# Understanding the Genetic Basis of Schizophrenia and other Mental Disorders by using RNA-Sequencing Data

*Thesis proposal*

**Cong Lu**
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA, 15213
congl@stat.cmu.edu

November 2014

**Committee:**

*Dr. Kathryn Roeder, Chair*
*Dr. Bernie Devlin*
*Dr. Joel Greenhouse*
*Dr. Maxwell G'Sell*
*Dr. George C. Tseng*

**Abstract**

RNA-sequencing data are replacing microarray data as a new and powerful tool to measure gene expression. At the same time, RNA-seq brings new challenges in statistical analyses of differential expression analysis and gene association networks. We explore solutions to the problem of modeling the count data in such statistical analyses using RNA-seq data. As a result of analyzing three data sets, we conclude that an appropriate transformation of the count data provides a key to the problem. To address the hidden confounders, surrogate variable analysis (SVA) has also been studied in data and a simulation. It has been shown to remove the effects of both hidden confounders and gene co-expression, which makes it unsuitable to use in conjunction with gene network analysis. In the analyses of differential expression, none of the choices for modeling the distribution of RNA-seq counts has emerged as a superior choice. Thus all of these problems merit further investigation.

The primary goal of the thesis is to use RNA-seq data obtained from the CommonMind study, together with the results of a genetic association study, to predict genes that increase risk of schizophrenia. This valuable resource, the largest of its kind, includes gene expression measurements obtained from more than 600 brains (including subjects with schizophrenia, bipolar disorder, as well as controls). We also will incorporate results from the largest GWAS of schizophrenia to date by utilizing a map between SNPs and genes based on the eSNPs identified by the CommonMind study. To succeed this task depends on each of the open questions described above as well as many others that will be addressed in this thesis.

# Contents

# 1 Introduction

## 1.1 Statistical Approaches to Gene Expression Studies

Genetic studies help researchers understand genetic basis of diseases, which are guided by statistical analyses such as association tests and network analysis. Many models of such studies have been developed based on microarray data. With the emergence of next-generation sequencing technology, microarray data is being replaced by RNA-sequencing (RNA-Seq) [37] data. Therefore, new models are needed to fit the new data. We propose new models for association tests and network analysis of gene expression studies using RNA-seq data

**Association Tests−Differential Expression and eQTL mapping:** In the area of genetic studies, there are two main statistical problems of interest: differential expression (DE) and eQTL mapping. Differential expression identifies the genes whose expression levels change between two or more groups, such as conditions, tissues or treatments. Using a multiple testing procedure, we can detect genes that are significantly differently expressed between different groups. eQTL mapping tests the association between a gene's expression and a genetic marker. By doing this, we identify the genetic markers (or SNPs) associated with expression of a particular gene.

**Network Analysis−Gene Association Network:** In gene expression studies, another important problem is to capture the dependence between genes , or co-expression. When genes are co-expressed we conjecture that they may have selected function. There are many methods for modeling the dependence between genes. One of them is to infer a gene-gene association network from gene expression levels in specific tissues.

## 1.2 New Data: RNA-Sequencing Data

Compared to microarray technology, RNA-seq has several advantages, such as an increased dynamic range, a lower background level, and the ability to detect and quantify the expression of previously unknown transcripts and isoforms. By obtaining tens of millions of short sequence reads from the transcript population and mapping these reads to the reference genome, RNA-seq generates highly reproducible results with less technical variation [23].

Models for tests of significance and network analysis were developed for continuous approximately normally distributed microarray data; however, recent developments in next-generation sequencing present new challenges via RNA-sequencing data. Modeling the experimental error presents a new challenge in that the expression levels generated from RNA-seq produces count data that does not appear to follow any of the popular models such as Poisson or negative binomial.

There are two different types of replications in RNA-seq experiments: biological replicates and technical replicates. Biological replicates are the "true" replicates, which measure different biological samples. Technical replicates are just repeated measurements from the same biological sample. We would like to investigate the biological differences between conditions/tissues/treatments, while accounting for the technical variations.

### 1.2.1 Sources of Noise in RNA-Seq Data

**Between-sample Difference−Sequencing Depth:** Due to the random sampling nature of RNA-seq, the transcripts at a low expression level need to be measured by a large number of sequences. That is because with a low sequencing depth, the short reads in weakly expressed genes would have either 0 count or very small number of counts, which brings inaccurate measurement and large variation of estimates. Therefore, to get accurate enough measurements on weakly expressed genes, sufficient depth is required. This problem has been widely discussed [2], and it is suggested that about 700 million reads would be required [23] in a mammalian genome to obtain accurate quantification of $> 95\%$ of expressed transcripts. Choosing an appropriate sequencing depth is a concern of upstream analysis.

In a downstream analysis, a problem occurs when we have samples of different sequencing depth. These counts are not directly comparable. Hence we should estimate the depth and then reduce the effect of it on our analysis. Sequencing depth is usually estimated as the total number of mapped sequences (short reads). Nevertheless, according to the the limit of the sequencing depth, for any given sequencing depth, low count transcripts are often excluded from the analysis because these reads show low consistency between technical replicates and hence cannot be considered reliable.

**Within-sample Difference−Gene Length and Others:** First, the sequencing process is not an unbiased random sampling due to the effect of transcript length. As a longer transcript tends to have larger chance to be sampled, after mapping the short reads to the genome, counts for transcripts are not uniformly distributed. Therefore, from the raw count data we expect to observe larger counts for a longer gene and lower counts for a short gene although they are truly expressed at the same level.

Second, there will be some sequence polymorphisms between biological replicates. This can result in a gain or loss in reads from different biological replicates depending on their concordance with the reference genome. For example, a short read with a SNP cannot be matched with the reference sequence and hence might not be counted.

### 1.2.2 New Challenges in RNA-seq Data

To model the noise in RNA-seq data, first we need to develop new normalization strategies that specifically address each of them. Furthermore, besides the need for new normalization, RNA-seq data brings new challenges to the downstream analysis including differential expression (DE) analysis, eQTL mapping, and estimation of a gene network. Previous approaches will not work without adaptation because these array-based analyses relied on the normal distribution assumption. The normal assumption is clearly violated for RNA-seq counts. This suggests either development of new models for count data or making appropriate transformations to achieve approximate normality.

# 2 Literature Review

## 2.1 Normalization of RNA-Seq Data

The sampling process of generating RNA-seq data is as follows: for gene $i$, $i = 1, 2, ...G$, let $\mu_i$ be the true expression value for each gene $i$ and let $L_i$ be the gene length. Then the probability that a read is from gene $i$ is $p_i = \mu_i L_i / \sum_{i=1}^{G} \mu_i L_i$. The underlying idea of the normalization strategies is that most of the genes are non-DE genes, and all of the non-DE genes are equivalently expressed, which means they should have similar read counts across samples.

According to the normalization issues mentioned in the Introduction, we need some normalization strategies to address the effects of both sequencing depth and transcript length. To adjust the effects of both, Mortazavi et al. (2008) proposed a natural normalization, which is RPKM/FPKM (Reads/Fragments Per Kilobase per Million mapped reads) [23]. It is based on the idea that the distribution of $p_i$'s are approximately the same across genes. With RPKM/FPKM, expressions of a gene across samples are assumed to be comparable, also have different genes from one sample at comparable scales. However, sometimes we do not need to normalize the count data with respect to transcript length: in some studies, such as DE analysis and computation of gene correlation, genes from different samples need not be compared directly. Thus transcript length has no impact on the results. Therefore, a simpler measure, like Count per Million (CPM) or Transcript per Million(TPM) [35] [13], that adjusts based on sequencing depth, is utilized. For both RPKM/FPKM and CPM/TPM, the sequencing depth is estimated as total count $n_i^N = \frac{n_i T_i}{T}$. Besides total count, there are other options for simple statistics, such as $n_i^N = \frac{n_i}{Q_{3,i}} \bar{Q}_3$ for upper quartile, $n_i^N = n_i Median_i$ for median, $n_i^N = \frac{n_i q_i}{q}$ for quantile.

In addition to normalization by estimated sequencing depth, there are other model-based methods. Robison et al. (2010) proposed a normalization method for DE analysis called Trimmed Mean of M-values (TMM) [29]. First, it computes the log fold changes (M) and log total counts (A) for all genes. Then the genes with extreme M and A values are discarded,and it computes a weighted mean of M's for the rest of the genes, in which the weights are the inverse of the approximate asymptotic variances. Anders and Huber (2010) proposed a robust way (implemented by $DESeq$) to estimate the sequencing depth, which takes the median of the ratios of observed counts to estimate the size factors [1]. Hanse et al. (2012) proposed a gene-specific normalization, which gives each gene a different normalization factor. Gene-specific biases were also considered, including GC content, and gene length. The basic model assumption is $Y_{g,i}|\mu_{g,i} \sim Poission(\mu_{g,i})$, and $\mu_{g,i} = exp\{h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j})\}$, for gene $i$ in sample $j$, where $h_i(\theta_{g,i})$ is the part of true expression, and $\sum_{j=1}^{p} f_{i,j}(X_{g,j})$ accounts for the biases, including GC content. This gene-specific normalization has the potential to over-fit.

The choice of normalization has a great influence on the subsequent statistical analyses; different normalization strategies have been evaluated both in simulations and real data. For example, RPKM is pointed out to be insufficient for removing the gene length bias [3] [24] [8]. And in the context of DE analysis, only $DESeq$(via the geometric mean) and TMM are able to maintain a reasonable false-positive rate without any loss of power [8]. Generally, RNA-seq data do not require sophisticated normalization [37].

## 2.2 Modeling of Count Data in Association Tests

In the context of DE analysis, to appropriately model the variation of RNA-seq data, many models have been proposed. Besides t-test, most of them are based on parametric assumptions and use Poisson or negative binomial (NB) distribution to model the counts. In addition, nonparametric models and appropriate transformations have also been suggested.

### 2.2.1 Simple t-test

To detect differentially expressed genes in groups, classic parametric tests are often used. For example, Baggerly et al. (2003, 2004) and Lu et al. (2005) used differences in proportions or Fisher's exact test and Ryu et al. (2012) adapted two-sample t-test. We cannot expect these tests to perform very well as, however, with small sample sizes because the normal assumption would fail. Therefore, choosing Poisson or negative binomial distribution is a better choice for the fragment counts of RNA-seq data.

### 2.2.2 Poisson Model

Poisson model is a natural choice for count data; however, the variation from RNA-seq data is usually much larger than the mean causing over-dispersion.

If there are only technical replicates and all of the variation arose from the random sampling variance of sequencing, Poisson has been shown to be a good fit. However, typically we do not have technical replicates available. The expense of current sequencing technology limits access to technical replicates. As a result, we cannot decompose the variation due to technical random sampling variation and biological variation.

To conquer the problem of over-dispersion, Baggerly et al. explored using beta-binomial (2003), and more generally over-dispersed logistic (2004). More recently, Lu et al. (2005) showed that the negative binomial distribution is more flexible than the beta-binomial distribution.

### 2.2.3 Negative Binomial Models

Negative binomial models were proposed to model variation of the counts because it allows variance larger than the mean. Lu et al. (2005) firstly proposed the gamma-Poisson (i.e. negative binomial) model by letting Y be an negative binomial random variable with mean $\mu$ and dispersion $\phi$ as $Y \sim NB(\mu, \phi)$. We have $E(Y) = \mu$ and $Var(Y) = \mu + \alpha\mu^2$. However, for either the Poisson or negative binomial, the estimation of parameters will be very difficult with a small sample size.

To address the difficulty of estimating parameters, Robinson and Smyth (2007, 2008 and 2010) proposed an adjusted version of negative binomial models that assumes all genes/transcripts are independent. With this assumption, we can estimate the common dispersion for all genes/transcripts, and then use an empirical Bayes estimate to smooth the gene-specific estimators towards a common parameter and yet obtain different estimates for variances. Anders and Huber (2010) proposed a similar negative binomial model, in which they computed a pooled estimate of the dispersion parameter for each gene. Then they used local regression to find the mean-variance relationship, and employed the conservative approach of selecting the largest among the fitted value and the individual dispersion estimate for each gene.

However, the adjusted negative binomial models are not perfect as the assumption of independence does not always hold; we know some fragments (exon, transcript, or gene) might be co-expressed and have a correlated pattern. The negative binomial models can be implemented in $edgeR$ [20] and $DESeq$ [1] Bioconductor packages.

### 2.2.4 Nonparametric Models

Besides the parametric models including the Poisson and negative binomial, some nonparametric models have been proposed for DE analysis using RNA-seq data. For example, Li and Tibshirani (2013) proposed a resampling nonparametric method that uses the Wilcoxon statistic to account for different library sizes [16]. They favor this approach because the Poisson or negative binomial models are heavily influence by outliers in the data. And they showed that when the assumption of parametric models do not hold, the models have low accuracy in estimating False Discovery Rate (FDR). When the sample size is large in the simulations, the resampling method was shown to remove the technical effects from the sequencing depth, and leave the biological effects for downstream DE analysis. However, when the sample size is small, the nonparametric model has much lower power to detect DE genes.

### 2.2.5 Variance Modeling at the Observational Level (VOOM)

All of the above are the DE models that fit the counts of RNA-seq data appropriately. Besides developing new models for RNA-seq data, we can also transform counts and fit the established microarray models. Law et al. (2013) proposed such a pipeline that (1) transforms counts by VOOM [13], and (2) fits the adjusted version of an empirical Bayes model for microarray implemented by $limma$ [30] [31]. The idea is to simply treat the log-counts per million (log-CPM) as analogous to the log-intensity values from a microarray experiment. And differences in log-CPM between samples can be interpreted as log-fold-changes. The challenge is that the log-CPM do not have constant variances, a solution is to use weighted least squares in the $limma$ pipeline.

Simulation studies of VOOM shows that it outperforms the negative binomial based methods in $edgeR$ and $DESeq$, and also performs better than other methods with respective to type I error rate, power, false discovery rate and computing speed.

## 2.3 Account for Hidden Confounders: Surrogate Variable Analysis

In genetic studies, besides the primary variable of interest, some other covariates are usually measured and included in the association tests. However, it is not possible to measure all the variables related to gene expression. In addition to the known and measured variables, there might be some unknown and unmeasured factors that contribute to the heterogeneity of expression levels of particular genes. Their effect cannot be ignored because the unmodeled factors, either biological or technical, can bias the expression or induce extra variability in the gene expression and decrease the power to detect the association between gene expression and the primary variable. Or they might introduce spurious signals of the association as the variation on the gene expression could be confounded by a unmodeled factor and the primary variable.

Leek and Storey (2007) proposed Surrogate variable analysis (SVA) [14] which was designed to address this problem. SVA aims to identify and estimates the components of expression hetero-

geneity (EH), and parse the signal and noise more accurately and reproducibly. SVA has become a popular way to address the problem of hidden confounders.

## 2.4 Gene Network Estimation

Gene network modeling is a fundamental tool to help understand biological pathways; there are multiple types of networks based on gene expression data. We focus on reconstructing static undirected association networks by inferring network edges and identifying gene modules as potential functional groups.

To estimate such gene network using gene expression data, in general, there are three following approaches: (1) Correlation-based co-expression network: Pearson correlation has been a popular measure compared to the other alternatives such as Euclidean distance, because correlation is invariant to linear transformations. Either hard (Carter et al. 2004) [6] or soft (Langfelder and Horvath, 2008) [12] thresholding have been proposed to produce a binary or weighted network, respectively. Correlation-based co-expression network is easy to interpret and is not computationally intensive. However, this approach cannot infer the conditional independence structure, which may lead to false positive of edges. (2) Mutual Information (MI) based network: MI is a more general way to measure gene relationship than correlation as it measures and captures the nonlinear relationship between two genes [7]. However, Steuer et al. (2002) showed that MI tends to produce almost identical results as the correlation measure [32]. (3) Gaussian Graphical Models (GGM)/Partial Correlation: Assuming that the gene expression levels follow a multivariate normal distribution, the conditional independence structure can be recovered by estimating the precision matrix (the inverse of covariance matrix, i.e. $\Omega$) of the expression data. We have genes $i$ and $j$ to be conditionally independent when the corresponding element $\Omega_{i,j}$ to be zero. This problem of estimating a sparse precision matrix is equivalent to the neighborhood selection method that is based on penalized regression to select genes with non-zero partial correlations. For instance, Meinshausen and Buhlmann (2006) applied lasso for the neighborhood selection of each gene [22]. Peng et al. (2009) proposed a joint sparse regression method for estimating the inverse covariance matrix [26]. Besides, there are many methods proposed to estimate the inverse covariance matrix directly using penalized maximum likelihood approaches (Friedman, Hastie and Tibshirani, 2008 [9]; Cai, Liu and Luo, 2011 [5]; Cai and Zhou, 2012 [4]; Ma, Xue and Zou, 2013 [19]). In addition to capturing the conditional independence structure, GGM is able to infer some edges not reconstructed by the pairwise correlation/MI. This is the case when a gene interacts with a group of other genes but not have a strong marginal relationship with any single one of the group. However, compared to the co-expression network, GGM implemented by neighbor selection method or maximum likelihood approach will usually cause a computational burden.

# 3 Preliminary Work

In Section 3.1, we first give a brief description for three available data sets. Then, in Section 3.2 we conduct simulation studies to discuss hidden covariates, specifically on how surrogate variable analysis [14] affects on network estimation. Section 3.3 compares models of differential expression in two data sets. Finally, in Section 3.4, we discuss transformations of RNA-seq data, estimate the gene association network with different transformations and compare the results.

## 3.1 Data Description

**BrainSpan** The BrainSpan transcriptome data set (Kang et al., 2011) measures 16 brain regions of human subjects that were sampled in 57 postmortem brains, with a wide age range from six weeks post-conception to 82 years of age [11].

BrainSpan data include both gene expression and sample covariates. Gene expression data are measured in RNA-seq and microarray for the same tissues, which makes it possible to directly compare the different technologies for downstream analysis. RNA-seq data are available in the measure RPKM at gene-level and exon-level. Sample covariates include age, sex, PMI (Postmortem Interval in hours) and PH. The count measurement of RNA-seq data at both gene-level and transcript-level will be available soon.

**CommonMind** CommonMind is a consortium of seven organizations that generate large scale data from human subjects with neuropsychiatric disorders. The data were generated from the Anterior Cingulate Cortex region of human brains. The data were collected from 3 sites: Hospital Mount Sinai, University of Pittsburgh, and University of Pennsylvania.The aggregation of samples collected from multiple organizations has made the data set the largest RNA-seq on brains ever assembled to date, with more than 600 human subjects are classified into schizophrenia, bipolar disorder and control groups.

CommonMind also provides RNA-seq data of gene expression and sample covariates. Expression data are available in two measurements (count and FPKM) and different levels (gene, transcript and exon). Sample covariates include disease state, age, sex, BMI and ethnicity. Covariates measured during the sequencing, such as RIN (RNA Integrity Number), PMI, PH and intergeneric rate, are also available.

$CHD8$ **Knock-Down Data** According to autism spectrum disorder (ASD), from whole-exome sequencing studies, to date nine "high-confidence" ASD risk genes have been identified. Of these genes, $CHD8$ has the strongest association with ASD risk. $CHD8$ has also been shown to regulate expression of some other genes by binding their promoters and some other ways.

We have 12 samples from human neural stem cells (hNSCs), among which 4 have $CHD8$ knockdown at C, 4 have $CHD8$ knockdown at G and 4 normal samples as controls. For each gene at each sample, gene expression data is available in the measurement of counts.

## 3.2 Simulation Study of SVA

Before moving to the downstream DE analysis and network estimation, we discuss whether surrogate variable analysis (SVA) is able to appropriately account for hidden confounders. To better

understand how SVA works under situations that have both confounders and gene co-expression, we conduct the following simulation study. To simulate the co-expression, we generate 1,000 genes in total, made up of 10 blocks, each of which contains 100 genes. We generate 200 samples from a multivariate normal distribution that has correlation within one block to be 0.4, and the correlation between blocks is set to be 0.1. Then we add the effects for case-control status (DX) and the confounder (CONFOUND). In total we have 100 cases and 100 controls, and the first 20 in blocks 1 and 2 are affected by the case/control state. The effect of the confounder is on genes 16 - 55 in each of the blocks, and on half of the cases and half of the controls.

First, we look at the performance of SVA in DE analysis. SVA estimates 11 significant surrogate variables (SVs); we fit a model including these variables as well as Disease Status (DX). We compare estimated coefficients, p-values and standard error of coefficients for DX with the true model, which includes the confounder. In Figure 1, we notice that the coefficient is estimated in an unbiased manner (a), but the standard error is underestimated (b), leading to inflated p-values (c). This suggests that SVA tends to identify too many genes as being differentially expressed.

Next, we explore the feasibility of applying SVA to remove the effects of the hidden confounder when estimating a gene network. Figure 2 shows the hierarchical clustering dendrograms of the residuals from previous two models. We notice that in the true generating model (top left), the structure of 10 blocks can be detected clearly. However, in the model fitting SVs (bottom left), we find the co-expression structure has been removed in the residual. As a result, we conclude that SVA not only removes the effects of the hidden confounder, but also removes co-expression structure among the genes. For this reason, we will not include SVs in our further network analysis. Furthermore, we observe a clear pattern of separation on the 11 estimated SV, among which one SV is highly correlated with the hidden confounder (with correlation 0.89) and the other 10 SVs are lowly correlated (with correlation smaller than 0.29). The clustering results on the residual removing the correlated SV (bottom right in Figure 2) seems to recover the co-expression structure of the true generating model. If we can distinguish between SV(s) accounting for confounders and SV(s) describing co-expression, we might be able to improve surrogate variable analysis in estimating a gene network.

## 3.3 Differential Expression Analysis

In RNA-seq data, the measurement RPKM was widely accepted [23] because both sequencing depth and transcript length have been adjusted. Gene quantification and differential expression can be analyzed using Cufflinks [33] and CuffDiff [34], respectively, based on the RPKM measure. However, recently the RPKM measure has been shown to be inconsistent among samples [35] which makes it unreliable. As an alternative, a simpler measure Count per Million (CPM) or Transcript per Million(TPM) that is only adjusted by sequencing depth, is preferred [35] and adopted for the DE analysis [13]. Next, we compare some popular differential expression (DE) models based on the measure CPM and count data.

### 3.3.1 Models of DE on CommonMind Data

We use CommonMind data and apply two models to detect the DE genes: (1) a negative binomial model [27] [28] [21] implemented by $edgeR$ [20] and (2) a linear model fitted by weighted least

squares regression on transformed data, which is implemented by VOOM [13] and $limma$ [31] [30].

From CommonMind data, after removing the outlier samples, which are significantly different from all of the others, there are in total 610 samples with both gene expression and covariates recorded, in which 262, 55 and 293 samples are from Scizophrenia, Bipolar and control groups, respectively.

**Gene Filtering** According to the nature of RNA-seq data, the transcripts and genes with low counts have extremely high variance, and it is almost impossible to differentiate their signal from noise. As a result, the first step of DE analysis using RNA-seq data is to filter out the genes with very low counts. For example, edgeR [20] recommend keeping only the genes with CPM greater than 1 in more than half of the samples.

**Sample Covariates** For DE analysis, besides the disease status (Schizophrenia or control), the effects of other related covariates on gene expression cannot be ignored. From dozens of covariates in CommonMind data, we use forward stepwise method to select features that contribute to the heterogeneity of gene expression. As a result, we make the decision to include $Sex$, $Age$, $PMI$, $RIN$, $RIN^2$, $Batch$ and $Site$.

**DE analysis** As pointed out in the literature, the Poisson models fail due to the over-dispersion. In Figure 3 we observe over-dispersion especially on the genes with relatively low counts.

By fitting the negative binomial model [27] [28] [21], after normalize the count data by TMM [29], we first estimate one common dispersion parameter $\phi$ for all the genes. Then based on the estimated common parameter, we estimate a tag-wise dispersion parameter for each gene. With the estimated parameters, we test associations between gene expression and Schizophrenia for all the genes, and eventually get a list of 1288 genes differentially expressed at the significance level 0.05 by False Discovery Rate (FDR) adjustment for multiple testing. In the linear model by $limma$, we transform the count data by applying VOOM and follow the pipeline of the empirical Bayes model implemented by $limma$. As a result, 1057 genes are detected to be differentially expressed at the same significance level 0.05.

We compare the results of the above two models. In Figure 4, we notice that most of the detected genes in the two models agree with each other. 74% of the DE genes detected by $limma$ are also significant by the negative binomial model, while 61% of the DE genes by negative binomial model are significant in $limma$ model. However, the remaining significant genes by only one approach differ notably, with 276 only by $limma$ and 507 only by the negative binomial model.

### 3.3.2 Models of DE on Knock-down Data

We apply two models: Poisson and negative binomial models to the knockdown data, to observe how they perform in detecting differentially expressed genes between the knock-down groups and the control group.

After filtering out extremely lowly expressed genes, we have 14,841 genes left, which then are normalized by TMM [29]. We apply Poisson and negative binomial models on both treatment C and treatment G. The distributions of the DE p-values are shown in Figure 5. We notice that

the large number of small p-values suggests a strong effect of $CHD8$ knockdown on many genes while a portion of small p-values could be due to the failure of model assumption.

Next, given a list of genes bound by $CHD8$ from previous research, we analyze gene enrichment to see if the above models could truly detect genes bound (probably regulated) by $CHD8$. Among all the 14,841 genes, 8,140 are bound by $CHD8$ and 6,701 are not. The genes bound by CHD8 are expected to express differently when a depletion of CHD8 happens, and they should be detected by an appropriate DE model.

To compare the two groups of 8,140 and 6,701 genes, we use a Wilcoxon rank test, in which the ranks of genes are based on the DE p-values acquired from the above models. First, by using the p-values from the negative binomial model, the two groups of genes are significantly different from comparing treatment C and controls, with p-value $9.7 \times 10^{-5}$. However, from comparing treatment G and controls, the p-value of the Wilcoxon test is 0.0681, which suggests depletion of $CHD8$ at G has no different impact on the genes bound by $CHD8$. This contradicts with previous conclusions. However, we get significantly small p-values for the Wilcoxon test based on Poisson model of DE. These results suggest that negative binomial model may not capture the variation of count data well, at least for this specific data set, possibly leading to inaccurate signals of DE analysis in general.

## 3.4 Estimation of Gene Association Network

As we move away from using microarray to RNA-seq, continuous data is being replaced by discrete count data; to estimate the gene association network, the previous models for microarray data will probably fail for the following reasons. In a co-expression network based on either pairwise correlation, the correlation coefficients will be different in RNA-seq from microarray, as show in Figure 6 $(a)$ from BrainSpan. An estimated network based on pairwise Mutual Information(MI) results in the same situation. As for a network based on partial correlation(Gaussian Graphical Models), the assumption of normality fails in the count data. As a result, to address the problems, we need to either transform the count data appropriately to have similar correlation structure with microarray data to meet the model assumptions, or we have to develop a brand-new model for count data.

### 3.4.1 Transformation of Count Data

**Gaussian Transformation** The first possible transformation suggested by VOOM is to transform the count data by taking the logarithm. We take the log-transformation of RPKM from BrainSpan data and compare the pairwise correlations to those obtained from microarray data. Figure 6 shows an example of correlation coefficients between gene AQP4 and the other genes, in which we can see the pairwise correlations are more close to the ones from microarray data in Figure 6 $(b)$, with a higher correlation 0.8987 from log-transformation compared to 0.8278 from untransformed RPKM data.

Besides the log-transformation, Liu et al. (2009) proposed Nonparanormal [18] [42] transformation, generally on any type of distributions to relax the normality assumption in estimation of network. After nonparannormal transformation, each gene follows a standard normal distribution. However, Figure 6 $(c)$ shows it does preserve correlation structure of microarray data as well as log-transformation.

**Poisson Transformation** Both of the above try to make the data meet the normality assumption with transformation; however, some other attempts of Poisson transformation were also proposed. Li et al. (2011), Witten at al. (2010) and Witten et al. (2011) adopted a power transformation to make the overall distributions across all genes approximately Poisson [17] [38] [40]. As we see in the DE analysis using RNA-seq data, the Poisson model naturally captures the variation of count data. However, we can always observe problems of over-dispersion relative to the Poisson model. Witten at al. (2011) showed that even when the data are truly generated from a negative binomial model with over dispersion, clustering methods based on the Poisson model performed well on the transformed data [40].

The idea of this Poisson Transformation is based on a test of the goodness of fit. Let $X_{ij}$ be the counts of gene $j$ for sample $i$, and $X'_{ij} = {X_{ij}}^{\alpha}$, where $\alpha \in (0, 1]$ is chosen such that $\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(X'_{ij}-X'_{i.}X'_{.j}/X'_{..})^2}{X'_{i.}X'_{.j}/X'_{..}} \approx (n-1)(p-1)$. In Figure 7, we notice that most genes have variance larger than the mean, while after transformation almost half of the genes below the mean. Besides, after transformation we observe a much lower common dispersion and many more genes with zero dispersion that follows Poisson distributions. But we also observe the pairwise correlation in Figure $6(d)$ , Poisson transformation is not as good as the other two transformations to keep the correlation structure of microarray data.

**Empirical Comparison** For all of the above transformations, Figure 8 shows examples of gene expression distributions before and after the transformations. The correlation between microarray and transformed RNA-seq data in Pearson correlation coefficients is 0.7740, 0.7681 and 0.7634 for log-transformation, power-transformation and nonparanormal-transformation respectively. All is better than the correlation between microarray and untransformed RPKM data, which is 0.7531. Empirically, all transformations are better than original data in keeping the correlation structure of microarray data. And log-transformation is slightly better than the other two alternatives.

### 3.4.2 Sparse Canonical Correlation Analysis

To justify if a transformation is necessary and which transformation is more appropriate for RNA-seq data, we propose applying sparse canonical correlation analysis (sparse CCA) to microarray and transformed RNA-seq data. With sparse CCA, we conclude that log-transformed RNA-seq data has the most similar structure to microarray data. And with this result, we will proceed to adopt log-transformation in the microarray framework of estimating gene network in Section 3.4.3.

**Review of Canonical Correlation Analysis (CCA)** Let $X \in \mathbb{R}^{n\times p}$ be a matrix comprised of n observations on p variables, and $Y \in \mathbb{R}^{n\times q}$ be a matrix comprised of n observations on q variables. CCA was introduced by Hotelling (1936) to find maximally correlated linear combinations between the two sets of variables. More explicitly, we would like to find $a \in \mathbb{R}^q$ and $b \in \mathbb{R}^p$ that maximize $a^T\sum_{YX}b$, subject to $a^T\sum_{YY}a = 1$ and $b^T\sum_{XX}b = 1$, where$\sum_{(.,.)}$ is a correlation matrix. In practice, the population correlation matrices will be replaced with sample correlation matrices, which are $\hat{\sum}_{YX} = S_{YX} = Y^TX/(n-1)$, $\hat{\sum}_{XX} = S_{XX} = X^TX/(n-1)$, $\hat{\sum}_{YY} = S_{YY} = Y^TY/(n-1)$, assuming the columns of $X$ and $Y$ have been centered and scaled. Then the problem turns to $max_{a,b}a^TS_{YX}b$, subject to $a^TS_{YY}a = 1$ and $b^TS_{XX}b = 1$.

**Sparse CCA** However, when $min\{p, q\} > n$, the classical CCA does not work because the singular value decomposition method by Hotelling (1936) requires sample covariance matrices to be invertible, which is not possible in high dimensional genomic data. Motivated by this problem, a number of studies imposed a sparsity constraint on the leading canonical correlation directions (Waaijenborg, Verselewe de Witt Hamer and Zwinderman (2008); Witten, Tibshirani and Hastie (2009); Parkhomenko, Tritchler and Beyene (2009)) [36] [39] [25]. As a result, only a subset of the $p$ variables in $X$ and a subset of the $q$ variables in $Y$ correspond to nonzero coefficients in $a$ and $b$, for each of the leading canonical directions.

Here we adopt the diagonal sparse CCA criterion by Witten, Tibshirani and Hastie (2009) [39], which treats the covariance matrices as diagonal and relaxes the equality constraints for convexity: $max_{a,b} a^T Y^T X b$, subject to $a^T a \leq 1$, $b^T b \leq 1$, $p_1(a) \leq c_1$, $p_2(b) \leq c_2$, where $p_1$ and $p_2$ are convex penalty functions. We apply the sparse CCA on the Brainspan data set, in which $Y_1$, $Y_2$, $Y_3$ , $Y_4$ and $X$ are defined to be RNA-seq data with RPKM, log-transformed RPKM, nonparanormal-transformed RPKM, power-transformed RPKM and microarray respectively. The matrices $X$, $Y_1$, $Y_2$, $Y_3$ and $Y_4$ have matched rows that are samples and matched columns that are genes.

In Figure 9, we observe in the first 20 leading canonical variables, the original RPKM has lower correlations than the three transformations. Log-transformation and power-transformation are almost the same, and both are better than nonparanormal. We compare the selected genes (with nonzero coefficients) in $X$ and $Y$'s; in total we selected about 160 genes out of 10,800. For the log-transformation, the overlap of selected genes is as high as 50% while for the original RPKM and the other transformations it is below 40%. Furthermore, we look into the correlation of the weights for selected genes between microarray and RNA-seq, and it is much higher (0.84) for the log-transformation than the others which range from 0.15 to 0.37.

The above results suggest that we need to do some transformation on the original RNA-seq data prior to performing network analysis. We believe the same conclusion also applies to the measure in CPM because for each gene, $CPM = l \times RPKM$, where $l$ is the gene length. We also conclude that log-transformation is better than the other available counterparts.

### 3.4.3 Gaussian and Poisson Graphical Model

After comparing different transformations, next we explore reconstructing gene association networks based on microarray data and transformed RNA-seq data. To capture the conditional independence structure, we adopt Gaussian Graphical model on microarray data, log-transformed data and nonparanomral-transformed data. For the power-transformed data, as the data is approximately Poisson distributed, we apply a Poisson version of GMM, of which the details are discussed below.

There are different approaches to implement GGM, and here we use neighborhood selection based on lasso-regularized regression [22], by regressing gene $i$ on the other genes: $X_i = \sum_{j \neq i} \beta_{ij} X_j + \epsilon_i$. The Poisson Graphical Model is simply replacing the above by a Poisson regression using the log link function. In practice, we use a slightly different version−Partial Neighborhood Selection (PNS) (Liu et al. 2014) that starts with a subset of candidate connected genes instead of all the available genes to obtain a more manageable dimension. This is because (1) partial correlation computed from all the available genes will contain genes that are not biological correlated, leading to spurious dependency and false positive edges in the network; (2) as our ultimate goal is to detect disease-specific genes, the dependency among risk genes for a particular

disease is more essential than of non-risk genes, suggesting analysis restricted to a group of risk genes as candidate neighbors; and (3) it has been shown that ignoring some components of high dimensional parameter will lead to better estimation accuracy [15].

Therefore, before applying the PNS to construct network edges, we first select a subset of risk genes for ASD by using the p-values for each gene obtained from the Transmission And De novo Association (TADA) test (He et al., 2013) [10], which integrated multiple sources of exome genome sequencing data. Firstly, we exclude genes with p-value $p_i > t$, where $t$ is some threshold. Second, we use hard-thresholding based on pairwise correlation to exclude the genes that are not connected to any others. Third, since genes with large p-values (low risks) may be closely connected to risk genes but have been removed in step 1, we retrieve the neighbors of the genes we obtained in step 2. After the above 3 steps, we will have a group of candidate genes.

Next we apply lasso regression on one gene each time against the other candidate genes to select neighbors. The penalized parameter $\lambda$ is chosen to obtain a network that follows power-law distribution (Zhang and Horvath, 2005) [41]. As a result, compared to the other transformations, we have gene network estimated from log-transformation to be the most similar to the counterpart from microarray data, with the smallest distance (in Frobenius norm) between the adjacency matrices.

# 4 Research Plan

## 4.1 Core microarray and RNA-seq analysis in BrainSpan

A major area for further study relies on the BrainSpan data for which both microarray and RNA-seq data are available from the same tissue samples. This valuable source of data provides us with great opportunities for improved understanding of the two technologies.

- Assuming that the correlation between genes should be similar regardless of technology, sparse CCA has provided a direction for determining an appropriate transformation. Although this approach has lead to insights about optimal transformations, there are some weaknesses to this approach. Specifically, the genes included in each sparse representation are not necessarily matches. Hence we have future plans to refine and improve this type of analysis, moving from CCA to PCA performed on a function of both data types.

- In principle RNA-seq is supposed to provide additional data for genes with expression in the lower end of the range. This advantage is particularly important for transcription factors (TF), which have a big impact on other genes even when their expression level is low. We aim to investigate differences in the pattern of correlation observed for this type of gene in both types of data. Such insights could aid in modeling genes with low expression.

- One limitation of DAWN has been the inability to include genes with poor expression in network analysis due to lack of knowledge about correlation patterns. We aim to use results obtained from (2) to improve our analysis of gene networks in DAWN. One simple option is to use expression from both sources to derive networks and particularly to patch in missing edges.

## 4.2 Case-control RNA-seq analysis in CommonMind

- From the simulation study shown in Figure 2, we notice that SVA tends to account for both the hidden confounder and co-expression structure in the gene expression matrix. Using RNA-seq control samples, we aim to model the biological correlation, which will help to differentiate gene co-expression structure from the effects of hidden confounders. The aim is to remove the confounding effects only and preserve the biological structure. This will improve surrogate variable analysis (reduce false discoveries) and facilitate analysis of true biological structure.

- Aim to develop a high dimensional approach akin to sparse CCA to compare paired samples of cases and controls to identify covariates (genes) that predict similarities and differences between the two groups with respect to their networks.

## 4.3 Network analysis to discover Schizophrenia genes and subnetworks

In the DAWN framework (Liu et al. 2014), p-values for Autism were used in partial neighborhood selection (PNS). As for CommonMind data, we are going to incorporate the results of p-values from DE analysis to the PNS procedure, estimating gene association network in schizophrenia.

Since the real relationship between genes may change over time or across disease states, we are going to estimate networks in both cases and controls separately. The difference between them will help understanding how gene relationships change according to schizophrenia and then the biological pathway of the disease.

# List of Figures

# 5 Bibliography


[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(R106), 2010.

[2] Ali Bashir, Vikas Bansal, and Vineet Bafna. Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC genomics*, 11(1):385, 2010.

[3] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[4] T Tony Cai, Harrison H Zhou, et al. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

[5] Tony Cai, Weidong Liu, and Xi Luo. A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[6] Scott L Carter, Christian M Brechbühler, Michael Griffin, and Andrew T Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

[7] Carsten O Daub, Ralf Steuer, Joachim Selbig, and Sebastian Kloska. Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, 5(1): 118, 2004.

[8] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14 (6):671–683, 2013.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[10] Xin He, Stephan J Sanders, Li Liu, Silvia De Rubeis, Elaine T Lim, James S Sutcliffe, Gerard D Schellenberg, Richard A Gibbs, Mark J Daly, Joseph D Buxbaum, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics*, 9(8):e1003671, 2013.

[11] Hyo Jung Kang, Yuka Imamura Kawasawa, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li, André MM Sousa, Mihovil Pletikos, Kyle A Meyer, Goran Sedmak, et al. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, 2011.

[12] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[13] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Preprint 2013*, 2013.

[14] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[15] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2004.

[16] Jun Li and Robert Tibshirani. Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536, 2013.

[17] Jun Li, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, page kxr031, 2011.

[18] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

[19] Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

[20] Davis J. McCarthy Mark D. Robinson and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[21] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, page gks042, 2012.

[22] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[23] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[24] Alicia Oshlack, Matthew J Wakefield, et al. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4(1):14, 2009.

[25] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[26] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486), 2009.

[27] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[28] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[29] Mark D Robinson, Alicia Oshlack, et al. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[30] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[31] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[32] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, 2002.

[33] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[34] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[35] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012.

[36] YX Rachel Wang, Keni Jiang, Lewis J Feldman, Peter J Bickel, and Haiyan Huang. Inferring gene association networks using sparse canonical correlation analysis. *arXiv preprint arXiv:1401.6504*, 2014.

[37] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[38] Daniela Witten, Robert Tibshirani, Sam G Gu, Andrew Fire, and Weng-Onn Lui. Ultra-high throughput sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC biology*, 8(1):58, 2010.

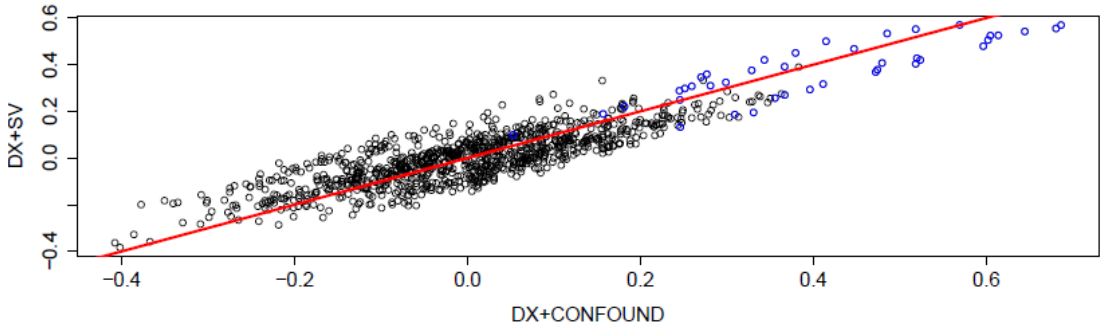
[39] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[40] Daniela M Witten et al. Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518, 2011.

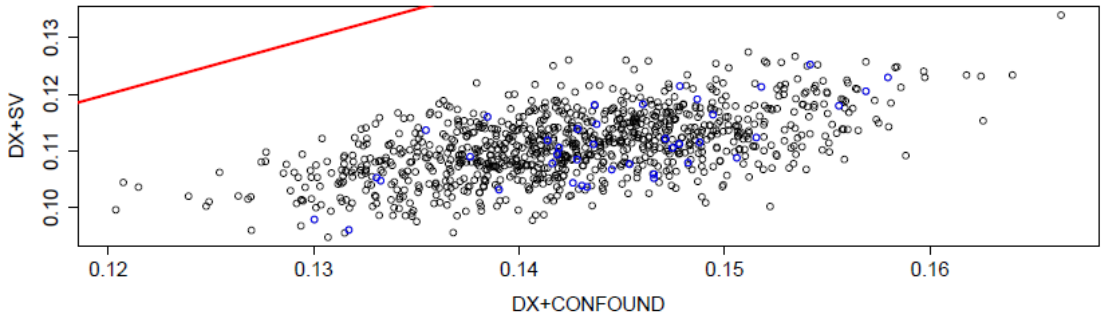
[41] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[42] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13(1):1059–1062, 2012.

(a) Coefficients




(b) Standard Error




(c) p-values

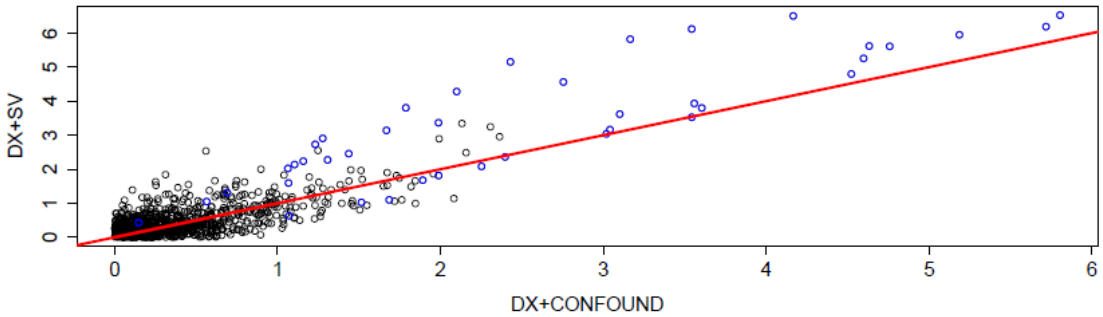



Figure 1: Simulation Study of SVA: Coefficient and p-value

Figure 2: Simulation Study of SVA: Co-expression Structure

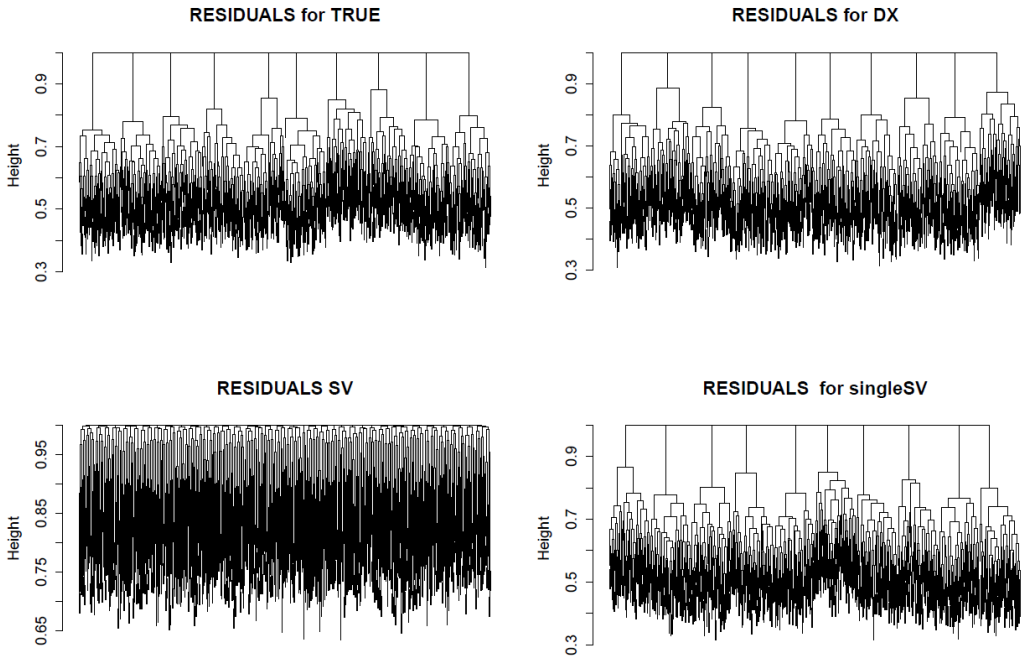



Figure 3: Over-dispersion of CommonMind data

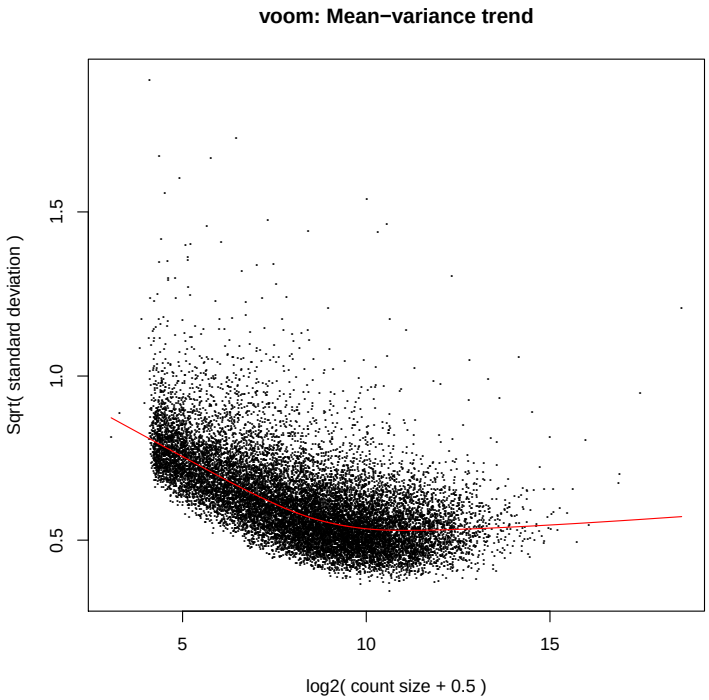


Figure 4: Comparison of DE Models

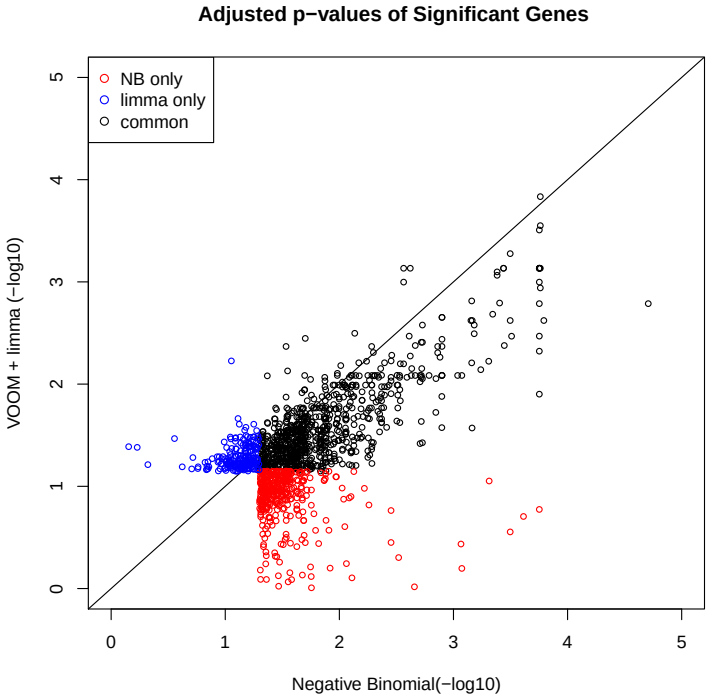



Figure 5: p-values for DE on CHD8 Knock-down Data

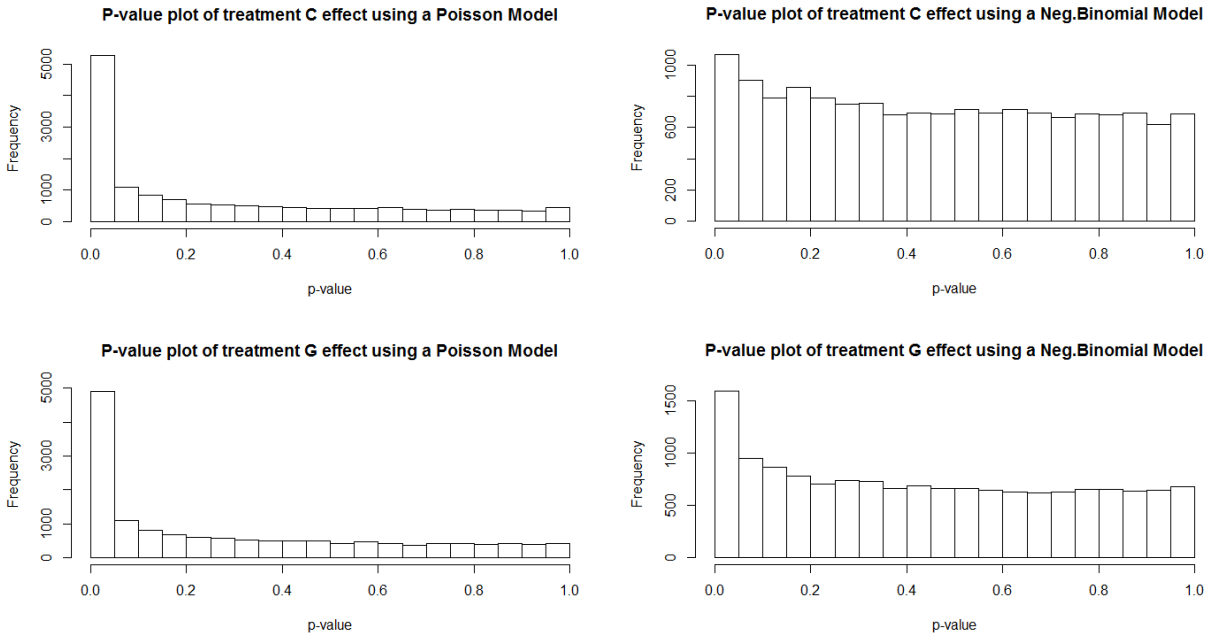


Figure 6: Correlation between Microarray and RNA-seq Correlation Coefficients

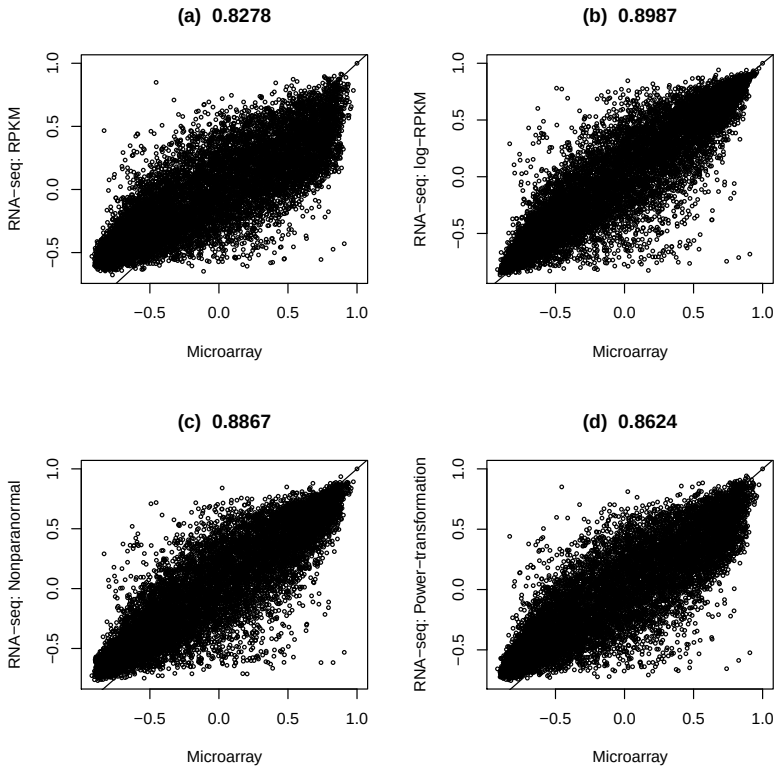


Figure 7: Mean vs Variance: before and after power transformation

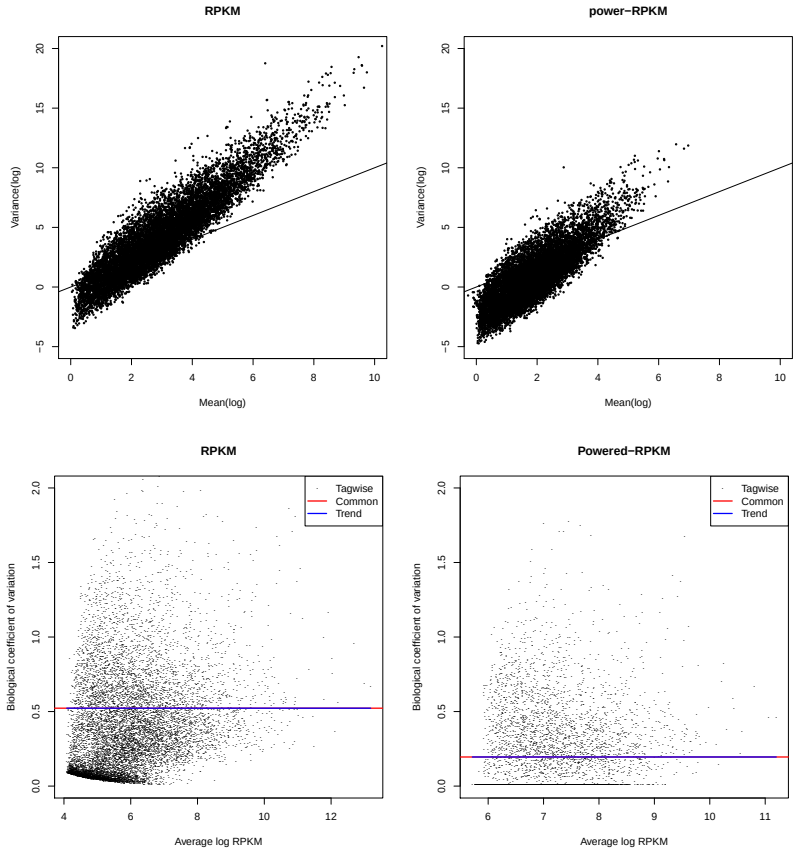


Figure 8: Distribution of Gene Expression after Transformations

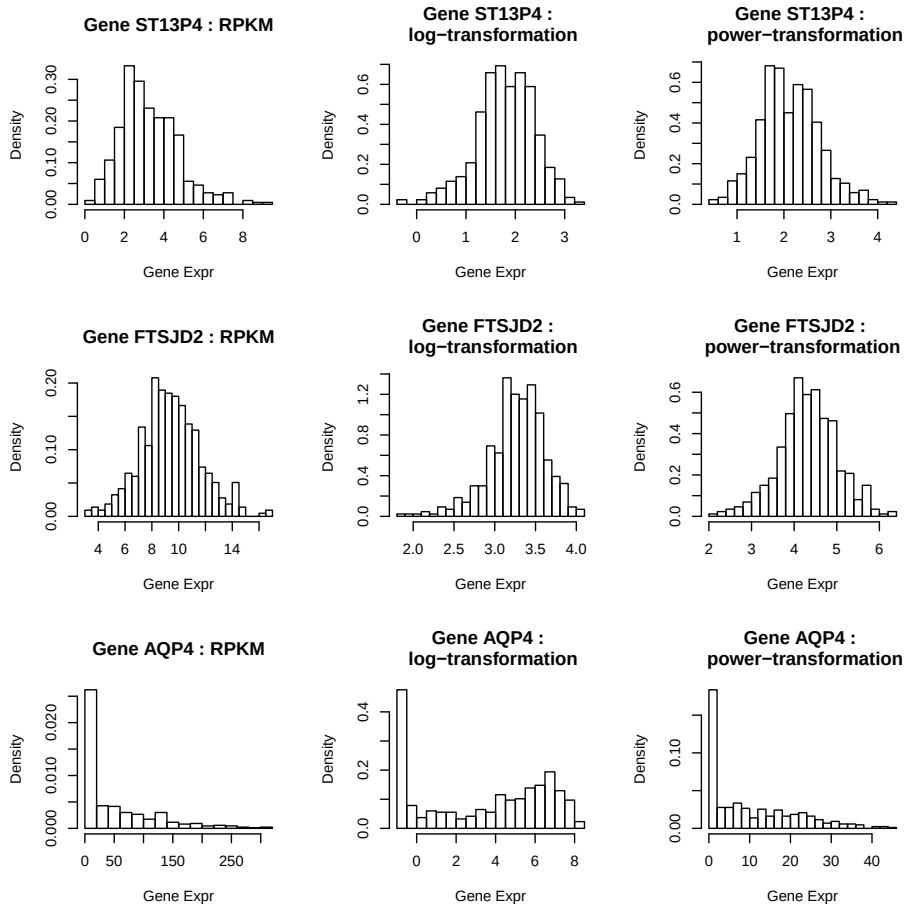



Figure 9: sparse CCA on Transformations