

12-2019

## **MOVING BEYOND THE SINGLE GENE: INTEGRATIVE GENE SET ANALYSIS FOR RNA-SEQ**

ANDREW RICHARD ASCHENBRENNER

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/uthsph\\_dissertsopen](https://digitalcommons.library.tmc.edu/uthsph_dissertsopen)



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

MOVING BEYOND THE SINGLE GENE: INTEGRATIVE GENE SET ANALYSIS FOR  
RNA-SEQ

by

ANDREW RICHARD ASCHENBRENNER, BS

APPROVED:



---

YUN-XIN FU, PhD



---

STEPHEN DAIGER, PhD



---

MOMIAO XIONG, PhD



---

DEAN, THE UNIVERSITY OF TEXAS  
SCHOOL OF PUBLIC HEALTH

Copyright  
by  
Andrew Aschenbrenner, BS, PhD  
2020

## **DEDICATION**

To Ricardo Cerros Jr

MOVING BEYOND THE SINGLE GENE: AN INTEGRATIVE APPROACH TO GENE  
SET ANALYSIS IN RNA-SEQ

by

ANDREW ASCHENBRENNER  
BS, University of California, Irvine, 2009

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS  
SCHOOL OF PUBLIC HEALTH  
Houston, Texas  
December, 2019

## **ACKNOWLEDGEMENTS**

I would like to thank my advisor and committee chair Professor Yun-Xin Fu for his patience and giving my ideas when I had none or at least no good ones (happens more often than you hope). His help in setting deadlines for me enabled scenarios to practice self-discipline, while being flexible when I didn't meet them. I would like to extend my gratitude to my committee members Professor Stephen Daiger and Professor Momiao Xiong for entertaining helpful discussions about biology and helpful suggestions that were simple enough for me to understand. I would also like Professor David Loose and Professor Patricia Mullen through the CPRIT fellowship, offering suggestions and guidance whenever they could. Additionally, a thank you is deserved of CPRIT pre-docs and post-docs with whom I had a number of conversations and helped me through difficult times, which include Mary, Heidi, Krit, Frank, Sara, Mandana, and Thanh.

I would also like to thank many of my friends and colleagues who let me complain about my academic woes and my startling opinions of universities, usually over a frosty beverage (or 10). Despite whatever was happening or how depressed I was by school, you were always there and instantly made me feel better. I realized that I should always make time for people because they are the most important parts of life.

A special thanks to my college roommate and one of my best friends Ricardo Cerros Jr, an Army Ranger who gave his life for his country fighting the War in Afghanistan in the Logar Province in October 2011. He sacrificed his life jumping on a grenade to save his Commanding Officer's life. Vicious Rick, as we used to call him (a joke since he was the nicest and most honorable person you would ever meet, is the standard of righteousness in this

world that I live by. Every day since you fell in battle, I try to live up to the man you always were. You have and always will motivate me to be a better person. You cross my mind from time to time and I reminisce on the college days at the Anthill Pub, playing Guitar Hero, and having you teach me new crazy exercises fit for military personnel. I laughed when you said I should join the Air Force (I mean, Chair Force) because I was fit enough already. You were the greatest person I'll ever know. Thanks for giving up your life so I could study.

Finally, I'd like to thank my parents, in particular my mom. You always gave me pep talks and listened to my woes. My conversations helped ease my stress and it made me feel better that you were on my side, even on the bad days. To my dad, thanks for your patience and providing me with the resources to succeed. I appreciated our conversations about career, the future, and thinking economically about my choices. Thanks for teaching me to always follow the money.

# MOVING BEYOND THE SINGLE GENE: AN INTEGRATIVE APPROACH TO GENE SET ANALYSIS IN RNA-SEQ

Andrew Aschenbrenner, PhD  
The University of Texas  
School of Public Health, 2020

Dissertation Chair: Yun-Xin Fu, PhD

RNA-seq is the next-generation sequencing technology for gene expression and while many tools have been developed to assess differential expression, most focus on gene-level statistics. Gene-level statistics implicitly ignore any dependence among genes. In order to directly incorporate correlation into testing for differential expression, genes are sorted into networks using Ingenuity Pathway Analysis (IPA) and the gene expression of each network is modeled using Generalized Estimating Equations (GEE). Since the gene network data often exhibits correlation structures containing positive and negative values, a new intermediate correlation structure is developed. This structure provides a compromise between an exchangeable (one parameter for all gene pairs) and unstructured (one parameter for each gene pair). A log-linear regression model and Wald test are proposed for differentially expressed gene networks via hypothesis testing. Additionally, a statistical test to determine whether a given gene network is independent or correlated is given. Numerical studies via simulations of correlated negative binomial data are used to compare different correlation structures of GEE based on type I error and statistical power. Also, these simulations are used to benchmark the test of gene network independence. Models that incorporate correlation into estimation



are able to conserve type I error, while independent correlation structures do not. Positive correlations unaccounted for in the independence model lead to increases in type I error, while negative correlations lead to decreases in type I error. Power between models is roughly the same. Goodness-of-fit tests reveal that the correlated negative binomial data is a better fit to the actual data than the univariate negative binomial distribution. Data analysis consisted of two stages: (i) analyzing gene networks with GEE to find differentially expressed networks and (ii) performing a single gene analysis on these differentially expressed networks. A RNA-seq dataset from the Cancer Genome Atlas (TCGA) of breast cancer patients was analyzed, adjusting for relevant clinical covariates. The top genes from each network are scrutinized for biological relevance using PubMed searches and other knowledge-based databases such as OMIM. These genes show a mix of genes with no citations to genes with many citations. This implies that this data analysis approach finds both novel genes as well as genes that have been well-studied in the field of breast cancer research.

## TABLE OF CONTENTS

List of Tables .....	ii
List of Figures .....	iii
Background .....	1
Literature Review.....	1
Public Health Significance.....	6
Specific Aims.....	7
Journal Article.....	10
Incorporating correlation structure into the analysis of RNA-seq data .....	10
Statistical Applications in Genetics and Molecular Biology .....	10
JOURNAL ARTICLE .....	37
A gene network analysis of primary breast tumors in RNA-seq .....	37
Genetic Epidemiology .....	37
Conclusion .....	53
Appendices.....	57
References.....	67

## LIST OF TABLES

Table 1: Example Correlation Structures.....	15
Table 2: Type I error rates for different GEE models.....	23
Table 3: Power analysis for different GEE models. ....	27
Table 4: Type I error rate and power benchmarking for test of gene set independence.....	29
Table 5: Results of Gene Network Analysis.....	46
Table 6: Single gene analysis of statistically significant gene networks. ....	48

## LIST OF FIGURES

Figure 1: QQ plots for log counts per million (CPMs) for all of the genes in the RNA-seq dataset. ....	44
Figure 2: Patient demographics for samples in the TCGA dataset .....	45

## **LIST OF APPENDICES**

Appendix A: Gene Network and Single Gene Analysis Results .....	57
---	----

## BACKGROUND

### **Literature Review**

Cancer is second leading cause of death in the United States and is expected to overtake heart disease as the leading cause of death in the next few years (Rl, Kd, & Jemal, 2015). Surveillance Epidemiology and End Results (SEER) estimates that 1,658,370 people will be diagnosed with cancer and 589,430 people will die from cancer in 2015; these deaths will account for 25% of the total deaths in the United States (Howlader et al., 2014). Moreover, cancer survival has increased by about 4% in the past two decades.

Cancer is a group of diseases that, if left untreated, are fatal for many patients. For example, patients with breast cancer who are not treated have a median survival time of 2.7 years (5-year survival rate: 18%, 10-year survival rate: 3.6%) (R. H. Riffenburgh & Johnstone, 2001; Robert H Riffenburgh, 2000). Many of these tumors can be treated effectively, improving survival and in some cases, enabling patients to manage their cancer throughout their lives (Howlader et al., 2014). However, other tumors cannot be managed with current therapies and will ultimately lead to death. Prevention efforts such as mammograms (x-ray pictures of the breast) can identify a tumor at an earlier stage, giving greater lead time, but provide no reduction in mortality rates (Miller et al., 2014). To reduce mortality rates, there is a need to understand the molecular etiologies of tumors.

RNA-sequencing (RNA-seq) is a next-generation sequencing technology that measures the expression of genes with great accuracy and with genome-wide coverage. Analyzing RNA-seq data is typically done for each gene separately using a statistical test or regression model with a multiple testing adjustment such as the false discovery rate (FDR). This

procedure identifies genes that are differentially expressed across two or more populations (such as tumor vs normal). Both experimental and numerical studies have shown these single gene approaches have low sensitivity (~50%), implying that many genes are missed. To improve sensitivity, networks, instead of single genes, will need to be analyzed.

Oncogenesis (the development of cancer) is largely the result of activated oncogenes (cancer-causing genes) and deactivated tumor suppressor genes (Weinberg, 2013). The interplay between these genes regulate cell proliferation in healthy humans, but when homeostasis is disrupted the result is the creation of a tumor. As the cancer progresses, additional genetic damage is acquired through mutations or epigenetic silencing that alters gene expression (Nussbaum, McInnes, & Willard, 2015). Identifying genes which have their expression altered during both the initiation and progression of cancer is both clinically important and biologically relevant. Discovering these differentially expressed genes helps to generate hypotheses about tumor genetics, providing evidence and the first step for the development of future clinical interventions.

A number of different genes have been identified in different types of cancers. Cancer genes can be specific to a cancer type or shared among many cancer types. As of 2013, a total of 125 genes (either oncogenes or tumor suppressor genes) have been found (Vogelstein et al., 2013). Additional experiments of gene finding have yielded similar genes in different tumors, leading the authors to the conclusion that most of the genes in cancer have been found. However, estimates of explained genetic variation argue that there may be much to learn in cancer genetics. For example, genes found in breast cancer have only explained 12.5% of the total genetic variation, implying there is still much to discover (So, Gui, Cherny, &

Sham, 2011). One possible reason for the lack of predictiveness of genetics is that the statistical methods for determining important genes do not accurately portray the biology. Specifically, in gene expression studies, genes are assumed to be the unit of analysis and each gene is independently tested, ignoring any relationship among genes. The importance of gene clusters is well-established, yet this assumption of independence, and a focus on individual genes, persists.

Single gene analyses in RNA-seq differential expression fall into two categories: count modeling and variance modeling. The goal of single gene methods is to address the phenomenon that different genes have different variances (of gene expression). Implicitly, these single gene analyses assume statistical independence among genes, reducing this covariance to zero. Count modeling focuses on modeling the raw counts of RNA fragments measured to quantify gene expression for each gene in an RNA-seq experiment. While the Poisson distribution is the most natural for count data, it only accounts for technical variation (one sample repeated multiple times) (Marioni, Mason, Mane, Stephens, & Gilad, 2008). To allow for biological variation (multiple patients as opposed to one patient repeated multiple times), a negative binomial distribution is used because it allows for “overdispersion” or the modeling of extra variation beyond that predicted by the Poisson distribution (McCarthy, Chen, & Smyth, 2012a). In the negative binomial model, the overdispersion parameter,  $\phi$ , controls the amount of variance beyond Poisson through the equation  $V = \mu + \phi\mu^2$  ( $\mu$  is the mean).



Competing methods for count modeling mostly differ in how they estimate this dispersion parameter. For example, trending dispersion as a function of the mean (Love, Huber, & Anders, 2014a), squeezing between the gene-specific and common (pooling across genes) dispersion (McCarthy et al., 2012a). Additionally, methods focus on other complicating data features such as outliers (Zhou, Lindsay, & Robinson, 2014) or excess zeroes (Van De Wiel et al., 2013). Variance modeling strategies differ from count modeling in that variance modeling ignores the correct distribution of the count data in favor of modeling the mean-variance relationship (Law, Chen, Shi, & Smyth, 2014). The idea here is to use established methods in microarray data analysis for RNA-seq. Additionally, the normal distribution, which is used heavily in microarrays, is much easier to work with and is better understood than a negative binomial model. However, these methods have not considered addressing gene covariance in the context of RNA-seq which could contribute a nontrivial amount to the variance estimation. This represents an opportunity to develop methods for addressing gene covariance in RNA-seq data analyses.

Methods that incorporate intergene correlation (broadly labeled gene set analyses) can improve detection of differentially expressed genes. For single gene analyses, sensitivities ranges from 50 % – 65%, when data is simulated under an independence assumption (Love, Huber, & Anders, 2014b; Zhou et al., 2014). Current simulation studies have shown that when there is unaccounted intergene correlation there is a greater number of false positives among the most significant genes (H. Zhang, Xu, Jiang, Hu, & Luo, 2015). Gene set analyses can improve the sensitivity and also the interpretability of these analyses. Gene set analyses can be partitioned into two types of tests based on the null hypothesis that they test: self-contained

and competitive. Self-contained tests compare if a collection of genes is differentially expressed between two groups (e.g. tumor versus normal), while a competitive test compares a collection of genes against all other genes not in that set (Goeman, Bu, Zurich, & Zu, 2007). An example of a competitive test is enrichment testing, where single gene test statistics are aggregated, and an example of a self-contained test is network analysis.

Addressing intergene correlations has been used extensively in differential expression analyses of microarrays. One of these competitive gene set approaches is gene set enrichment analysis (GSEA) (Tamayo, Steinhardt, Liberzon, & Mesirov, 2012), (Subramanian et al., 2005). GSEA addresses correlation among genes by organizing them into sets (such as through gene ontology, statistical clusters, or pathways from previous biological knowledge) and determines enrichment scores. Enrichment is a procedure which aggregates test statistics from a single gene analysis into one global statistic (through some mathematical function) to determine if grouping genes improves detection of differential expression. Correlation among genes is taken into account using resampling procedures and develops an empirical null distribution for hypothesis testing (Tamayo et al., 2012). GSEA splits one formal statistical modeling task into two stages. This is similar to a situation in longitudinal data analysis where repeated measurements are assumed (incorrectly) to be independent and then adjusted for post-hoc (Hardin & Hilbe, 2002). This discretization into two stages results in an efficiency (the rate of convergence to the standard error) loss when estimating standard errors and suggests that directly assessing correlation in one statistical model can improve the detection of differentially expressed genes by estimating standard errors that are closer to the true standard errors. Efficiency is important in genomics because the sample size is likely never sufficient

and thus, greater efficiency will translate to greater statistical power and better detection of differentially expressed genes.

A better approach is to organize genes into gene networks and address correlation among genes within each network. Multivariate (multiple response) regression models can be useful in this context because they are able to model gene expression levels of multiple genes in one regression, which allows for a direct assessment of the correlation among these genes. The methods of multivariate regression models vary based on the distribution of the response. For example, a multivariate linear model can be used with the assumption that the areas are distributed according to multivariate normal (Huang & Lin, 2013). This type of model has shown improvements in sensitivity over the GSEA procedure when analyzing microarray data and the authors argue that this translates to RNA-seq. Currently, no methods are available for the assumption of a multivariate negative binomial distribution (for count data in RNA-seq). This is potentially important because many of the single gene approaches rely on a negative binomial assumption. Mimicking this assumption for multivariate regressions seems logical for increasing sensitivity in RNA-seq analyses.

### **Public Health Significance**

Cancer is a major public health threat in the United States and learning more about the genetics of cancer is the first step to better treatment and management of the disease. Current RNA-seq analyses suffer from low sensitivity, preventing researchers from identifying important genes. Multivariate regression allows for direct assessments of correlation among genes and provides a useful framework for more sophisticated RNA-seq analyses. While multivariate regression for count data has been used in longitudinal analysis for clinical trials

data, it has not been recognized for their usefulness in RNA-seq analyses (Solis-Trapala & Farewell, 2005). Recognizing that longitudinal approaches can be repurposed for RNA-seq opens up many possibilities to improve sensitivity in RNA-seq analyses. Furthermore, the creation of more sensitive RNA-seq approaches can be applied broadly across all different cancer types. Using large databases, this allows for relatively quick scans to find important genes across a number of different cancers. By focusing on increasing sensitivity for a general method, a number of cancer genes can be discovered which have been missed by previous methodologies.

### **Specific Aims**

Despite the enormous investment in cancer genomics research, the majority of somatically disrupted genes are unknown. One of the reasons that researchers are unable to discover these genes is because statistical methods for locating these genes ignores any covariance between the gene expressions for a given pair of genes in a given population (e.g. tumor samples). The consequence of using these statistical models is that they incorrectly estimate the variance of gene expression, leading to decreased sensitivity and specificity. Relaxing this assumption, using a network of genes instead of single genes, I assume that genes are correlated within a network, but are (approximately) independent between networks. Developing a statistical model that directly incorporates this correlation among genes will likely increase specificity and sensitivity, compared to single gene methods, and improve the detection of these unknown genes being disrupted in cancers.

I propose that using gene networks, instead of single genes, and leveraging prior biological information in these networks via a multivariate regression model will enable the discovery of

new cancer genes. This is because a gene network approach will have greater sensitivity (while controlling specificity) than a single gene approach. I will develop a multivariate regression model (multiple responses) in the context of RNA-sequencing (RNA-seq) data, modeling the gene expression of each individual gene in a network as the response with tumor status (tumor vs. normal tissue) as the primary exposure. I expect statistical methods that integrate gene dependence into analyses will have greater statistical power in the presence of gene dependence and equivalent statistical power among genes with no dependence.

**Aim 1: Develop an algorithm for identifying differentially expressed gene networks**

A multivariate (multiple responses) regression model will be developed for gene networks in RNA-seq count data. This method will extend the univariate negative binomial model to enable an entire network of genes to be analyzed in one regression analysis while incorporating correlations among different genes in a gene network. This proposed method is expected to improve the sensitivity while controlling specificity at current levels.

**Aim 2: Empirically evaluate the proposed algorithm with current single-gene approaches**

Simulations will be conducted to assess the sensitivity and specificity of the proposed algorithm with single-gene approaches in the presence (absence) of correlation among genes. This will help to identify the strengths and weaknesses of various approaches and help to improve our methods.

**Aim 3: Identify differentially expressed gene networks and genes in primary breast tumors**

RNA-seq data of primary breast tumors will be downloaded from the Cancer Genome Atlas (TCGA) and genes will be organized into networks using Ingenuity Pathway Analysis (IPA),

a literature-driven database. These networks will be analyzed using the algorithm in Aim 1 and significance will be assessed at the false discovery rate (FDR) of 0.05. Although this analysis focuses on primary breast tumors, it is applicable to all other tumor types in TCGA (as well as for RNA-seq data for other diseases). The newly identified genes by this method will be carefully scrutinized and their biological significances will be assessed.

I propose a new statistical method to locate differentially expressed gene networks from RNA-seq data. Using prior biological information and gene structure, I expect the proposed method will have greater statistical power than single gene approaches. This implies that analyzing data with a gene network approach will identify new cancer genes to explain more of the total genetic variation in different cancers. The results will add to the current knowledge of disrupted genes in different cancers, providing a foundation for the development of new treatments.

## **Incorporating Correlation Structure into the Analysis of RNA-seq Data**

### **Statistical Applications in Genetics and Molecular Biology**

Author List: Aschenbrenner, Andrew; Loose, David; Mullen, Patricia Dolan; Fu, Yun-Xin

#### **Abstract:**

RNA-sequencing (RNA-seq) experiments are becoming the standard for gene expression studies. Typically, analyzing these datasets involves using a count distribution such as negative binomial (NB) with genewise hypothesis testing. However, these approaches ignore the correlation among genes. Simulation studies show that ignoring correlation among genes results in a lack of conservation of type I error; greater error for positive correlation and less error for negative correlation. Additionally, we present a general framework for finding correlation structures with varying degrees of complexity, creating a compromise between exchangeable and unstructured correlation structures. This allows researchers to customize the estimated correlation structure that is incorporated into the data analysis and reducing the chance of a misspecified correlation structure.

#### **Introduction**

RNA-sequencing (RNA-seq) has proven to be a useful tool for investigations of gene expression due to its high-resolution view of the entire transcriptome. Each RNA-seq experiment follows a general workflow: random fragmentation of mRNA from samples, reverse transcription of fragmented mRNAs to create complementary DNAs (cDNAs), PCR amplification and sequencing of cDNAs to a list of subsequences or reads, mapping of these reads to genes via a reference genome, and creation of a flat-file table that contains counts for

each gene, for each sample (Garber, Grabherr, Guttman, & Trapnell, 2011). Each count in a given gene correlates with that gene's expression level. This resulting table of gene expression counts is often the subject of downstream statistical analyses.

Initially, data analysis focused on individual gene expression. For example, several count distributions have been used to model the gene expression (due to gene expression data is presented as a matrix of counts), for example using the Poisson (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008), (Marioni et al., 2008) or the negative binomial distribution (McCarthy, Chen, & Smyth, 2012b), (Love et al., 2014b), (Lund, Nettleton, McCarthy, & Smyth, 2012). However, this approach has a few shortcomings. A typical analysis resulted in a large list of differentially expressed genes, giving results which were difficult to interpret. Also, individual genes do not act alone, but work in tandem with other genes in complex pathways. This implies that a univariate statistical analysis may inappropriately reflects the biological reality, thus using pathways as the unit of analysis is preferable. In this setting, a multivariate statistical model is more suitable. Some examples include the N-statistic, a multivariate nonparametric method, (Larson & Owen, 2015) (Baringhaus & Franz, 2004) and ROAST (Wu et al., 2010), which uses multivariate linear regression to test for differential expression. In fact, multivariate techniques have been shown to achieve higher power and conserve type I error when compared to aggregating univariate analyses (Rahmatallah, Emmert-Streib, & Glazko, 2016).

While these techniques address using the pathway as the unit of analysis, they do not explicitly incorporate the relationship between genes in a given pathway. For RNA-seq data, there are some gene pairs with non-negligible correlation as well a number of gene pairs with



negligible correlation. This is important to add such features to the data analysis and failure to do so may lead to errors in statistical inference. Additionally, the correlation structures are typically not simple and it is important to choose a statistical model with sufficient flexibility. Certain models such as the multivariate negative binomial (Solis-Trapala & Farewell, 2005) have a restrictive implied correlation structure, only allowing for positive correlation, equal among all genes. Indeed, most multivariate parametric models for count data fall into similar categories.

To address the concern of correlation among genes, we propose to use Generalized Estimating Equations (GEE) (Li & Chan, 2006) (K.-L. Liang & Zeger, 1986; Wang, 2014). In GEE, a “working” correlation structure is used, which allows for a flexible correlation structure between the genes in a pathway. A new intermediate correlation structure is also developed to handle a mixed correlation structure, which is more characteristic of RNA-seq data. Simulations are conducted to compare different correlation structures and how correlation among genes impacts the type I error and statistical power. Also, a statistical test for independence of genes in a pathway is formulated and simulations are used to identify the statistical power and type I error in the presence of different correlation structures.

Using appropriate correlation matrices for a gene network helps to safeguard against errors in statistical inference and makes a more statistically sound model. Using an unstructured correlation matrix may be impractical for most datasets because of the number of parameters that need to be estimated. Often, the sample size is insufficiently large to use unstructured and the estimates are unreliable. Alternatively, exchangeable uses only one parameter, which may be insufficient for more complicated structure, which are more likely in

real datasets. An intermediate correlation, with more parameters than exchangeable and fewer parameters than unstructured, helps to overcome these challenges in the estimation and incorporation of appropriate correlation structures in RNA-seq data analysis. Increased flexibility in estimated correlation structures allows for less of a chance of a misspecified correlation matrix. Furthermore, a general procedure for determining correlation matrices allows researchers to customize their correlation estimation based on their dataset.

## Methods

### *Generalized Estimating Equations (GEE)*

Generalized Estimating Equations (GEE) is an estimation technique for analyzing non-normal correlated data, for example clustered data. A gene network can be interpreted as a cluster of genes with connections between genes denoting some correlation while unconnected genes having approximately zero correlation.

Let  $y_i$  be the gene expression measurements for each of  $g$  genes in a given gene network for  $i = 1, \dots, n$  patient samples such that  $y_i$  is a  $g \times 1$  vector. In contrast to the typical gene expression matrix format, this notation is for a long format where the columns of such a matrix are stacked on top of each other. Additionally, let  $\mu_i = E(Y_i)$ , be the  $g \times 1$  mean vector for each of the  $g$  genes in a given gene network. Under this notation,  $y_{ij}$  refers to the gene expression of sample  $i$  for gene  $j$  and similarly,  $\mu_{ij}$  represents the mean parameter for sample  $i$  and gene  $j$ . In this scenario, we assume that patient samples are independent, but genes within a gene network are correlated with one another (to some degree). Instead of using an assumed probability distribution and deriving its likelihood, GEE uses the quasi-likelihood, which only requires specifying the relationship between the mean and the variance. For RNA-seq data,

the negative binomial distribution is the most common for modeling the expression of a single gene. We can use the same mean-variance structure as the univariate negative binomial distribution:

$$E(Y_{ij}) = \mu_{ij} \quad (1)$$

$$v(\mu_{ij}) = \mu_{ij} + \omega\mu_{ij}^2 \quad (2)$$

Where,  $\omega$  represents the negative binomial dispersion parameter. Using this relationship, the covariance matrix of  $y_i$  (a  $g \times g$  matrix) can be calculated as  $V_i = \phi A_i^{1/2} R_i A_i^{1/2}$ , where  $\phi$  is a scale parameter to be estimated,  $A_i$  is a  $g \times g$  diagonal matrix with  $v(\mu_{ij})$  as the diagonal elements, i.e.  $Diag(v(\mu_{i1}), v(\mu_{i2}), \dots, v(\mu_{ig}))$ , and  $R_i$  is a  $g \times g$  matrix that represents the “working correlation” structure, which describes the relationship between the genes within a gene network. The correlation structure is chosen by the researcher and is “working” in the sense that is an initial guess, which may be misspecified. Simple correlation structures such as exchangeable (all correlation values are the same). Alternatively, an unstructured correlation structure, in which all correlation parameters are unique, requires a large sample size relative to the size of the gene network ( $\sim 30$  genes per network) and is not feasible except for very small gene networks with large sample sizes.

Table 1: Example Correlation Structures

Correlation Structure	Example Correlation Matrix	Estimator
Independent	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	NA
Exchangeable	$\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$	$\hat{\alpha} = \frac{1}{(N' - p)\phi} \sum_{i=1}^K \sum_{j \neq k} e_{ij} e_{ik}$ $N' = \sum_{i=1}^K g * (g - 1)$
Unstructured	$\begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & 1 & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & 1 \end{pmatrix}$	$\hat{\alpha}_{jk} = \frac{1}{(K - p)\phi} \sum_{i=1}^K e_{ij} e_{ik}$

*Framework for Correlation Structures*

An exchangeable correlation structure represents one parameter being estimated and the correlation between each gene pair is the same. Similarly, an unstructured correlation structure estimates a different correlation parameter for each different gene pair. These structures represent two extremes for correlation structures. Exchangeable structures may not be as useful if the true correlation matrix between genes contains a mix of signs. An unstructured correlation matrix is limited in applicability because the sample size needs to be sufficiently larger than the number of genes in a network in order to obtain adequate estimates of these correlations. Using an unstructured correlation for larger networks can result in an

unstable correlation matrix, which leads to numerical divergence and consequently reduced power.

In practice, Pearson residuals are used to estimate the individual correlation parameters. For unstructured, each correlation parameter is estimated by  $\hat{\alpha}_{jk} = \frac{1}{(K-p)\phi} \sum_{i=1}^K e_{ij}e_{ik}$ . In contrast, exchangeable estimates its one correlation parameter as  $\hat{\alpha} = \frac{1}{(N'-p)\phi} \sum_{i=1}^K \sum_{j \neq k} e_{ij}e_{ik}$   $N' = \sum_{i=1}^K g * (g - 1)$ . Exchangeable can be interpreted as an average of the all the unstructured parameters. It is useful to identify a procedure for estimating intermediate correlation structures; one with more parameters than exchangeable, but fewer than unstructured. The main idea is to sort Pearson residuals into  $k$  quantiles and average across these quantiles. For example, for  $k = 3$ :

1. Calculate the mean Pearson residual products  $\hat{e}_{jk} = \frac{1}{n} \sum_{i=1}^n e_{ij}e_{ik}$ .
2. Sort products into positive and negative. For each, sort into three quantiles.
3. For each of these quantiles calculate the average to get parameter estimates  $\hat{\alpha}_m$ ,  $m = 1, \dots, 6$ .

The result of this estimation procedure will result in six different parameters; three of which are positive and three which are negative. Additionally, we need to map each parameter to its correct location (gene pair) in the correlation matrix. To do this, the parameter which minimizes the distance of each parameter and the mean Pearson residual products is selected for that particular pair:

$$\min|\hat{\alpha}_m - \hat{e}|$$

For  $k=1$ , this procedure reduces to the exchangeable model (an average of all the Pearson residuals) and if  $k$  is equal to the number of off-diagonal elements, this reduces to the unstructured model. A general procedure can be defined by sorting Pearson residual products into  $k$  quantiles.

### *Estimation and Hypothesis Testing*

The primary purpose of using this technique is to assess differential expression at the gene network level. To do this, a regression model is fit with a log link function:

$$\log \mu_{ij} = \beta_0 + \beta_1 X_1 + \sum_{k=2}^l \beta_k X_k \quad (3)$$

$X_1$  denotes the group for each of samples (for example,  $X_1$  might be a binary vector with zero value denoting normal samples and one value denoting tumor samples),  $X_k$  are additional covariates which can be adjusted for. Estimating the regression coefficients will give the differential expression ( $\hat{\beta}_1$ ) adjusted for any additional covariates. In order to estimate the regression coefficients, the following estimating equation should be solved:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \quad (4)$$

where  $D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$ . Since  $V_i$  is also a function of the parameters  $\omega$  and  $\phi$ , these will be estimated separately and replaced in the estimating equation with their estimated values.  $\omega$  is the negative binomial dispersion parameter and in the context of single gene RNA-seq analyses, estimates have been developed, for example the edgeR package (McCarthy et al., 2012b;

Robinson, McCarthy, & Smyth, 2010). The estimate of  $\omega$ ,  $\hat{\omega}$ , is obtained by taking the average of each tagwise dispersion estimate for each gene in a given gene network. Additionally, the scale parameter is estimated as:

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^n \sum_{j=1}^g \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}} \quad (5)$$

$N = n * g$  and  $p$  is the number of regression coefficients,  $\beta$ . Both  $\hat{\omega}$  and  $\hat{\phi}$  replace  $\omega$  and  $\phi$  in  $V_i$  and the equation in (4) is solved for  $\beta$ , yielding estimates  $\hat{\beta}$ . The covariance matrix for  $\hat{\beta}$  can be calculated as:

$$\text{Cov}(\hat{\beta}) = V_0^{-1} V_1 V_0^{-1} \quad (6)$$

$$V_0 = \sum_{i=1}^n D_i^T V_i^{-1} D_i \quad (7)$$

$$V_1 = \sum_{i=1}^n D_i^T V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i \quad (8)$$

where  $\text{Cov}(Y_i) = \hat{r}_i \hat{r}_i^T$ ,  $\hat{r}_i = Y_i - \hat{\mu}_i$ . According to (K.-Y. Liang & Zeger, 1986),  $\hat{\beta}$  is distributed asymptotically normal with mean  $\beta$  and covariance matrix given in (6). Based on these asymptotic results, a Wald test can be used to test for differential expression at the gene network level:

$$Z = \frac{\hat{\beta}}{\sqrt{\text{Cov}(\hat{\beta})}} \sim N(0, 1) \quad (9)$$

here, the test statistic is for a particular  $\beta$  and the denominator represents the standard error, which can be obtained from the corresponding element of the diagonal of the covariance matrix. The covariates in the regression model may also be tested if these parameters are of interest, but in general, the primary focus is  $\hat{\beta}_1$ , the differential expression effect.

The aforementioned procedure details how to test for differential expression for a given gene network. For a list of gene networks, this process can be repeated multiple times, however, additional false positives may be incurred. False discovery rates (FDRs) can be used in lieu of the p-values obtained from the Wald test to control false positives (Benjamini & Hochberg, 1995; Storey, 2010).

#### *Test for Gene Set Independence*

For a given gene network, researchers may be interested if the genes in a particular gene network are correlated or are independent. Under the null hypothesis (a gene network is independent), we expect that each pairwise correlation coefficient is approximately zero. Formally, this leads to the null hypothesis:  $H_0: \mathbf{R} = \mathbf{I}$ . In order to test this hypothesis, the estimated correlation matrices from the expression matrix of the gene network are compared to an estimated correlation matrix from a simulated dataset, where the underlying correlation is zero. A dataset simulated under the null hypothesis is preferable as a comparison rather than the identity matrix because realistically true independent gene networks may still have a small degree of correlation and this helps not to overinflate the statistical power of the test.



The test statistic we propose to use is the distance test, initially used to compare correlation matrices across groups (for example, tumor versus normal) in the context of co-expression analyses(Choi & Kendzierski, 2009). For this application, the perspective is to estimate the correlation matrix of the gene network and then compare it to the estimated correlation matrix of a null dataset. This test statistic is:

$$D = \sqrt{\frac{1}{g * (g - 1)} \sum_{i=1}^{g*(g-1)} (\rho_{i1} - \rho_{i0})^2} \quad (10)$$

Where  $g$  is the number of genes in the gene network,  $\rho_{i1}$  is a vector of the correlation coefficients for all the gene pairs from the dataset, and  $\rho_{i0}$  is the correlation vector from the null dataset. Also, setting  $\rho_{i0} = 0$  reduces the test statistic  $D$  to the case where the identity matrix is used under the null hypothesis. Since the test statistic does not follow a simple distribution, p-values are obtained through the following permutation approach. Class labels are assigned to the correlation matrix from the data and the correlation matrix from the simulated dataset where the underlying correlation is zero. These labels are shuffled  $k$  times, for example,  $k = 1,000$ . For each shuffle, the test statistic is calculated. The p-value is calculated as the number of these shuffled test statistics that are greater than the observed test statistic (non-shuffled version) divided by  $k$ .

## Numerical Studies

In order to compare the efficacy of GEE for RNA-seq data analysis and the measure the impact correlation has on statistical inference, we conduct a simulation study. For this simulation study, the GEE method will be compared against different correlation structures: independence, exchangeable, intermediate, and unstructured. In the context of the GEE method, two types of analyses are pursued: a type I error analysis and a power analysis. The type I error analysis uses simulated data to estimate the type I error rate of each particular method. The power analysis estimates the statistical power for each method over an array of alternative hypotheses. Additionally, these analyses are also performed for the statistical test of gene set independence.

### *Simulation of Data*

For each gene network, a  $g \times n$  expression matrix is generated with  $g$  genes in the given network and  $n$  samples. The expression matrix is a matrix of counts with mean  $\boldsymbol{\mu}$  (a vector of means corresponding to each individual gene in the gene network), dispersion  $\boldsymbol{\omega}$  (a vector of dispersions), and  $\mathbf{R}$ , a correlation matrix for the genes in the gene network. This expression matrix can be generated using the following algorithm:

1. Generate data from a multivariate normal distribution with sample size  $n$ , mean vector  $\mathbf{0}$ , and correlation matrix  $R$ .
2. For each column, calculate the standard normal cdf,  $\Phi$ , of the data values.
3. Calculate the quantile function of the negative binomial distribution with mean  $\mu$  and dispersion  $\omega$ .

This algorithm uses an inverse cdf technique in order to generate correlated negative binomial random variables that has been used to generate correlated random variables for a number of other distributions such as Poisson. This technique is attractive because of its simplicity, relative speed, and a fairly accurate empirical correlation matrix.

### *Type I Error Analysis*

To simulate gene networks under the null hypothesis the means are kept the same for both groups (e.g tumor and normal,  $\mu_{tumor} = \mu_{normal}$ ). We investigate nine different correlation scenarios for each positive, negative, and mixed correlation structures with low, medium, and high strength (magnitude) of correlation. Pairwise correlation values are generated from a uniform distribution with range according to strength of correlation. Weak correlations are generated as  $U(0, 0.1)$ , medium correlations are generated as  $U(0.2, 0.4)$ , and strong correlations are generated as  $U(0.5, 0.8)$ . Correlation matrices which are not positive definite are replaced with the nearest positive definite matrix. In the case of strong negative correlation, a few values are chosen to be highly negative while the rest are zero. This ensures positive definiteness while still having some highly negative values. For each of these correlation scenarios, we simulate data for gene networks of size 10, 30, and 50 genes. For each of these

Table 2: Type I error rates for different GEE models.

Correlation	Independent	Exchangeable	Intermediate	Unstructured
(+) weak	0.077 (10 Genes)	0.045	0.043	0.040
	0.150 (30 Genes)	0.053	0.051	0.047
	0.195 (50 Genes)	0.043	0.052	0.059
(+) medium	0.249	0.052	0.054	0.057
	0.468	0.056	0.052	0.049
	0.561	0.049	0.050	0.051
(+) strong	0.434	0.049	0.052	0.053
	0.628	0.077	0.054	0.045
	0.718	0.087	0.057	0.054
(-) weak	0.008	0.065	0.064	0.064
	0.006	0.075	0.065	0.067
	0.04	0.053	0.052	0.051
(-) medium	0.008	0.065	0.064	0.060
	0.007	0.066	0.065	0.060
	0.010	0.059	0.055	0.052
(-) strong	0.008	0.060	0.058	0.060

	0.003	0.045	0.055	0.060
	0.010	0.055	0.057	0.062
(mix) weak	0.044	0.045	0.047	0.048
	0.053	0.053	0.052	0.057
	0.072	0.068	0.059	0.057
(mix) medium	0.078	0.054	0.049	0.045
	0.089	0.051	0.052	0.056
	0.120	0.051	0.059	0.071
(mix) strong	0.096	0.045	0.049	0.051
	0.143	0.058	0.056	0.057
	0.140	0.05	0.05	0.05

datasets, type I error is estimated by the percentage of p-values  $< 0.05$ . This factorial design helps to identify how different correlation structures and gene network size impact type I error.

When using the independence correlation structure, the type I error is not conserved. For positive correlation, the type I error is increased, while for negative correlation, the type I error is decreased, relative to the p-value threshold level. In the case of mixed correlations, the type I error is increased, but by a smaller degree than for positive correlation. For both the exchangeable and unstructured correlation structure, type I error levels are roughly conserved for all the different correlation structures. For a gene network of larger size, the errors from independence correlation structure are larger. For example, for positive correlation the type I error is increased for a larger gene network, while for negative correlation the type I error is

smaller. There is a similar pattern for mixed correlation structures. For larger gene networks, both the exchangeable and unstructured conserve the type I error.

### *Power Analysis*

To simulate gene networks under the alternative hypothesis, the means in one group are larger than the other, by a multiplicative factor (e.g. tumor and normal,  $FC * \mu_{tumor} = \mu_{normal}$ , where  $FC$  is the multiplicative factor). Similar to the type I error analysis, the same correlation structures and gene network sizes are used. Additionally, an empirical p-value threshold is used to determine significance, to guard against inflating power from higher type I error rates. Usually, a threshold level of  $\alpha = 0.05$  is used to determine significance for p-values. However, some methods may have more or less type I error than the 0.05 level, implying the method is overly liberal or overly conservative. In these cases, a different threshold value should be estimated to ensure that type I error is conserved. Let  $p$  denote the p-value,  $N$  denote the number of tests conducted (for example, number of simulations), and  $\alpha'$  denote the new threshold value. The updated threshold level can be calculated by solving the following equation for  $\alpha'$ :

$$\frac{1}{N} I(p < \alpha') = \alpha \quad (11)$$

Where  $I(\cdot)$  represents the indicator function. The left hand side can be interpreted as the percentage of p-values below the threshold  $\alpha'$ . This equation can be numerically solved, for

example, by using the function `uniroot()` in R. Using this new threshold, we estimate statistical power as the percentage of p-values  $< \alpha'$ .

For each individual correlation structure, power across the different models is roughly the same. This implies that any unadjusted power gains from independent model are mostly errors, resulting from a greater type I error. When using the correct p-value threshold level, all of the models will likely give similar results. However, the correct p-value threshold is unknown, making the independence model a riskier choice than models which adjust for correlation. Additionally, all of the models have greater power for larger gene networks. This observation is consistent with other work on gene networks in both RNA-seq and microarrays.

#### *Test of Gene Set Independence: Type I Error and Power*

It is important to benchmark new statistical tests to evaluate their efficacy of achieving their intended goal. To accomplish this, 1000 datasets are generated with network size 10, 30, and 50. The test statistic  $D$  is applied to each dataset and p-values are obtained using 1000 permutations. For type I error analysis, datasets are generated under the null hypothesis (i.e. no difference in correlation structure), while for power analysis, datasets are generated under nine different correlation scenarios (positive, negative, and mixed; weak, medium, and strong).

Table 3: Power analysis for different GEE models.

Correlation	Method	DE = 1.2	DE = 1.5	DE = 1.8	DE = 2	DE = 3
(+) weak	Independent	0.407 (10 genes)	0.961	1	1	1
		0.679 (30)	0.998	1	1	1
		0.771 (50)	1	1	1	1
	Exchangeable	0.409	0.960	1	1	1
		0.657	0.998	1	1	1
		0.753	1	1	1	1
	Intermediate	0.410	0.960	1	1	1
		0.669	0.999	1	1	1
		0.763	1	1	1	1
	Unstructured	0.409	0.960	1	1	1
		0.608	0.998	1	1	1
		0.596	0.998	1	1	1
(+) medium	Independent	0.209	0.679	0.935	0.985	1
		0.221	0.701	0.950	1	1
		0.259	0.800	0.982	1	1
	Exchangeable	0.197	0.640	0.920	0.977	1
		0.202	0.691	0.927	1	1
		0.213	0.659	0.918	0.975	1
	Intermediate	0.201	0.662	0.924	0.981	1
		0.207	0.697	0.941	1	1
		0.210	0.699	0.949	1	1
	Unstructured	0.180	0.628	0.913	0.975	1
		0.207	0.683	0.938	1	1
		0.196	0.598	0.898	0.960	1
(+) strong	Independent	0.204	0.503	0.797	0.894	0.998
		0.152	0.427	0.682	0.793	1
		0.133	0.469	0.768	0.852	1
	Exchangeable	0.125	0.393	0.684	0.801	0.993
		0.105	0.309	0.587	0.745	0.990
		0.094	0.272	0.487	0.614	0.975
	Intermediate	0.123	0.402	0.701	0.823	0.996
		0.111	0.423	0.756	0.856	1
		0.102	0.487	0.802	0.902	1
	Unstructured	0.120	0.353	0.658	0.796	0.993
		0.127	0.37	0.666	0.777	0.995
		0.107	0.343	0.575	0.678	0.914
(-) weak	Independent	0.534	0.997	1	1	1
		0.996	1	1	1	1
		1	1	1	1	1
	Exchangeable	0.531	0.997	1	1	1
		0.997	1	1	1	1
		1	1	1	1	1
	Intermediate	0.532	0.998	1	1	1
		0.995	1	1	1	1
		1	1	1	1	1
	Unstructured	0.517	0.997	1	1	1
		0.982	1	1	1	1



		1	1	1	1	1
(-) medium	Independent	0.710	1	1	1	1
		0.996	1	1	1	1
		1	1	1	1	1
	Exchangeable	0.713	1	1	1	1
		0.997	1	1	1	1
		1	1	1	1	1
	Intermediate	0.716	1	1	1	1
		0.998	1	1	1	1
		1	1	1	1	1
	Unstructured	0.693	1	1	1	1
		0.982	1	1	1	1
		0.998	1	1	1	1
(-) strong	Independent	0.717	1	1	1	1
		0.997	1	1	1	1
		0.999	1	1	1	1
	Exchangeable	0.714	1	1	1	1
		0.993	1	1	1	1
		0.998	1	1	1	1
	Intermediate	0.716	1	1	1	1
		0.995	1	1	1	1
		0.999	1	1	1	1
	Unstructured	0.669	1	1	1	1
		0.982	1	1	1	1
		0.995	1	1	1	1
(mix) weak	Independent	0.483	0.990	1	1	1
		0.875	1	1	1	1
		0.973	1	1	1	1
	Exchangeable	0.480	0.990	1	1	1
		0.877	1	1	1	1
		0.968	1	1	1	1
	Intermediate	0.482	0.992	1	1	1
		0.879	1	1	1	1
		0.973	1	1	1	1
	Unstructured	0.454	0.985	1	1	1
		0.815	1	1	1	1
		0.934	1	1	1	1
(mix) medium	Independent	0.380	0.954	1	1	1
		0.791	1	1	1	1
		0.907	1	1	1	1
	Exchangeable	0.382	0.948	1	1	1
		0.789	1	1	1	1
		0.906	1	1	1	1
	Intermediate	0.384	0.956	1	1	1
		0.792	1	1	1	1
		0.913	1	1	1	1
	Unstructured	0.443	0.943	1	1	1
		0.772	1	1	1	1
		0.832	1	1	1	1
(mix) strong	Independent	0.392	0.938	1	1	1

		0.689	1	1	1	1
		0.707	1	1	1	1
	Exchangeable	0.396	0.940	1	1	1
		0.665	1	1	1	1
		0.678	1	1	1	1
	Intermediate	0.399	0.951	1	1	1
		0.672	1	1	1	1
		0.692	1	1	1	1
	Unstructured	0.400	0.955	1	1	1
		0.677	1	1	1	1
		0.718	1	1	1	1

Table 4: Type I error rate and power benchmarking for test of gene set independence.

	10 Genes	30 Genes	50 Genes
Type I Error	0.046	0.052	0.040
(+) weak	0.114	0.334	0.655
(+) medium	1.000	1.000	1.000
(+) strong	1.000	1.000	1.000
(-) weak	0.140	0.159	0.185
(-) medium	0.512	0.433	0.653
(-) strong	0.542	0.832	0.995
(mix) weak	0.116	0.328	0.641
(mix) medium	1.000	1.000	1.000
(mix) strong	1.000	1.000	1.000

## Discussion

We have applied GEE to analyze gene networks for RNA-seq and developed a framework for identifying intermediate correlation structures that align more closely with real RNA-seq data. Using an independent correlation structure results in an increased type I error when the true underlying correlation matrix contains positive correlations. Similarly, when the true underlying correlation matrix is negative, type I error is decreased. Incorporating correlation into the data analysis acts as a safeguard to ensure type I error is conserved. The developed framework for estimating intermediate correlation structures allows researchers to incorporate correlation structures of varying complexity into the data analysis. The tradeoff is simpler correlation structures will have quicker convergence and more reliable estimation, while more complicated structures may risk divergence or be slower to converge. These choices should be made within the context of sample size; a larger sample size may be able to afford a more complicated correlation structure. Using this framework, researchers are no longer forced to choose between two extremes when incorporating correlation into the data analysis, which leads to a more accurate model of the data.

The development and incorporation of an intermediate correlation structure unlocks the potential of analyzing clustered data with mixed correlation structures. Typically, GEE analyses of clustered data use an exchangeable structure due to its simplicity. This paradigm likely exists because of its initial application to longitudinal datasets, where measurements at follow-up times are positively correlated with one another. While the usage of a robust variance estimator could adjust for a misspecified correlation structure post-hoc, this would be less efficient and conceptually sound. Creating better initial guesses makes the technique more

statistically sound and beneficial when the sample size is lower. Additionally, there are likely other datasets with count (or other data type) outcomes which could benefit from the intermediate correlation structure.

Another approach to modeling correlation in gene networks for RNA-seq is to use random effects via a mixed model. Mixed models have been applied for count data in analysis of microbiomes (X. Zhang et al., 2017). The fundamental difference between mixed models and GEE is a difference in interpretation. Estimates from a mixed model are individual specific, while estimates from GEE indicate the population average. Also, these estimates are usually different, implying that they could lead to the discovery of different gene networks.

The GEE method also has some shortcomings. The numerical algorithm for fitting the regression models can sometimes diverge leading to inaccurate estimates and inferences. We observed that this happens when the correlation structure is misspecified, specifically for high positive correlation; an issue that has been previously reported (Sutradhar & Das, 1999). These divergence issues are often exacerbated for larger gene networks. Model divergence can also occur in cases where the number of genes in a network are large relative to the sample size, specifically for unstructured correlation. These issues limit the applicability of the method for the purpose of analyzing RNA-seq, however, the intermediate structure attempts to mitigate these issues. Nevertheless, there are scenarios where the usage of GEE may not be correct.

There are a couple of extensions of this particular model that researchers could pursue in the context of RNA-seq analysis. First, we could consider the regression coefficients of interest ( $\beta$ ) to be only positive. The individual gene coefficients are rarely all of the same sign; many gene networks have a mix of up-regulated and down-regulated genes. This mixing

likely shrinks the coefficient towards zero and underestimates the true coefficient. Restricting it to only be positive may be closer to the true value. Second, methods could be developed to average expert opinion and/or network connections with RNA-seq data. The current model supports entering in a user-defined correlation structure, but suffers some of the same issues as the unstructured correlation model. Incorporating this information in the estimation process is intuitively appealing and may offer improved power.

**Acknowledgements:**

Andrew Aschenbrenner was funded by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP160015) and the NIH Pre-doctoral Traineeship in Biostatistics (grant # 2T32GM074902-06). We would like to thank Stephen Daiger and Momiao Xiong for their helpful comments on written drafts.

## References

- Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1), 190–206. [https://doi.org/10.1016/S0047-259X\(03\)00079-4](https://doi.org/10.1016/S0047-259X(03)00079-4)
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B ( Methodological ), Vol . 57 , No . 1 Published by : Wi. *Journal of the Royal Statistical Society Series B*, 57(1), 289–300.
- Choi, Y., & Kendzierski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21), 2780–2786. <https://doi.org/10.1093/bioinformatics/btp502>
- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477. <https://doi.org/10.1038/nmeth.1613>
- Larson, J. L., & Owen, A. B. (2015). Moment based gene set tests. *BMC Bioinformatics*, 16(1), 1–17. <https://doi.org/10.1186/s12859-015-0571-7>
- Li, Y. P., & Chan, W. (2006). Analysis of longitudinal multinomial outcome data. *Biometrical Journal*, 48(2), 319–326. <https://doi.org/10.1002/bimj.200510187>
- Liang, K.-L., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear

models. *Biometrika*, 73(1), 13–22.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.  
<https://doi.org/10.1186/s13059-014-0550-8>

Lund, S. P., Nettleton, D., McCarthy, D. J., & Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. In *Statistical applications in genetics and molecular biology* (Vol. 11).  
<https://doi.org/10.1515/1544-6115.1826>

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9), 1509–1517. <https://doi.org/10.1101/gr.079558.108>

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

Mortazavi, A., Williams, B. a, McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628.  
<https://doi.org/10.1038/nmeth.1226>

Rahmatallah, Y., Emmert-Streib, F., & Glazko, G. (2016). Gene set analysis approaches for RNA-seq data: Performance evaluation and application guideline. *Briefings in Bioinformatics*, 17(3), 393–407. <https://doi.org/10.1093/bib/bbv069>



- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Solis-Trapala, I. L., & Farewell, V. T. (2005). Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in Medicine*, 24(16), 2557–2575. <https://doi.org/10.1002/sim.2121>
- Storey, J. D. (2010). False Discovery Rates. *Princeton University, Princeton, USA*, (January), 1–7. <https://doi.org/10.1198/016214507000000941>
- Sutradhar, B., & Das, K. (1999). On the Efficiency of Regression Estimators in Generalised Linear Models for Longitudinal Data. *Biometrika*, 86(2), 459–465.
- Wang, M. (2014). Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. *Advances in Statistics*, 2014, 1–11. <https://doi.org/http://dx.doi.org/10.1155/2014/303728>
- Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M. L., Visvader, J. E., & Smyth, G. K. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17), 2176–2182. <https://doi.org/10.1093/bioinformatics/btq401>
- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1), 1–10. <https://doi.org/10.1186/s12859-016-1441-7>

## JOURNAL ARTICLE

### **A Gene Network Analysis of Primary Breast Tumors in RNA-seq**

#### **Genetic Epidemiology**

Author List: Aschenbrenner, Andrew; Loose, David; Mullen, Patricia Dolan; Fu, Yun-Xin

#### **Abstract:**

Differential expression of RNA-seq data can be useful for generating hypotheses about genes that important in diseases. In contrast to traditional approaches, analyzing data at the gene network level first allows researchers to incorporate correlation among genes in the gene network into the data analysis. This helps to conserve type I error, preventing potential false positives or false negatives. Using a breast cancer dataset from the Cancer Genome Atlas (TCGA), gene networks are analyzed and subsequently, the individual genes from significant networks are analyzed. This approach leads to the identification of unreported differentially expressed genes as well as previously reported genes. Usage of a gene network approach initially in data analysis can act as a safeguard against potential statistical inferential errors and is advised when analyzing data.

#### **Introduction**

Breast cancer continues to be a threat to public health with new cases totaling approximately 250,000 and 41,000 deaths per year (Siegel, Miller, & Jemal, 2016). While there have been many research efforts dedicated to uncovering the biology of breast cancer and how this can translate into treatments, a number of tumors remain untreatable due the lack of complete knowledge of how different breast tumors work. To close these gaps in knowledge, additional investigations must be undertaken to elucidate the molecular changes

and functions in breast tumors. One such method for doing this is through gene expression studies.

Gene expression studies measure the amount of mRNA in different genes for a number of samples. In the context of breast cancer, often these samples will be derived from tumor and normal breast tissue. Comparing expression levels between tumor and normal samples help to identify which genes may expression levels that are different between these two types of samples, generating hypotheses about which genes may be important in tumors (a process called differential expression analysis). Genes which are expressed differently in the two types of samples may have a particular function that the tumor is trying to exploit. For example, genes that control regulate cell replication may be overexpressed in tumor tissue.

Recently, gene expression studies use RNA-sequencing in order to measure the mRNA in samples (Shendure, 2008). This technology allows for a coverage of the entire genome with greater accuracy compared to the older technology of microarrays. While the strength of gene expression studies is in the broad scan of a large number of genes, the large number of genes tested for differential expression can equally be a curse. Such an analysis may be prone to errors such as false positives or negatives because of the paradigm of testing many genes simultaneously and ignoring structures within the data, such as correlations among genes. While correcting for multiple comparisons helps to adjust for testing many hypotheses, correlations among genes are typically ignored in analyses. The consequence of ignoring such features of the data can result in a greater number of false positives or negatives (Efron & Tibshirani, 2007), (Schaalje & Butts, 1993), (Gatti, Barry, Nobel, Rusyn, & Wright, 2010). This means that some genes will seem to be important but further investigations will be a waste

of resources (false positive), while some genes may be clinically relevant, but remain undiscovered (false negatives). False negatives can be particularly nefarious because the undiscovered connections of these genes with the disease (such as breast cancer) can prevent the development of new treatments. To correct for these limitations, we use a gene network approach to analyze gene expression data. A network is collection of genes that are connected to each other. Generally, genes within a given network are hypothesized to have some amounts of correlation with other genes in the network. Analyzing these networks instead of each gene separately allows researchers to consider the interrelationships between genes when conducting a differential expression analysis. Additionally, using networks instead of genes aids in the issue of multiplicity (testing many genes at once) and leads to a balance between fine resolutions of differential expression in genes, while still incorporating the interdependence genes may have with one another.

## **Methods**

The RNA-seq data used was obtained from the Cancer Genome Atlas (TCGA) Research Network: <http://cancergenome.nih.gov/>. Files were aggregated into a single expression table with genes on the rows and samples on the columns. Corresponding patient clinical data was downloaded and matched to the samples in the expression table for analyses with covariates. A total of  $n = 96$  matched samples are used in analyses, where each tumor sample is matched on patient to a control sample. That is, each patient has a tumor sample and a sample of normal tissue. Matched samples represent a fraction of the samples available but

are used because of the availability of tissue sample controls and a matched design helps to adjust for patient-to-patient heterogeneity.

Prior to data analysis, the expression data in raw count form must be normalized and sorted into gene networks. The raw counts can take a large range of integer values, leading to a scale problem. Additionally, gene counts are proportional to the gene length, with longer genes having higher counts potentially biasing statistical inference. To overcome such a problem, we transform the data through a common scaling procedure called reads per kilobase of transcript, per million (RPKM) (Mortazavi et al., 2008):

$$RPKM = \frac{n_g \times 10^9}{l_g \times N}$$

Where  $n_g$  is the number of reads mapped to gene  $g$ ,  $l_g$  is the length of gene  $g$ , and  $N$  is the total number of reads mapped for a particular sample (i.e. the sequencing depth of that sample).

To analyze gene expression data at the gene network level, expression data from 20,000 genes were sorted into 200 networks through the use of Ingenuity Pathway Analysis (IPA) (QIAGEN Inc., <https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis>) (Kramer, Green, Pollard, & Tugendreich, 2014). We employ Generalized Estimating Equations (GEE) with an intermediate correlation structure (cite methods paper). Let  $y_{ij}$  be the measurement of gene expression for  $i^{th}$  gene and the  $j^{th}$  sample. Here, we assume that the patient samples are independent, but the genes within a gene network are correlated. Instead of assumed probability distributions, only a relationship between the mean and variance is needed for GEE. For RNA-seq data, we use the mean-variance relationship from the negative

binomial distribution because of the popularity of the negative binomial model for individual genes:

$$E(Y_{ij}) = \mu_{ij}$$

$$v(\mu_{ij}) = \mu_{ij} + \omega\mu_{ij}^2$$

Where  $E(Y_{ij})$  denotes the expected value,  $\mu_{ij}$  is the mean,  $v(\mu_{ij})$  is the variance, as a function of  $\mu_{ij}$ , and  $\omega$  is the overdispersion parameter. The overdispersion parameter is calculated as the mean of individual genewise dispersion parameters using edgeR (McCarthy et al., 2012b). This type of model is beneficial because it allows for a number of different “working” correlation structures to describe the relationship between different genes. This enables flexibility of capturing the true relationship between genes, when other multivariate models may have rigid correlation structures that do not work well with RNA-seq data. This method uses an intermediate correlation structure, where six different parameters are estimated; three being positive and three being negative. This structure is tailor-made for RNA-seq data to model the real data’s correlation having values that are positive, negative, and close to zero. Additionally, a regression model is fit to test the hypothesis of differential expression (i.e.  $H_0: \beta_1 = 0$ ):

$$\log \mu_{ij} = \beta_0 + \beta_1 X_1 + \sum_{k=2}^m \beta_k X_k$$

The use of regression models over two-sample tests improves the analysis because it allows for the incorporation of covariates into the model, given estimates of differential expression

that are adjusted for these covariates. Based on asymptotic results (K.-L. Liang & Zeger, 1986), a Wald test is used to test for differential expression:

$$Z = \frac{\hat{\beta}}{\sqrt{\text{Cov}(\hat{\beta})}} \sim N(0, 1)$$

Since multiple gene networks are tested simultaneously, p-values should be converted to false discovery rates (FDRs) to adjust for an increase in type I error.

Once a list of significantly differentially expressed gene networks is obtained, a separate univariate analysis is run on each network. For each gene network, it is likely not all genes in a given network are important. Using a single analysis will help to separate out the important signals for each of these gene networks. A popular method for univariate analysis of RNA-seq data is the R package edgeR (McCarthy et al., 2012a; Robinson et al., 2010). This method focuses on developing a generalized linear model approach to assess differential expression at the gene level. The gene expression level is assumed to be negative binomial (equivalently, a Poisson-Gamma mixture) and the differential expression hypothesis is tested via the likelihood ratio test. The dispersion parameter can be estimated using an adjusted profile likelihood, introducing a penalty via the Fisher information. Additionally, edgeR allows for genewise dispersion estimation, where the dispersion estimate is squeezed between the estimate of an individual gene and the weighted average of the gene and its (physical) neighbors. Such an estimation approach enables different genes to have different dispersion estimates.

## Results

The data quality is assessed for each of the patient samples used in the differential expression analysis. While TCGA has rigorous quality standards in the preparation of each sample, we want to confirm the similarity between samples. In order to remove any effects of differences in gene expression, normal samples were compared to other normal samples, while tumor samples were compared to other tumor samples. QQ plots were constructed for the log of counts per million (CPMs) for each pair of samples. The logarithm helps with the right skewed distribution of counts (large variations between different genes for a sample), while CPMs helps to alleviate the bias of sequencing depth. Additionally, presence of zeroes in the data cause a kink in the graphs because of the use of pseudo-counts (adding values to zero so the log is defined). For all pairs of samples, the QQ plots follow a straight line confirming high similarity among samples and thus, high data quality.

Clinical data is collected for each of the patient samples and are used as covariates in the differential expression analysis. Using clinical data is useful because it helps to explain more of the variation between gene expression levels as opposed to only comparing tumor and normal tissues. Clinical data was chosen by the sufficient availability of data and covariates relevant to breast cancer. The chosen covariates are menopause status, race, vital status, stage of cancer, age at diagnosis, and molecular subtypes. Most of the samples are white women with stage II primary breast tumors, and generalizability of the results from the differential expression analysis should be interpreted within this context.

Gene expression data is sorted into 200 non-overlapping gene networks, totaling approximately 6,000 genes. The data is then fit to the following model:



$$\log\mu_{ij} = \text{Intercept} + \text{Tumor} + \text{Vital Status} + \text{Stage} + \text{Molecular Subtype} + \text{Race} + \text{Menopause Status} + \text{Age}$$

To test for differential expression, a Wald test is used to test the coefficient for tumor status of tissue is zero. P-values are transformed into false discovery rates (FDR) and gene networks are declared significant if they fall below the  $\text{FDR} < 0.05$  threshold. For each of these significant

Figure 1: QQ plots for log counts per million (CPMs) for all of the genes in the RNA-seq dataset.

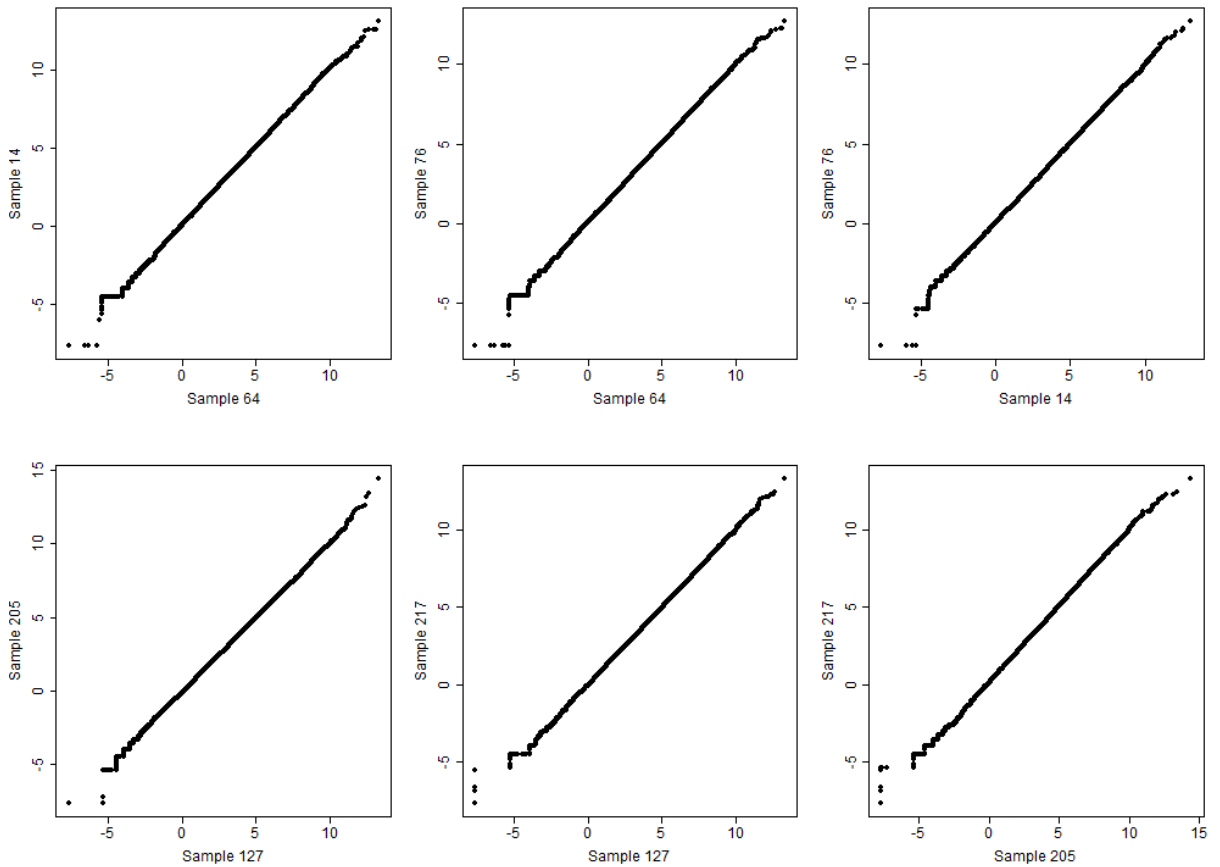


Figure 2: Patient demographics for samples in the TCGA dataset

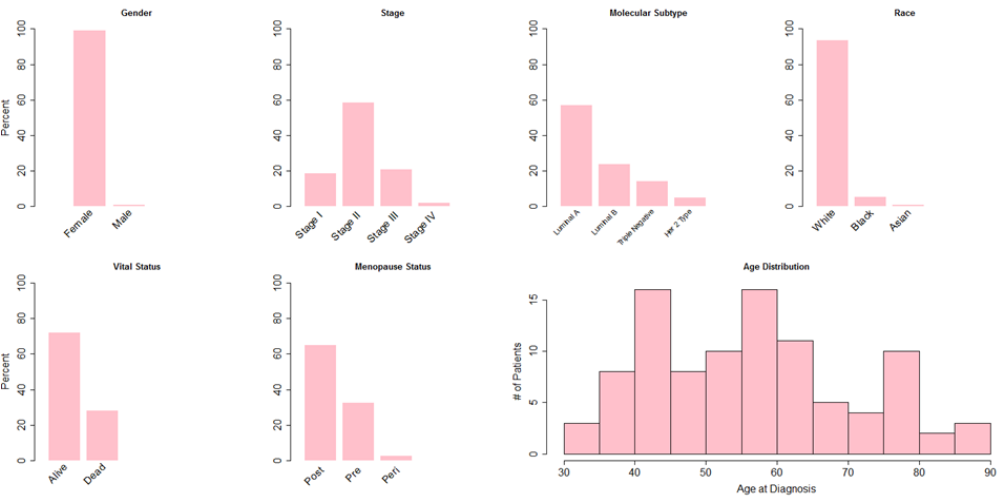


Table 5: Results of Gene Network Analysis

Top Genes in Network	Network Function	FDR
<i>BTBD8, SH3BGRL, IFT46, SRCIN1, HSPBP1</i>	Cellular Assembly and Organization, Cell Morphology, Cellular Function and Maintenance	1.95e-302
<i>PLEKHA5, ERP29, NCKAP1, SHROOM2, ABL1</i>	Cell Morphology, Cellular Development, Cellular Growth and Proliferation	8.85e-92
<i>DC34, FBXO31, STON2, ILIRL1, DUOX2</i>	Cellular Assembly and Organization, Cellular Function and Maintenance, Cellular Movement	1.02e-13
<i>ADAM9, MAPK1, PSMB8, ADAR, ERCC1</i>	Cancer, Dermatological Diseases and Conditions, Hereditary Disorder	1.14e-11
<i>PEG3, HSD17B2, ID3, APC, GALNT6</i>	Embryonic Development, Organismal Development, Tissue Development	8.22e-08
<i>SP5, ZPI, CRTC2, HIVEP2, IQSEC1</i>	Cell Morphology, Endocrine System Disorders, Organ Morphology	1.03e-07
<i>DPYSL4, MYH16, GOLGA4, ZNF292, NUDT13</i>	Hereditary Disorder, Neurological Disease, Organismal Injury and Abnormalities	8.96e-06
<i>RASSF3, PGAM5, HERC2, UBE3A, PMS1</i>	DNA Replication, Recombination, and Repair, Cell Cycle, Cellular Assembly and Organization	2.75e-05
<i>PRKCE, NEFL, DPP6, ATL1, NRK</i>	Cellular Assembly and Organization, Cellular Compromise, Cellular Function and Maintenance	2.77e-05
<i>TYROBP, EEF1A1, TNFSF13, RPS18, SIRPB1</i>	Cancer, Cell Death and Survival, Organismal Injury and Abnormalities	5.70e-05

gene networks, a single gene analysis is performed using the ‘edgeR’ package. For each gene in a gene network, FDRs are calculated and only genes with  $FDR < 0.05$  are kept.

Significant genes are scrutinized for biological significance by using PubMed to search for each gene in conjunction with breast cancer. Among the top 10 genes, 4 have no citations, while 6 have one or more citations. This implies that the data analysis method finds a mix of genes with known changes in gene expression as well genes with unreported changes in gene expression. These unreported genes include *BTBD8*, *IFT46*, *PLEKHA5*, and *SHROOM2*. A full list of the significant gene networks and the associated significant single genes is available in the appendix.

Table 6: Single gene analysis of statistically significant gene networks.

Gene	Log Fold Change	FDR	Citations
<i>BTBD8</i>	2.68	8.83e-08	0
<i>SH3BGRL</i>	-2.16	2.51e-07	1
<i>IFT46</i>	-1.93	2.60e-05	0
<i>SRCIN1</i>	-1.54	5.90e-04	9
<i>HSPBP1</i>	1.15	1.56e-02	1
<i>PLEKHA5</i>	3.06	4.75e-11	0
<i>ERP29</i>	-2.54	1.95e-08	15
<i>NCKAP1</i>	2.02	4.79e-06	4
<i>SHROOM2</i>	-1.99	4.79e-06	0
<i>ABL1</i>	1.79	3.11e-05	21

## Discussion

We have successfully analyzed RNA-seq gene expression data from patients with primary breast tumors using a gene network analysis that directly incorporates correlation among genes within a gene network. Additionally, we used a popular R package for single gene analysis, edgeR, to identify the differences between genes within the top networks. This

novel analysis led to identification of genes previously unidentified to be differentially expressed in breast cancer patients and helps provide more evidence for differential expression of other genes which have already been identified in the literature.

Using a novel gene network approach we are able to identify gene networks and genes within those networks that help researchers learn more about new roles in breast cancer. The top network helps uncover a previously unknown gene in breast cancer and its possible function in the tumor process. Additionally, using the network perspective we are able to offer a more complete story on genes that whose function are already established. Network analyses can lead to more information than more traditional analyses. From these results, we suggest that a network analysis such as the one used in this article be used in for future RNA-seq datasets.

**Acknowledgements:**

Andrew Aschenbrenner was funded by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP160015) and the NIH Pre-doctoral Traineeship in Biostatistics (grant # 2T32GM074902-06). We would like to thank Stephen Daiger and Momiao Xiong for their helpful comments on written drafts.

## References

- Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1), 107–129. <https://doi.org/10.1214/07-AOAS101>
- Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., & Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11(1), 574. <https://doi.org/10.1186/1471-2164-11-574>
- Kramer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4), 523–530. <https://doi.org/10.1093/bioinformatics/btt703>
- Liang, K.-L., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012a). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10), 4288–4297. <https://doi.org/10.1093/nar/gks042>
- McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012b). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10), 4288–4297. <https://doi.org/10.1093/nar/gks042>
- Mortazavi, A., Williams, B. a, McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628. <https://doi.org/10.1038/nmeth.1226>



- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Schaalje, G. B., & Butts, R. A. (1993). Some Effects of Ignoring Correlated Measurement Errors in Straight Line Regression and Prediction. *Biometrics*, 49(4), 1262–1267. <https://doi.org/10.2307/2532270>
- Shendure, J. (2008). The beginning of the end for microarrays? *Nature Methods*, 5(7), 585–587. <https://doi.org/10.1038/nmeth0708-585>
- Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, 66(1), 7–30. <https://doi.org/10.3322/caac.21332>

## CONCLUSION

This dissertation has developed a statistical methodology for analyzing gene networks in RNA-seq data. In this approach, we propose to analyze networks as the unit of choice in contrast to a single gene, which is the current standard. Such an approach is useful because it directly incorporates correlation between genes within a network into hypothesis testing for a differential expression analysis. This will help to avoid false positives and false negatives when analyzing RNA-seq data, allowing researcher more precise information with which they can follow-up these hypotheses in laboratory settings. This will lead to a greater understanding of breast cancer (or other diseases or condition that wish to be studied) and this knowledge may provide the basis of investigations into new therapies.

I proposed using generalized estimating equations (GEE) for analyzing gene networks from RNA-seq data. This approach allows for researchers to directly incorporate the correlation between genes in a gene network into the data analysis. Additionally, a framework for estimating correlation structures is developed to identify correlation structures with varying degrees of complexity. This allows researchers to customize the correlation matrix to be incorporated in the data analysis, relative to sample size and gene network size. A Wald test is used to test hypotheses and multiple hypotheses are adjusted via false discovery rates.

Using a simulation algorithm to generate data from a correlated negative binomial distribution, data is simulated for different correlation structures under both the null hypothesis of no difference in the mean and the alternative hypothesis of difference in the mean. Different assumed correlation structures were used in analyzing simulated data including independent, exchangeable, intermediate (six parameters), and unstructured. In the presence of correlation,

using an independent correlation structure increases the type I error for positive correlation or decreases the type I error for negative correlation, relative to the significance threshold level. On the other hand, methods incorporating correlation conserved type I error. This particular method can be seen as a safeguard against type I errors.

Using this technique, 200 networks from a 96 sample breast cancer dataset from the Cancer Genome Atlas (TCGA) Research Network (<http://cancergenome.nih.gov/>) were analyzed. Networks with a FDR < 0.05 were kept and ranked according to smallest FDR. A single gene analysis was then conducted on each of these networks via edgeR and the top genes are reported. Among the top ten genes, four were unreported, while six had previously been reported in the literature. The unreported genes include *BTBD8*, *IFT46*, *PLEKHA5*, and *SHROOM2*. The reported genes include *SH3BGRL*, *SRCIN1*, *HSPB1*, *ERP29*, *NCKAP1*, and *ABLI*. The mix of unreported and reported differentially expressed genes implies that this study generates new hypotheses as well as providing additional evidence for existing hypotheses about the genetic etiology of primary breast tumors. This study as well as future gene expression studies help to create a more complete picture of gene functions in relation to primary breast tumors.

The statistical method presented here should be viewed as a supplementary method to statistical methods for RNA-seq gene expression. It provides a different perspective of the data which can lead to previously unidentified gene networks and genes within those networks. However, this method is not uniformly superior to other methods tested. With many methods for assessing differential expression in RNA-seq, researchers should probably use many methods to analyze their data. Each method gives a different perspective, which can lead to

new insights. Also, critically thinking about the parameters of a researcher's data (sample size, expected fold change, number of tests conducted, etc.) may impact the results and using methods that fit datasets may improve quality of the results.

This method also has some limitations. First, the networks gathered from IPA may not be the “true” networks. One reason for this is technical: a user cannot upload the entire list of the human genome onto IPA's server at one time, so networks may be incomplete. Also, these networks reflect the experimental literature and it cannot include what is not known. Using networks as the unit of choice force a tradeoff when interpreting the results. Fewer hypotheses are tested, while the researcher gives up the specificity of which gene is differentially expressed. Consequently, a differentially expressed network may contain equivalently expressed genes or some genes that up-regulated with some genes that are down-regulated. This may lead to unexpected results and may limit the interpretability of the estimated log fold change.

While the focus of this dissertation was on GEE and marginal models (a population averaged approach), there are other methodological approaches that could be pursued. A mixed modeling approach could be considered under the generalized linear mixed model (GLMM) framework. Using random effects might be more comparable and useful when looking at single gene techniques. Examples of these techniques could be mixed Poisson, mixed Negative Binomial, and Dirichlet Negative Multinomial. Also, since networks often contain both up-regulated and down-regulated genes it may be more useful to estimate total magnitudes rather than coefficients with signs. This would be a measure of the networks total disruption rather than making statements about whether it's overexpressed or underexpressed.

The networks used in this dissertation were derived from IPA but genes could be organized in different ways which could give different results. For example, using gene ontologies (GOs). There are a number of ways to sort genes into networks, but it isn't clear which is superior. Analyses could be conducted using many methods of gene sorting. When using networks for the analyses the overall structure and connectedness of the network was ignored. Incorporating this network structure into correlation estimation (in conjunction with estimates from the data) could improve performance of the technique. Additionally, this technique is general and can be used on different RNA-seq datasets. In particular, analyzing the gene expression of metastatic breast cancer samples may be useful, though there are much fewer samples. RNA-seq datasets of metastatic cancer samples are also available from the Cancer Genome Atlas (TCGA) Research Network: <http://cancergenome.nih.gov/>.

RNA-seq data continues to be a challenge to analyze in a faithful manner. The large number of genes, samples, and complex systems with which genes interact with each other makes it difficult to test for differential expression while including these complicating features. Correlation among these genes is important, as noted by many researchers, and new insights can be gleaned when incorporating such features. One of the reasons why progress on cancer therapies has been relatively stunted may very well be the oversimplification of data analyses.

## APPENDICES

### Appendix A: Gene Network and Single Gene Analysis Results

Gene Network Results Table

<b>Top Genes in Network</b>	<b>Network Function</b>	<b>Network FDR</b>
<i>BTBD8, SH3BGRL, IFT46, SRCIN1, HSPBP1</i>	Cellular Assembly and Organization, Cell Morphology, Cellular Function and Maintenance	1.95e-302
<i>PLEKHA5, ERP29, NCKAP1, SHROOM2, ABL1</i>	Cell Morphology, Cellular Development, Cellular Growth and Proliferation	8.85e-92
<i>CDC34, FBXO31, STON2, IL1RL1, DUOX2</i>	Cellular Assembly and Organization, Cellular Function and Maintenance, Cellular Movement	1.02e-13
<i>ADAM9, MAPK1, PSMB8, ADAR, ERCC1</i>	Cancer, Dermatological Diseases and Conditions, Hereditary Disorder	1.14e-11
<i>PEG3, HSD17B2, ID3, APC, GALNT6</i>	Embryonic Development, Organismal Development, Tissue Development	8.22e-08
<i>SP5, ZP1, CRTC2, HIVEP2, IQSEC1</i>	Cell Morphology, Endocrine System Disorders, Organ Morphology	1.03e-07
<i>DPYSL4, MYH16, GOLGA4, ZNF292, NUDT13</i>	Hereditary Disorder, Neurological Disease, Organismal Injury and Abnormalities	8.96e-06
<i>RASSF3, PGAM5, HERC2, UBE3A, PMS1</i>	DNA Replication, Recombination, and Repair, Cell Cycle, Cellular Assembly and Organization	2.75e-05
<i>PRKCE, NEFL, DPP6, ATL1, NRK</i>	Cellular Assembly and Organization, Cellular Compromise, Cellular Function and Maintenance	2.77e-05
<i>TYROBP, EEF1A1, TNFSF13, RPS18, SIRPB1</i>	Cancer, Cell Death and Survival, Organismal Injury and Abnormalities	5.70e-05
<i>QPCT, UBAP2L, PBLD, NNAT, KRT19</i>	Skeletal and Muscular Disorders, Cellular Assembly and Organization, Cellular Function and Maintenance	8.51e-05
<i>BRWD1, SUPT3H, EID3, RPUSD4, MED22</i>	Cellular Function and Maintenance, Cell Signaling, Gene Expression	1.88e-04
<i>PDLIM7, NOC4L, DNAH14, SSSCA1, DNAH6</i>	RNA Post-Transcriptional Modification, Connective Tissue Disorders, Developmental Disorder	2.21e-04

<b>Top Genes in Network</b>	<b>Network Function</b>	<b>Network FDR</b>
<i>CCL3, VCL, ATG7, ATG4B, EGF</i>	Cellular Movement, Cancer, Organismal Injury and Abnormalities	6.92e-04
<i>PDCD6, CALCB, MDGA2, CALCRL, EIF3G</i>	Cell Signaling, Nucleic Acid Metabolism, Small Molecule Biochemistry	9.36e-04
<i>HLA-C, ZNF81, VIPR1, TAPBP, HLA-A</i>	Cellular Function and Maintenance, Hematological System Development and Function, Cell Death and Survival	1.23e-03
<i>ZNF330, MEOX1, CLASP1, WNT7A, IL17C</i>	Dermatological Diseases and Conditions, Organismal Injury and Abnormalities, Cell Death and Survival	1.23e-03
<i>FKBP15, NXNL1, NANOS1, GRAMD4, COMMD7</i>	Developmental Disorder, Hereditary Disorder, Organismal Injury and Abnormalities	2.05e-03
<i>VPS26A, ANXA6, RBP3, GPR161, LTB4R</i>	Cell-To-Cell Signaling and Interaction, Cellular Movement, Hematological System Development and Function	2.05e-03
<i>LRRC1, MAD2L2, POLK, DAAMI, RAB31</i>	Cancer, Endocrine System Disorders, Gastrointestinal Disease	2.22e-03
<i>FLII, NCOA2, CBX6, RBP5, STAR</i>	Lipid Metabolism, Small Molecule Biochemistry, Endocrine System Disorders	3.09e-03
<i>LGALS9C, FCER2, CD40, PRDM1, MANF</i>	Cancer, Hematological Disease, Immunological Disease	3.12e-03
<i>MMP13, ESCO2, SMAD3, MYOF, KDR</i>	Cancer, Organismal Injury and Abnormalities, Cardiovascular System Development and Function	5.00e-03
<i>NASP, KANK2, ADARB2, PSMD11, PSMD4</i>	Protein Degradation, Protein Synthesis, Cellular Assembly and Organization	5.54e-03
<i>RPS8, C11orf58, TUFM, FAU, HBB</i>	Cancer, Cell Death and Survival, Organismal Injury and Abnormalities	5.56e-03
<i>ZNF333, CRYZL1, CAPN6, C12orf4, CAPN3</i>	Connective Tissue Disorders, Developmental Disorder, Hereditary Disorder	1.08e-02
<i>NAT10, STAU1, ZNF592, OVOL1, IPO4</i>	Cellular Assembly and Organization, Cellular Function and Maintenance, Dermatological Diseases and Conditions	1.17e-02
<i>DSTN, RAPH1, HEATR5B, MYO1D, SLC45A3</i>	Cellular Assembly and Organization, Cellular Function and Maintenance, Cell Morphology	1.17e-02
<i>CENPO, PHF6, H2AFZ, HIST1H2BA, APLF</i>	Cellular Assembly and Organization, DNA Replication, Recombination, and Repair, Cell Cycle	1.42e-02

Top Genes in Network	Network Function	Network FDR
<i>CEP170, PRKD1, CXorf21, MYH6, STEAP1</i>	Cardiovascular System Development and Function, Organ Development, Organ Morphology	1.78e-02
<i>TRIM5, RASGRP1, KEAP1, PPME1, JUNB</i>	Post-Translational Modification, Protein Degradation, Cell Morphology	1.98e-02
<i>NR1H4, PLAA, KCNJ2, TADA3, DLEU2</i>	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry	1.98e-02
<i>ITGB6, CACNA1C, OBSCN, PTPRA, HAX1</i>	Cardiac Arteriopathy, Cardiovascular Disease, Organismal Injury and Abnormalities	2.00e-02
<i>COL12A1, IL1RL2, SIPR2, DUOX1, MAML1</i>	Hematological System Development and Function, Immune Cell Trafficking, Inflammatory Response	2.00e-02
<i>TYRP1, RSPO3, FCHO2, TM9SF3, TFE3</i>	Amino Acid Metabolism, Small Molecule Biochemistry, Vitamin and Mineral Metabolism	2.16e-02
<i>CSTF1, CSTF2, ZNF502, GSPT1</i>	RNA Post-Transcriptional Modification, RNA Damage and Repair, Molecular Transport	2.16e-02
<i>ZNF398, RRP1, RNH1, NR0B1, JMJD4</i>	Amino Acid Metabolism, Post-Translational Modification, Small Molecule Biochemistry	2.48e-02
<i>YWHAB, C3AR1, SMAD2, RLIM, SKI</i>	Cellular Growth and Proliferation, Cardiovascular System Development and Function, Cardiovascular Disease	2.52e-02
<i>ZFP64, GNL2, DLEU7, CCL22, NME1</i>	Cardiovascular Disease, Cardiovascular System Development and Function, Nervous System Development and Function	2.85e-02
<i>NOD2, KDELRI, SNRNP40, ATF7, SELE</i>	Connective Tissue Disorders, Hereditary Disorder, Inflammatory Disease	3.11e-02
<i>TGSI, GOLPH3, YARS2, MRPL40, NAALADL2</i>	Hereditary Disorder, Neurological Disease, Organismal Injury and Abnormalities	3.22e-02
<i>CD8B, CD8A, ITGB7, CDC42SE1, MX2</i>	Cellular Assembly and Organization, Cell Signaling, Cell-mediated Immune Response	3.27e-02
<i>NLGN4X, EMILIN2, HRH1, OTUB2, NME4</i>	Cell Death and Survival, Cancer, Organismal Injury and Abnormalities	3.44e-02
<i>SFTPA1, RPS7, SIRPA, PAX4, EEF2K</i>	Cell Death and Survival, Skeletal and Muscular System Development and Function, Cell Morphology	3.52e-02
<i>RASSF2, JAG2, TLR5, FZD2, RIN1</i>	Cell-To-Cell Signaling and Interaction, Hematological System Development and Function, Immune Cell Trafficking	4.05e-02



Top Genes in Network	Network Function	Network FDR
<i>FOXC1, PBX1, HNRNPA2B1, SF11, LETM2</i>	Cell Cycle, Cellular Assembly and Organization, Embryonic Development	4.05e-02

#### Single Gene Analysis Results

Gene	logFC	FDR
BTBD8	2.68	1.00e-07
SH3BGRL	-2.16	3.00e-07
IFT46	-1.93	2.60e-05
SRCIN1	-1.54	5.90e-04
HSPBP1	1.15	1.56e-02
PLEKHA5	3.06	0.00e+00
ERP29	-2.54	0.00e+00
NCKAP1	2.02	4.80e-06
SHROOM2	-1.99	4.80e-06
ABL1	1.79	3.11e-05
CDC34	3.54	0.00e+00
FBXO31	-1.71	4.47e-05
STON2	-1.80	4.86e-05
IL1RL1	1.93	1.01e-04
DUOX2	-1.70	1.01e-04
ADAM9	-6.41	0.00e+00
MAPK1	2.91	0.00e+00
PSMB8	-2.17	0.00e+00
ADAR	2.38	2.00e-07
ERCC1	2.23	3.00e-07
PEG3	-5.90	0.00e+00
HSD17B2	-4.17	0.00e+00
ID3	-2.27	1.20e-06
APC	1.66	1.09e-05
GALNT6	1.89	3.79e-05
SP5	-7.02	0.00e+00
ZP1	4.42	0.00e+00
CRTC2	-3.16	0.00e+00

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
HIVEP2	-2.74	1.00e-07
IQSEC1	-1.79	1.14e-03
DPYSL4	-8.12	0.00e+00
MYH16	-5.84	0.00e+00
GOLGA4	-4.71	0.00e+00
ZNF292	-3.28	0.00e+00
NUDT13	-2.14	4.61e-05
RASSF3	-7.29	0.00e+00
PGAM5	-2.95	0.00e+00
HERC2	2.11	1.20e-06
UBE3A	-1.57	3.78e-04
PMS1	1.43	1.14e-03
PRKCE	-4.16	0.00e+00
NEFL	3.90	0.00e+00
DPP6	-2.87	0.00e+00
ATL1	-2.71	0.00e+00
NRK	-2.46	0.00e+00
TYROBP	-9.10	0.00e+00
EEF1A1	-6.20	0.00e+00
TNFSF13	-3.46	0.00e+00
RPS18	-1.61	8.02e-04
SIRPB1	-1.62	1.07e-03
QPCT	-5.93	0.00e+00
UBAP2L	4.08	0.00e+00
PBLD	-3.39	0.00e+00
NNAT	2.22	1.90e-06
KRT19	1.92	1.01e-05
BRWD1	2.02	6.20e-06
SUPT3H	1.92	1.25e-02
EID3	-1.25	1.65e-02
RPUSD4	1.29	3.90e-02
MED22	1.07	4.90e-02
PDLIM7	-4.45	0.00e+00
NOC4L	2.09	8.30e-06
DNAH14	1.35	3.18e-03

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
SSSCA1	1.31	6.19e-03
DNAH6	-1.25	8.18e-03
CCL3	-7.18	0.00e+00
VCL	-3.31	0.00e+00
ATG7	2.35	2.00e-06
ATG4B	1.49	3.09e-05
EGF	1.78	7.26e-04
PDCD6	2.52	1.00e-07
CALCB	2.49	2.20e-06
MDGA2	-1.90	1.09e-05
CALCRL	1.94	1.09e-05
EIF3G	2.21	1.09e-05
HLA-C	-3.47	0.00e+00
ZNF81	-2.85	2.00e-07
VIPR1	2.79	7.00e-07
TAPBP	2.72	1.20e-06
HLA-A	-2.42	2.40e-06
ZNF330	-6.45	0.00e+00
MEOX1	-4.88	0.00e+00
CLASP1	-1.74	5.47e-05
WNT7A	-2.30	7.86e-05
IL17C	-2.23	1.89e-04
FKBP15	-3.37	0.00e+00
NXNL1	3.09	0.00e+00
NANOS1	-2.73	5.00e-07
GRAMD4	2.06	2.70e-06
COMMD7	-2.17	1.86e-05
VPS26A	-6.62	0.00e+00
ANXA6	2.62	0.00e+00
RBP3	2.27	1.00e-07
GPR161	1.67	1.54e-04
LTB4R	1.65	1.54e-04
LRRC1	-5.93	0.00e+00
MAD2L2	-5.99	0.00e+00
POLK	-2.16	2.70e-04

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
DAAM1	1.36	3.03e-03
RAB31	-1.26	1.12e-02
FLII	-5.87	0.00e+00
NCOA2	2.04	2.10e-06
CBX6	1.69	2.24e-04
RBP5	2.05	3.96e-04
STAR	1.59	6.40e-04
LGALS9C	-3.05	0.00e+00
FCER2	2.86	0.00e+00
CD40	-3.38	1.00e-07
PRDM1	-2.65	1.00e-07
MANF	1.21	1.12e-02
MMP13	-7.47	0.00e+00
ESCO2	-2.58	5.00e-07
SMAD3	1.86	2.80e-06
MYOF	-2.24	8.40e-06
KDR	-2.02	2.05e-04
NASP	-5.88	0.00e+00
KANK2	2.38	6.00e-07
ADARB2	2.01	1.64e-05
PSMD11	1.61	2.40e-05
PSMD4	1.80	4.32e-04
RPS8	-1.99	4.42e-05
C11orf58	1.52	9.51e-04
TUFM	-1.54	6.65e-03
FAU	1.51	7.38e-03
HBB	-1.56	8.95e-03
ZNF333	-5.76	0.00e+00
CRYZL1	-6.42	0.00e+00
CAPN6	4.06	0.00e+00
C12orf4	2.71	1.00e-07
CAPN3	2.19	6.50e-06
NAT10	-6.40	0.00e+00
STAU1	-6.58	0.00e+00
ZNF592	-3.99	0.00e+00

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
OVOL1	-3.38	0.00e+00
IPO4	2.08	9.30e-04
DSTN	-7.11	0.00e+00
RAPH1	-4.22	0.00e+00
HEATR5B	1.54	2.65e-05
MYO1D	1.76	6.10e-05
SLC45A3	2.24	6.10e-05
CENPO	3.49	0.00e+00
PHF6	-2.67	5.00e-07
H2AFZ	2.42	1.10e-06
HIST1H2BA	2.45	2.10e-06
APLF	1.90	8.00e-06
CEP170	-4.32	0.00e+00
PRKD1	-3.75	0.00e+00
CXorf21	3.36	0.00e+00
MYH6	1.84	5.00e-06
STEAP1	-2.04	2.20e-05
TRIM5	-5.00	0.00e+00
RASGRP1	2.27	6.30e-06
KEAP1	2.19	7.30e-06
PPME1	-2.01	3.33e-05
JUNB	-1.73	3.33e-05
NR1H4	-4.61	0.00e+00
PLAA	-3.73	0.00e+00
KCNJ2	-3.92	0.00e+00
TADA3	2.14	1.10e-06
DLEU2	-1.98	1.09e-04
ITGB6	-6.12	0.00e+00
CACNA1C	3.37	0.00e+00
OBSCN	2.00	1.20e-06
PTPRA	-1.93	6.50e-06
HAX1	-2.14	1.01e-04
COL12A1	1.97	5.40e-06
IL1RL2	2.21	5.90e-06
S1PR2	1.60	3.95e-03

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
DUOX1	-1.36	3.95e-03
MAML1	1.30	3.95e-03
TYRP1	-7.50	0.00e+00
RSPO3	-5.87	0.00e+00
FCHO2	-5.54	0.00e+00
TM9SF3	-5.74	0.00e+00
TFE3	-6.16	0.00e+00
CSTF1	-2.16	8.00e-07
CSTF2	-1.93	1.78e-05
ZNF502	1.37	4.28e-03
GSPT1	1.24	6.95e-03
ZNF398	-5.34	0.00e+00
RRP1	-3.53	0.00e+00
RNH1	2.57	0.00e+00
NR0B1	-2.08	1.40e-06
JMJD4	2.03	1.84e-05
YWHAB	-4.92	0.00e+00
C3AR1	-3.28	0.00e+00
SMAD2	2.10	1.40e-06
RLIM	1.65	2.37e-04
SKI	-1.83	2.37e-04
ZFP64	-3.82	0.00e+00
GNL2	-3.55	0.00e+00
DLEU7	-3.49	0.00e+00
CCL22	2.56	1.00e-07
NME1	2.41	7.00e-07
NOD2	2.99	0.00e+00
KDELRL1	2.90	0.00e+00
SNRNP40	2.77	0.00e+00
ATF7	2.11	8.00e-07
SELE	-2.37	8.00e-07
TGS1	-5.18	0.00e+00
GOLPH3	5.35	0.00e+00
YARS2	-2.30	7.80e-06
MRPL40	-1.84	1.19e-04

<b>Gene</b>	<b>logFC</b>	<b>FDR</b>
NAALADL2	-1.47	4.89e-04
CD8B	-8.01	0.00e+00
CD8A	-7.47	0.00e+00
ITGB7	-6.65	0.00e+00
CDC42SE1	-4.16	0.00e+00
MX2	-3.44	0.00e+00
NLGN4X	-5.70	0.00e+00
EMILIN2	-5.35	0.00e+00
HRH1	2.03	5.00e-05
OTUB2	-1.59	5.78e-04
NME4	1.62	6.97e-04
SFTPA1	-2.79	2.79e-05
RPS7	-2.00	5.27e-04
SIRPA	-1.20	1.20e-02
PAX4	1.19	1.20e-02
EEF2K	1.19	1.21e-02
RASSF2	-6.10	0.00e+00
JAG2	-3.88	0.00e+00
TLR5	-3.07	0.00e+00
FZD2	-3.27	0.00e+00
RIN1	-3.44	0.00e+00
FOXC1	-1.76	1.94e-03
PBX1	1.67	1.94e-03
HNRNPA2B1	1.29	8.31e-03
SFI1	-1.39	8.31e-03
LETM2	1.31	8.31e-03

## REFERENCES

- Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1), 190–206. [https://doi.org/10.1016/S0047-259X\(03\)00079-4](https://doi.org/10.1016/S0047-259X(03)00079-4)
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B ( Methodological ), Vol . 57 , No . 1 Published by : Wi. *Journal of the Royal Statistical Society Series B*, 57(1), 289–300.
- Choi, Y., & Kendzierski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21), 2780–2786. <https://doi.org/10.1093/bioinformatics/btp502>
- Efron, B., & Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1), 107–129. <https://doi.org/10.1214/07-AOAS101>
- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477. <https://doi.org/10.1038/nmeth.1613>
- Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., & Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11(1), 574. <https://doi.org/10.1186/1471-2164-11-574>
- Goeman, J. J., Bu, P., Zurich, E. T. H., & Zu, C.-. (2007). *Gene expression Analyzing gene expression data in terms of gene sets : methodological issues*. 23(8), 980–987. <https://doi.org/10.1093/bioinformatics/btm051>
- Hardin, J. W., & Hilbe, J. M. (2002). *Generalized Estimating Equations*. Retrieved from



<https://books.google.com/books?id=wx17ajq8ymYC&pgis=1>

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., ... Cronin, K.

(2014). SEER Cancer Statistics Review, 1975-2011. Retrieved from National Cancer Institute website: [http://seer.cancer.gov/csr/1975\\_2011/](http://seer.cancer.gov/csr/1975_2011/)

Huang, Y.-T. T., & Lin, X. (2013). Gene set analysis using variance component tests. *BMC Bioinformatics*, 14, 210. <https://doi.org/10.1186/1471-2105-14-210>

Kramer, A., Green, J., Pollard, J., & Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4), 523–530.

<https://doi.org/10.1093/bioinformatics/btt703>

Larson, J. L., & Owen, A. B. (2015). Moment based gene set tests. *BMC Bioinformatics*, 16(1), 1–17. <https://doi.org/10.1186/s12859-015-0571-7>

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29. <https://doi.org/10.1186/gb-2014-15-2-r29>

Li, Y. P., & Chan, W. (2006). Analysis of longitudinal multinomial outcome data. *Biometrical Journal*, 48(2), 319–326. <https://doi.org/10.1002/bimj.200510187>

Liang, K.-L., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Love, M. I., Huber, W., & Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.

<https://doi.org/10.1186/s13059-014-0550-8>

Love, M. I., Huber, W., & Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.

<https://doi.org/10.1186/s13059-014-0550-8>

Lund, S. P., Nettleton, D., McCarthy, D. J., & Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. In *Statistical applications in genetics and molecular biology* (Vol. 11).

<https://doi.org/10.1515/1544-6115.1826>

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.

*Genome Research*, 18(9), 1509–1517. <https://doi.org/10.1101/gr.079558.108>

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

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

Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., & Narod, S. a. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *Bmj*, 348(feb11 9), g366–g366.

<https://doi.org/10.1136/bmj.g366>

Mortazavi, A., Williams, B. a, McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and

- quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628.  
<https://doi.org/10.1038/nmeth.1226>
- Nussbaum, R. L., McInnes, R. R., & Willard, H. F. (2015). *Thompson & Thompson Genetics in Medicine*. Retrieved from  
<https://books.google.com/books?id=4yV1CQAAQBAJ&pgis=1>
- Rahmatallah, Y., Emmert-Streib, F., & Glazko, G. (2016). Gene set analysis approaches for RNA-seq data: Performance evaluation and application guideline. *Briefings in Bioinformatics*, 17(3), 393–407. <https://doi.org/10.1093/bib/bbv069>
- Riffenburgh, R. H., & Johnstone, P. A. S. (2001). Survival patterns of cancer patients. *Cancer*, 91(12), 2469–2475. [https://doi.org/10.1002/1097-0142\(20010615\)91:12<2469::AID-CNCR1282>3.0.CO;2-U](https://doi.org/10.1002/1097-0142(20010615)91:12<2469::AID-CNCR1282>3.0.CO;2-U)
- Riffenburgh, Robert H. (2000). Survival of Patients With Untreated. *Journal of Surgical Oncology*, 73(November 1999), 273–277.  
<https://doi.org/10.1001/archopht.1987.01060020020004>
- RI, S., Kd, M., & Jemal, A. (2015). Cancer statistics , 2015 . *CA Cancer J Clin*, 65(1), 21254.  
<https://doi.org/10.3322/caac.21254>.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Schaalje, G. B., & Butts, R. A. (1993). Some Effects of Ignoring Correlated Measurement Errors in Straight Line Regression and Prediction. *Biometrics*, 49(4), 1262–1267.  
<https://doi.org/10.2307/2532270>

- Shendure, J. (2008). The beginning of the end for microarrays? *Nature Methods*, 5(7), 585–587. <https://doi.org/10.1038/nmeth0708-585>
- Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, 66(1), 7–30. <https://doi.org/10.3322/caac.21332>
- So, H. C., Gui, A. H. S., Cherny, S. S., & Sham, P. C. (2011). Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. *Genetic Epidemiology*, 35(5), 310–317. <https://doi.org/10.1002/gepi.20579>
- Solis-Trapala, I. L., & Farewell, V. T. (2005). Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in Medicine*, 24(16), 2557–2575. <https://doi.org/10.1002/sim.2121>
- Storey, J. D. (2010). False Discovery Rates. *Princeton University, Princeton, USA*, (January), 1–7. <https://doi.org/10.1198/016214507000000941>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a, ... Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Sutradhar, B., & Das, K. (1999). On the Efficiency of Regression Estimators in Generalised Linear Models for Longitudinal Data. *Biometrika*, 86(2), 459–465.
- Tamayo, P., Steinhardt, G., Liberzon, a., & Mesirov, J. P. (2012). The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 1–22. <https://doi.org/10.1177/0962280212460441>

- Van De Wiel, M. a., Leday, G. G. R., Pardo, L., Rue, H., Van Der Vaart, A. W., & Van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, *14*(1), 113–128.  
<https://doi.org/10.1093/biostatistics/kxs031>
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558.  
<https://doi.org/10.1126/science.1235122>
- Wang, M. (2014). Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. *Advances in Statistics*, *2014*, 1–11.  
<https://doi.org/http://dx.doi.org/10.1155/2014/303728>
- Weinberg, R. (2013). *The Biology of Cancer, Second Edition*. Retrieved from <https://books.google.com/books?id=MzMmAgAAQBAJ&pgis=1>
- Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M. L., Visvader, J. E., & Smyth, G. K. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, *26*(17), 2176–2182. <https://doi.org/10.1093/bioinformatics/btq401>
- Zhang, H., Xu, J., Jiang, N., Hu, X., & Luo, Z. (2015). PLNseq: a multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Statistics in Medicine*, (April 2014), n/a-n/a. <https://doi.org/10.1002/sim.6449>
- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, *18*(1), 1–10. <https://doi.org/10.1186/s12859-016-1441-7>
- Zhou, X., Lindsay, H., & Robinson, M. D. (2014). Robustly detecting differential expression

in RNA sequencing data using observation weights. *Nucleic Acids Research*, 42(11), 1–10. <https://doi.org/10.1093/nar/gku310>