SLC25A26	99	0.033	0.02026	solute carrier family 25 (S-adenosylmethionine carrier), member 26
3	ENSG00000173848	NET1	95	0.032	0.87420	neuroepithelial cell transforming 1
4	ENSG00000143882	ATP6V1C2	65	0.022	0.18060	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C2
5	ENSG00000205246	RPSAP58	62	0.021	0.25290	ribosomal protein SA pseudogene 58
6	ENSG00000188612	SUMO2	56	0.019	0.10030	small ubiquitin-like modifier 2
7	ENSG00000178233	TMEM151B	46	0.015	0.00009	transmembrane protein 151B
8	ENSG00000008226	DLEC1	35	0.012	0.04607	deleted in lung and esophageal cancer 1
9	ENSG00000110092	CCND1	34	0.011	0.20620	cyclin D1
10	ENSG00000213199	ASIC3	28	0.009	0.32670	acid sensing (proton gated) ion channel 3

11	ENSG00000173406	DAB1	27	0.009	0.67540	Dab, reelin signal transducer, homolog 1 (Drosophila)
12	ENSG00000171530	TBCA	26	0.009	0.57790	tubulin folding cofactor A
13	ENSG00000083457	ITGAE	23	0.008	0.21320	integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)
14	ENSG00000005882	PDK2	22	0.007	0.09860	pyruvate dehydrogenase kinase, isozyme 2
15	ENSG00000164587	RPS14	20	0.007	0.08469	ribosomal protein S14
16	ENSG00000075826	SEC31B	18	0.006	0.09235	SEC31 homolog B (S. cerevisiae)
17	ENSG00000122033	MTIF3	16	0.005	0.02986	mitochondrial translational initiation factor 3
18	ENSG00000154556	SORBS2	15	0.005	0.06234	sorbin and SH3 domain containing 2
19	ENSG00000157764	BRAF	14	0.005	0.22110	B-Raf proto-oncogene, serine/threonine kinase
20	ENSG00000204366	ZBTB12	14	0.005	0.01791	zinc finger and BTB domain containing 12
21	ENSG00000012822	CALCOCO1	13	0.004	0.14440	calcium binding and coiled-coil domain 1
22	ENSG00000204301	NOTCH4	13	0.004	0.67870	notch 4
23	ENSG00000205758	CRYZL1	13	0.004	0.56880	crystallin, zeta (quinone reductase)-like 1
24	ENSG00000122484	RPAP2	12	0.004	0.12760	RNA polymerase II associated protein 2
25	ENSG00000104728	ARHGEF10	11	0.004	0.21490	Rho guanine nucleotide exchange factor (GEF) 10
26	ENSG00000138385	SSB	11	0.004	0.96180	Sjogren syndrome antigen B (autoantigen La)
27	ENSG00000197417	SHPK	11	0.004	0.01575	sedoheptulokinase
28	ENSG00000070269	TMEM260	9	0.003	0.18020	transmembrane protein 260
29	ENSG00000101473	ACOT8	9	0.003	0.84160	acyl-CoA thioesterase 8
30	ENSG00000154099	DNAAF1	9	0.003	0.20460	dynein, axonemal, assembly factor 1
31	ENSG00000160551	TAOK1	9	0.003	0.75440	TAO kinase 1
32	ENSG00000091009	RBM27	8	0.003	0.91360	RNA binding motif protein 27
33	ENSG00000100632	ERH	8	0.003	0.92620	enhancer of rudimentary homolog (Drosophila)
34	ENSG00000101639	CEP192	8	0.003	0.18820	centrosomal protein 192kDa
35	ENSG00000120314	WDR55	8	0.003	0.00000	WD repeat domain 55
36	ENSG00000132507	EIF5A	8	0.003	0.36200	eukaryotic translation initiation factor 5A
37	ENSG00000138430	OLA1	8	0.003	0.00027	Obg-like ATPase 1
38	ENSG00000164039	BDH2	8	0.003	0.00809	3-hydroxybutyrate dehydrogenase, type 2
39	ENSG00000171928	TVP23B	8	0.003	0.00229	trans-golgi network vesicle protein 23 homolog B (S. cerevisiae)
40	ENSG00000196715	VKORC1L1	8	0.003	0.69310	vitamin K epoxide reductase complex, subunit 1-like 1
41	ENSG00000116171	SCP2	7	0.002	0.17110	sterol carrier protein 2
42	ENSG00000161010	C5orf45	7	0.002	0.85500	chromosome 5 open reading frame 45
43	ENSG00000165660	FAM175B	7	0.002	0.32490	family with sequence similarity 175, member B

44	ENSG00000074181	NOTCH3	6	0.002	0.51180	notch 3
45	ENSG00000078304	PPP2R5C	6	0.002	0.26430	protein phosphatase 2, regulatory subunit B', gamma
46	ENSG00000127022	CANX	6	0.002	0.01363	calnexin
47	ENSG00000133739	LRRCC1	6	0.002	0.22340	leucine rich repeat and coiled-coil centrosomal protein 1
48	ENSG00000154134	ROBO3	6	0.002	0.05006	roundabout, axon guidance receptor, homolog 3 (Drosophila)
49	ENSG00000159884	CCDC107	6	0.002	0.39380	coiled-coil domain containing 107
50	ENSG00000163933	RFT1	6	0.002	0.32330	RFT1 homolog (S. cerevisiae)

7.6 List of COP Hubs after Transformation

Table 7.6.1: Genes Involved in Most COPs on Average across Thresholds after Transformation

	Ensembl ID	Associated Name	Description	p-value	Avg #COPs
1	ENSG00000237984	PTENP1	phosphatase and tensin homolog pseudogene 1 (functional)	0.39090	51
2	ENSG00000106610	STAG3L4	stromal antigen 3-like 4 (pseudogene)	0.00100	44
3	ENSG00000130876	SLC7A10	solute carrier family 7 (neutral amino acid transporter light chain, asc system), member 10	0.38950	44
4	ENSG00000215908	CROCCP2	ciliary rootlet coiled-coil, rootletin pseudogene 2	0.02624	44
5	ENSG00000168818	STX18	syntaxin 18	0.87590	42
6	ENSG00000106682	EIF4H	eukaryotic translation initiation factor 4H	0.22450	40
7	ENSG00000197785	ATAD3A	ATPase family, AAA domain containing 3A	0.60220	39
8	ENSG00000184313	MROH7	maestro heat-like repeat family member 7	0.25480	22
9	ENSG00000109536	FRG1	FSHD region gene 1	0.64450	19
10	ENSG00000157426	AASDH	aminoadipate-semialdehyde dehydrogenase	0.06103	17
11	ENSG00000144589	STK11IP	serine/threonine kinase 11 interacting protein	0.90580	16
12	ENSG00000146828	SLC12A9	solute carrier family 12, member 9	0.02037	16
13	ENSG00000239857	GET4	golgi to ER traffic protein 4 homolog (S. cerevisiae)	0.21970	16

7.7 List of Primary and Secondary Risk Genes

Table 7.7.1: Potential Primary and Secondary Risk Genes in Schizophrenia Gene Co-expression Network

	Associated Name	Description	Genetics-based p-value	FDR-ctrl. Posterior Probability	Risk Gene Neighbors (#)	Global Neighbors (#)	Risk Gene Neighbors (Fraction)	Type	Fixed Hidden State?	Seed Gene?
1	MNT	MAX network transcriptional repressor	0.06100	2.113E-08	5	5	1.000	Primary		
2	CHERP	calcium homeostasis endoplasmic reticulum protein	0.06342	1.303E-11	4	7	0.571	Primary		
3	TAF1C	TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa	0.03001	1.050E-08	3	5	0.600	Primary		
4	PRKD2	protein kinase D2	0.01235	4.805E-09	3	6	0.500	Primary		
5	ZMIZ1	zinc finger, MIZ-type containing 1	0.00976	5.264E-14	5	9	0.556	Primary		
6	RAI1	retinoic acid induced 1	0.00000	9.305E-08	1	3	0.333	Primary	Yes	Yes
7	LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein	0.02409	8.612E-09	3	5	0.600	Primary		
8	PCCB	propionyl CoA carboxylase, beta polypeptide	0.00000	1.147E-08	1	5	0.200	Primary	Yes	Yes
9	CRYZ	crystallin, zeta (quinone reductase)	0.04943	9.346E-12	4	7	0.571	Primary		
10	QRICH2	glutamine rich 2	0.02388	7.787E-09	5	5	1.000	Primary		
11	TSPAN2	tetraspanin 2	0.02686	1.795E-10	4	6	0.667	Primary		
12	CKAP2	cytoskeleton associated protein 2	0.00129	2.943E-11	3	6	0.500	Primary		
13	TUBB2A	tubulin, beta 2A class IIa	0.00493	1.116E-07	2	4	0.500	Primary		
14	ATAT1	alpha tubulin acetyltransferase 1	0.00000	2.187E-11	2	6	0.333	Primary	Yes	Yes
15	PRPSAP2	phosphoribosyl pyrophosphate synthetase-associated protein 2	0.00006	2.255E-08	2	4	0.500	Primary		
16	PEX19	peroxisomal biogenesis factor 19	0.05562	1.114E-11	4	8	0.500	Primary		
17	MAPK7	mitogen-activated protein kinase 7	0.00000	1.242E-07	2	3	0.667	Primary	Yes	Yes
18	SRR	serine racemase	0.00000	2.775E-08	1	3	0.333	Primary	Yes	Yes
19	PGM2	phosphoglucomutase 2	0.00060	1.229E-09	4	5	0.800	Primary		
20	C11orf80	chromosome 11 open reading frame 80	0.00220	7.038E-08	2	4	0.500	Primary		
21	CCDC57	coiled-coil domain containing 57	0.07278	1.885E-13	4	8	0.500	Primary		
22	KRBA2	KRAB-A domain containing 2	0.00062	1.351E-09	4	6	0.667	Primary		
23	BCL9L	B-cell CLL/lymphoma 9-like	0.09670	6.312E-10	5	6	0.833	Primary		
24	SEPT10	septin 10	0.00000	2.058E-09	1	4	0.250	Primary	Yes	Yes
25	ZFP69	ZFP69 zinc finger protein	0.00333	1.013E-07	3	4	0.750	Primary		
26	FAM114A1	family with sequence similarity 114, member A1	0.03814	2.624E-10	3	6	0.500	Primary		
27	BRD2	bromodomain containing 2	0.00000	7.058E-10	2	4	0.500	Primary	Yes	Yes
28	DDAH2	dimethylarginine dimethylaminohydrolase 2	0.00000	1.129E-15	1	8	0.125	Primary	Yes	Yes
29	ARFGAP3	ADP-ribosylation factor GTPase activating protein 3	0.00639	2.751E-09	4	5	0.800	Primary		
30	CRMP1	collapsin response mediator protein 1	0.37980	2.377E-14	5	9	0.556	Primary		
31	NLGN2	neuroligin 2	0.35080	1.163E-12	4	8	0.500	Primary		
32	ENSG00000105663	Unknown	0.15490	3.897E-13	6	8	0.750	Primary		
33	KLF11	Kruppel-like factor 11	0.30700	8.474E-08	3	5	0.600	Primary		
34	PTPN23	protein tyrosine phosphatase, non-receptor type 23	0.24070	1.702E-09	4	6	0.667	Primary		
35	ITGA6	integrin, alpha 6	0.27210	5.674E-08	3	5	0.600	Primary		
36	RNF219	ring finger protein 219	0.14960	9.938E-10	3	6	0.500	Primary		
37	ARHGAP33	Rho GTPase activating protein 33	0.15490	1.104E-09	5	6	0.833	Primary		
38	FKBP10	FK506 binding protein 10, 65 kDa	0.68650	1.340E-08	3	6	0.500	Primary		

39	MRC2	mannose receptor, C type 2	0.13390	8.835E-10	4	6	0.667	Primary	
40	CENPQ	centromere protein Q	0.00317	2.333E-09	2	5	0.400	Secondary	
41	SPA17	sperm autoantigenic protein 17	0.00002	2.208E-10	1	6	0.167	Secondary	Yes
42	NKAIN1	Na+/K+ transporting ATPase interacting 1	0.00019	7.833E-10	2	5	0.400	Secondary	
43	MICALL1	MICAL-like 1	0.02546	5.149E-12	3	7	0.429	Secondary	
44	NINL	ninein-like	0.00516	0.000E+00	5	12	0.417	Secondary	
45	ANKMY2	ankyrin repeat and MYND domain containing 2	0.07881	5.534E-10	2	8	0.250	Secondary	
46	ENSG00000108292	Unknown	0.00134	1.873E-09	1	5	0.200	Secondary	
47	DUSP16	dual specificity phosphatase 16	0.00016	3.363E-08	1	5	0.200	Secondary	
48	SOBP	sine oculis binding protein homolog (Drosophila)	0.00453	7.252E-11	2	7	0.286	Secondary	
49	COMMD2	COMM domain containing 2	0.02992	9.572E-09	1	6	0.167	Secondary	
50	RPL22	ribosomal protein L22	0.03012	7.585E-12	2	7	0.286	Secondary	
51	PILRB	paired immunoglobin-like type 2 receptor beta	0.03325	0.000E+00	3	11	0.273	Secondary	
52	SIN3B	SIN3 transcription regulator family member B	0.08825	1.553E-11	2	7	0.286	Secondary	
53	IVD	isovaleryl-CoA dehydrogenase	0.01107	4.227E-09	1	5	0.200	Secondary	
54	NDUFA2	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa	0.00154	3.717E-11	1	9	0.111	Secondary	
55	SMC2	structural maintenance of chromosomes 2	0.04768	1.577E-08	1	5	0.200	Secondary	
56	ENSG00000141140	Unknown	0.00101	5.028E-08	1	4	0.250	Secondary	
57	PCTP	phosphatidylcholine transfer protein	0.05615	3.634E-10	1	6	0.167	Secondary	
58	GIGYF1	GRB10 interacting GYF protein 1	0.02037	2.576E-12	1	7	0.143	Secondary	
59	GPD1L	glycerol-3-phosphate dehydrogenase 1-like	0.00001	6.394E-12	1	6	0.167	Secondary	Yes
60	ATAD2	ATPase family, AAA domain containing 2	0.09316	2.969E-08	1	7	0.143	Secondary	
61	DPYSL5	dihydropyrimidinase-like 5	0.00034	0.000E+00	4	10	0.400	Secondary	
62	PCNT	pericentrin	0.00131	4.163E-16	3	11	0.273	Secondary	
63	APOA1BP	apolipoprotein A-I binding protein	0.01591	5.480E-09	1	6	0.167	Secondary	
64	MTMR10	myotubularin related protein 10	0.00787	3.216E-09	1	6	0.167	Secondary	
65	LMBRD1	LMBR1 domain containing 1	0.00005	1.834E-08	1	5	0.200	Secondary	
66	SEPT2	septin 2	0.01460	1.449E-10	2	6	0.333	Secondary	
67	NMD3	NMD3 ribosome export adaptor	0.00199	6.343E-08	1	6	0.167	Secondary	
68	KDELC2	KDEL (Lys-Asp-Glu-Leu) containing 2	0.04949	1.701E-08	2	5	0.400	Secondary	
69	GPATCH8	G patch domain containing 8	0.01116	1.226E-10	2	6	0.333	Secondary	
70	MAPT	microtubule-associated protein tau	0.00114	6.732E-13	2	7	0.286	Secondary	
71	GM2A	GM2 ganglioside activator	0.07415	2.418E-08	1	6	0.167	Secondary	
72	ECI2	enoyl-CoA delta isomerase 2	0.00050	4.209E-08	1	5	0.200	Secondary	
73	L3MBTL3	l(3)mbt-like 3 (Drosophila)	0.07808	4.883E-10	2	6	0.333	Secondary	
74	TRIM13	tripartite motif containing 13	0.02148	3.854E-12	3	7	0.429	Secondary	
75	EMP2	epithelial membrane protein 2	0.04659	3.098E-10	3	7	0.429	Secondary	
76	KCTD7	potassium channel tetramerization domain containing 7	0.08272	2.598E-08	3	7	0.429	Secondary	
77	MDM4	MDM4, p53 regulator	0.10800	3.578E-08	2	5	0.400	Secondary	
78	ANXA11	annexin A11	0.23360	1.526E-09	2	6	0.333	Secondary	
79	IL6ST	interleukin 6 signal transducer	0.76910	1.972E-08	2	6	0.333	Secondary	
80	SLA	Src-like-adaptor	0.40050	8.528E-11	1	7	0.143	Secondary	
81	HN1	hematological and neurological expressed 1	0.48060	1.033E-10	2	7	0.286	Secondary	
82	TUBB2B	tubulin, beta 2B class IIb	0.31210	4.710E-11	3	7	0.429	Secondary	
83	LGALS8	lectin, galactoside-binding, soluble, 8	0.16750	4.552E-08	2	5	0.400	Secondary	

7.8 Complete DAWN Network

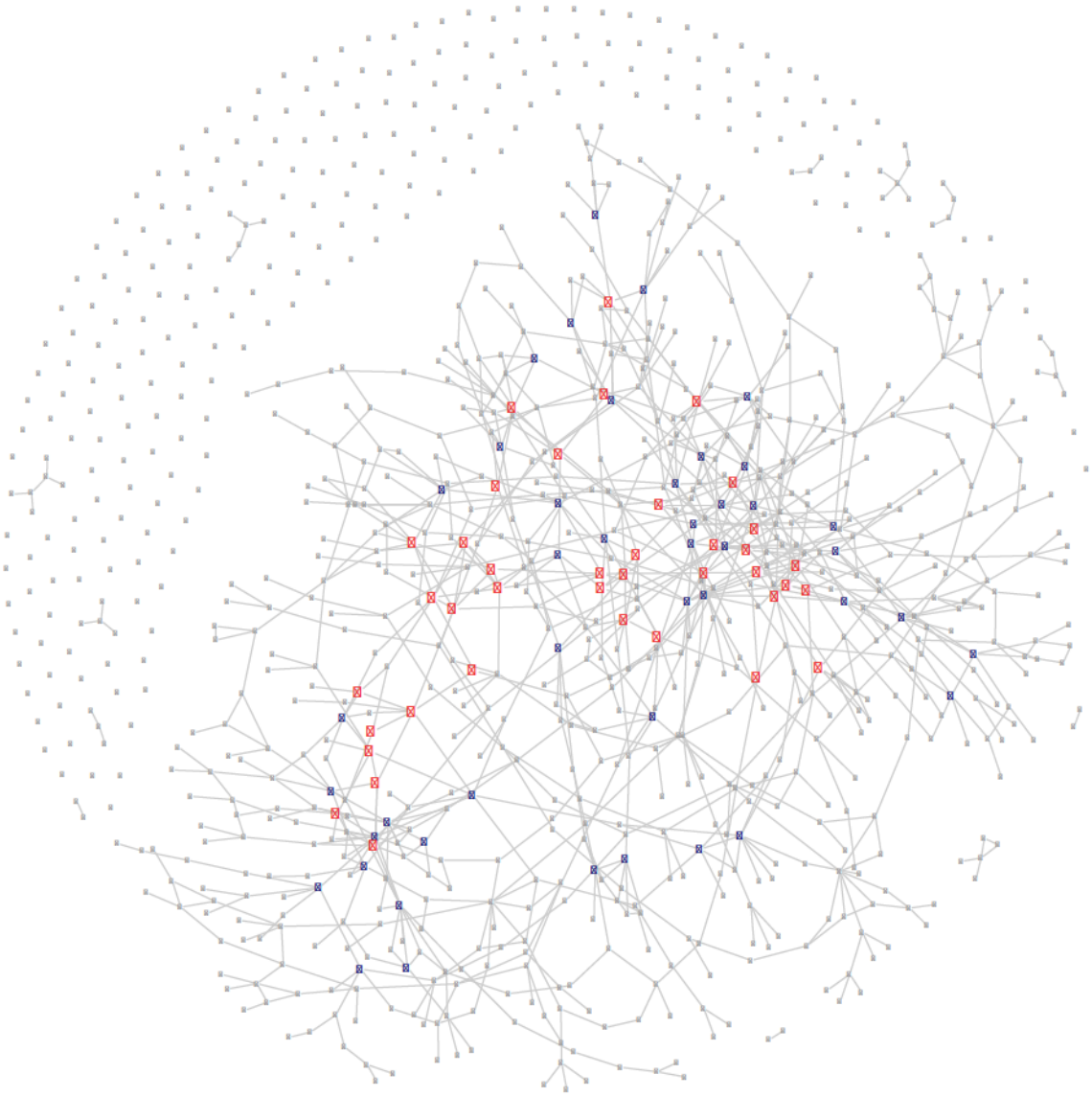


Figure 7.8.1: *Complete DAWN Network.* All genes and edges selected by DAWN are shown. Gene names are omitted due to space limitation. Primary risk genes are colored in red, secondary risk genes are colored in blue, and non-risk genes are colored in gray.

7.9 Code

All computing is performed in `R` [16]. Code is available for those with access on *Uber Genno*. Table 7.9.1 shows the scripts used for different sections. Scripts for DAWN's source code are written by Dr. Li Liu [8] and denoted with * in Table 7.9.1. Note that `source_scalef.R` contains a function to compute the $SF - R^2$ that is not available in `source_DAWN_PNS_HMRF.R`, the latest DAWN source code as of May 2015. In addition, `source_modified_dawn_main.R` contains a slightly modified version of the function `DAWN_main_addTF` from `source_DAWN_PNS_HMRF.R`. The modified version corrects a small numerical problem in the original function that produces `NaN` for posterior probabilities when the input p-values are too small. All other scripts for the analysis are written by the author.

Table 7.9.1: Code by Section

Section	`R` Script
2.1, 2.2, 2.3	`mapping.R`
3.1	`data_prep.R`
3.2	`data_regress.R`
3.3, 3.4	`data_transform_part1.R`
3.5	`data_transform_part2.R`
4.1	`dawn_pick_lambda.R`
	`source_DAWN_PNS_HMRF.R`*
	`source_scalef.R`*
4.2, 4.3	`dawn_main.R`
	`source_DAWN_PNS_HMRF.R`*
	`source_modified_dawn_main.R`*

THE END