RNA-Seq Analysis Strategies and Ethical Considerations Involved in Precision Medicine

By

Janelle Rose Noel-MacDonnell
M.S., University of Kansas Medical Center, 2013
B.S., University of Colorado Denver, 2011

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Brooke L. Fridley, Ph.D.

Jeremy Chien, Ph.D.

Byron J. Gajewski, Ph.D.

Devin C. Koestler, Ph.D.

Jo A. Wick, Ph.D.

Date Defended: August 19, 2016

The Dissertation Committee for Janelle Rose Noel-MacDonnell
certifies that this is the approved version of the following dissertation:

RNA-Seq ANALYSIS STRATEGIES AND ETHICAL CONSIDERATIONS INVOLVED IN PRECISION MEDICINE

____

Chairperson Brooke L. Fridley, Ph.D.

Date approved: August 19, 2016

Abstract

RNA-Seq has become the most recently and widely accepted method to evaluate gene expression. Though with RNA-Seq being a fairly green technology, analytical methods for its output data have not been fully investigated as they have for preceding technology; such as those methods used in analyses of microarray data. This is likely the result of the potential breadth of information that can be obtained from the different applications of RNA-Seq. Analyses of RNA-Seq data include: detecting differentially expressed genes, transcriptome profiling, and interpretation of gene functions. As with any advanced technology medical or otherwise, the longer it is available, the price of the technology, in general, decreases and the technology itself becomes more refined. This has been true for genomic sequencing—costs per sample have continued to decrease; and the accuracy and precision of results has improved greatly. Synchronously, more physicians have opted to have more of their patients' genetic material sequenced. This has caused both challenges in the development of accurate, efficient, and consistent statistical methods; and much debate regarding the ethics involved in genomic sequencing. To provide insight into two statistical challenges that are common with analyzing RNA-Seq data, we conduct extensive simulation studies. These simulations studies include: 1) investigation of fitting complex models which account for pairedness across subject's measurements in terms of the power gained and control of Type I error rate; and 2) evaluation of clustering performance of various clustering methods in transformed RNA-Seq data. In addition to investigating the aforementioned statistical challenges, we develop a protocol for a survey study which has the potential to provide insight into cancer patients' opinions towards genomic sequencing as there is much ethics related controversy that surrounds the topic.

Dedication

This thesis is dedicated to my loving grandpa and grandma, Joseph "Papa" and Jennie "Nana" Furia. Their love for life, family, and community fueled them to continually work hard and enjoy the simple things in life. Watching and learning from them helped build the foundation of my life. They instilled in me the importance of education and faith at an early age, and supported my wildest dreams with every ounce of their being. They never failed to remind me that blessings and beauty can be found in any situation, and that some of the most rewarding outcomes arise from the challenges you face in life. In my grandpa's passing, he made sure that I would never give up on my goals or myself; and that his wife, my grandma, would continue to encourage and support me through the rigorous process of this degree.

Acknowledgements

The completion of this dissertation and doctoral degree would not have been possible without a “village-sized” group of individuals. I am truly grateful for the guidance from my committee members; support from my family, husband, and friends; backing from faculty and staff in the Department of Biostatistics at the University of Kansas Medical Center (KUMC); and encouragement from many prior academic mentors. While these last five years have been some of the most challenging mentally, and at times also physically; I truly wouldn’t trade a moment of any of it for the countless memories, experiences, friendships, late nights, and early mornings which have enriched my life tremendously.

First, I would like to thank my advisor and dissertation committee chair, Dr. Brooke L. Fridley. When I was assigned to you as Graduate Research Assistant (GRA), I didn’t quite know what to expect—I was excited to work in a different domain, genomics, and slightly terrified. After our first meeting, I knew that a great opportunity was ahead of me and that I had to absorb as much information and knowledge from you as I could. Your vast knowledge of numerous areas (e.g., genomics, clinical trials, grants, management, and teaching) never cease to make me want to learn more. Over the last three and a half years, I have been able to grow so much through your mentorship and supervision. You have encouraged me when I have spun my wheels, and have always provided thoughtful advice in how to proceed. The experiences that I have had under your leadership have equipped me with a breadth of tools that will not only be useful in my future career but throughout my entire life. Thank you for pushing me to new levels when I was hesitant and needed a push. I am very grateful to have had the opportunity to work with an advisor that challenges my knowledge of academic topics, makes me think independently, and tasks me to be innovated in my methods. Also, I am much appreciative of how you have developed the culture of comradery between all members of your lab; it has truly been a highlight of my time. I sincerely want to thank you for all of the teaching moments, memories, tasty snacks at lab meetings, and great conversations about sports. GO ROYALS!

Secondly, I thank my wonderful parents, Daniel and Roberta Noel, grandmother, Jennie "Nana" Furia, and brother and sister-in-law, Johnny "J.J." and Ashley Noel. Words cannot express how much you have contributed to the years I have been in graduate school. From miles away, you have found ways to make sure your presence was felt on the best or toughest of days. Each of you have truly went out of your way to be sources of motivation, from wakeup calls in the morning when I was up late working/studying the night before, to late night FaceTime sessions when I needed some company and hadn't been home, back to Trinidad, in a while, or even to sending me care packages. Thank you for always lending an interested ear when I've "geeked" out over one of my studies or an article I've read that might not make much or any sense to you. You all have played some of the most critical roles in my journey in pursuing a Ph.D. degree and I am forever grateful to call you family. Thanks for rooting for me (as my dad says), all of your prayers, and believing in me. I love you.

Also secondly, I want to thank my gem of a husband and best friend Drew MacDonnell. There are not enough words to fully express my appreciation towards you during this time. You have been the support of all the earthly support systems combined when it comes to this journey, and have taken the brute of the ups and downs that others haven't witnessed. That being said, I am forever grateful for your never ending, unconditional support towards me reaching my goals. You took a big leap of faith and moved out to Kansas, the land of corn and tornados as you once referred to it when we were dating, to see me through completion of a master's degree. Now, as your wife, we close a chapter as I complete a doctoral degree in a place that has captured both of our heats. You have been a rock--my rock. Throughout this entire process you have probably seen every possible emotion that I'm capable of having and have always been there to lend a shoulder or give some tough love. I thank you for all of the sacrifices you have made for us to remain in Kansas City. I feel incredibly blessed to have you as my partner in life. Thank you for being patient through this process and helping me to keep my sanity. Last but not least, thank you for taking care of me, our fish, and our fur child when I haven't been able to. I love you mucho!

Additionally, I wish to thank my dissertation committee members: Dr. Byron Gajewski, Dr. Jo Wick, Dr. Devin Koestler, all form the Department of Biostatistics at KUMC, and Dr. Jeremy Chien, my non-departmental committee member from the Department of Cancer Biology at KUMC. Each of you have played instrumental roles in my time in graduate school from developing my professional, research, and statistical skills; to motivating and inspiring me in the ways you balance life, family, your careers, and everything in between. You all are some of the most amazing, academic, role models that I have had the privilege to get to know. I know that many students and future students will feel the same way that I do. Thank you for mentoring me throughout these years and allowing me to bounce ideas off you any time and in any setting. Dr. Gajewski, thank you for your willingness to meet on short notice to discuss problems and for sparking my interest in sampling methods. Dr. Wick, I'd like to thank you for being there for me in so many ways since my first day at KUMC. Thank you for your honesty during tough conversations and for always advocating for any role that I wanted to pursue on or off campus. Dr. Koestler, I want to thank you for all of your knowledge regarding topics of professional development and potential career options; as well as, thoroughness in answering any questions—statistics related or not. Dr. Chien, thank you for allowing me to use your data in one of my studies and for providing insightful considerations on various projects. Thank you all from the bottom of my heart for your support and guidance.

To my former colleagues, Dr. Wei (Will) Jiang, Dr. Yu (Joyce) Jiang, Dr. Yang Lei, Dr. Milan Bimali, and Dr. Lili Garrard, thank you all for setting the standard's bar high and for your continued support as you have moved on in your careers. I'm consider myself very lucky to have been able to work alongside you in many courses. Drs. Jiang, Lei, and Garrard, thank you for all of the gal pal support, coffee walks, vent sessions, and midnight ice cream runs. Also, thank you all for the continued advice and encouragement in navigating the closing of this chapter and the beginning of a new one. Dr. Garrard, when you started the program I had no clue that you and I would become such great friends nor did I think that joining this program would I acquire another sister. Thanks for being like a sister to me by allowing me to bounce ideas off of you about life, school, marriage, or anything on my mind. Your friendship is second to none and is valued more than you know. And to my current colleagues, Stefan Graw, Richard Meier, Alex Karenovich, Eric Dai, Alvin Beltramo, thanks for going through this whole process together with me. Thanks for always being available to help out with making graphs look nice, troubleshooting codes, and great casual conversation. I'm so thankful for all of you.

A very special thanks goes out to Dr. Matthew Mayo (Chair of the Department of Biostatistics), whom admitted me to the program. I wouldn't be here today if it weren't for the acceptance letter that I received five years ago. Thank you, for that. Also, I'd like to thank you for always putting us students first and ensuring that we have the highest level of faculty and staff possible. On that note, to the faculty and staff of the Department of Biostatistics at KUMC, I extend my gratefulness. Thank you department faculty for aiming to provide us students with a great education so that we are fully capable to succeed once our grad school tenure comes to an end. Department staff, you all have been wonderful to have around. Thank you for all the help in setting up rooms, reservations, technical support, conversation, and tracking down faculty members. Collectively, I want to thank the Department of Biostatistics for the making my time in the department so enjoyable.

Drs. Mike Werle and Robert Klein: I am so glad that I was able to work with both of you. I'd like to sincerely thank you for believing and trusting in me to run the Student Research Forum (SRF) in '14 and '15. Over the course of those two years, I was able to take away many valuable lessons that have aided in completion of this degree. Also, thank you for always checking in on my degree progress during that time.

To my Kansas City family, Sarah and Alex Jones, Jim Xu, Laura Hays, Ruth Schmidt, Kim Gregory, the Dunnes, the Dragons, the Vanderhorst, and folks from Dream Factory of Greater Kansas City, thank you for embracing myself, and Drew, as family. Our time in Kansas City would not have been as gratifying without all of you. Each of you have gone above and beyond to ensure that we always felt welcomed and loved while away from Colorado. Sara, Laura, and Ruth, you have brought so much light and joy to my life. Thank you for all of our Core Group meetings, and for always praying for me and listening to my struggles. Thank you for allowing me to be part of your lives.

To the some of the best Colorado friends a girl could ask for, Jen De Groot, Dr. Shannon Ortiz, Jeremy Begley, Zack Newman, and David Brandt thank you for being some of my biggest fans from a far. David, thank you for talking me into giving the University of Kansas a shot when I was applying to graduate schools and for making a point to get together when you were in Kansas City. You were really the catalyst for this degree becoming a reality out here in Kansas City. Jen, Shannon, Jeremy, and Zack, there is something so special about the friendships we share. Thanks for taking a chance and friending me in childhood. Whenever we are able to catch up it hardly seems like we skip a beat. I am so grateful that I have nearly 80 years of combined friendship between all of you, and a lifetime of memories to be made ahead. Knowing that I have such a strong network of friends that have always supported my decisions has been one of the greatest blessings. Thank you, thank you, thank you.

Lastly, to all of the mentors I have had along the way both in a personal and professional settings: Debbie Harnish, Patricia Johnson, and Carrie Gongaware for cultivating my love for mathematics early on; Geneva Villegas and Randy Begano for all lessons you have taught me about physical strength and mental endurance through three-a-day practices; Diane Castner for teaching me the art of being disciplined through years of piano lessons, Mr. Babnick and Mr. Fredricks for instilling in me more than how the concept of "measure twice, cut once" applies to construction, but in the day to day; all my former educators from Colorado School District #1 for providing me with the educational foundation that has been the basis of all of my adult learning; and many others not named here. Your continued encouragement and support have molded me into the person I am today, and have helped me reach many goals. Thank you for impacting my life in such positive ways and for your lasting support.

Table of Contents

List of Figures

CHAPTER 1 ***Introduction***

CHAPTER 2 ***A Comparison of Paired and Unpaired Methods for Differential Expression Analysis of RNA-Seq Data***

CHAPTER 3 ***An Assessment of Transformations and Clustering Methods Using RNA-Seq data***

CHAPTER 4 ***Ethical Considerations for Precision Medicine—A Survey Protocol to Investigate Patient's Opinion Towards Genomic Sequencing***

CHAPTER 5 ***Discussion***

List of Tables

CHAPTER 1 ***Introduction***

CHAPTER 2 ***A Comparison of Paired and Unpaired Methods for Differential Expression Analysis of RNA-Seq Data***

CHAPTER 3 ***An Assessment of Transformations and Clustering Methods Using RNA-Seq data***

CHAPTER 4 ***Ethical Considerations for Precision Medicine—A Survey Protocol to Investigate Patient's Opinion Towards Genomic Sequencing***

CHAPTER 5 ***Discussion***

CHAPTER 1

Introduction

It is highly unlikely that Charles Darwin, Gregor Mendel, and Frederick Miescher, the first fathers of genetics, fully understood the greatness of their historic scientific discoveries. Beginning in 1859, the field of genetics was established with Charles Darwin's discovery of natural selection, where generations of organisms were shown to reproduce and survive through evolution, mutation, migration, and genetic drift; Gregor Mendel's experiment which revealed heritability in 1865; and Frederick Miescher's detection and isolation of DNA for the first time in 1869 (Darwin, 1872, Fisher, 1930, Dahm, 2005, Bateson and Gregor, 1913). Nearly a century later, James D. Watson and Francis H. Crick made another significant, well-known, discovery. With some help from X-ray diffraction images contributed by Rosalind Franklin, Watson and Crick discovered that the molecular structure of DNA was a three-dimensional, double helix (Wilkins, 1963, Watson and Crick, 1953, Heather and Chain, 2016).

Crick furthered his research in 1970 through his documentation of the *Central Dogma of Biology* which explains the transfer of genetic information from the three major molecular components, DNA, RNA, and protein, which are responsible for structure and function in any living organism (Crick, 1970). Figure I-1 contains a simplified version of the *Central Dogma of Biology*. The general information transfers that can occur are: DNA transcribed into RNA, RNA translated into proteins, DNA and RNA replicate into copies of themselves, RNA reversed transcribed in DNA, and the rare phenomena of DNA to protein depending on the cellular environment (Crick, 1970). As the *Central Dogma of Biology* became widely accepted across the life sciences' research community, the race to expand its three branches began with the overall goal to better understand the molecular basis of life. Thus, it became highly important to be able to identify genetic transcripts (i.e., read the sequence of nucleic acid), quantify genes and their expression values, and understand functional responsibilities of genes and proteins. To address these important topics, a collective approach needs to be taken where biological, statistical, and informatics techniques are heavily utilized together.

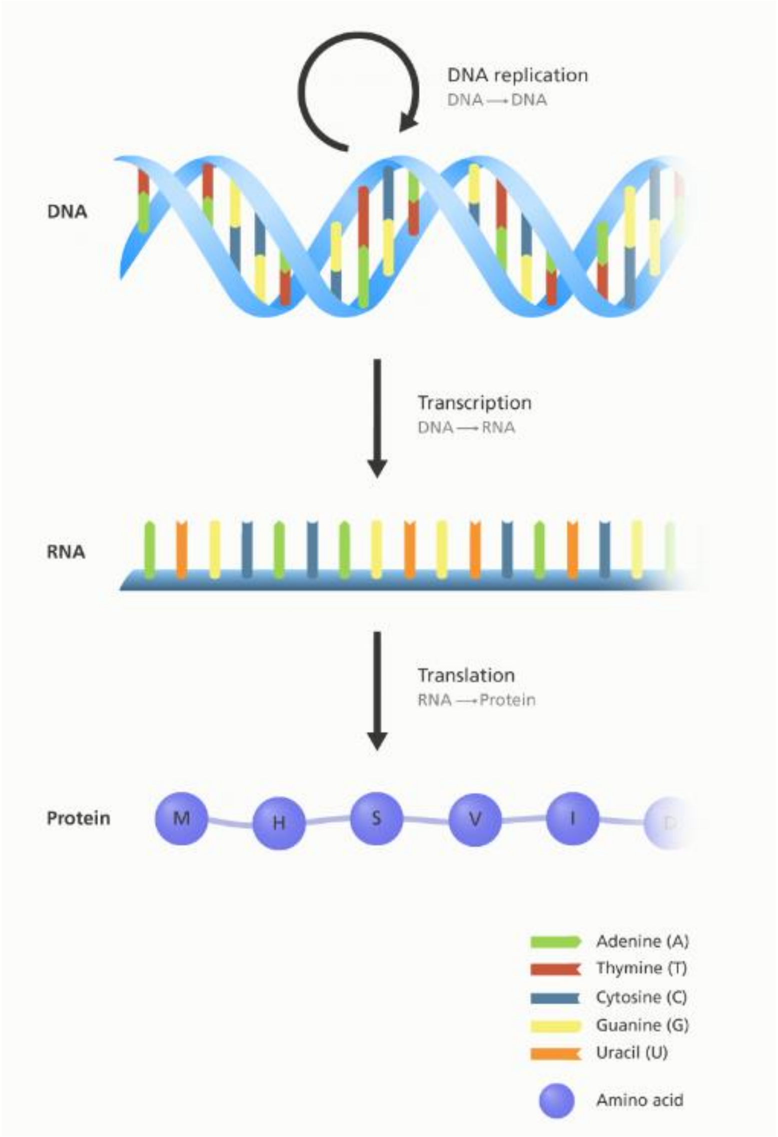


Figure I-1. The Central Dogma of Biology (Your Genome, 2016)**.** The Central Dogma of Biology is the flow of genetic information. DNA is made into RNA through a process called transcription; and RNA is made into proteins through translation. Both DNA and RNA have the ability to replicate itself. Both DNA and RNA are made up of four nucleic acid bases. In DNA, the nucleic acid bases are Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), which pair A to T and G to C. In RNA, the nucleic acid bases are the same with the Thymine being exchanged for Uracil (U), where the new pairings become A to U and G to C (Your Genome, 2016) *Figure credit: Genome Research Limited*.

The advent of genomic sequencing, along with its advancements, have been an integral part to better understand the molecular components of life. Initially said by Frederick Sanger, “knowledge of sequences could contribute much to our understanding of living matter” (Sanger, 1980). Ideas and efforts amongst many in the molecular biology and chemistry research communities were focused on developing techniques to read the nucleic acid sequence present in DNA. The mid 1960s gave way to the first-generation of sequencing, sequencing capable of reading up to approximately one killobase (kb), through paralleling work completed by Robert Holley and Frederick Sanger and their respective colleagues in fragments of RNA, and contributions of sequencing DNA fragments from Maxam and Gilbert (Heather and Chain, 2016, Holley, 1965, Maxam and Gilbert, 1977). This first-generation of sequencing was termed Sanger sequencing (Heather and Chain, 2016). It was this first-generation of sequencing that set the stage for second- and third-generations of sequencing which currently have the capability to sequence vast amounts of genetic material by running multiple samples at the same times, and even single molecule real time (SMRT) sequencing (Heather and Chain, 2016, Van Dijk et al., 2014).

The improvements of sequencing led to additional development of innovative technologies that have the ability to determine genetic expression levels. Microarrays were invented in the 1990s to conduct gene expression studies on a large-scale and was used religiously by the science community to solve a multitude of scientific problems (Zhao et al., 2014). However, the mid 2000s gave rise to an updated method to quantify gene expression. That next generation sequencing (NGS) method was RNA-sequencing which has come to be known as RNA-seq. While obtained gene expression is the end result between each of these technologies, the basis behind each of them is very different. Microarray technologies utilizes relative mRNA which is measured using pre-defined probe sets via fluorescence to determine expression value; whereas, RNA-Seq experiments measure gene expression levels from the total number of reads that fall into the exons of a gene. Hence, the output data from these two technologies is dissimilar. Gene expression values from microarrays are continuous and have a tendency to follow a Gaussian distribution, while gene expression values from RNA-Seq are count in nature and follow either over-dispersed Poisson or Negative Binomial distributions.

1.1 RNA-Seq Studies

With microarray technology having tenure amongst most of the biological research community, there have been debates about adoption of the newer RNA-Seq technology. Prior to the mid 2000s, many, Schena (1995), DeRisi et al. (1997), Brown and Botstein (1999), Neilsen et al. (2002), and Monti et al. (2005) to name a few, worked extensively to quantify patterns in gene expression present in particular disease states, environmental or biological conditions, and different tissue types. Since 2008, RNA-Seq has rapidly become a forerunner in next-generation sequencing (NGS) when it comes to analysis of high-throughput gene expression analysis (Reeb, 2013). The saturation of the current literature discussing studies that use RNA-Seq and its applications is proof of its rise in popularity. The RNA-Seq platform itself has the ability to address many applications outside of obtaining determining gene expression values. Specifically, scientists have used the RNA-Seq platform for discovery of novel transcripts and isoforms, RNA editing, alternative splicing, allele-specific expression, and exploration of non-model-organism transcriptomes (Anders et al., 2013, Wang et al., 2009, Mortazavi, 2008).

With the adoption and wide-spread use of RNA-Seq, new analysis methods have been developed. As the implication of these analysis have the potential to play roles in treatment plans for patients or future diagnostics, it is important that sound, accurate statistical methods need to be implemented. That is, the statistical analysis should consider both experimental design and the unique characteristics of "omic" study type (Reeb, 2013). All "omic" studies (i.e., genomics, proteomics, metabolomics, etc) have unique characteristics that are solely unique to said study. The potential amount of information that can be derived from these types of studies is highly impressive. Concurrently, the amount of physical data that is output from these types of studies requires much storage as sequencers for a single sample can produce more than 500 gigabases for a single run depending on the platform used (Trapnell et al., 2012). However, concerns about the difficulties involved in analyzing the massively complex gene expression datasets often containing expression information for 60K+ gene IDs have also been published. Some of the questions that arise are centered around the challenges that come with the analyses of RNA-Seq data, or benefit gained from the abundance of data that is provided using RNA-Seq. In order to move forward with the advancement of the area of genomics, statisticians and bioinformaticians need to work together seamlessly to insure that all of the analyses that are taking place are correct and computationally efficient. Such analyses takes much practice, patience, careful revision, and understanding of both biological processes and statistical methodologies.

To investigate some of the statistical challenges and difficulties that arise when working with RNA-Seq data, two extensive simulation studies were conducted. Our first study was motivated by the poor overlap similarly found differentially expressed genes when comparing commonly used differential expression methods when using paired measurement data. We sought to determine if the basic models that were fit within the differential expression methods controlled Type I error rate or has sufficient power when data were of a paired structure. In our second study, we aim to evaluate clustering performance of RNA-Seq data that were subjected to a variety of data transformations to make them “look” more normally distributed. In planning these two studies, an interest in the ethics behind personalized medicine via genomic sequencing was sparked.

1.2 Ethics in Precision Medicine

According to the National Human Genome Research Institute (NHGRI) through the National Institutes of Health (NIH), medical science will take on an extremely personalized view in the next 50 years. Their hope was to use genome-based research to develop “highly effective diagnostic tools”, “better understand the health needs of people based on their individual genetic make-ups”, and “design new and highly effective treatments for disease” (National Human Genome Research Institute et al., 2010). Moreover, the goal is to have individualized analysis based on a given person’s genome to gather information to regarding the types of preventative measures that can be prescribed, lifestyle changes, and even molecular understanding of diseases such as diabetes, heart disease, or cancer which are make up a large portion of the amount of medical expenditures in any given year in the United States (National Human Genome Research Institute et al., 2010). Formally defined, precision medicine, or also synonymously termed personalized or individualized medicine, is the tailoring of disease treatments and/or interventions to the unique characteristics, both genotypic and phenotypic, that an individual has (Ciardiello et al., 2014). However, with the idea of using genomic sequencing to tailor medical treatment, concerns over the ethics behind such approach to medical care ensue. There are numerous concerns regarding how incidental finds should be handled, identification of the individual, and many others to be mentioned in Chapter IV. As advances in personalized medicine through way of genomic sequencing continue, it will be crucial to understand cancer patients’ opinions regarding the topic. Thus, a protocol was developed to carry out a pilot survey study which contains a 22-item survey with questions regarding patient’s demographic information and their opinion towards genomic testing.

The studies involved in this dissertation are motivated by the need for accurate, efficient, and consistent statistical methods to analyze RNA-Seq data; as well as, the ethical concerns that arise with genomic sequencing. Chapter 2 contains a comparison study of paired and unpaired methods for Differential Expression Analysis of RNA-Seq. An empirical study is completed comparing the number of similar genes found between sets of overlapping methods. Additionally, we conducted a simulation study to examine consequences of improperly analyzing data structures common in RNA-Seq studies. Chapter 3 is comprised of a lengthy simulation study which assesses data transformations and clustering methods for RNA-Seq data. Clustering is completed under the assumption that the number of clusters is known or unknown. Chapter 4 describes the setup our patient’s opinion survey study which can potentially be extended to a large-scale, national survey. This dissertation concludes with an overall discussion and possible future work that are motivated or can be extended by these works (Chapter 5).

CHAPTER 2

A Comparison of Paired and Unpaired Methods for Differential Expression Analysis of RNA-Seq Data

Janelle R. Noel, Rama Raghavan, Joe Usset, Prabhakar Chalise, Byunggil Yoo, Jeremy Chien, Brooke L. Fridley

To be published in PLOS ONE

2.1 Abstract

Discovery of differentially expressed (DE) genes is imperative for the understanding of the genomic basis of complex diseases and phenotypes. Thus, the development of powerful computational methods with control of the Type I error rate for analysis is crucial. In this study, we applied multiple DE analysis methods to an RNA-Seq study involving paired ovarian tumor samples pre- and post- treatment with carboplatin from 11 subjects. Our objective was to investigate how much statistical power is gained by using a paired analysis method for RNA-Seq data when a generalized linear model is fit with either subject effect modeled as a covariate or as a random effect which can be difficult for small sample sizes. Moreover, we wanted to gain insight into whether fitting a more complex model, which accounts for pairedness across subject's measurements for small sample size (i.e., n = 11), is more beneficial than ignoring the paired data structure and proceeding with an unpaired analysis method. Additionally, we sought to see how results changed between various distributional models for RNA-Seq count data—Negative Binomial, Poisson, of Gaussian. To accomplish these objectives, we compared the results between a number of DE methods that do and do not account for the paired nature of the study to assess the power gained and/or increase and Type I error rates using this ovarian study and an extensive simulation study. Results from our empirical study found that the DE methods do not select the same set of DE genes, with only a few DE genes found to be in common between the different analyses. To investigate the root cause of poor overlap of determined DE genes, a simulation study was conducted with the objective to examine Type I error rates and power. The simulation study contained multiple scenarios which varied the following: from which distribution the data were simulated, the sample size in each group (i.e., N = 100, 150, and 200 samples), and the level of correlation / dependency between the measurements (i.e., $\rho =$

$0, 0.3,$ and 0.5). Data were simulated under the null hypothesis to assess the control of the Type I error rate, assuming correlated (paired) or uncorrelated (unpaired) data measurements from Bivariate Normal, Bivariate Poisson, and Bivariate Negative Binomial distributions. Following the simulation of the data, two types of models where then fit to determine DE genes; method that account for the correlation or paired-ness of the data and ones that do not. The simulation results demonstrated that Type I error was controlled for all paired and unpaired scenarios where data were simulated from the Bivariate Normal distribution. However, this was not the case for data simulated using Bivariate Poisson and Bivariate Negative Binomial Distributions. Type I error rate was only controlled at the 0.05 level when data were unpaired and analyzed using a Generalized Linear Model (GLM). Concurrently, fitting the more complex Generalized Linear Mixed Model (GLMM) resulted in controlled Type I error rate in the Bivariate Poisson and Bivariate Negative Binomial paired data when measurements were correlated at $\rho = 0.3$ and $\rho = 0.5$, respectively. Furthermore, empirical power was calculated for those scenarios for which Type I error rate was controlled (Type I error < 0.05). Overall results suggest that data structure should not be ignored when conducting analyses, especially if study sample size is lower. While control of the Type I error rate was not affected by sample size, the power to conduct analyses that reached control of the Type I error rate did vary with sample size, specifically power increased with larger sample size regardless of simulated correlation and distribution framework as expected. Additionally, our findings showed that if Type I error rate was controlled beyond < 0.05, that power would be loss in comparison to those scenarios controlled at right around 0.05. In conclusion, our results advocate that it is more beneficial to fit the more complex model to account for pairedness of subjects' measurements.

*Keywords***:** Differential Expression, Next-Generation Sequencing, RNA-Seq, Type I Error Rate, Simulation Study, Power, Correlated data, Paired study designs

2.2 Introduction

RNA Sequencing (RNA-Seq) studies can be used to address many critical research questions. Most commonly, RNA-Seq studies address questions relating to the relative abundance of read expression counts that are present for a given gene. One of the most fundamental analysis that is performed on RNA-Seq count data is differential expression (DE) analysis. As its name suggests, DE analysis is the comparison of gene expression values among samples from varying experimental conditions. Some examples of experimental conditions that might lead to genes that are differentially expressed include comparisons of: normal tissue verses tumor tissue; different tissue types; and tissue samples before, during or after a given treatment or exposure. Inherently, one could assume that the different experimental conditions have the potential to produce expression differences across samples for a given gene; however, some genes may remain unaffected (Hardcastle, 2016). At face value, analysis for DE seems rather simple. Although, according to Trapnell et al. (2013), two major challenges exist: (1) obtaining gene and isoform expression values accurately from raw sequencing data, and (2) handling the variation that is present across biological replicates within an experiment. The resulting goal of DE analysis is to determine in a gene-wise fashion those genes that are differentially expressed according to a specified cutoff of a given evaluation criteria (i.e., p-value or False Discovery Rate (FDR)) when ranked (Love et al., 2014, Trabzuni et al., 2014).

Quantification of gene expression via read counts is based on how many reads absolutely or probabilistically align with the reference genome (Conesa et al., 2016). These read counts can be modelled through the use of a discrete distribution; such as the Poisson distribution or the Negative Binomial distribution (Anders, 2010, Robinson, 2007). However, it should be noted for RNA-Seq analysis that it is important to know the background of the samples that have been processed. In RNA-Seq type experiments, typically more than one sample is obtained—rather multiple samples, or replicates, are obtained for a given condition. These replicates can be technical replicates meaning that they are from the same organism; or they can be biological replicates meaning that they are from different individuals. This is relevant in selecting the distribution that best fit the data. The Poisson distribution is typically chosen when the data are comprised of technical replicates (Marioni et al., 2008). Though, for biological replicates, the Negative Binomial distribution is more appropriate as the overdispersion parameter can be tuned to account for the variation between people. Additionally, data from RNA-Seq studies tends to be collected in a paired data structure that is represented by varying (e.g., samples from different tissue types) or contrasting conditions (e.g., before and after treatment) (Chung et al., 2013). Microarray data also have technical and biological replicates. However, when fitting and analyzing microarray data the Gaussian framework is applied as expression values are continuous.

Working with count distributions is often less appealing than normal distributions as the mathematical theory restricts "performance and the usefulness of RNA-Seq analysis methods" (Law et al., 2014). Even with the rapid advancement of technology, limitations exist in the range of statistical tools available to handle count distributions as compared to normal distributions (Law et al., 2014). While current statistical tools used to analyze RNA-Seq count data have attempted to incorporate many types of count distributions and models, no tool has been universally adopted. The arguments when analyzing RNA-Seq data, in general, have been a continued debate as to whether or not the data "needs" to be analyzed using discrete distributions or is it possible to use a Normal distribution which is continuous. Though, as any type of sequencing experiment is highly costly, small sample sizes may be ideal for researchers (Hardcastle and Kelly, 2010). However, these small sample sizes can cause major challenges and issues for statisticians. Statisticians are constantly working to improve the statistical theory to allow for analyses that are hindered by the type of study data, sample size availability, and model / data assumptions. There is evidence in statistical literature showing that correct modeling of the mean-variance relationship inherent in a count data generating process is key to designing statistically powerful methods of analysis (McCullagh and Nelder, 1989). Additionally, in conducting DE analysis, consideration needs to be given regarding the multiple testing that occurs—separate hypothesis is test for each of the thousands of genes (Trabzuni et al., 2014).

Count data alone are fairly straightforward; however, the attributes of sequencing count data make it unique. In sequencing data, sequencing depth and library size are taken into consideration when preforming analysis. Additionally, transcripts are not independent from one another; and across the genome much of the information is shared (Trabzuni et al., 2014). However, the Generalized Linear Model (GLM) has the capability to adapt and handle the for mentioned sequencing attributes and ability to handle complex experiments (Anders et al., 2013). Linear Mixed Models (LMMs) have also been used in the analysis of gene expression data. One caveat to fitting a LMM is that they are limited as they do not allow for the differences in expression variability that are typically present in RNA-Seq data (Trabzuni et al., 2014).

When it comes to the analysis of paired structured count data, many have tried using many variations of the above listed discrete distributions. In earlier years prior to the invention of the microarray or next-generation sequencing, Farwell and Sprott (1988) and Lee (1996) considered the use of a mixture model to handle the nuances of count data. However, these early attempts at testing paired structured data assume independence of the paired data which is conditioned on the samples mean (Chung et al., 2013). In general, the Poisson model can be utilized when samples are independent of one another, rather no replicates are present. However, in recent years it has been recognized that the paired nature of the data should be analyzed accordingly. The Negative Binomial model and the Bivariate Poisson model have been proposed to be used when handling paired samples as they can account for correlation between observations (Chung et al., 2013, Karlis and Ntzoufras, 2006, Khafri et al., 2008).

In the sections that follow, we investigate how much statistical power is gained by using a paired analysis method for RNA-Seq data when a generalized linear model is fit with either subject effect modeled as a covariate or as a random effect which can be difficult for small sample sizes. Moreover, we wanted to gain insight into whether fitting a more complex model, which accounts for pairedness across subject's measurements for small sample size (i.e., n = 11), is more beneficial than ignoring the paired data structure and proceeding with an unpaired analysis method. Additionally, we seek to know how results change between various distributional models for RNA-Seq count data—Negative Binomial, Poisson, of Gaussian. To do so, we conduct multiple comparisons between seven differential expression methods that do and do not account for the paired nature of the study to assess the power gained and/or increase in Type I error rates using an ovarian cancer study and a simulation study. Summaries are then provided for both the empirical and simulation studies.

2.3 Materials and Methods

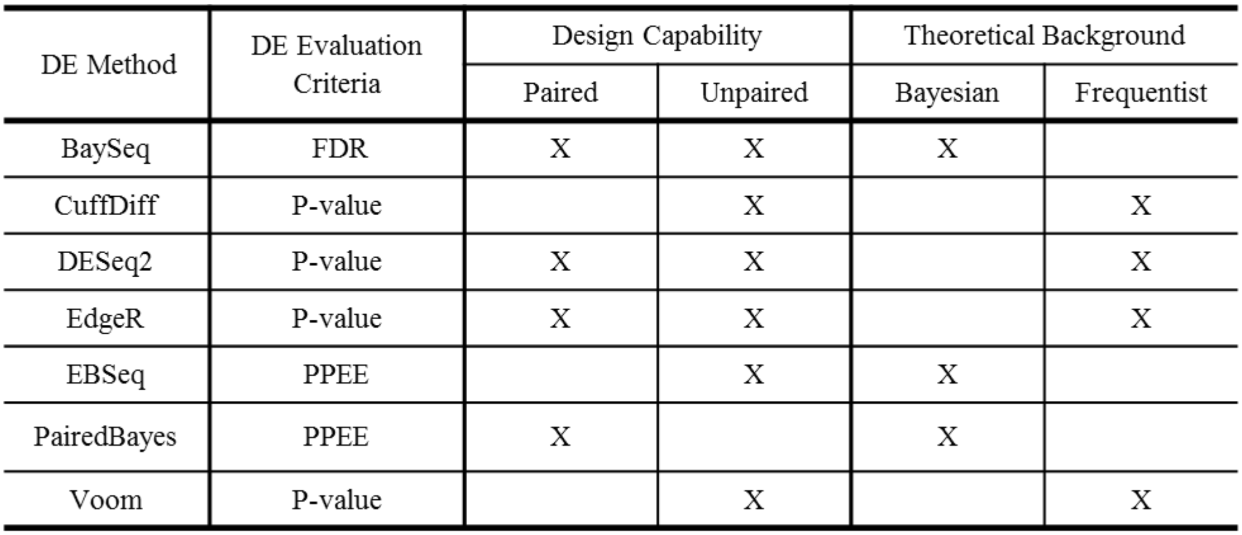
To build upon the current knowledge in the literature, we address the aims of this study through an empirical study of several differential expression (DE) methods to find genes that are differentially expressed. Conducting analyses that assess the differences expression, rather differences among the counts in transcripts or exons, across varying experimental conditions is a fundamental building greater understanding about the human genome (Robinson et al., 2010). A consensus has yet to be reached regarding which of the developed tools for evaluating is best. This likely is due to the general complexities of RNA-Seq data, and in turn model variations that serve as the basis in each of the DE methods. Often DE methods are capable of addressing data that are paired and/or unpaired. Hence, it is important when running the methods to correctly specify the fit models to insure that they are consistent with the structure of the data. This leads us to a second investigation. Additionally, we explore the model basis for testing for differential expression for each of the DE methods. We wanted to look at the shear basic models that are fit to test for differentially expressed genes without influence from any other tuning factors that are implemented by "black box" programs or packages. This is accomplished through a simulation study which simulates paired and unpaired data structures from varying distributions and varied levels of pairedness, and investigates control of the Type I error rate when data structure is considered in test for differential expression. All analyses for this study were conducted in R statistical software (R Development Core Team, 2016).

2.3.1 Empirical Study

To date, many studies have been completed for comparing DE methods. Searching for genes that are differentially expressed when different experimental conditions have been exhibited, has been said to be the most popular use of transcriptome profiling (Soneson and Delorenzi, 2013). In 2013, several similarly themed articles were published that compared analyses methods for differential expression by Rapaprot et al. (2013), Soneson and Delorenzi (2013), Guo et al. (2013), and Seyednasrollah et al. (2013). However, there has been limited results on comparison of methods that account for the paired-ness in the study design. In our study, we chose to compare DE methods based on whether or not the method can handle both paired (i.e., measurements taken from the same subject at different time points) and unpaired samples (i.e., assuming that all samples and measurements are independent from one another). Currently, there are nearly 15 developed methods that can conduct differential expression analysis each of which use different normalization techniques, read count distribution assumptions, methods for estimating the over-dispersion parameter in the negative binomial distributional model, or in the or type of test used to determine differential expression.

In our empirical study, we investigated seven commonly used methods to determine differentially expressed genes between two groups of samples (e.g, tumor-normal, treatment – no treatment). These seven methods consisted of BaySeq, CuffDiff, DESeq2, EdgeR, EBSeq, PairedBayes, and Voom. Most of these methods are commonly used in practice and are contained within packages in Bioconductor (Gentleman et al., 2004). Methods were selected based on their ability to handle either or both paired (i.e., matched pairs) and unpaired (i.e., independent) data. Additionally, they were selected to reflect both Frequentist and Bayesian theoretic backgrounds. Table II-1 provides a summary of the aforementioned method’s design and theoretical attributes. In Table II-1, it can be noticed that the DE evaluation criteria are not the same for all DE methods.

By default, the evaluation criteria are the p-value, False Discovery Rate (FDR), and posterior probability of equal expression (PPEE). Recall, the p-value is the probability of obtaining an observed or greater result assuming that given null hypothesis holds true. FDR is a metric that evaluates “the proportion of errors committed by falsely rejecting the null hypotheses” when multiple test are conducted (Benjamini and Hochberg, 1995). From Bayesian statistics, the posterior probability is the probability of observations being assigned to relative groups given the data (Gelman, 2013). The cutoffs for each of the method's evaluation criteria were determine to reflect those value which would traditionally be used for such type of analysis. It should also be mentioned each method was carried out using only the default codes and functions. Additionally, it should be noted that baySeq, DESeq2, edgeR, and EBSeq apply filters to remove those genes that are not expressed; also, in Cuffdiff genes with low expression are removed (Leng et al., 2013).



DE Method	DE Evaluation Criteria	Design Capability		Theoretical Background	
		Paired	Unpaired	Bayesian	Frequentist
BaySeq	FDR	X	X	X	
CuffDiff	P-value		X		X
DESeq2	P-value	X	X		X
EdgeR	P-value	X	X		X
EBSeq	PPEE		X	X	
PairedBayes	PPEE	X		X	
Voom	P-value		X		X

Table II-1. Summary of Differential Expression methods. Differential Expression (DE) methods are summarized based upon their design capabilities (i.e., the ability to handle data that are paired in nature and those that are truly independent from one another); as well as, the theoretical backgrounds behind each method. The DE methods use False Discovery Rate (FDR), p-value, or Posterior Probability of Equal Expression (PPEE).

Following the analyses that used solely defaults, we extended our analysis in converting our evaluation criteria to be similar across all methods so that we might make stronger conclusions about the number of DE genes that are determined by each of the methods. To accomplish this, we needed to find an evaluation criteria that would be suitable across all DE methods. For our scenarios, it was decided to change our p-values and PPEE to a FDR. Converting PPEE is fairly simple as the way in which it is calculated is actually an estimate for the FDR (Leng et al., 2013). It is also possible to convert p-values to FDR using theory developed my Benjamini and Hochber (1995), Efron et. al. (2001) , Storey (2010), and Storey et. al. (2015). This procedure has been simplified by the development of the *qvalue* package in Bioconductor (Storey et al., 2015, Gentleman et al., 2004). A list of p-values can be supplies to the *qvalue()* function within the package. Calculations are carried out to determine the local FDR (lFDR) which can be used as an estimation of FDR. Specifically, the lFDR in an extension of the FDR developed by Benjamini and Hochberg (1995) which allows for a posterior probability for each feature level (Chong et al., 2015). Details on this calculation can be found in Appendix A.

2.3.1.1 Ovarian Tumor Study

The study data that we empirically evaluated came from Dr. Jeremy Chien in the Department of Cancer Biology at the University of Kansas Medical Center. In Dr. Chien's study 11 matched pair ovarian cancer samples were obtained. These matched samples were obtained from the same patient taken pre- and post-treatment of intravenous Carboplatin. Carboplatin is a chemotherapy medication that damages genetic material in cells making it harder for repair of any genetic material (Rozencweig et al., 1983). Each patient was consented for tumor collection and DNA testing. Tissue samples were flash frozen to preserve their attributes and further processed by cryosection at 20-30 micron sections. RNA from these sections was extracted through the use of Trizol. Furthermore, the Illumina RNA-Sequencing kit was used to generate the sequencing libraries. Sequencing was completed using Illumina HiSeq 2000 using eight samples per lane. Running eight samples per lane helped to combat variations due to lane effects which include "any errors that occur from the point at which the sample is input to the flow cell until data are output from the sequencing machine" (Auer and Doerge, 2010). Output for each of the samples contain read counts for approximately 63K Ensembl gene IDs. Figure II-1 displays the relationship between the log-transformed mean and variance of this RNA-Seq data from the ovarian tumor study. As we expect from RNA-Seq data, we observe that our data are overdispersed with respect to the mean and variance relationship.

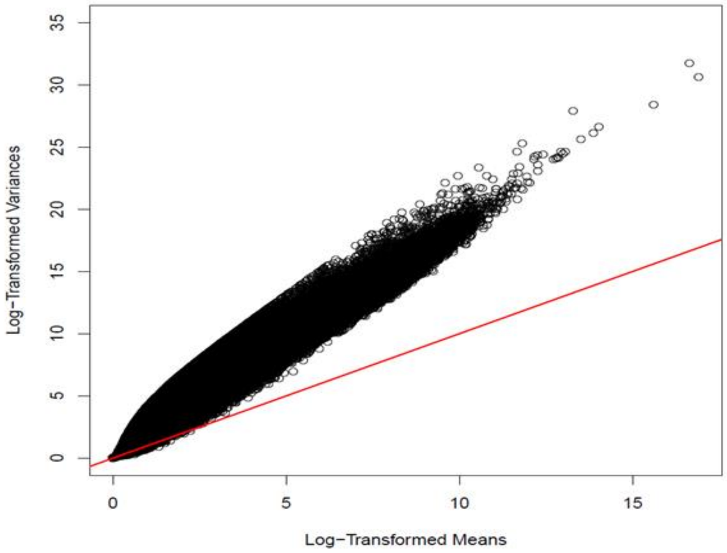


Figure II-1. Comparison of log-transformed mean and log-transformed variance across samples per Ensemble gene ID from the empirical study data. Log-mean and log-variance of gene express counts were calculated and plotted verses each other to show overdispersion present in the data. The red 45 degree line is representative of equal mean and variance (i.e., equal dispersion).

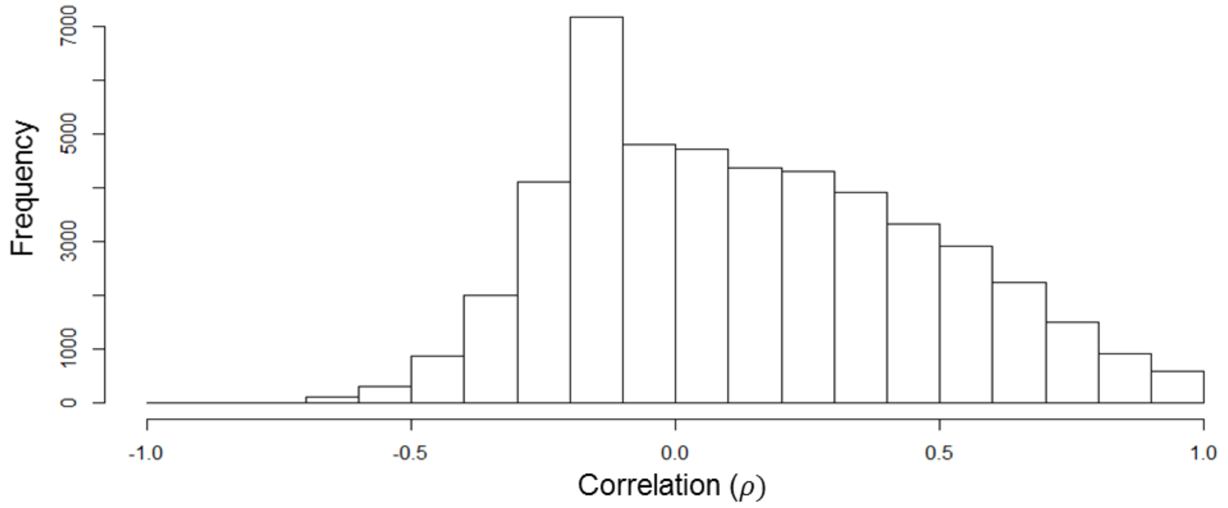


Figure II-2. Distribution of gene-wise correlation from the empirical study data. Histogram displays the frequency of correlations that were found in data from the empirical study. Correlations were calculated for individual genes between pre- and post-treatment samples.

The data from the ovarian tumor study are paired in nature as the RNA-Seq expression measurements were taken prior to the patient being treated with Carboplatin and post completion of the study. The correlation between all 63K+ Ensembl gene IDs range from -1 to 1. Seventy-five percent of the gene correlations were seen between -0.25 and 0.5 (Figure II-2). In our empirical study, we sought to conducted differential expression analysis for this study using methods that do or do not account for the pairedness in the measurements. Additionally, we broke the relationship between the pairedness of the samples and analyzed them as if they were independent observations (i.e., ignoring the paired-nature of the data). This relationship is broken so that we can gain a better understanding of the consequences that arise when the data structure is not considered in the statistical analysis. The structure of the data for the paired and unpaired are given below.

For this empirical study, let $\boldsymbol{X}$ be the G by N matrix where x_{gi} is the raw RNA-Seq expression count for the g^{th} gene ($g = 1, \dots, G$) and the i^{th} sample ($i = 1, \dots, N$). Here, $G =$ 62,897 Ensembl gene IDs and $N = 22$ samples which reflects the 11 patients that have two measures recorded to reflect pre- and post-treatment expression levels

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{G1} & x_{G2} & \cdots & x_{GN} \end{bmatrix}.$$

In the paired analyses, we define an additional $22\ x\ 1$ vector that defines when sample were taken say $\boldsymbol{q}$. Values for $\boldsymbol{q}$ are assigned as such:

$$q_i = \begin{cases} 0\ \ if\ measurement\ taken\ pre - treatment \\ 1\ if\ measurement\ taken\ post - treatment \end{cases}.$$

Thus, $\boldsymbol{q} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. Additionally, we have a subject vector, $\boldsymbol{s} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 11 \\ 1 \\ 2 \\ \vdots \\ 11 \end{bmatrix}$. Using $\boldsymbol{X}, \boldsymbol{q}$ and $\boldsymbol{s}$ an appropriate model or design matrix can be defined. The unpaired analysis does not utilize $\boldsymbol{q}$ as we assume that there is no dependency between the treatment groups and treat measurements as if they were independent.

2.3.1.2 Methods Used in Differential Expression Analysis

Testing for differentially expressed genes from two or more conditions requires conducting a statistical test for each gene $G = 1, \dots, g$. The most simple hypothesis for differential expression is assuming that one experimental condition which yields a “prior to” condition and an “after” condition. Rather, in looking at differential expression, we test the null hypothesis that the expression of a gene remains equal when comparing two or more conditions (i.e., equally expression regardless of experimental condition)

$$H_o: \mu_{g,\ A} = \mu_{g,\ B}$$

where $\mu_{g,\ A}$ or $\mu_{g,\ B}$ is the mean expression of the g^{th} gene in condition *A* or *B*. Rejecting this null hypothesis results in the conclusion that the gene of interest is differentially expressed. The goal with any method that conducts differential expression analysis is to minimize the number of type I errors (controlling at a given alpha level), while have the most power to detect a true difference.

baySeq

baySeq employs an empirical Bayesian method for identifying differential expression in sequence count data; and has the capability of considering more than just pairwise comparisons by borrowing information across the dataset unlike its counterparts, edgeR, DEGSeq, and DESeq (Hardcastle and Kelly, 2010). Specifically, the empirical Bayesian method is used to estimate posterior likelihoods for patterns of differential expression by gene (Hardcastle, 2016). These posterior probabilities are assessed through the consideration of a parametrically defined distribution for which some prior distribution exists (Hardcastle and Kelly, 2010). Determination of DE genes is based upon similarity of their prior distributions—genes that are similar will have the same prior distribution, while genes that are different will have different prior distributions (Hardcastle and Kelly, 2010). In this approach, new estimates for the prior probabilities for each model can be obtained by an iterative procedure starting at an initial choice (Hardcastle and Kelly, 2010). Once convergence of this iterative process is reached, the estimate is assumed to be found. Through the use of this numerical Bayesian method the structure of the original data is maintained (Hardcastle and Kelly, 2010). Details of this method can be found in Hardcastle and Kelly (2010).

Cuffdiff

Cuffdiff, a program within Cufflinks, is a differential expression analysis method that models variability of RNA-Seq library fragments by using individual transcript and across all replicates (Trapnell et al., 2012, Trapnell et al., 2013). The method seeks to test statistical significance of the observed change in gene expression between two or more samples for a given condition. Statistical significance is tested through the use of a mixture model containing a Beta and Negative Binomial Distribution, a Beta Negative Binomial distribution model, which assumes that the number of reads per transcript is proportional to its abundance (Trapnell et al., 2013, Trapnell et al., 2012). The Beta distribution accounts for the uncertainty in the transcript fragment counts while the Negative Binomial distribution considers the overdispersion that is present across counts (Trapnell et al., 2013). An advantage that Cuffdiff has is that upstream analysis from Cufflinks has the ability to remove the source bias that is a result of the protocol used in the library preparation prior to completion of assessing genes that are differentially expressed (Trapnell et al., 2012).

DESeq2

DESeq2 is another RNA-Seq method which tests for differential expressions under the Negative Binomial Generalized Linear Model (GLM) framework for paired samples (Love et al., 2014, Love et al., 2016). DESeq2 is the improved version of DESeq (Anders and Huber, 2010) with the addition of its capability to use shrinkage to estimate fold change and dispersion (Love et al., 2014). The implementation of the DESeq2 method follows below. Read counts are modeled as $NB{\sim}\left(\mu_{gi}, \alpha_g\right)$ where $\mu_{gi} = s_{gi}q_{gi}$. q_{gi} is a quantity that is proportional to the concentration of cDNA fragments scaled by s_{gi} for the g^{th} gene and the i^{th} sample (Love et al., 2014). The link function for the GLM is $\log_2 q_{gi} = \sum_r \beta_{gr} x_{ir}$. Following closely to the empirical Bayes procedure mentioned in baySeq, empirical Bayes shrinkage is used to obtain the new estimates for dispersion and fold change. Utilizing a shrinkage type estimation is highly beneficial for moderating "noisy estimates" that may be the result of controlled experiments with small sample size (Love et al., 2014). To test for differential expression, a Wald test is used to compare β coefficients.

EdgeR

EdgeR is a Bioconductor package that performs differential expression analysis between two or more groups through the use empirical Bayes estimation (Robinson et al., 2010, Robinson, 2008, Robinson, 2007). One constraint that this software implements is that replicated measurements must be present for at least one of the groups (Robinson et al., 2010). EdgeR utilizes read counts from multiple unpaired or paired samples that are compiled into FASTQ files and later processed into BAM files. EdgeR is highly flexible in that it can account for samples that are independent or paired/matched. Raw read counts are model using an overdispersed Poisson model which can be written as a Negative Binomial(Robinson et al., 2010). Formally, the count data are model as

$$Y_{gi} \sim NB(M_i p_{gj}, \phi_g)$$

for gene g and sample i; where M_i is the library size or total number of reads, p_{gj} is the relative abundance of gene g in experimental group j, if appropriate, to which sample i belongs, and ϕ_g the dispersion for the gene g (Robinson et al., 2010). It follows that the mean and variance for this parameterization are $\mu_{gi} = M_i p_{gj}$ and $\mu_{gi}(1 + \mu_{gi}\phi_g)$, respectively. It should be noted that when the dispersion parameter is equal to zero, the model becomes Poisson. The combination of the overdispersed Poisson or Negative Binomial distribution to model the data and the empirical Bayes estimation procedure help to account for the technical and biological variability present across genes and allow for information to be borrowed between genes (Robinson, 2007).

EBSeq

EBSeq is another empirical Bayesian approach to determining differential expression. Though this method is not limited to solely DE in genes, it has the ability to identify DE isoforms when inputs are estimates of isoform expression (Leng et al., 2013). Prior to testing, expression values are normalized using Median Normalization which accounts for the variability across samples (Anders and Huber, 2010). Following the empirical Bayes process mentioned above in baySeq and DESeq2, posterior likelihoods are estimated to determine genes that are differentially expressed (Seyednasrollah, 2013). Specific details can be found in Leng et al. (2013).

pairedBayes

The last of the Bayesian methods investigated is paired Bayes. As the name suggests, this method utilizes a Bayesian hierarchical approach that is capable of handling paired gene expression data—both within and between sample variation are accounted for in this method (Chung et al., 2013). The over-arching goal is to determine differential expression through the estimation of the posterior probability of a given gene (Chung et al., 2013). The Poisson-Gamma mixture model that is used has priors assigned to some of its parameters. Following Chung et al. (2013), we start with the paired design we have two observations $\left(Y_{gi1}, Y_{gi2}\right)$ where Y_{gi1} is the observed expression level before treatment, Y_{gi2} is the observed expression level after treatment, $g = 1, \dots, G$ genes and $i = 1, \dots, n$ samples. When conditioning Y_{gi1} and Y_{gi2} on their true baseline relative expression to the library size (i.e., λ_{gi}), and the expression level fold chance after treatment (i.e., $\mathcal{X}_g$), the basis of our mixture model, the Poisson portion, takes shape as:

$$Y_{gi1}|\lambda_{gi}, \mathcal{X}_g \sim Poisson\left(N_{i1}\lambda_{gi}\right)$$

$$Y_{gi2}|\lambda_{gi}, \mathcal{X}_g \sim Poisson\left(N_{i2}\lambda_{gi}\mathcal{X}_g\right)$$

where N_{i1} and N_{i2} are the sizes of the libraries. Here, the goal is to test if there is any treatment effect, or rather, where $\mathcal{X}_g \neq 1$. Furthermore, the Gamma portion of the mixture model which is used to account for overdispersion takes on the form

$$f_\lambda\left(\lambda_{gi}\right) = \frac{\beta_g^{\alpha_g}}{\Gamma\left(\alpha_g\right)} {\lambda_{gi}}^{\alpha_g - 1} e^{-\beta_g \lambda_{gi}}$$

where the shape and rate parameters are α_g and β_g, respectively. In this two-component mixture model we are able to describe the fold change distribution through a latent variable z_g.

$$z_g = \begin{Bmatrix} 0\; with\; probability\; of\; equal\; expression\; (EE), \pi_0 \\ 1\; with\; probability\; of\; differential\; expression\; (DE), \pi_1 \end{Bmatrix}.$$

Within the model hierarchy above, prior distributions are assigned to many of the parameters. Those priors are as follows: $\log(\mathcal{X}_g)|\left(z_g = 0\right) \sim Normal(0, \sigma_o^2)$; $\log(\mathcal{X}_g)|\left(z_g = 1\right) \sim Normal(\mu_1, \sigma_1^2)$; non-informationve priors for α_g and β_g; $(\pi_0, \pi_1\ ) \sim Dirichlet(1,1)$; μ_1 has an improper prior; $p(\sigma_0^2) \propto\ {}^{1}\!/_{\sigma_0^2}$; and $p(\sigma_1^2) \propto\ {}^{1}\!/_{\sigma_1^2}$. It should be noted that joint independency exists between all the parameters.

Voom

Variance modeling at the observation level, termed Voom, is a method located within the *limma* software package (Smyth, 2005, R Development Core Team, 2016) which aims to conduct differential expression while considering the mean-variance relationship that exists among counts (Law et al., 2014). Voom does so by applying precision weights to normalized counts while considering the trend of the mean-variance (Law et al., 2014). Estimation of the mean-variance trend of log transformed reads is completed non-parametrically (Law et al., 2014). Once the estimates are obtained, they are used to predict the variance of each of the log counts per million (cpm) values (Law et al., 2014). The predicted variance is then incorporated into inverse weights for corresponding log-cpm values (Law et al., 2014). A summary of the procedure to find the associated weights continues as such: using the normalized log-cpm values gene-wise linear models are fitted; residual standard deviations for each gene are produced and a robust trend is fitted; results from the linear model and standard deviation trend produce predicted count and count size, respectively; and weights for a given observation are specified by the inverse squared predicted standard deviation for said observation (Law et al., 2014). From here, the information, log-cpm values and associated weights, can be put into the *limma* pipeline to for differential expression.

2.3.2 Simulation Study

In this section, we describe the setup of the conducted simulation which investigates controlling Type I error rate when testing for differential expression of paired and unpaired data with methods that account for the study design / repeated measurements nature of the data. Conduction of this type of simulation study to determine the validity of a statistical model for differential expression has become the most popular method used (Reeb, 2013). The following simulation study extends from a recent study conducted by Reeb and Steibell (2013). In our simulation study, we simulate both paired and unpaired data from Normal, Poisson, and the Negative Binomial distributions. These distributions were specifically selected as each have been used to simulated gene expression data at some point in history. The Normal distribution has been used in the past to simulate continuous microarray data; whereas, the Poisson, and Negative Binomial distributions have been used to simulate RNA-Seq data. Though, some researchers apply data transformations to the count data with the aim to make the count data more normal which allows for the use of methods that suitable for continuous data. One of the data transformation that is used is log("expression count value" + 1).

Recently, the most common way to simulate RNA-Seq data has been through use of variations of the Negative Binomial distribution. In addition to simulating data from different distributions, we further consider different sample sizes (i.e., N = 100, 150, and 200 subjects). As an aside, it should be mentioned that smaller sample sizes were attempted in our simulation study prior to settling on the aforementioned sample sizes. The initial smaller sample sizes we used were N = 5, 10, and 25. However, when fitting the Generalized Linear Mixed Models (GLMM) using these smaller sample sizes caused convergence issues.

In addition to simulating data with different sample sizes, we varied levels of pairedness through variations in correlation (i.e., $\rho = 0, 0.3$, and 0.5). Correlation of $\rho = 0$ means that the data have no link with each other and can be considered to represent an "unpaired design". When correlation is present, data have the potential to be deemed as paired in nature depending on the context of the research study. Each scenario undergoes analysis to test whether a treatment effect is present (i.e., differential gene expression between the two conditions/groups). Paired (or repeated measures or correlated) data are analyzed using both paired and unpaired statistical methods. Similarly, unpaired simulated data are also analyzed using both paired and unpaired statistical methods. In doing so, we further address the aims of this study to determine how well the Type I error rate is controlled and statistical power considering relationships between the data structure and capabilities of statistical methods. Data corresponding to a single gene for N subjects were simulated for each scenario, with 1,000 datasets simulated for each simulation scenario. Though it should be mentioned that the original goal of this study was to address the same aim mentioned above in small sample sizes (i.e., Is there a loss of control of Type I error rate or loss of power?). Due to convergence issues, sample sizes were increased. In Section 2.5, we discuss some of the future work that may combat the convergence issue.

2.3.2.1 Data Simulation

For purposes of this study, data were simulated gene-wise (i.e., one gene at a time) for N = 100, 150, and 200 subjects for paired and unpaired data structures. Utilizing information from the ovarian tumor study conducted at KUMC, we observed that nearly 70% of correlation between the paired samples were between the vales of -0.25 and 0.5. Hence, it was decided to use correlations values of $\rho = 0$, 0.3, and 0.5 as more positive correlation was exhibited in the genes pre- and post-treatment in the empirical study. Those data that are simulated to have $\rho = 0$ are considered as unpaired data, while those simulated at $\rho = 0.3$ and 0.5 are referred to as paired data in this study. To simulate the desired correlations, we used Cholesky Decomposition developed by André-Louis Cholesky and Trivariate Reduction depending on which distributional framework is being used (Rencher and Christensen, 2012, Mardia, 1970). Data for all distributions were simulated bivariately to account for the pairedness between pre- and post-treatment observations we wish to induce for each gene. Under the null, genes are simulated to have equal means between repeated measurements. Conversely, when investigating power, genes are simulated with unequal means between pre-and post-treatment measurements—one of the treatment measurements is simulated with the addition of a given effect size.

Normal Distribution

Following a similar approach used to simulate pairwise single nucleotide variants (SNPs), we are able to simulate paired gene expression data while using the Normal distribution and assistance from a Cholesky Decomposition matrix (Liu et al., 2010, Rencher and Christensen, 2012). To do so, we first simulate a vector of two groups of independent Standard Normal random variables, Z. The groups follow to identify those expression values pre- and post-treatment. We then define a 2x2 correlation matrix (R), $R_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, to reflect the pairedness that we would like to simulate. In our study, we have $R_{\rho=0.3} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$ and $R_{\rho=0.5} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ which are positive definite (i.e., R is symmetric and $x^T Rx > 0$ for all x) (Rencher and Christensen, 2012). $R_{\rho=0.3}$ and $R_{\rho=0.5}$ undergo Cholesky Decomposition to obtain a lower triangular matrix, L, which can be multiplied by Z to obtain the correlated, or paired, random variables X. L is calculated as $\begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$, which implies $L_{\rho=0.3} = \begin{bmatrix} 1 & 0 \\ 0.3 & \sqrt{1-0.3^2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.3 & 0.9539 \end{bmatrix}$ and $L_{\rho=0.5} = \begin{bmatrix} 1 & 0 \\ 0.5 & \sqrt{1-0.5^2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.866 \end{bmatrix}$. Thus, the correlated random variables are calculated as $X_\rho = LZ \rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$. The end result when simulating data under the null is a bivariate normal distribution $\boldsymbol{Z} \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$. Conversely, simulation of the unpaired data (ρ = 0) was done so solely using the Standard Normal distribution (i.e., $N \sim (0,1)$) for each of the group of subjects. For those data simulated until the alternative hypothesis that the mean gene expression values are different, an effect shift, Δ, of either 0.3 or 0.5 was added to measurements to one of the simulated conditions; such as,

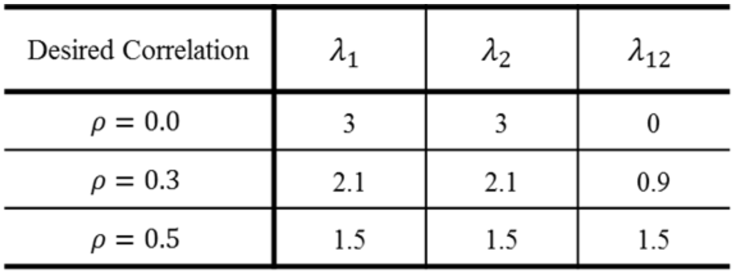
$$\begin{bmatrix} Z_{11} \\ Z_{12} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} 0 \\ \Delta \end{bmatrix}, \boldsymbol{\Sigma} \right].$$

Poisson Distribution

Simulation of paired and unpaired data from the Poisson distribution for two groups has most commonly been done through the use Trivariate Reduction proposed by Mardia (1970). Many other researchers have implemented this type of simulation in their count data simulation studies (Barbiero and Ferrari, 2014, Yahav and Shmueli, 2011, Johnson et al., 1997). This elegant method of simulation relies on the theoretical property that a sum of independent Poisson random variables is also distributed as a Poisson (Casella and Berger, 2002). Following Mardia’s Trivariate Reduction, we begin by generating three independent Poisson (i.e., $Z_1 \sim Poisson(\lambda_1),\ Z_2 \sim Poisson(\lambda_2),$ and $Z_{12} \sim Poisson(\lambda_{12})$). These variables are combined to generate two new dependent random variables--$X_1 = Z_1 + Z_{12}$ distributed $Poisson(\lambda_1 + \lambda_{12})$ and $X_2 = Z_2 + Z_{12}$ distributed $(\lambda_2 + \lambda_{12})$. Correlation between these two new dependent random variables becomes

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\lambda_{12}}{\sqrt{(\lambda_1 + \lambda_{12})(\lambda_2 + \lambda_{12})}}$$

(Yahav and Shmueli, 2011). However, prior to any data simulation using this approach, we needed to determine the rates, the λs, for each of the three Poisson random variables to relate this simulation to simulations from the other distributions used. Through some basic algebra, the rates were determined. Rates used in this simulation are found in Table II-2. By plugging in the respective rates into the three Poisson random variables, we achieve the desired correlation within our pre- and post-treatment data for 1,000 genes. It should be mentioned that the rate values are not unique. There are many other rate values that would satisfy the correlations that are utilized throughout our simulation study.



Desired Correlation	λ_1	λ_2	λ_{12}
$\rho = 0.0$	3	3	0
$\rho = 0.3$	2.1	2.1	0.9
$\rho = 0.5$	1.5	1.5	1.5

Table II-2. Rates for Poisson random variables. Table contains a summary of the rates that are required to be used in the simulation of the three Poisson random variables to achieve desired correlation through the Trivariate Reduction.

Similarly to simulating data under the alternative for the normal distribution, an effect shift was added to a portion of the simulated Poisson distributed data. It follows $Z_1 \sim Poisson(\lambda_1 + \Delta),\ Z_2 \sim Poisson(\lambda_2),$ and $Z_{12} \sim Poisson(\lambda_{12})$ were generated. Though Mardia's Trivariate reduction, we are left with $X_1 \sim Poisson(\lambda_1 + \Delta + \lambda_{12})$ and $X_2 \sim Poisson(\lambda_2 + \lambda_{12})$. The same rates from Table II-2 were used for this simulation of data under the null.

Negative Binomial Distribution

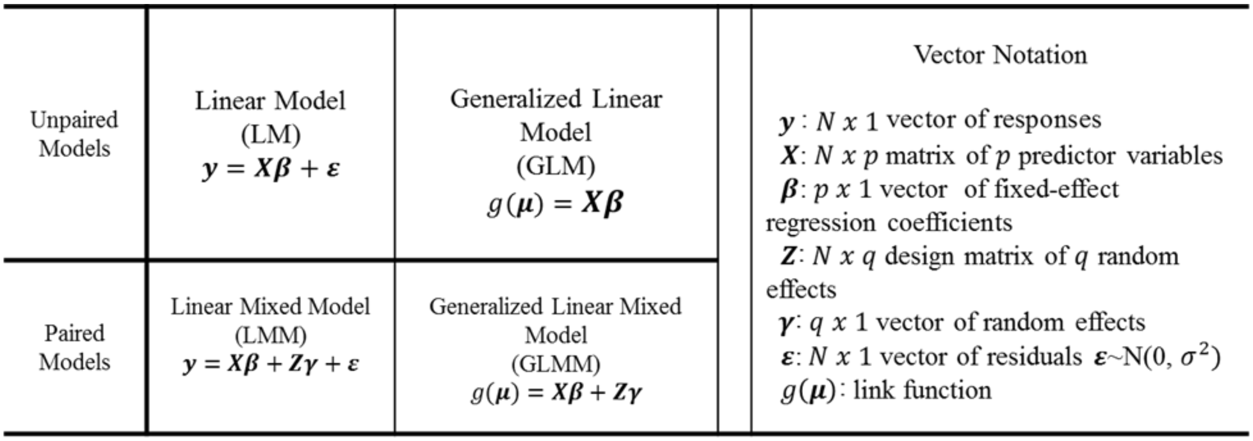
While simulation of Bivariate Negative Binomial data can also be accomplished using Trivariate Reduction, the method is not as straightforward due to number of parameters that are involved in calculating the mean and variance. To simulate our Bivariate Negative Binomial data, we use an approach that uses conditional sampling that is based on decomposition of two dimensional distribution of bivariate copulas (Erhardt and Czado, 2008). The use of the copulas aid in setting up the dependency (i.e., correlation) between the simulated random variables; as well as, help to obtain multivariate count distributions (Erhardt and Czado, 2008). The desired correlations remain the same as in above simulations (i.e., ρ = 0, 0.3, and 0.5). Specific details of this simulation approach can be Erhardt and Czado (2008). Fortunately, Erhardt and Czado have created an R package, *corcount*, that easily allows for the implementation of this type of simulation of bivariate data (Erhardt, 2009).

2.3.2.2 Application of Statistical Analysis Methods to Simulated Data

Completion of the simulation study looking at the statistical analysis of the paired verses unpaired data can be broken into two sections: 1) control of Type I error rate and 2) power. Once simulated data were generated, we were able to conduct analyses to evaluate how well the various statistical tests controlled the Type I error rate in settings where observations were uncorrelated and correlated. Additionally, if the Type I error rate ended up being controlled for a given scenario, we proceeded to determine the empirical power.

In this portion of the simulation, we assume the null hypothesis where the means of pre- and post-treatment expression values are the same (i.e., $H_o\colon \mu_{g,\ A} = \mu_{g,\ B}$). For data simulated from the Bivariate Normal distribution, we use the T-Test or Linear Model (LM) and the paired T-Test or Linear Mixed Model (LMM) (assuming equal variances) to test if there is a difference in the mean gene expression. For data simulated from the Bivariate Poisson and Bivariate Negative Binomial distributions, we use Generalized Linear Models (GLM) and Generalized Linear Mixed Models (GLMM). In the GLM, expression count values and treatment group were the response and predictor variables, respectively. In the GLMM, the response variable was also expression counts; however, treatment group and paired sample relation were modeled as fixed and random effects, respectively. The use of the generalized-type linear models allow us to describe how the mean depends on the linear predictor through some link function, g (i.e.,

$g(\mu_i) = \eta_i$, and how variance depends on the mean. In our scenarios where the generalized models are used, the log link function is used. All statistical test, were set to test for a difference in treatment effect. A summary of the tested models (i.e., LM, LMM, GLM, and GLMM) and be found in Table II-3. Control of the Type I error rate was established if the empirically calculated Type I error rate was $< \ \alpha$ where $\alpha = 0.05$. By setting $\alpha = 0.05$, we are in acceptance that five percent of the time we end up with a false positive result--rejecting the null hypothesis given that the null hypothesis is true.



			Vector Notation
Unpaired Models	Linear Model (LM) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$	Generalized Linear Model (GLM) $g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$	$\boldsymbol{y}$: $N\ x\ 1$ vector of responses $\boldsymbol{X}$: $N\ x\ p$ matrix of p predictor variables $\boldsymbol{\beta}$: $p\ x\ 1$ vector of fixed-effect regression coefficients $\boldsymbol{Z}$: $N\ x\ q$ design matrix of q random effects $\boldsymbol{\gamma}$: $q\ x\ 1$ vector of random effects $\boldsymbol{\varepsilon}$: $N\ x\ 1$ vector of residuals $\boldsymbol{\varepsilon}$~N(0, σ^2) $g(\boldsymbol{\mu})$: link function
Paired Models	Linear Mixed Model (LMM) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$	Generalized Linear Mixed Model (GLMM) $g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$	

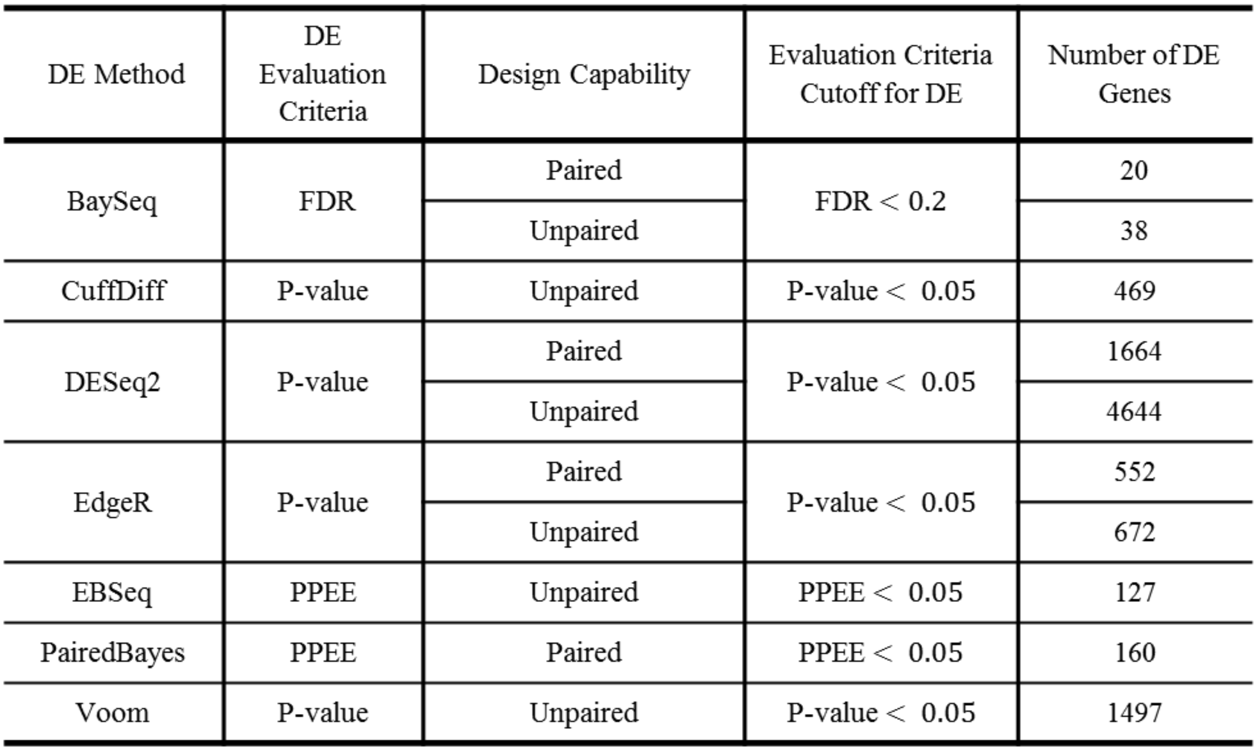
Table II-3. Unpaired and paired models fit in simulation study. Table summarizes the models that are fit to conduct the paired and unpaired simulation study and corresponding Type I error rate and power analyses. Models are presented in vector notation.

Once results for all of the scenarios are evaluated to determine their Type I error rate is controlled, we can continue by conducting an empirical power analyses. Power analyses can only be conducted if scenarios have control over the Type I error rate. The empirical simulation of power is nearly identical to the setup of the simulation for the empirical Type I error rate. However, in the empirical simulation, one of the treatment measurements, either pre- or post-treatment is simulated to have a shift applied to it's mean expression values. Essentially, an effect size is introduced between treatment measurements in the simulations.

2.4 Results

2.4.1 Comparison of Differential Expression Methods in Empirical Study

In comparing the results from all seven of the DE methods, we begin by looking at how many genes were determined to be DE based upon set cutoff values for each method's default evaluation criteria. All methods determined that the ovarian cancer samples contained genes that are DE in both a paired and unpaired design capability (Table II-4). The number of DE genes range from 20 genes to ~ 4,600 genes, with the most and fewest DE genes being determined by DESeq2 and BaySeq, respectively. An FDR of < 0.2 was selected as an equivalent evaluation criteria cutoff for DE genes for comparing methods with evaluation criteria of p-value and PPEE of < 0.05. The cutoff value for methods using the FDR criteria needed to be set higher as it is more stringent due to the way it accounts for multiple testing. We also observe that when comparing methods that are capable of both paired and unpaired designs that the unpaired design calls an increased number of DE genes verses the paired design (e.g., EdgeR unpaired design determines that there are 552 DE genes, while EdgeR paired design found 672 DE genes) (Table II-4).



DE Method	DE Evaluation Criteria	Design Capability	Evaluation Criteria Cutoff for DE	Number of DE Genes
BaySeq	FDR	Paired	FDR < 0.2	20
		Unpaired		38
CuffDiff	P-value	Unpaired	P-value < 0.05	469
DESeq2	P-value	Paired	P-value < 0.05	1664
		Unpaired		4644
EdgeR	P-value	Paired	P-value < 0.05	552
		Unpaired		672
EBSeq	PPEE	Unpaired	PPEE < 0.05	127
PairedBayes	PPEE	Paired	PPEE < 0.05	160
Voom	P-value	Unpaired	P-value < 0.05	1497

Table II-4. Number of genes found to be differentially expressed (DE) based on evaluation criteria cutoff. Each of the seven differential expression (DE) methods were executed for their respective design capabilities. The number of DE genes were determined by those genes which met the respective evaluation criteria cutoff. The DE methods use False Discovery Rate (FDR), p-value, and Posterior Probability of Equal Expression (PPEE) for their evaluation criteria.

Furthermore, to assess how well each of these seven methods performed in terms of selecting the same DE genes, comparisons of the intersection of similar DE Ensembl gene IDs were completed in both a pairwise and a multi-way manner. We began by making comparisons between those methods that were capable of paired and unpaired designs which include BaySeq, DESeq2, and EdgeR. Both DESeq2 and EdgeR contain an overlap of the DE genes that are selected; however, no overlap exists in the BaySeq comparison (Figure II-3). We also see that EdgeR resulted in the greatest proportion of DE genes found between the paired and unpaired designs (Figure II-3).

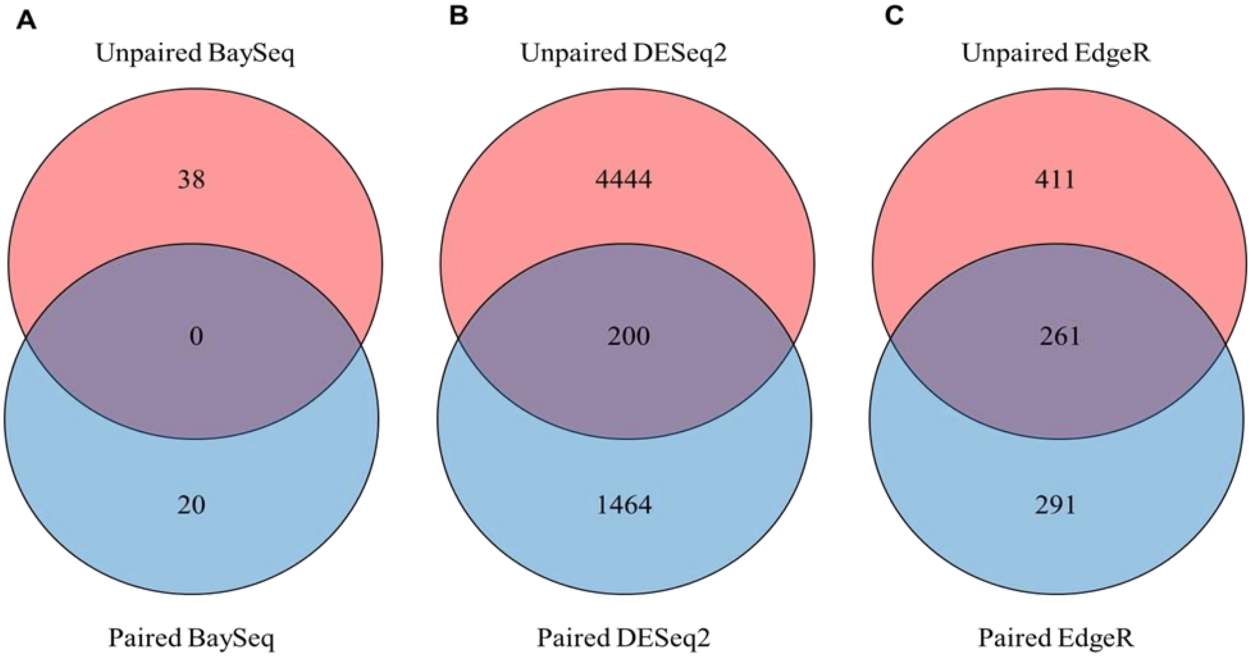


Figure II-3. Comparison of Differential Expression (DE) methods capable of paired and unpaired designs. The Venn diagrams above are contain of the number of DE genes that were determined by each method. The overlapping portion of the Venn diagrams represent the number of DE genes that were selected by both design context—unpaired verses paired. A) depicts DE genes found using BaySeq, B) depicts DE genes found using DESeq2, and C) depicts DE genes found using EdgeR. Criteria for DE genes in A) was FDR<0.2, and for B) and C) criteria was p-value<0.05.

Without altering the evaluation criteria to be similar across all methods, we proceeded in making comparisons across all unpaired and paired methods. Paired methods initially have less DE genes found (Table II-4). As the number of intersecting methods increase, we see that the number of DE genes that are the same between all methods reduce in number significantly for all multi-way comparison scenarios. The number of same DE genes found between all paired methods is ten genes, while only two genes were found in the unpaired methods (Figure II-4).

Another comparison was also made to consider the number of same DE genes found between the overlap of Bayesian and Frequentists methods. This resulted in zero and 14 genes, respectively for the Bayesian and Frequentists methods (Figure II-5). Lastly, combinations of five DE methods were compared. After combinatorically looking at all combinations of any five DE methods while using default evaluation criteria, it was found that the greatest number of same DE genes to overlap was 35 DE genes (Figure II-6). The five methods that determined 35 of the same DE genes were paired EdgeR, unpaired EdgeR, Paired DESeq2, Cuffdiff, and Voom. Although, when we continued to compare more than five DE methods, we noticed that with each additional DE method added to the comparison that there was a decrease in the number of genes that were determined to same as was also previously mentioned. When comparing all of the DE methods together, no similar DE genes were found.

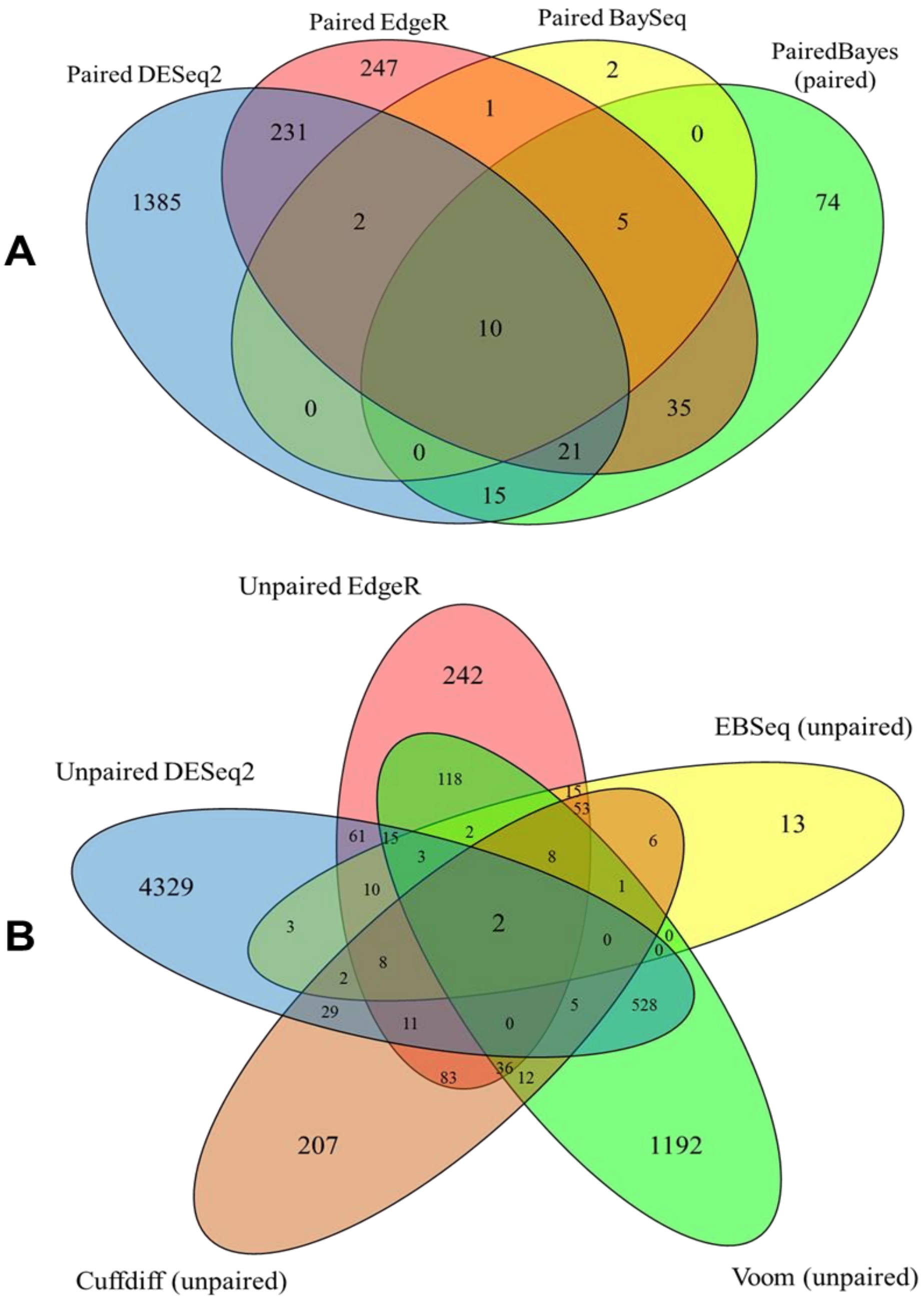


Figure II-4. Comparison of Differentially Expressed (DE) genes found in unpaired and paired methods. The Venn diagrams above contain the number of Differentially Expressed (DE) genes that were determined by each method. The overlapping portions of the Venn diagrams represent the number of DE genes selected to be the same between the compared DE methods. A) contains comparisons of DE gens found using paired designs; and B) contains comparisons of DE genes found using unpaired designs minus results from unpaired BaySeq. Default evaluation criteria were used for all DE methods.

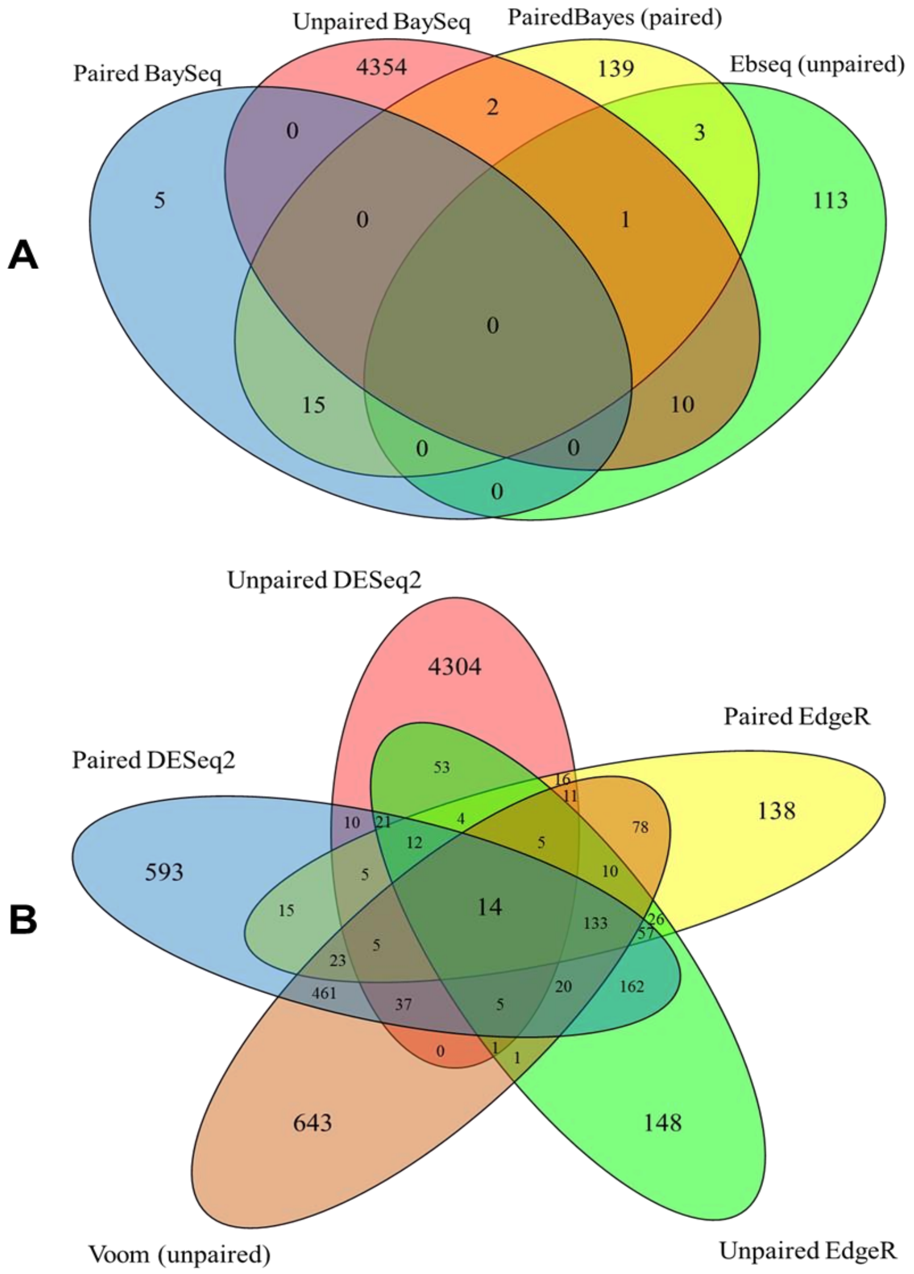


Figure II-5. Comparison of Differentially Expressed (DE) genes found Bayesian and Frequentist theoretical backgrounds. The Venn diagrams above contain the number of Differentially Expressed (DE) genes that were determined by each method. The overlapping portions of the Venn diagrams represent the number of DE genes selected to be the same between the compared DE methods. A) contains comparisons of DE gens found using Bayesian methods; and B) contains comparisons of DE genes found using Frequentist methods without results from CuffDiff. Default evaluation criteria were used for all DE methods.

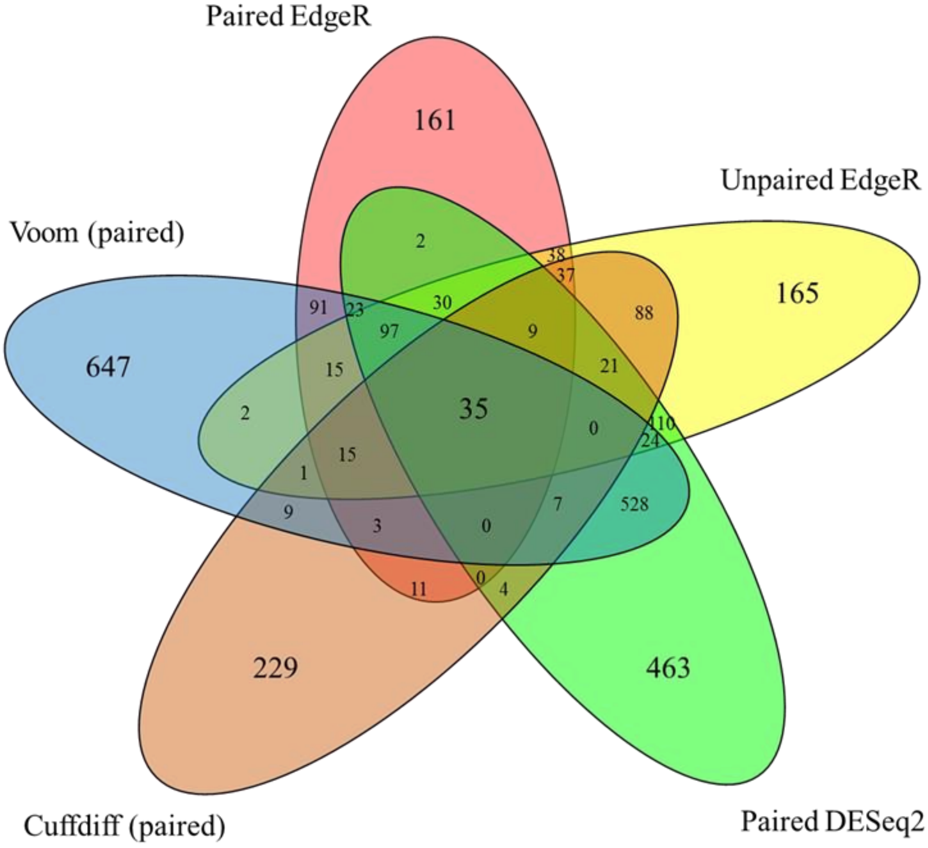
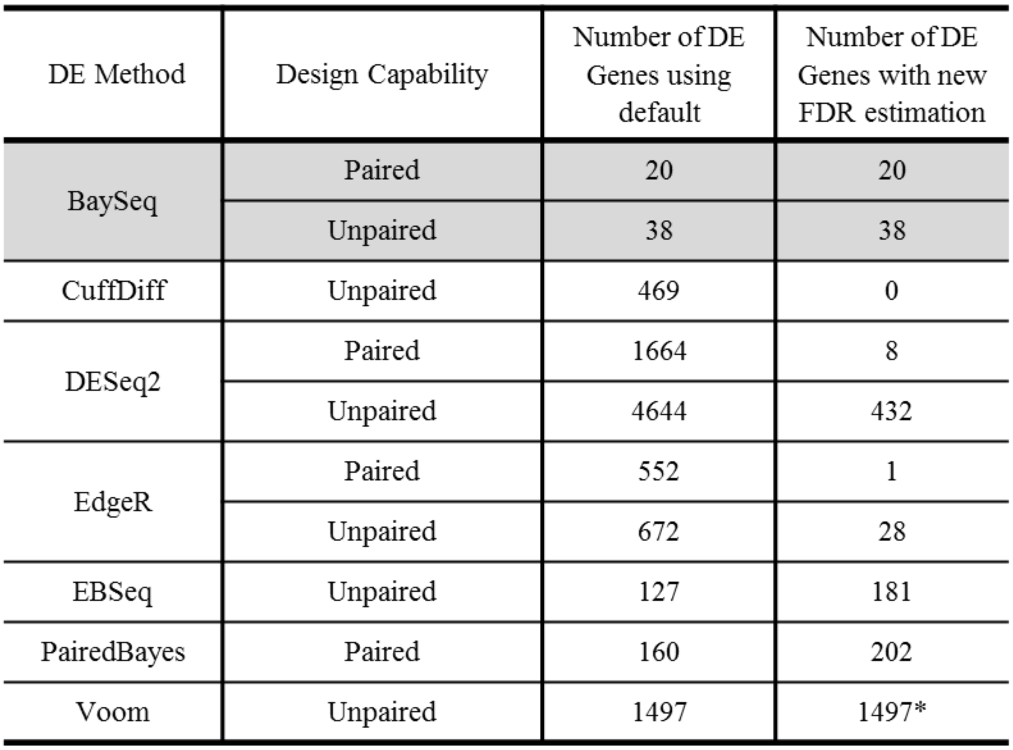


Figure II-6. Comparison of Differential Expression (DE) methods which find the most overlapping DE genes. The Venn diagrams above contain the number of Differentially Expressed (DE) genes that were determined by each method. The overlapping portions of the Venn diagrams represent the number of DE genes selected to be the same between the compared DE methods. The five methods in this figure produce the highest number of similar DE genes between comparison of all combinations of five DE methods.

In addition to conducting comparisons based off of default evaluation criteria. We converted all evaluation criteria to estimated FDR. While we would expect the number of DE genes that meet the criteria to decrease somewhat in using FDR, we notice that there are some extreme differences. The pattern that the unpaired designs contained larger numbers of DE genes verses that of the paired designs remains the same. Similar comparisons to those found using the default evaluation criteria were investigated. However, the results ended up being very poor in terms of methods selecting the same DE genes after re-evaluation. Most multi-way comparisons resulted in few to no DE genes that were selected to be the same. This was also the case for some two-way comparisons.

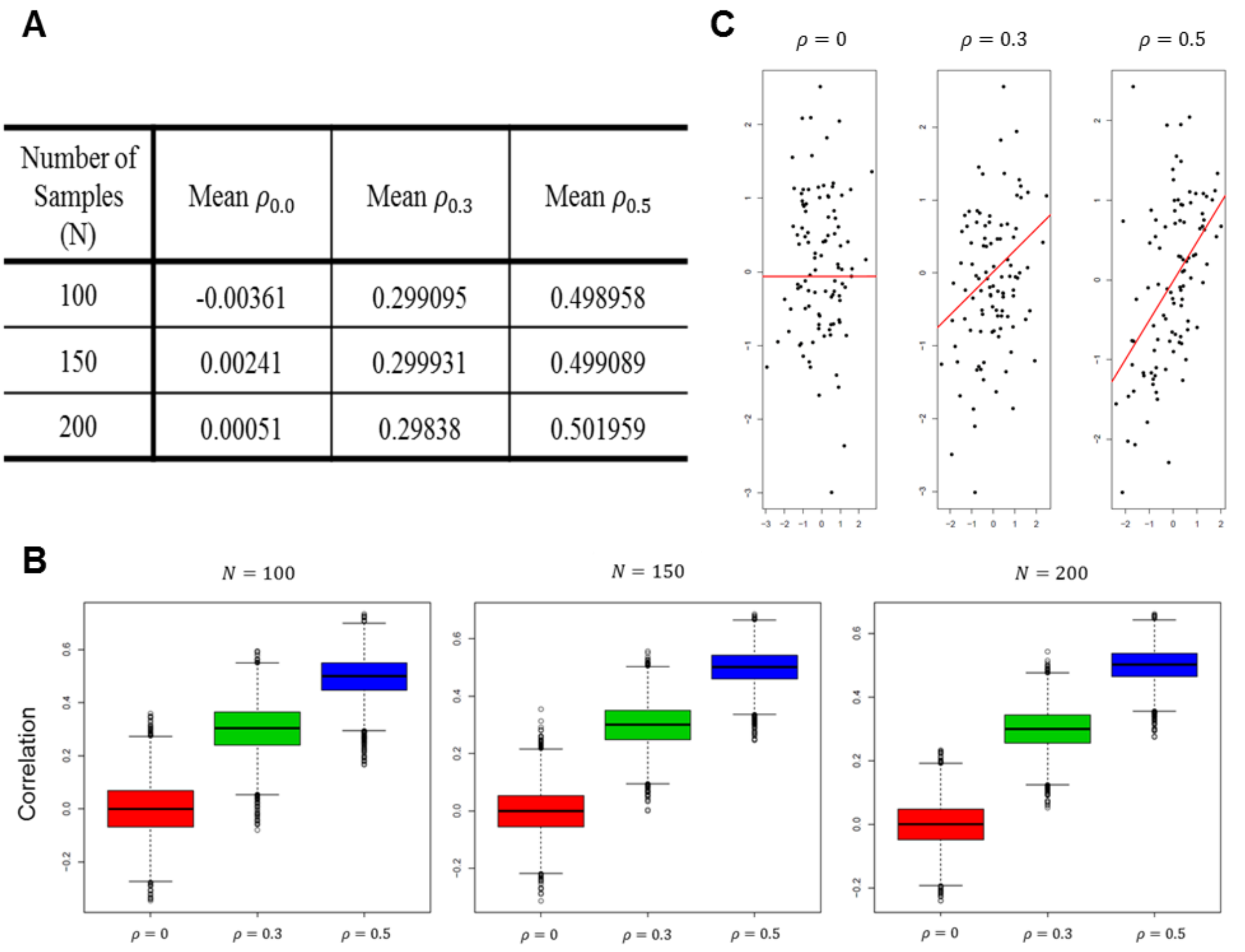


DE Method	Design Capability	Number of DE Genes using default	Number of DE Genes with new FDR estimation
BaySeq	Paired	20	20
	Unpaired	38	38
CuffDiff	Unpaired	469	0
DESeq2	Paired	1664	8
	Unpaired	4644	432
EdgeR	Paired	552	1
	Unpaired	672	28
EBSeq	Unpaired	127	181
PairedBayes	Paired	160	202
Voom	Unpaired	1497	1497*

Table II-5. Summary of genes that meet re-evaluation criteria of False Discovery Rate (FDR) < 0.2. Each of the seven Differential Expression (DE) methods were executed for their respective design capabilities. All evaluation criteria, p-values, and Posterior Probability of Equal Expression (PPEE), were converted to False Discovery Rates (FDR). No conversion took place for BaySeq. * Denotes that all DE genes met criteria.

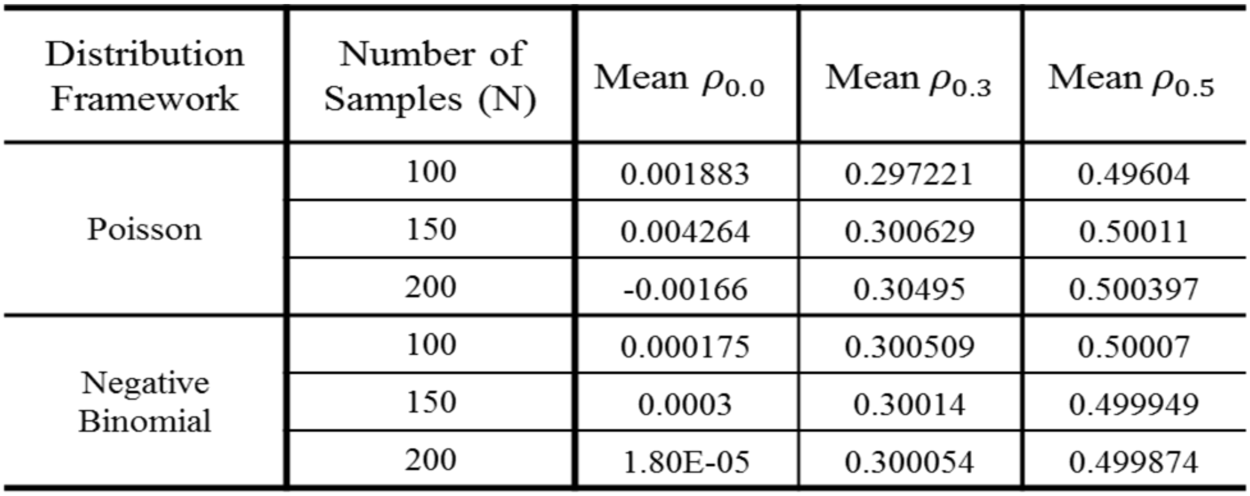
2.4.2 Simulated Data

In our simulation study, data were simulated in paired and unpaired structures from the following Bivariate distributions: Normal, Poisson, and Negative Binomial distributions. Prior to conducting any analyses, we verified that our simulated data contained the desired correlations between our repeated measures of $\rho = 0$, 0.3, and 0.5. Each of the simulated data scenarios under the null resulted in having average correlations matching those which were desired (Figure II-7(A) and Table II-6) with acceptable standard deviations (Figure II-7(B) and Figure II-A3). As there are no extreme deviations from the average correlations when compared to the desired correlations, we can assume that the distribution from which the data are simulated does not affect the outcome of the correlated data as long as the data simulation algorithm is set up correctly. An example for one gene with similar correlations to $\rho = 0$, 0.3, and 0.5 is plotted in Figure II-7(C). The solid red line that is plotted through the panels of Figure II-7(C) depict the relationship between the pre- and post-treatment simulated expression values.



Number of Samples (N)	Mean $\rho_{0.0}$	Mean $\rho_{0.3}$	Mean $\rho_{0.5}$
100	-0.00361	0.299095	0.498958
150	0.00241	0.299931	0.499089
200	0.00051	0.29838	0.501959

Figure II-7. Correlation summary for simulated data from the Bivariate Normal distribution under the null. Figure contains numeric and graphical summary of the correlations found in the simulation study for $\rho = 0, 0.3,$ and 0.5. Data are simulated from the Bivariate Normal distribution. A) contains a table summarizing the average correlations from the simulated data for N = 100, 150, and 200 for correlations $\rho = 0, 0.3,$ and 0.5. B) depicts the variability of correlations in simulated data for N = 100, 150, and 200 and $\rho = 0, 0.3,$ and 0.5. C) Simulated data for one gene are plotted for N = 100 samples for $\rho = 0, 0.3,$ and 0.5. Correlation trend lines are plotted in red.



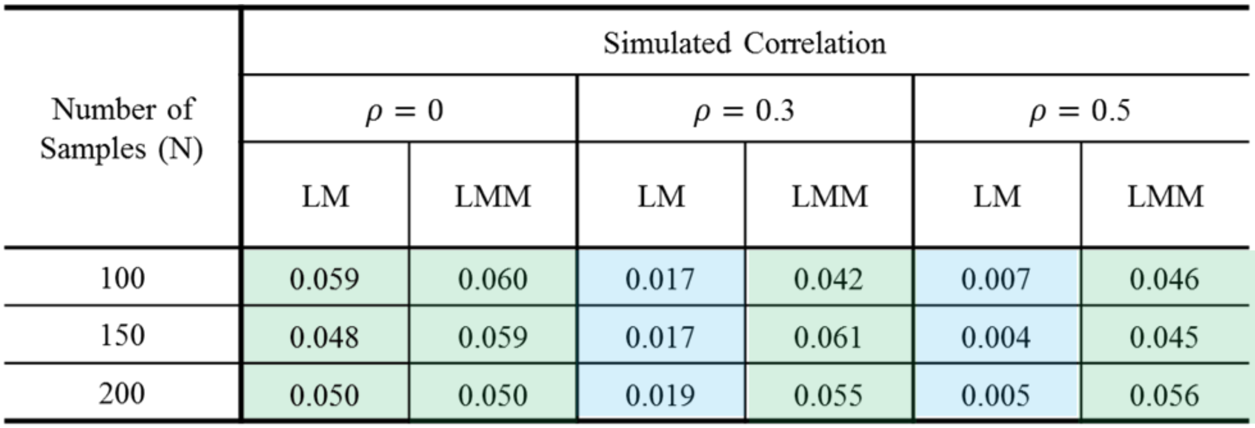
Distribution Framework	Number of Samples (N)	Mean $\rho_{0.0}$	Mean $\rho_{0.3}$	Mean $\rho_{0.5}$
Poisson	100	0.001883	0.297221	0.49604
	150	0.004264	0.300629	0.50011
	200	-0.00166	0.30495	0.500397
Negative Binomial	100	0.000175	0.300509	0.50007
	150	0.0003	0.30014	0.499949
	200	1.80E-05	0.300054	0.499874

Table II-6. Correlation summary for simulated data from the Bivariate Poisson and Negative Binomial distributions under the null. Table contains a summary of average correlations from the simulated data N = 100, 150, and 200 for simulated correlations of $\rho = 0, 0.3,$ and 0.5.

2.4.3 Comparison of Paired Verses Unpaired Analyses Techniques

With confirmation that our simulated data contained the correlations that we desired to allow for both paired and unpaired data structures, we proceeded with our analyses. Our objective of our simulation study was to determine if paired and unpaired analysis techniques controlled the Type I error rate when corresponding data structures were paired and unpaired, respectively. Likewise, we investigated whether or not Type I error rate was controlled if analyses were conducted where the data structure and analyses methods were not the same. Furthermore, if Type I error rate was controlled for a given scenario, we continued our simulation and analyzed empirically how well powered are analyses were. Rather, we wanted investigate the potential power gained from using a paired analysis method when the simulated study data are also paired; and to see if results varied based on the distribution that was used to simulate the data.

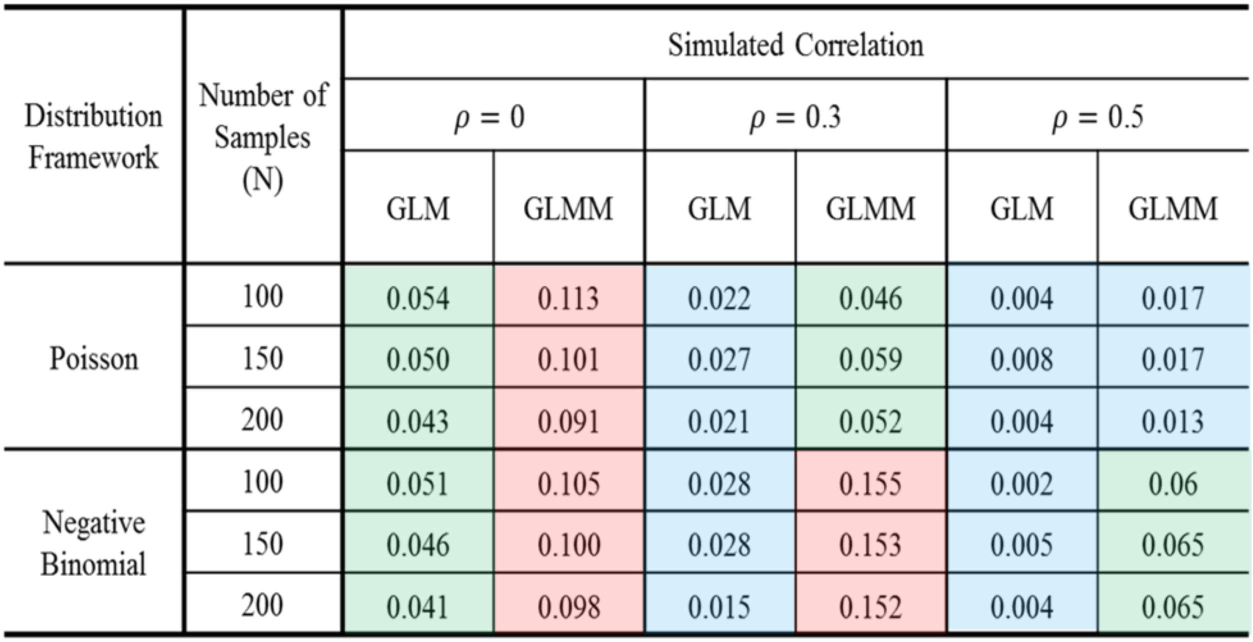
Beginning with our simulation results from the Bivariate Normal distribution simulated data, we observe that the empirically calculated Type I error rate is relatively controlled for all scenarios at the $\alpha = 0.05$ level (Table II-7). As correlation, or pairedness, is introduced into the simulated data, $\rho = 0.3$ and $\rho = 0.5$, we see that when using the LM to test for mean differences in gene expression values that the Type I error rate is over controlled or conservative in nature (Table II-7). For $\rho = 0.5$, the Type I error rate becomes very small under analysis using the LM framework—much less than the control threshold that was previously set to be < *0.05*. However, analyses of the simulated paired data while using a LMM seems to provide control of the Type I error rate. Although it should be noted that Type I error rate control is slightly missed for $\rho = 0.3$ and sample sizes of N = 150 and N=200; as well as, for $\rho = 0.5$ and N=200 (Table II-7).



Number of Samples (N)	Simulated Correlation					
	$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
	LM	LMM	LM	LMM	LM	LMM
100	0.059	0.060	0.017	0.042	0.007	0.046
150	0.048	0.059	0.017	0.061	0.004	0.045
200	0.050	0.050	0.019	0.055	0.005	0.056

Table II-7. Empirical Type I error rates from paired and unpaired analyses using the Bivariate Normal distribution under the null. Table contains a summary of empirical Type I error rates from the simulation study were the Bivariate Normal distribution was used to simulate study data. Error rates were calculated from 1,000 simulations. Cells shaded in green and blue have Type I error rate controlled near 0.05 and $<$ 0.05, respectively.

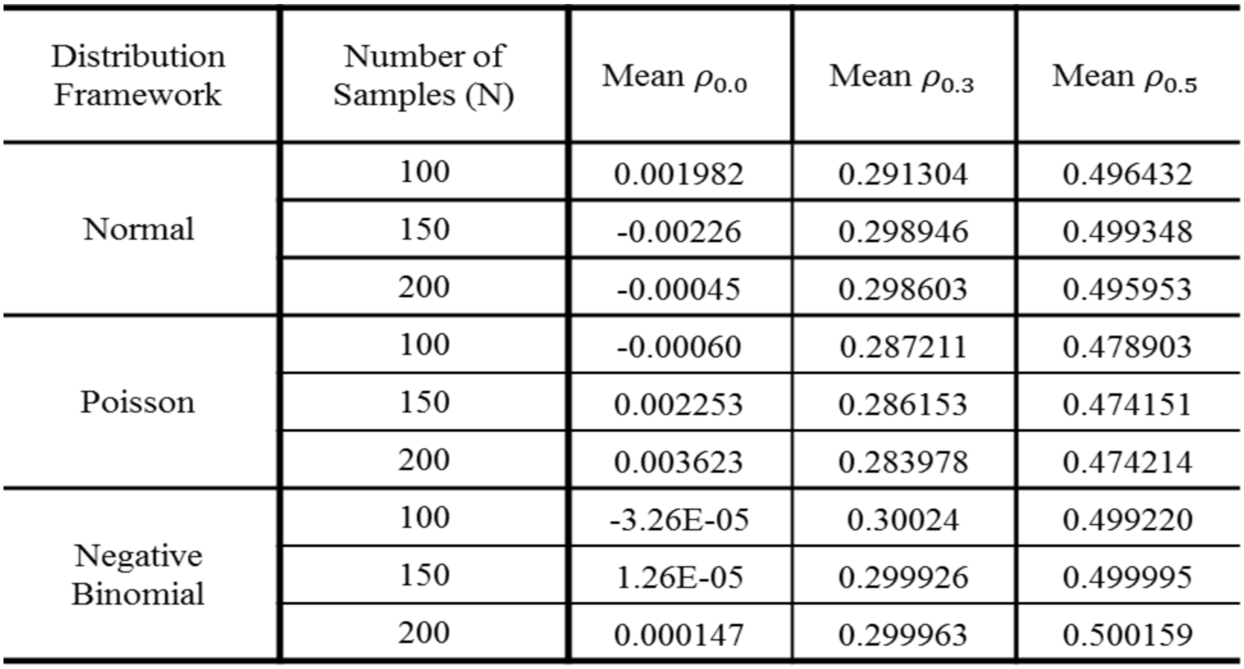
Next, we examine the results for data that were simulated from discrete distributions. Recall from above, GLMs and GLMMs are fit to model the dependency of both the linear predictors and variance with respect to the mean in our unpaired and paired analyses, respectively. Some of the same trends regarding control of the Type I error rate exist in the paired and unpaired analyses of the simulated count data. Particularly, in the Poisson and Negative Binomial results we observe that when the simulated correlation was $\rho = 0$ and analysis was completed using a GLM, Type I error rate was controlled at the < 0.05 level (Table II-8). Though, no control was observed for $\rho = 0$ when the GLMM was implemented—all empirically calculated Type I error rates are greater than 0.05. Additionally, we observe that control of the Type I error rate occurs $\rho = 0.3$ and $\rho = 0.5$ for in the scenarios using GLM for the Poisson and the Negative Binomial distributions used to simulate the paired data (Table II-7). Though, the same issue still arises that was seen in the Normal results found in Table II-6. The empirically calculated Type I error rate becomes very small in the aforementioned scenarios where $\rho = 0.5$. Control over the Type I error rate also exists for $\rho = 0.5$ using GLMM for analyses of data simulated from Poisson, and it is nearly controlled for the Negative Binomial. All other scenarios fail to control for Type I error rate.



Distribution Framework	Number of Samples (N)	Simulated Correlation					
		$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
		GLM	GLMM	GLM	GLMM	GLM	GLMM
Poisson	100	0.054	0.113	0.022	0.046	0.004	0.017
	150	0.050	0.101	0.027	0.059	0.008	0.017
	200	0.043	0.091	0.021	0.052	0.004	0.013
Negative Binomial	100	0.051	0.105	0.028	0.155	0.002	0.06
	150	0.046	0.100	0.028	0.153	0.005	0.065
	200	0.041	0.098	0.015	0.152	0.004	0.065

Table II-8. Empirical Type I error rate from paired and unpaired analyses using the Bivariate Poisson and Bivariate Negative Binomial distributions under the null. Table contains a summary of empirical Type I error rates from the simulation study. Bivariate Poisson and Bivariate Negative Binomial distributions were evaluated for N = 100, 150, and 200 for simulated correlations $\rho = 0, 0.3,$ and 0.5. Error rates were calculated from 1,000 simulations. Cells shaded in green, blue, and red have Type I error rate controlled near $0.05, < \ 0.05, and \ >$ 0.05, respectively.

For those scenarios in which the Type I error rate was controlled, we calculated the empirical power for such tests. Under all distributional frameworks, simulated correlation and all N, there was at least one scenario which contained a Type I error rate that was controlled. Thus, we simulated data under the alternative hypothesis (i.e., $H_a\colon \mu_{g,\ A} \neq \mu_{g,\ B}$). Consideration was given for two mean shifts in all pre-treatment gene expression simulated values. The two mean shifts that were considered were 0.3 and 0.5. Similarly to the null scenarios, correlations of the simulated data were examined to see how closely they matched the desired correlations. The mean correlations in the simulated data resembled the desired correlations for both mean shifts (Table II-9 and Table II-A1). Here, the mean correlation again does not appear to be affected by distributional framework. Thus, we are confident that the applied mean shifts were implemented in the simulation correctly.



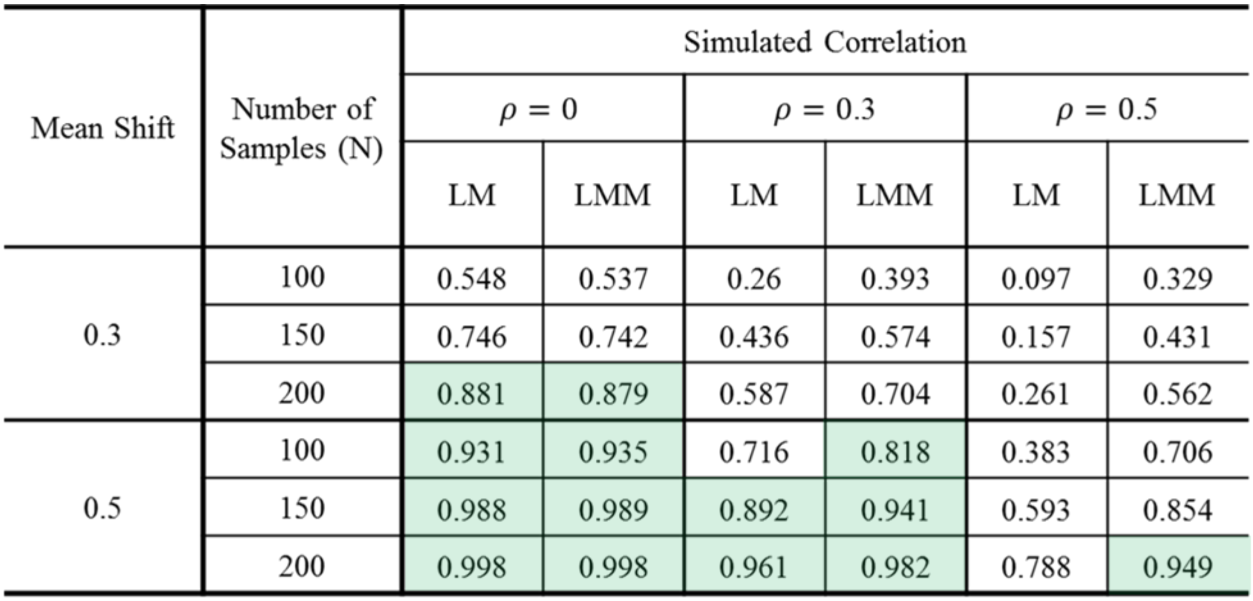
Distribution Framework	Number of Samples (N)	Mean $\rho_{0.0}$	Mean $\rho_{0.3}$	Mean $\rho_{0.5}$
Normal	100	0.001982	0.291304	0.496432
	150	-0.00226	0.298946	0.499348
	200	-0.00045	0.298603	0.495953
Poisson	100	-0.00060	0.287211	0.478903
	150	0.002253	0.286153	0.474151
	200	0.003623	0.283978	0.474214
Negative Binomial	100	-3.26E-05	0.30024	0.499220
	150	1.26E-05	0.299926	0.499995
	200	0.000147	0.299963	0.500159

Table II-9. Correlation summary for simulated data from the Bivariate Normal, Bivariate Poisson, and Bivariate Negative Binomial distributions under the alternative with one measurement having a shift of 0.3. Table contains a summary of average correlations for the simulated data for N = 100, 150, and 200 for simulated correlations $\rho = 0, 0.3,$ and 0.5. Data were simulated to reflect unequal means by the addition of a mean shift of 0.3.

As the desired correlations have been met on average, we continued with our empirical power simulation for those scenarios in which Type I error rate was controlled. Observed able in Table II-7, we see that all of the scenarios resulted in a Type I error rate that was controlled which allows us to determine the empirical power to conduct such test. Using the Normal framework, we first observe that the empirical power is highly variable among all simulations.

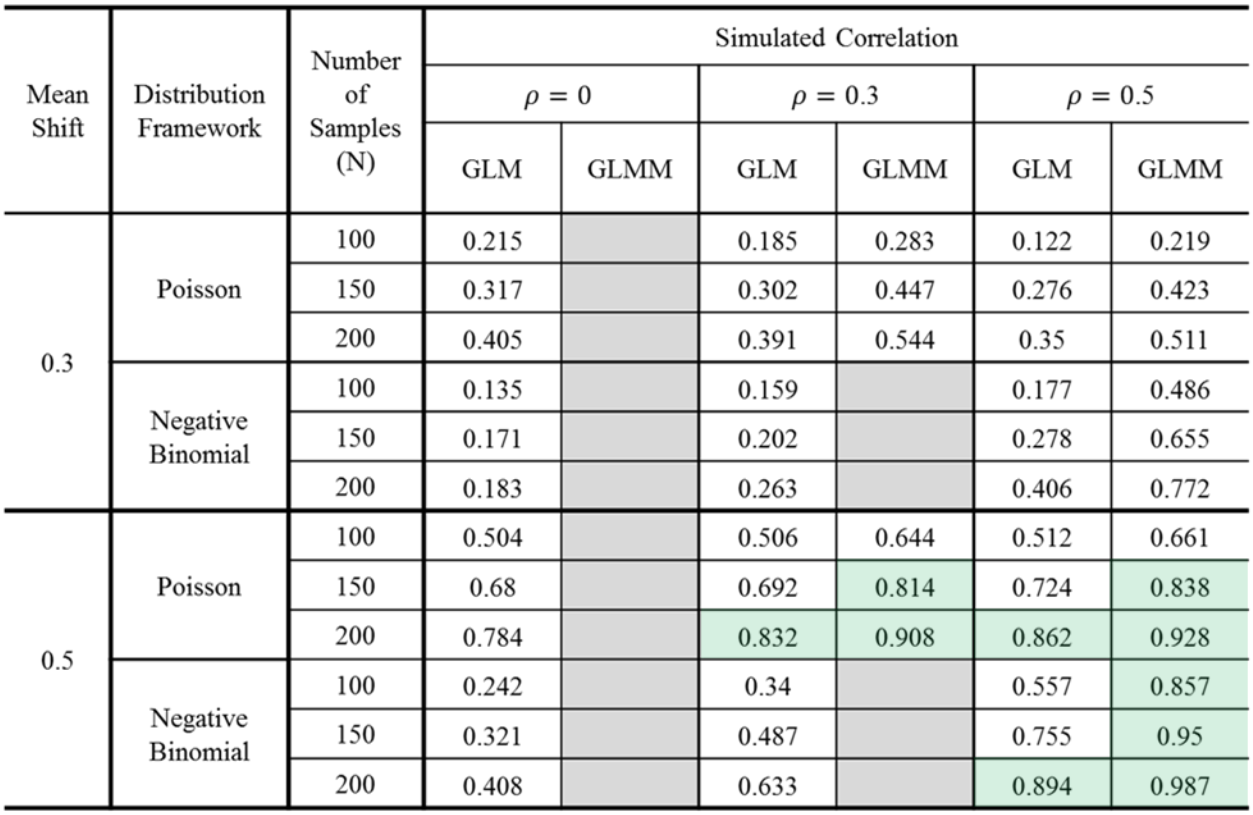
Most notably, we notice that as the number of samples increases for all simulated correlation values that the empirical power also increases (Table II-10). This is also the case when the mean shift was increased from 0.3 to 0.5 (Table II-10). Another comprehensive observation we observe is found when keeping the fitted model constant across the simulated correlation values, we notice that the empirical power decreases as simulated correlation increase (e.g., LM with $\rho = 0, 0.3,$ and 0.5 yields empirical power of 0.548, 0.26, and 0.097 for N = 100) (Table II-10). In the scenarios when $\rho = 0$, there are only minimal differences between statistical test that utilize a LM verses that of a LMM with respect to the calculated empirical power (Table II-10). This is not apply for $\rho = 0.3$ or $\rho = 0.5$. When comparing LM verses LMM, we see that the empirical power is greater for all sample sizes when fitting a LMM when the structure of the data are simulated in a paired way (Table II-10). Reaching 80% power is almost always obtained for a mean shift of 0.5 with exceptions at $\rho = 0.5$ for N = 100 and N = 150 samples.

In reviewing the simulation scenarios for the discrete distribution framework which controlled for Type I error rate above, it was determined that not all scenarios were controlled. Therefore, empirical power simulations only needed to be run for those scenarios in which Type I error rate was controlled. The results for the discrete distribution framework scenarios for both mean shifts are provided in Table II-11. The grayed out cells present in the table are representative of those scenarios where the Type I error rate was not controlled (Table II-11). According to the empirical simulation results using discrete distribution frameworks found in Table II-11, we observe that only a few scenarios reach 80% power. These scenarios that have empirically calculated power of at least 80% exist only for a mean shift of 0.5. When fitting a GLMM using the Poisson distribution, 81.4% and 90.8% power is reached when $\rho = 0.3$ for N = 150 and N = 200, respectively (Table II-11). Additionally, for $\rho = 0.5$ and GLMM model greater than 80% empirical power was achieved when N = 150 and N=200 (Table II-11). Concurrently for $\rho = 0.5$, the Poisson and GLM distribution framework and fit model resulted in 86.2% power (Table II-11). None of the scenarios for $\rho = 0$ reached 80% empirical power for either the Poisson or Negative Binomial distributional frameworks (Table II-11). Lastly, empirical power was obtained for all $\rho = 0.5$ scenarios when the model fit was a GLMM and the distribution from which the data were simulated was the Negative Binomial for all sample sizes (Table II-11). The only other combination of simulation components that yielded at least 80% empirical power was for N = 200, $\rho = 0.5$, GLM, and Negative Binomial distribution framework (Table II-11).



Mean Shift	Number of Samples (N)	Simulated Correlation					
		$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
		LM	LMM	LM	LMM	LM	LMM
0.3	100	0.548	0.537	0.26	0.393	0.097	0.329
	150	0.746	0.742	0.436	0.574	0.157	0.431
	200	0.881	0.879	0.587	0.704	0.261	0.562
0.5	100	0.931	0.935	0.716	0.818	0.383	0.706
	150	0.988	0.989	0.892	0.941	0.593	0.854
	200	0.998	0.998	0.961	0.982	0.788	0.949

Table II-10. Empirical power for paired and unpaired analyses using the Bivariate Normal distribution with varying mean shifts under the alternative. Table displays a summary of empirical power from the simulation study where the Bivariate Normal distribution was used to simulated the study data. Data sere simulated to reflect mean shifts of 0.3 and 0.5 applied to on of the treatment measurement's gene expression values. Cells shaded in green have reached power of at least 80%



Mean Shift	Distribution Framework	Number of Samples (N)	Simulated Correlation					
			$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
			GLM	GLMM	GLM	GLMM	GLM	GLMM
0.3	Poisson	100	0.215		0.185	0.283	0.122	0.219
		150	0.317		0.302	0.447	0.276	0.423
		200	0.405		0.391	0.544	0.35	0.511
	Negative Binomial	100	0.135		0.159		0.177	0.486
		150	0.171		0.202		0.278	0.655
		200	0.183		0.263		0.406	0.772
0.5	Poisson	100	0.504		0.506	0.644	0.512	0.661
		150	0.68		0.692	0.814	0.724	0.838
		200	0.784		0.832	0.908	0.862	0.928
	Negative Binomial	100	0.242		0.34		0.557	0.857
		150	0.321		0.487		0.755	0.95
		200	0.408		0.633		0.894	0.987

Table II-11. Empirical power for paired and unpaired analyses using the Bivariate Poisson and Bivariate Negative Binomial distributions with varying mean shifts under the alternative. Table displays a summary of empirical power from the simulation study were the Bivariate Poisson and Negative Binomial distributions were used to simulate the study data. Data were simulated to reflect mean shifts of 0.3 and 0.5 applied to one of the treatment measurement's gene expression values. Those scenarios that do not contain power information did not have control of the Type I error (< 0.05) in the null scenarios (Table II-7).

2.5 Discussion

Differential expression (DE) analysis has become a very popular type of analysis among researchers that work with RNA-Seq data. However, current comparison studies of DE methods do not seem to explain the rationale behind why certain methods were used verses other methods. To that, there is very little information regarding the impact that different data structures (i.e., paired and unpaired data) have within DE methods that are capable of accommodating varying data structures when selecting genes that are truly differentially expressed.

From our empirical study, our results provide insight to the considerations that research may need to make when conducting DE analyses. First, we recognize that comparison of DE methods, with default setting used, perform poorly in determining the same genes to be differentially expressed. This was the case for all types of comparisons that were made—comparisons between Frequentists and Bayesian theoretical backgrounds; as well as, looking at the comparisons between methods capable of handling all paired or unpaired data structures. Although, when comparing the paired verses unpaired DE methods, we found that more DE genes that were similar between the methods are found in the paired context. This is a result that was expected as the structure of our original ovarian tumor data was paired in nature. The study DE methods does not utilize statistical models that account for the pairedness by using models that allow for mixed effects—fixed or random. It is more common for the pairedness to be modeled as a fixed covariate. In terms of the theoretical background behind the DE methods, we observed that few DE genes overlapped between any two methods, let alone in the comparison of all Bayesian methods; these are crucial finding. It is expected that the Bayesian methods would find different DE genes as the Bayesian approaches can highly vary from one method to another due to set up of prior distributions and other model parameters. After summarizing the results, we think that it is fair to add to the challenges mentioned by Trapnell et al. (2013). An additional challenge of conducting DE analyses is that the varying attributes found in the different DE methods make it difficult to determine similar genes that are DE.

As the results from our empirical study using the ovarian cancer data were poor in terms of the number of similar overlapping DE genes that were found when comparing the many DE methods, we decided to conduct a simulation study to determine if our findings from the empirical study were inhibited by data structure and/or the model which was fit. Our simulation study was solely conducted from a Frequentist viewpoint and sought to determine how well varying scenarios through sample size, correlation values between repeated measures, and fit model controlled the Type I error rate. Furthermore, if the Type I error rate was controlled, we investigated if the statistical test performed had adequate power by introducing a mean shift in the simulated expression values (i.e., $H_o\colon \mu_{g,\ A} \neq \mu_{g,\ B}$). Both portions of this simulation study were done in an empirical matter, and we purposefully analyzed data structures in correct and incorrect fashions. The latter, helped to provide insights to the consequences that arise when performing incorrect analyses.

The results for the portion of the simulation which explored control of the Type I error rate were as expected, especially for the analyses using the Bivariate Normal distribution. For the Bivariate Normal scenarios, we observed that as the simulated correlation increased Type I error rate remained controlled right around the 0.05 value when a Linear Mixed Model (LMM) was fit and tested. This is likely due to the fact that in fitting the mixed model we are able to account for the pairedness between observations that is a result of the simulated correlation. We observe that the empirically calculated Type I error rate when fitting a Linear Model (LM) for data simulated using the Bivariate Normal distribution decreases drastically as the correlation in the simulated data increases. While having a small Type I error rate according to its statistical definition is ideal, here it is probably a result of not have enough power to detect a difference in some given difference in mean expression values which we will discuss later.

Similar trend exists in the Bivariate Poisson and Bivariate Negative Binomial distribution scenarios in terms of control of the empirically calculated Type I error rate. For the Poisson and Negative Binomial scenarios, when data had no correlations, or were unpaired, the fit Generalized Linear Model (GLM) had control over the Type I error rate for all sample sizes. Type I error rate was not controlled at the 0.05 level for the aforementioned scenarios when a paired model, a Generalized Linear Mixed Model (GLMM), which is designed to account for paired data was fit. Again as correlation increases, for all GLM scenarios, it is observed that the empirically calculated Type I error rate decreases likely for the same rationale explained for the Bivariate Normal distribution scenarios. These results are what one would expect. If data are unpaired or without correlation between observations (i.e., pre- and post-treatment), then the models fit to test such data likely should not contain paired capabilities.

The simulation results for all scenarios which fit a GLMM are not as aligned with what we would expect for $\rho = 0.3$ and $\rho = 0.5$. When considering the Poisson scenarios, it was observed that Type I error was controlled at $\rho = 0.3$; however, at $\rho = 0.5$ it becomes slightly less than 0.05 with empirically calculated Type I error rates ranging between 0.013 and 0.017. The results at $\rho = 0.3$ are as expected a mirror the correspond scenario from the Bivariate Normal. Though, at $\rho = 0.5$, not having control at the 0.05 level not as expected. For the Bivariate Negative Binomial scenarios, Type I error rate control was only nearly obtained when $\rho = 0.5$ using the GLMM—no Type I error rate control was attained at the 0.05 level for $\rho = 0.3$. These results are also not as expected as one would expect. Fitting a model that accounts for data pairedness to data that have a paired structure should result in control over the Type I error rate.

In those scenarios for which Type I error rate was controlled at 0.05 or lower, the portion of the simulation which calculated empirical power provided some logic as to the values which were obtained. For those conservative Type I error rate values significantly lower than 0.05 for scenarios that simulated data using the Bivariate Normal distribution, it was notice that the empirical power is extremely low. Hence, conducting statistical tests using a model which does not agree with the structure of the data has been shown to have low power. Empirical power increased with both an increase in mean shift and sample size which is typical in any simulation study of this type. When the sample size was large enough, empirical power was achieved regardless of which type of model was fit to any of the data structures in the Normal scenarios. However, when sample size was smaller, modeling the data structure correctly becomes more important in reaching a desired power to conduct such test. This was a result that we expected to observe across scenarios where Type I error rate was controlled.

With multiple testing of numerous genes, there is potential for statistical significance failing to lead to meaningful clinically relevant findings. Thus, it should be advised that once differential expression analysis is completed that supplementary investigations should be completed. Empirical Type I error rates that are significantly lower than 0.05 are said to be conservative as the probability of rejecting the null hypothesis when the null hypothesis is true is very low. This translates to a decrease of statistical power.

In general, several types approximations are considered for the statistical methods developed for RNA-Seq studies (Law et al., 2014). One of the prevalent issues we discovered when fitting the GLMM model when we decreased our sample size was that our models did not converge in their analyses. This is likely due to the fact that many statistical tests that are only asymptotically valid or theoretically accurate only when the dispersion is small which is not the case in RNA-Seq data types (Law et al., 2014). This is problematic as many RNA-Seq studies contain smaller sample sizes due to funding constraints. Thus, methods that allow for analysis of paired RNA-Seq data that have small sample sizes need to be investigated. Moreover, these smaller sample size studies would need further investigate Type I error rate and power for conducting tests; as well as, adapt models and method approaches to handle small sample sizes. Within the scope of our study, we tried to troubleshoot why our models in the paired Poisson and Negative Binomial context would not converge at sample sizes less than 100 samples. A simple fix that was suggested in many forums, was to increase the number of iterations in the optimizer. Implementing this method provided no increase in performance. Additional limitations for our simulation study exists-- we only investigated negative correlation present in our simulated data and only considered gene expression values measures pre- and post-treatment. Future studies may seek to extend our study by using smaller sample sizes, a greater range of correlations (negative and positive) within the data, and including data that are have additional replicates.

Other methods that can be used for future work include implementing (1) a sandwich estimator, or (2) use method of moments as the estimator. The sandwich estimator would be able to better handle the paired structure of the data through the use of a robust covariance matrix estimator or the empirical covariance matrix estimator (Kauermann and Carroll, 2000). We propose implementing the sandwich estimate from Kaurman and Carroll as it has already been implemented for Poisson type data (Kauermann and Carroll, 2000). Further adjustments would need to be made to adapt the estimator to be used with the Negative Binomial distribution.

Findings from our empirical study may be a result of our small sample size of N = 11 subjects, our approach of using only default stings in DE methods, and number of measures taken on each subject. Others have found higher numbers of DE genes that were similar across methods they compared. There is also evidence that by increasing the number of replications, other have seen through simulation that the percentage of DE genes that are called also increase (Robles, 2012). Additionally, this study, specifically the simulation study, is limited as only positive correlation values were considered when simulating the data.

In conclusion, we determined that differential expression analysis methods, when multiple are compared, lack precision in determining similar genes that are differentially expressed for small sample size and low number of replicates. Our results suggest that EdgeR and DESeq2 are most robust to incorrect specification of data structure in terms of determining differentially expressed genes. All-in all, we agree with the conclusion that was made by Rapaport et al. (2013) and have results that suggest that no individual DE method appears to be best in determining DE genes. However, taking in combination the results from our empirical study and the simulation study, our recommendation (as expected based on statistical theory) is to use analysis techniques that coincide with the study design as Type I error rate is more likely to be controlled. Additionally, we conclude that statistical test will have greater power when study design and statistical model for analysis align with one another.

While statisticians have the ability to fit numerous statistical models to RNA-Seq data, it is crucial for them to keep in mind how to interpret their findings so that researchers can know the clinical implications. This is also the case when RNA-Seq data are transformed. Furthermore, it is fundamental that any analysis complete be evaluated and validated in terms of its performance and accuracy. All-in-all with any type of research study, there needs to be a balance that exists between considering clinical relevance and statistical relevance.

CHAPTER 3

An Assessment of Transformations and Clustering Methods Using RNA-Seq Data

Janelle R. Noel, Joseph Usset, Ellen L. Goode, and Brooke L. Fridley

To be published in BMC Genomics

3.1 Abstract

The analysis of RNA-Seq data comes with some different and additional challenges, as compared to microarray based data. In contrast to microarray based mRNA data in which relative mRNA is measured for pre-defined probe sets via fluorescence, RNA-Seq experiments measure the gene expression levels from the total number of reads that fall into the exons of a gene. Therefore, the quality control, global biases, normalization and analysis methods for RNA-Seq data are quite different than those for microarray based data. In particular, where the assumption of normality was a reasonable assumption for microarray based data, RNA-Seq data tends to follow an over-dispersed Poisson or Negative Binomial distribution. Little research has been done to assess how cluster methods perform for analysis of RNA-Seq data and if transformation of the data can improve the performance. Hence, we conducted an extensive simulation study to assess the performance of combinations of data transformations and clustering methods with respect to clustering performance and accuracy in estimating the correct number of clusters. Data were simulated based on RNA-Seq data collected on 56 serous ovarian cancer tumor samples. In total, 192 unique scenarios were investigated with variations in data transformation, clustering method, number of simulated clusters, and size of clusters. Within these scenarios, considerations are given to whether or not the number of clusters found in the data are known or unknown. Each scenario’s performance was evaluated by the adjusted rand index, clustering error rate, and concordance index. Evaluation results revealed that data transformations which cause the data to look more normal in combination with model-based clustering methods perform better with respect to all performance evaluation metrics when the number of clusters is said to be known. The *K Unknown* simulation branch revealed the difficulty in algorithmically selecting the number of clusters present in a given dataset when no expert advice is available. Globally, we conclude that model-based clustering (MC) approach may be the best starting place for exploratory clustering analysis of RNA-Seq data types when the number of clusters is backed by prior knowledge.

*Keywords***:** Clustering; genomics; RNA-Seq; Negative Binomial; Simulation Study

3.2 Introduction

The analysis of RNA-Seq data comes with some different and additional challenges, as compared to microarray based data. In contrast to microarray based mRNA data, in which relative mRNA is measured for pre-defined probe sets using fluorescence, RNA-Seq experiments measure the gene expression levels from the total number of reads that map to the exons of a gene. Furthermore, RNA-Seq experiments have the potential, in theory, to answer many more research questions as compared to mRNA microarray studies, such as splicing, fusion detection and allelic specific expression (ASE). Additionally, the quality control, global biases, normalization and analysis methods for RNA-Seq data are quite different than those for microarray based data.

Microarray data are continuous verses that of sequencing data which are count-based. Microarray data can be simulated using continuous distributions with varied parameters and many have accepted that microarray data can be measured using the Normal, or Gaussian, distribution.; whereas, sequencing data needs to be simulated from discrete distributions such as the Poisson (or over-dispersed Poisson) or the Negative Binomial distribution. Within the last few years, several researchers have evaluated and compared clustering methods in microarray analysis (Jiang et al., 2004, Shannon, 2003, Quackenbush, 2001, Eisen et al., 1998, Sorlie et al., 2001, Makretsov et al., 2004, Allison et al., 2006, Qu and Xu, 2004). Applications of clustering in microarray data were completed through the use of unsupervised classification as no hypotheses or data assumptions were needed (Allison et al., 2006). One of the earliest studies using hierarchical clustering in microarray data showed that genes with common role and function in the cellular process would cluster together (Eisen et al., 1998). Others have found that unsupervised, hierarchical clustering has the capability to determine prognostic clusters, clusters that are based upon some marker of health status; as well as, identify subtypes of invasive cancers (Makretsov et al., 2004, Sorlie et al., 2001). Despite that other studies support genes that share common function cluster together, clustering outcomes can be greatly affected by the dependency of a particular method used relative to the clustering algorithm used for classification, normalization across and within experiments, and the measure of similarity or dissimilarity that is used (Quackenbush, 2001). Some have also argued that if you have some prior insight as to what cluster may be, that using supervised model-based clustering algorithms is superior (Qu and Xu, 2004). Knowledge of these variations have suggested that researchers should select a couple clustering methods to summarize their results (Jiang et al., 2004, Shannon, 2003).

A common practice in statistics is to apply transformations to a given set of data to make the analysis methodologies more efficient and induce better statistical properties. RNA-Seq data have been said to have three problematic properties when it comes to statistical analysis--a skewed distribution, variability among the read counts for individual genes, and likelihood of extreme values (Zwiener et al., 2014). Two of these problematic properties can be addressed using simple algebraic approaches. The skewness of the distribution can be addressed by using a data transformation. Likewise, the variability can be handled through many types of normalization procedures. While this study will not cover types of normalization, a comprehensive procedure can be found by Dillies et al. (2012).

Clustering analysis can be viewed as one of the first steps when exploring data. Clustering analysis is purely exploratory and for hypothesis building as clustering methods will form clusters even in data that is unrelated and completely independent (Quackenbush, 2001, Shannon, 2003). The challenge for clustering analysis lies in obtaining a “good” clustering method and in turn coming up with the “correct” number of clusters (Yeung, 2001). The collective goal of clustering methods is to accurately group data objects of interests based on some type of mathematical calculation of similarity or dissimilarity to assess whether the object belongs to a cluster (Eisen et al., 1998). Common measures of similarity that are used in clustering methods for a variety of data domains include: Euclidean distance, cosine similarity, Jaccard correlation coefficient, and relative entropy (Huang, 2008).

Clustering methods tend to fall into two categories—supervised clustering and unsupervised clustering. Supervised clustering methods use algorithms that cluster objects into some pre-defined category. Whereas, unsupervised clustering methods aim to discover categories by grouping objects by one of the similarity measures mentioned above (Allison et al., 2006). Nevertheless, both of these clustering methods’ categories seek to reduce the data to be able to better explain potential relationships that may exist. Selection of the most appropriate clustering method is not always straightforward and is often driven by the context of the specific study (Chalise et al., 2014). Within the different clustering methods, analyses have been completed to cluster based upon genes (i.e., gene-based clustering) or clustering based upon subject samples (i.e., sample-based clustering) (Liu and Si, 2014).

In the early era of the microarray, many researchers sought to apply clustering analysis to the gene expression data as the importance to identify specific patterns of gene expression and groups of genes or groups of participant samples for that could provide greater insight into biological function (Quackenbush, 2001). The pioneered reports of researchers executing clustering analysis in expression data date back to 1997 (Weinstien, 1997). Weinstien et al. (1997) implemented a hierarchical clustering approach to a set of targets that contained different compounds and ordered them based upon Pearson correlation coefficients relating activity and target patterns. As time progressed, clustering methods for microarray data included: hierarchical clustering, graph-theoretical approaches, model-based clustering, K-Means, density-based hierarchical clustering, and self-organizing maps (SOMs) (Jiang et al., 2004, Quackenbush, 2001). A comprehensive evaluation and comparison of clustering methods for microarrays was completed in 2006 by Thalamuthu et al.. They compared six different gene clustering methods and found that model-based clustering outperformed non-model based cluster methods in a simulation study and applied to real data (Thalamuthu et al., 2006).

When it comes to the analysis of RNA-Seq data the literature is saturated with studies regarding differential expression as it relates to varied experimental conditions. The trend in clustering analysis followed thereafter. After reviewing the available literature, it was apparent that the evaluation and comparison of clustering methods has only been completed in microarray analyses--microarray technology predates RNA-Sequencing by approximately 10-15 years. Moreover, little research has been done to assess how cluster methods perform for analysis of RNA-Seq data and if transformation of the data can improve the performance. The current literature contains three closely related studies to this topic that have looked at performance of clustering methods—one investigates classification and clustering of sequencing data using a variety of methods (Witten, 2011), another used clustering analysis to identify features of the gene space in RNA-Seq and microarrays (Sibru et al., 2012), and the other study provides an in depth look at model based clustering for RNA-Seq data (Si, 2013).

In this paper, we aim to assess four data transformations applied to count data (RNA-Seq data type) and up to five clustering methods to provide insight into clustering using RNA-Seq data. Using an extensive simulation study that contains 192 simulation scenarios, we investigate several previously purposed data transformations and clustering methods that have been used in microarray analysis. Data for this simulation were simulated from parameters obtained from an actual RNA-Seq dataset. We limited the number of genes in our simulated datasets to account for the significant computational resources that were needed for our methods.

All simulation scenarios fit into four parent categories depending on how the genes were selected to be included in the clustering analyses. The factors varied in the simulation studies include: (1) how genes were selected to be included in the clustering analyses (top 100 genes according to their median absolute deviation (MAD), or random sample of a 100 genes); (2) size of the clusters (equal cluster sizes or extremely unequal cluster sizes); (3) number of clusters; (4) data transformations; (5) clustering methods; (6) whether *K* was *known* or *unknown*. The simulation scenarios assessed the following data transformations: naïve, logarithmic base 2 (Log), Blom (Beasley et al., 2009), and variance stabilizing transformation (VST) (Durbin et al., 2002). Concurrently, using the transformed datasets the following clustering methods were assessed: Hierarchical Clustering (HC), Model-based Clustering (MC), Non-Negative Matrix Factorization (NMF), Recursively Partitioned Mixture Model Clustering (RPMM), and K-Means Clustering (KM). Each of the clustering methods carried out a sample based clustering approach (i.e., interested in clustering patient tumors to determine molecular subtypes).

In the following sections of this paper, we will describe how our data were simulated using Negative Binomial maximum likelihood estimation (MLE), provide details of our simulation study, and further discus the data transformations and clustering methods used. We then summarize the normality and performance findings from all simulation scenarios. To our knowledge, this is the first comprehensive assessment of clustering for RNA-Seq data.

3.3 Materials and Methods

To address the aims of this study, an extensive simulation study was conducted. This simulation study has four major components—1) simulating data that is similar to actual RNA-Seq data collected from a set of ovarian tumors, 2) implementing various data transformations, 3) utilizing many clustering methods when the number clusters within the data is either approximately known by an expert or completely unknown requiring the use of model-based algorithms or the Gap Statistic (Tibshirani, 2001), and 4) evaluating all simulation scenarios using the Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (C-Index). The schematic in Figure III-1 provides a brief overview of the entire simulation study. It should be noted that for scenarios in which the number of clusters is known and set *a priori* will be referred to as "*K Known*" scenarios; whereas, those scenarios where the number of clusters in the data is completely unknown will be denoted as the *"K Unknown"* scenarios. All analysis for this study were conducted in R statistical software (R Development Core Team, 2016).

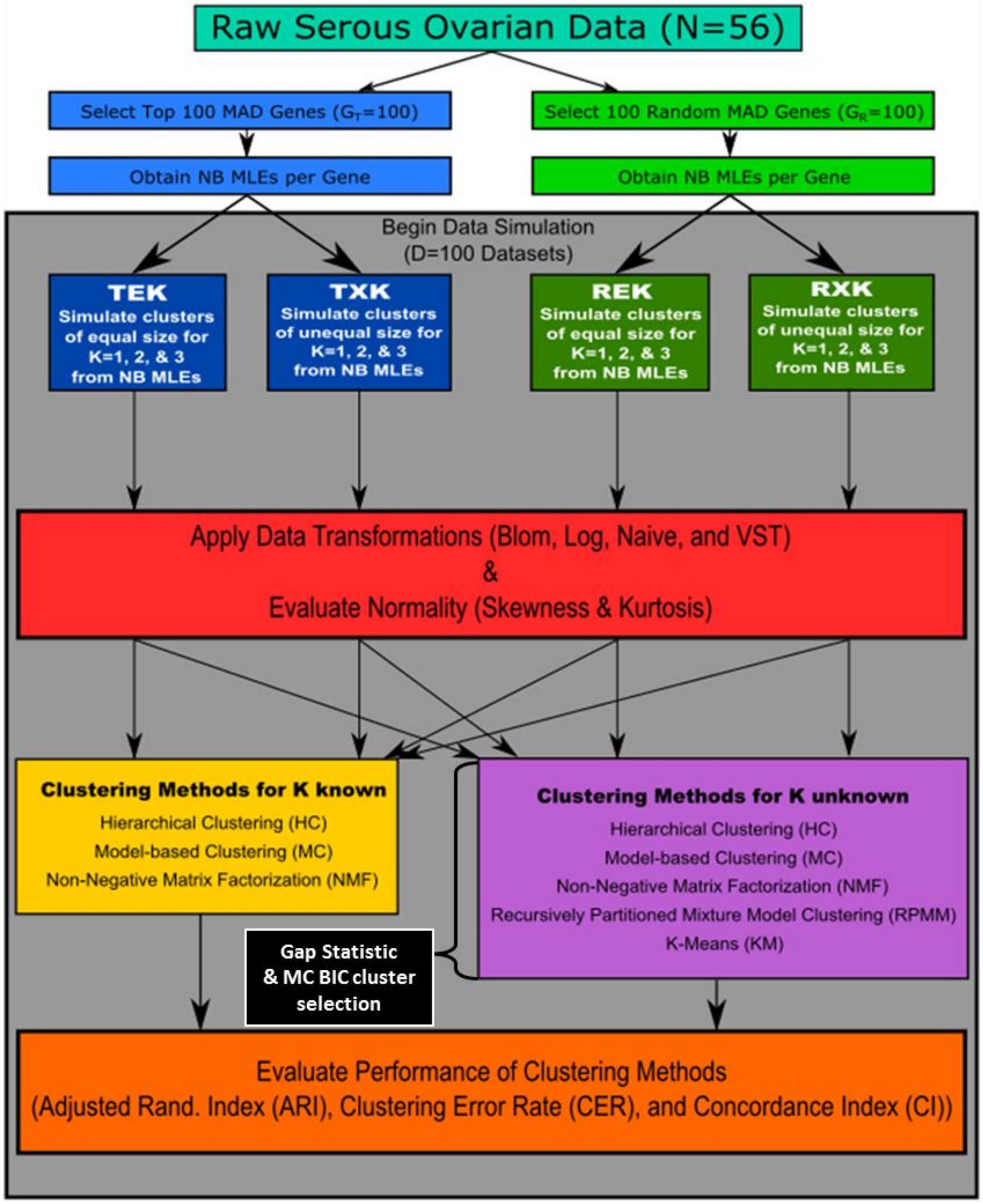

Raw Serous Ovarian Data (N=56)
Select Top 100 MAD Genes (G_T=100)
Select 100 Random MAD Genes (G_R=100)
Obtain NB MLEs per Gene
Obtain NB MLEs per Gene
Begin Data Simulation (D=100 Datasets)
TEK Simulate clusters of equal size for K=1, 2, & 3 from NB MLEs
TXK Simulate clusters of unequal size for K=1, 2, & 3 from NB MLEs
REK Simulate clusters of equal size for K=1, 2, & 3 from NB MLEs
RXK Simulate clusters of unequal size for K=1, 2, & 3 from NB MLEs
Apply Data Transformations (Blom, Log, Naive, and VST) & Evaluate Normality (Skewness & Kurtosis)
Clustering Methods for K known
Hierarchical Clustering (HC)
Model-based Clustering (MC)
Non-Negative Matrix Factorization (NMF)
Clustering Methods for K unknown
Hierarchical Clustering (HC)
Model-based Clustering (MC)
Non-Negative Matrix Factorization (NMF)
Recursively Partitioned Mixture Model Clustering (RPMM)
K-Means (KM)
Gap Statistic & MC BIC cluster selection
Evaluate Performance of Clustering Methods
(Adjusted Rand. Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI))

Figure III-1. Simulation schematic to assess aims of study. Prior to simulating our data, we began by obtaining Negative Binomial (NB) parameters from 100 top genes and 100 randomly selected genes based upon Median Absolute Deviation (MAD) of expression values. Data were then simulated for both an equal number of samples in each cluster and an unequal number of samples in each cluster for three classes of K ($K = 1, 2$ and 3) using the NB parameters for D = 100 datasets. Furthermore, data transformations were applied to all data sets; and K Known and K Unknown clustering methods were applied. Data transformations were evaluated according to normality measures (i.e., skewness and kurtosis) and clustering methods were assessed by common clustering accuracy metrics (i.e., ARI, CER, and CI).

Before assessing our data transformations and clustering methods, great consideration was given to the way in which the data are simulated to ensure that data are similar to what would be found in a real RNA-Seq experiment to ensure our results are robust and relevant. Most often researchers have simulated RNA-Seq count data from either an overdispersed Poisson distribution or a Negative Binomial distribution. The usage of both of these distributions can be found throughout the literature as they are able to deal with the unique challenges that arise when simulating RNA-Seq data. These difficulties include the nonnegative, integer-valued structure of RNA-Seq data; as well as, the highly variable total number of sequence reads across different samples (Witten, 2011). Recently more researchers have preferred the use of the Negative Binomial distribution when it comes to RNA-Seq studies. The Negative Binomial distributions allow for two distributional parameters to be controlled—the mean and shape parameters. Controlling the mean and shape parameters allow researchers to model the overdispersion, which typically exists in sequencing data. Overdispersion occurs when there is greater observed variance in the data than expected (e.g., under the Poisson distribution assumption the mean and variance are equal). In this simulation study we chose to simulate the data from a Negative Binomial distribution, in which we used parameter estimates for simulation based on the RNA-Seq data from an ovarian cancer study out of the Mayo Clinic (Rochester, MN) headed by Dr. Ellen L. Goode in the hope that our simulated data will better resemble that of “real-life”.

3.3.1 Mayo Clinic Ovarian Cancer Study

The Mayo Clinic data contains data collected on 56 patients with invasive epithelial ovarian cancer. Women who were eligible for the study needed to have a diagnosis that was less than one year prior and be ≥20 years of age. All participant samples were of Serous (SER) histology as confirmed by re-review by a gynecologic pathologist (GLK). RNA was extracted from these samples at the Tissue Microarray facility at the Mayo Clinic and sent to be sequenced by BGI Americas. Prep for the sequencing of the samples included riboZero treatment of 1 μg of RNA and using the Illumina TruSeq Stranded Total RNA kit to make libraries. After samples were prepped, sequencing was completed using the Illumina HiSeq 2000 with 100bp paired end reads, six samples were multiplexed per lane. The resulting FASTQ files were sent to Dr. Fridley's lab at The University of Kansas Medical Center (KUMC). At KUMC, the FASTQ files were aligned to the human genome (G = 63,152 ensemble gene IDs) using *TopHat2,* followed by application of *HTSeq* to generate gene count. To further understand the behavior of this study data, we calculated the mean and variance for each of the ~63K Ensemble genes across all participant samples. In computing the log transformations on the count data, an additional count of 1 was added avoid undefined values. Figure III-2 displays the relationship between the transformed mean and variance of the RNA-Seq data. It can be noticed that the data are over-dispersed as expected.

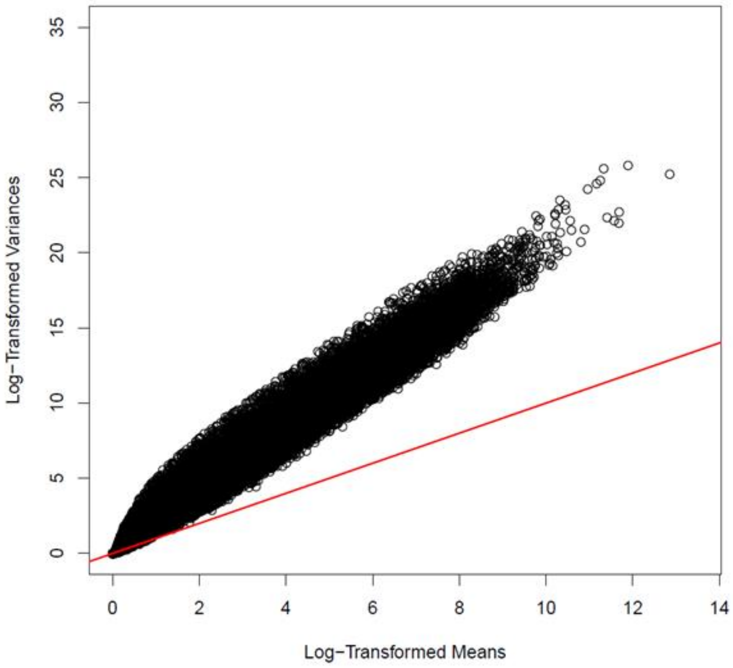


Figure III-2. Comparison of log-transformed mean and log-transformed variance across samples per Ensemble gene ID. It is common among RNA-Seq data that overdispersion will be present. That is, the variance is greater with respect to the mean. The red 45 degree line is representative of equal log-mean and log-variance.

3.3.1.1 Data Selection

The Mayo Clinic data contains gene abundance estimates for G = 63,152 Ensembl genes on N = 56 participants. Let $\boldsymbol{X}^*$ be the G by N matrix where ${x^*}_{gi}$ is the raw RNA-Seq count for the gth gene ($g = 1, \dots, G$) and the ith sample ($i = 1, \dots, N$).

$$\boldsymbol{X}^* = \begin{bmatrix} {x^*}_{11} & {x^*}_{12} & \cdots & {x^*}_{1N} \\ {x^*}_{21} & {x^*}_{22} & \dots & {x^*}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ {x^*}_{G1} & {x^*}_{G2} & \cdots & {x^*}_{GN} \end{bmatrix}$$

As with any sequencing dataset, data size is often a concern for statistical analysis and computational processing time. Additionally, we have the classic “small n, large p” phenomena that is often encountered in RNA-Seq studies (i.e., there is a much lower number of samples (i.e., small n) with respect to the large number of covariates or genes (i.e., large p) that are taken for a given sample). Hence, we decided to reduce the size of our data, specifically reduce the number of genes that whose attributes we would use to simulate our datasets. This was accomplished using two fairly intuitive ways: 1) selecting 100 of the top most variable genes according to the Median Absolute Deviation (MAD) (most common practice in selecting genes for clustering), and 2) selecting a random sample of 100 genes.

The top 100 most variable genes were selected by calculating each gene's median absolute deviation (MAD). The MAD is calculated by obtaining the median count value across all N samples, subtracting it from each of the sample's counts for a given gene, and further taking the median of all those differences. Rather, MAD is defined as:

$$MAD\left(\boldsymbol{x}_{g.}^{*}\right) = Median_i\left(\left|x_{gi}^{*} - Median\left(\boldsymbol{x}_{g.}^{*}\right)\right|\right).$$

The MADs for all of the genes were then ordered in decreasing value and subset to those 100 genes with the highest deviations. Here, let the subset of data be $\boldsymbol{X}_T$, a $G_T{}^*$ by N matrix where $G_T{}^* = 100$ Ensembl gene IDs, similar to $\boldsymbol{X}^*$. It follows:

$$\boldsymbol{X}_T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{G_T{}^*1} & x_{G_T{}^*2} & \cdots & x_{G_T{}^*N} \end{bmatrix},$$

where x_{gi} is the raw RNA-Seq count for the g^{th} top 100 most variable gene ($g = 1, \dots, G_T{}^*$) and the i^{th} sample ($i = 1, \dots, N$). The creation of the dataset that contains 100 randomly selected genes follows similarly to that of the dataset containing the top 100 MAD genes. Though, prior to obtaining a random sample of 100 genes, we filtered out the lower 50% MAD genes, ordered in decreasing order, as a majority of them have zero counts for all N samples.

Thus, from the residual 50%, 100 genes were randomly sampled using a random number generator within R (R Development Core Team, 2016). For this subset, let the data be $\boldsymbol{X}_R$, a ${G_R}^*$ by N matrix where ${G_R}^* = 100$ Ensembl gene IDs. It follows:

$$\boldsymbol{X}_R = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{{G_R}^*1} & x_{{G_R}^*2} & \cdots & x_{{G_R}^*N} \end{bmatrix},$$

where x_{gi} is the raw RNA-Seq count for the gth top 100 most variable gene ($g = 1, \dots, {G_R}^*$) and the ith sample ($i = 1, \dots, N$).

The resulting selections of genes are plotted in blue against the original Mayo Clinic data along with their corresponding expression levels in Figure III-3 A) and Figure III-3 B), respectively. The distinct way in which the data were selected can be observed. In Figure III-3 A) observe that all of the selected genes contain the highest log means and log variances. Additionally, notice the completely ransom selection of the genes in Figure III-2 B). For both types of data selections, there is a lack of distinct patterns of expression (Figure III-3: A and B). Though it should be noted that the distributions for the color breaks in the heatmaps in Figure III-3 are very different between the different ways the data were select. This was purposely done to allow for the variation in expression to be better depicted. Even though the scale for the color breaks differ from one another, genes with lower read counts are represented in red and those gene with higher counts are in blue.

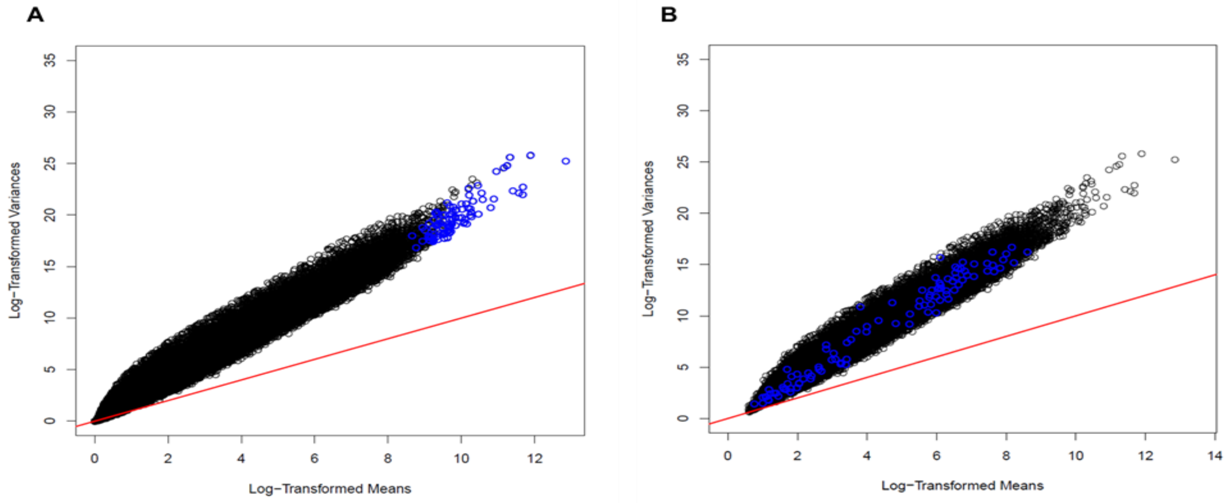


Figure III-3. Gene selection for both the top 100 genes and random 100 genes according to their Median Absolute Deviation (MAD). Highlighted in blue are the top 100 most variable genes in both panel A) and B). A) contains the top 100 MAD genes, and B) contains the random subset of 100 genes.

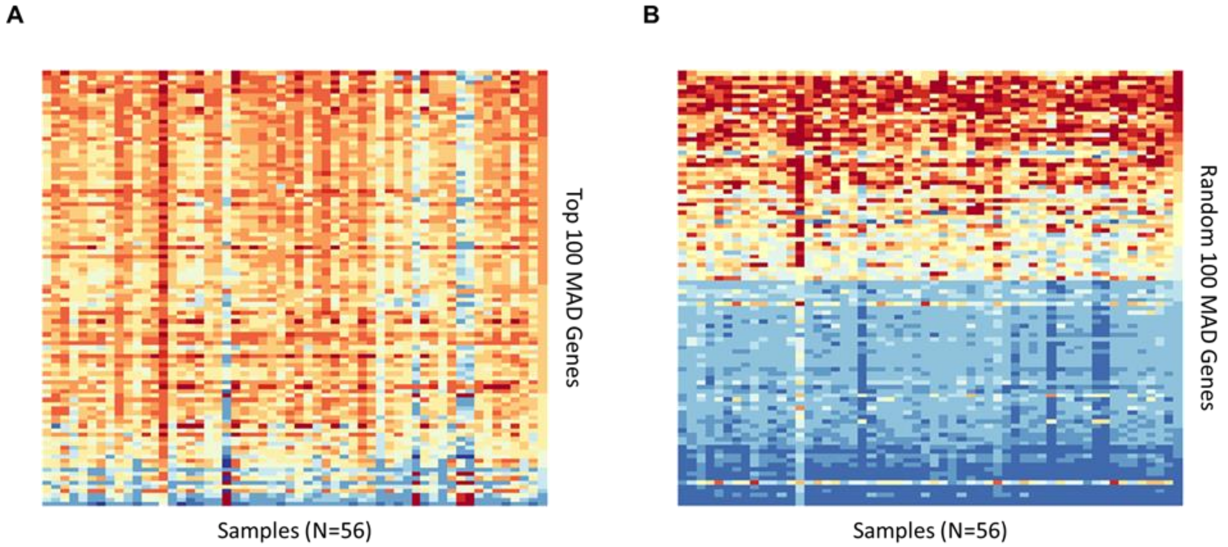


Figure III-4: Heatmaps of most variable genes. Each heatmap represents low, moderate, and high expression count values through the use of red, yellow, blue color scheme, respectively for one of the 100 simulated datasets. The color breaks that were used for the red, yellow, blue color scheme differed depending on the category of gene selection. Panel A) contains the top 100 MAD genes had color breaks corresponding the following gene expression values [0, 500, 1000, 5000, 9000, 14000, 25000, 50000, 75000, 100000, 500000, 1000000, 1789200]. Similarly panel B) contains the 100 randomly selected MAD genes at [0, 0.5, 1, 2, 3, 5, 10, 20, 50, 100, 500, 1000, 100000, 1789200].

3.3.1.2 Maximum Likelihood Estimators for the Negative Binomial Parameters

Vector Generalized Linear Models (VGLMs) are an inclusive class of models of various multivariate response types that are highly generalizable (Yee, 2003, Yee, 1996). VGLMs are able to handle problems that stem from uni- and multivariate distributions, categorical analysis, generalized estimating equations and many more (Yee, 2003). VGLMs are models of the form

$$f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B}) = h(\boldsymbol{y},\eta_1,\dots,\eta_M,\varphi)$$

for some known function $h(\cdot)$, where $\boldsymbol{B} = (\beta_1\beta_2 \dots \beta_M)$ is $p\ x\ M$, φ is an optional scaling parameter, and $\eta_j = \boldsymbol{\beta}'_j\boldsymbol{x} = \beta_{(j)1}x_1 + \cdots + \beta_{(j)g}x_g$ is the jth linear predictor (Yee, 2003). The only assumption for VGLMs is that the regression coefficients must be comprised of a set of linear predictors. Once the form of the model is established, the log-likelihood function can be obtained and Maximum Likelihood Estimates (MLEs) can be found for the parameters in the parent distribution through Iteratively Reweighted Least Squares (IRLS) using either the Newton-Raphson or Fisher-scoring algorithm(Green, 1984, Yee, 2003). Details of this process can be found in the Yee and Hastie paper from 2003. In 2015, Yee has made available an R package to that carries out the MLE process above for a specified distribution—the *VGAM* package. To obtain data that reflect that of “real-life”, we will utilize VGLMs to obtain MLEs from fitted NB models for each gene, and use those MLEs in the simulation of the datasets.

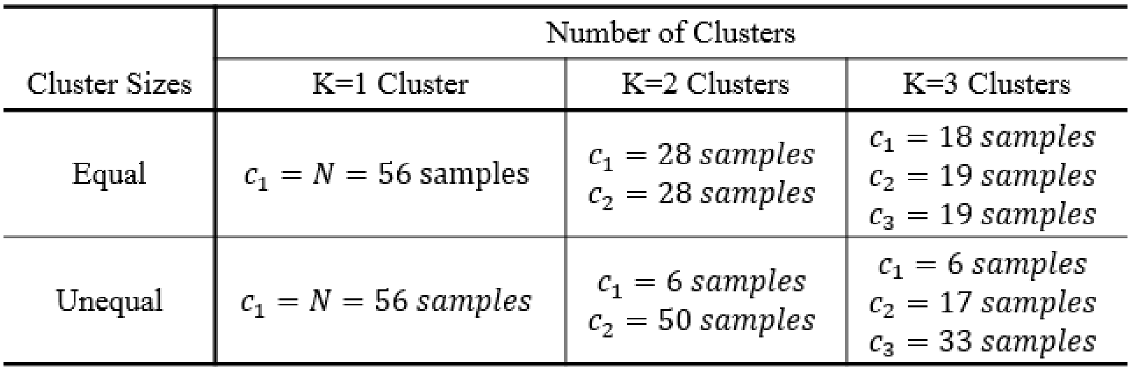
For each of the resulting datasets from the gene selections, $\boldsymbol{X}_T$ and $\boldsymbol{X}_R$ above, we fit 100 vector generalized linear models (VGLMs), one for each gene, using a Negative Binomial parameterization. The NB parameterization that was used to fit each of the models in both gene selection datasets was

$$P\big(X = x;\ \mu,\ k\big) = \binom{x+k-1}{x}\left(\frac{\mu}{\mu+k}\right)^x\left(\frac{k}{k+\mu}\right)^k, \mu > 0,\ k > 0$$

with mean μ, variance $\mu + \frac{\mu^2}{k}$, and dispersion parameter $\frac{1}{k}$. Additionally, the linear predictors for our VGLM are $\log(\mu) = \eta_1 = \boldsymbol{\beta}'_1\boldsymbol{x}$ and $\log(k) = \eta_2 = \boldsymbol{\beta}'_2\boldsymbol{x}$ where the log link is used here due to the range restrictions that are present for μ and k. Utilizing the IRLS method, the MLEs for μ and k, $\hat{\mu}$ and $\hat{k}$, were obtained for each gene, resulting in creation of $\hat{\boldsymbol{\mu}}_T = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_{G_T} \end{bmatrix}$ and $\hat{\boldsymbol{k}}_T = \begin{bmatrix} \hat{k}_1 \\ \vdots \\ \hat{k}_{G_T} \end{bmatrix}$ for those genes from the selection dataset containing the top 100 genes, and $\hat{\boldsymbol{\mu}}_R = \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_{G_R} \end{bmatrix}$ and $\hat{\boldsymbol{k}}_R = \begin{bmatrix} \hat{k}_1 \\ \vdots \\ \hat{k}_{G_R} \end{bmatrix}$ from the dataset containing 100 randomly selected genes.

3.3.2. Simulation Study

To address the specific aims proposed, an extensive simulation study was conducted. The simulation of the data in this study has many unique attributes to insure that our scenarios and results are generalizable. We further simulated these data to include exploring different numbers of clusters and clusters on different sizes. In our simulation study, we simulate data that considers one (i.e. no clusters) to three clusters (i.e., K=1 cluster, K=2 clusters, and K=3 clusters). Additionally, given the behavior of some clustering methods wanting to cluster so that cluster sizes are equivalent, we simulate data that is of equal cluster size and extremely unequal cluster sizes. The cluster sizes for K=2 clusters and K=3 clusters for equal and unequal cluster size can be found in Table III-1.



Cluster Sizes	Number of Clusters		
	K=1 Cluster	K=2 Clusters	K=3 Clusters
Equal	$c_1 = N = 56$ samples	$c_1 = 28\ samples$ $c_2 = 28\ samples$	$c_1 = 18\ samples$ $c_2 = 19\ samples$ $c_3 = 19\ samples$
Unequal	$c_1 = N = 56\ samples$	$c_1 = 6\ samples$ $c_2 = 50\ samples$	$c_1 = 6\ samples$ $c_2 = 17\ samples$ $c_3 = 33\ samples$

Table III-1. Cluster sizes for different number of clusters. Equal and Unequal cluster sizes were used in the simulation study. c_k for $k = 1, 2,$ or 3 designate the number of samples in a given cluster.

For organizational purposes of this simulation, we utilize the way data were selected to obtain the MLEs (i.e., top 100 MAD genes or Random 100 MAD genes) and the size of the clusters (i.e., equal or unequal cluster sizes) to define four parent categories of scenarios for all K. The 4 parent categories will be defined as: 1) Top 100 MAD Genes with Equal Cluster Sized (TE); 2) Random 100 MAD Genes with Equal Cluster Sizes (RE); 3) Top 100 MAD Genes with Unequal Cluster Sizes (TX); and 4) Random 100 MAD Genes with Unequal Cluster Sizes (RX).

3.3.2.1 Datasets Simulation

We generate 100 datasets (D=100 datasets) for each of the 4 parent categories previously mentioned. The data for these datasets were simulated from a NB distribution using the respective sets of MLEs (i.e., $\widehat{\boldsymbol{\mu}}_T$ and $\widehat{\boldsymbol{k}}_T$, or $\widehat{\boldsymbol{\mu}}_R$ and $\widehat{\boldsymbol{k}}_R$). Specifically,

$$x_{gi} \sim NB\left(\hat{\mu}_g, \hat{k}_g\right)$$

where $\hat{\mu}_g$ is the MLE of the mean and $\hat{k}_g$ is the MLE for the dispersion parameter for the gth gene. The data are simulated under the assumption that subjects are independent from one another and that the gene expression between genes is also independent. Though, in order to simulate data that resembled different clusters, we incorporated effect size shifts to $\hat{\mu}_g$ and $\hat{k}_g$ to a proportion of genes which would represent genes that were up-expressed in this cluster group. We set 10% of the genes in any dataset up-expressed for $K = 2$ and for $K = 3$ there would be a step progression for the percentage of genes that were up-expressed—10% for c_2 and 20% for c_3 of which 10% would be simulated with the same effect size as that of c_2. Figure III-5 depicts specifically how genes were simulated for each cluster and up-expressed genes.

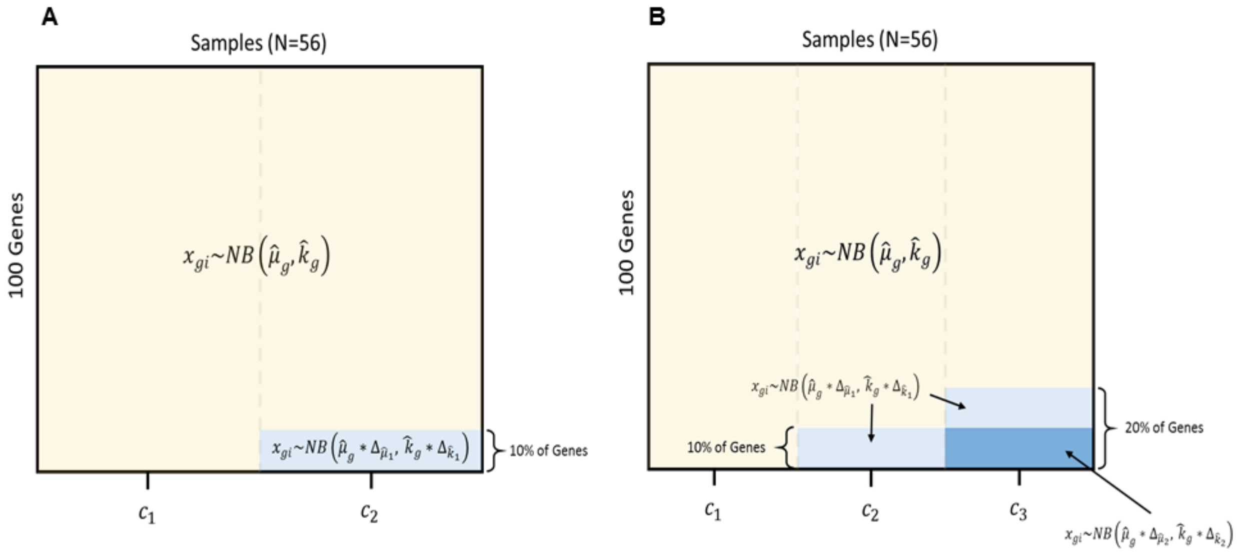

A
Samples (N=56)
100 Genes
$x_{gi} \sim NB\left(\hat{\mu}_g, \hat{k}_g\right)$
$x_{gi} \sim NB\left(\hat{\mu}_g * \Delta_{\hat{\mu}_1}, \hat{k}_g * \Delta_{\hat{k}_1}\right)$
10% of Genes
c_1
c_2
B
Samples (N=56)
100 Genes
$x_{gi} \sim NB\left(\hat{\mu}_g, \hat{k}_g\right)$
$x_{gi} \sim NB\left(\hat{\mu}_g * \Delta_{\hat{\mu}_1}, \hat{k}_g * \Delta_{\hat{k}_1}\right)$
10% of Genes
20% of Genes
c_1
c_2
c_3
$x_{gi} \sim NB\left(\hat{\mu}_g * \Delta_{\hat{\mu}_2}, \hat{k}_g * \Delta_{\hat{k}_2}\right)$

Figure III-5. Data simulation with clusters and up-expressed genes. Datasets were simulated so that clusters would be present (i.e., c_1, c_2, and c_3) through a shift to make certain percentages of genes up-expressed. A) depicts how data were simulated for $K = 2$, and B) depicts how data were simulated for $K = 3$.

3.3.2.2 Empirical Power Simulation to Determine Effect Size

The effect size shifts for the mean and dispersion parameters, parameters from the Negative Binomial distribution, were determined through an empirical pilot study. Data were simulated for the K=2, TE parent category using ranging combinations of shifts for each of the parameters. The rationale behind adding shifts to the mean and the dispersion parameters is to reflect the behavior that is present in simulating data from a Negative Binomial distribution—as mean values increases, so to do the variance, or overdispersion, values increase. Additionally, it should be noted that caution should be taken when specifying these shifts. The shifts applied to both the mean and overdispersion parameters needed to be substantial enough so that the clustering method would have the ability to distinguish clusters. Shifts too large would lead to the clustering method always obtaining the "truth" or correct cluster assignment for a given sample. Conversely, the shifts could not be too minimal which would result in no clusters being determined by the clustering method.

After determining a set range of shifts for the parameters, one hundred datasets were simulated for each unique combination of shifts. These datasets then underwent model-based clustering to determine which percentage of samples clustered identically with their simulated cluster. Adequate power was said to be achieved if the 100 datasets for a given combination of effect size shifts resulted in clusters that had samples that perfectly matched the simulated sample-cluster assignment at least 80% of the time. The empirical simulation showed that a mean shift of exp(3.375) and a dispersion shift of 1.01 would yield correct cluster assignment ~80% of the time. Additionally, it was determined for the K=3 simulation scenarios that in addition to the K=2 effect size shifts that exp(5.5) and 1.03 would be used for a percentage of genes in c_3 . For our simulation proposes, let $\Delta_{\hat{\mu}_1} = \exp(3.375)$, $\Delta_{\hat{\mu}_2} = \exp(5.5)$, $\Delta_{\hat{k}_1} = 1.01$, and $\Delta_{\hat{k}_2} = 1.03$ for the simulation of datasets.

3.3.2.3 Data Transformations

In a similar fashion to Zwiener et. al. (2014) and Witten (2011), we explore many RNA-Seq data transformations in regards to their performance in non-parametric and model-based approaches. The following sections describe the four data transformations that have been applied to all of the scenarios for of the simulated data prior to clustering. Even though, clustering methods can be executed using raw count RNA-Seq, clustering methods may run more efficiently and have higher accuracy in assigning samples to clusters when they are transformed (Shannon, 2003). The four data transformations assessed were: Naïve transformation, Log transformation, Blom transformation, and Variance Stabilizing Transformation (VST). The data transformations were evaluated in terms of skewness and kurtosis to assess which transformation yielded the "most normal" transformed RNA-Seq data. Skewness is the measure of symmetry and kurtosis measures flatness or peakedness for a distribution of values (Casella and Berger, 2002). When data are normally distributed, skewness equals zero (i.e., $Sk = 0$) and kurtosis equals three (i.e., $Kt = 3$) (Rencher and Christensen, 2012). $Sk > 0$ denotes positive skewness; whereas $Sk < 0$ denotes negative skewness (Rencher and Christensen, 2012). Similarly, $Kt > 3$ means that kurtosis is positive or more peaked; whereas $Kt < 3$ means negative kurtosis or a flatter distribution (Rencher and Christensen, 2012). Following the clustering analyses, the data transformations in combination with each different clustering method were further considered for how they may have played a role in the results.

Naïve Transformation

The naïve transformation is the untransformed or null, simulated RNA-Seq data. Denoted as

$$x_{gi}^{naive} = x_{gi},$$

the naïve transformed data contains all of the original attributes of the raw simulated data. This transformation often times does not yield accurate results in any type of statistical analysis when the dataset of interest includes highly variable data that span a wide range of values or are skewed. The naïve transformation will be used as a baseline to compare all other transformations with.

Log Transformation

Logarithm transformations are very useful when it comes to scaling a dataset that has a skewed wide range of data values; such as that of RNA-Seq data (Zwiener et al., 2014). Following suit from the popular Bioconductor Package, *edgeR (Robinson et al., 2010)*, the specific logarithm transformation that will be used is the log base 2 data transformation. The log base 2 transformation is applied to the data plus some constant *c* as follows:

$$x_{gi}^{log_2} = log_2(x_{gi} + c)$$

Here, we are using $c = 1$ to allow for the transformation of those zero count values to be non-infinite.

Blom Transformation

In the realm of statistical genomics, Blom transformations have become popularized as they allow for the data to be converted back to more or less the standard normal distribution (Beasley et al., 2009). This is accomplished through the use of an Inverse Normal, ϕ^{-1}, rank-based algorithm where $c = 3/8$.

$$x_{gi}^{Blom} = \phi^{-1}(rank(x_{gi}) - c)/(n - 2c + 1)$$

Variance-Stabilizing Transformation

The Variance-Stabilizing Transformation (VST) is carried out in the *DESeq2* R package in Bioconductor, and was initially proposed by Anders and Huber (Anders and Huber, 2010). VST allow for covariates with variances independent of the mean value to be obtained (Zwiener et al., 2014). The mean-variance relationship for the transformation can be written as

$$x_{gi}^{vst} = \int_0^{x_{gi}} \frac{1}{V(\mu_g)} d\mu_g,$$

where μ_i are the mean expression values and $V(\mu_g)$ is defined as the variance of the Negative Binomial distribution (i.e., $V(\mu_g) \coloneqq \mu_g + a_g {\mu_g}^2$) with dispersion parameter $a_g = a_0 + \frac{a_1}{\mu_g}$, where a_0 and a_1 are specific estimates based on the GLM (Zwiener et al., 2014). Furthermore, the delta method used with a Taylor expansion which considers squared Euclidean distances between pairs of samples (Durbin et al., 2002). Details of this transformation can be found in the *DESeq2* R package documentation (Love et al., 2014, Love et al., 2016). Utilization of the VST transformation lends itself while to RNA-Seq data as it based on the Negative Binomial Distribution (Witten, 2011).

3.3.2.4 Clustering Methods K Known and K Unknown Scenarios

In this section, we will describe each of the clustering methods that will be assessed in terms of how well they perform in terms of their accuracy in assigning samples to clusters. The clustering analysis portion of this simulation study can be divided into two branches. Branch 1 consist of those scenarios where the number of cluster(s) is considered "known" based upon expert's knowledge and literature. These scenarios will be denoted as "*K known*" scenarios. To know the number of clusters in a given dataset would be ideal. However, this is rarely the case. Thus, branch 2 consists of simulation scenarios where the number of cluster(s) is entirely unknown. Not knowing the number of clusters tasks the researcher to either make a subjective guess about the number of clusters based on graphical representations (i.e., a dendrogram or a scatterplot) of the data or use an algorithm based calculation (e.g., BIC, GAP statistics).

3.3.2.4.1 Clustering Methods for the K Known Scenarios

The *K Known* scenarios utilizes three clustering methods—Hierarchical Clustering (HC), Model-Based Clustering (MC) through the *mclust* package in R, and Nonnegative Matrix Factorization (NMF). For each of these clustering methods, the number of clusters, K, was purely determined by how many clusters were simulated in a given simulated dataset. Rather, if a dataset were simulated using $K = 2$ clusters, then the K fed to the clustering method would be two. By specifying a particular K for the clustering method, we force the method to partition the samples of the data into 2 clusters.

Hierarchical Clustering

One of the most common nonparametric clustering methods that is used in this study is Hierarchical clustering (HC). HC was developed by Eisen et al. in 1998 (Eisen et al., 1998). The HC utilizes all of samples and proceeds to divide them into smaller groups in an iterative manner. HC is a relatively simple type of clustering approach which provides a graphical representation of the results assuming that some hierarchical structure of the data exists (Shannon, 2003, Quackenbush, 2001). HC consists of two different variations—agglomerative, a type of bottom up approach, and divisible, a top-down approach. Furthermore, HC can be classified by the way in which clusters are formed or distance between clusters (Chalise et al., 2014). The formation of clusters is commonly termed as linkage which can be complete, average, and single. In this study, we use only the agglomerative variation with a complete linkage where all samples begin as their own cluster. Specifically, we let each sample be defined as S_i, it's own cluster, for $i = 1, \ldots, N$. From the individual clusters, pairwise distance comparisons are made in terms of the linkage. Since we use complete linkage in our method which seeks to maximize the distance between any pair (i.e, $\{S_i, S_{i+1}\}$) of individual clusters, the algorithm goes through all pairs and separates samples based upon furthest distance between them so that samples of c_1 are furthest from any sample in c_2 (Chen et al., 2002, Chalise et al., 2014). This procedure has been written into an R function, *hclust* in the basic *stats* package (R Development Core Team, 2016). Within this function, set up our parameters to reflect the prior type of HC clustering that we would like to use.

Model-Based Clustering

Model-Based Clustering (MC) comes in many different forms from methods that use mixture models (McLachlan et al., 2002, Yeung, 2001, Fraley et al., 2012, Farley and Raftery, 2002) to Bayesian model-based methods (Medvedovic and Sivaganesan, 2002, Medvedovic et al., 2004). Though, in clustering the incorporation of a "well-grounded" statistical model into a clustering method may serve to be beneficial in determining the best, most accurate clustering method (Yeung, 2001). In MC the data is assumed to be from some finite mixture of probability distributions (i.e., a mixture of Gaussian models) (Chalise et al., 2014, Yeung, 2001). Moreover, the likelihood of the mixture model can be written as:

$$L(\theta_1, \dots, \theta_K|\boldsymbol{X}) = \prod_{i=1}^{N} \sum_{c=1}^{K} \tau_c f_c(\boldsymbol{x_i}|\theta_c)$$

where K is the number of clusters or components in the data, $\boldsymbol{x_i}$ are the independent multivariate observations, f_c is the density of the some multivariate normal distribution distributional model with mean of $\boldsymbol{\mu_c}$ and covariance matrix $\sum_c$, θ_c are the parameters for the c^{th} component which can be thought of as the k$^{\text{th}}$ cluster, and τ_c is the probability that an observation belongs to the cth component-- τ_c has two restrictions $\tau_c \geq 0$ and $\sum_{c=1}^{K} \tau_c = 1$ (Yeung, 2001). Through, eigenvalue decomposition and an EM algorithm the number of clusters, K, is estimated. Within the eigenvalue decomposition, different parameterizations are used to define the model type that is being used. A wide range of model types exists: equal and unequal volume spherical models, unconstrained models, and elliptical models. Utilizing the *mclust* package in R, we are seamlessly able to implement this model-based clustering approach as proposed by Farley and Raftery in 2002 (Fraley et al., 2012, Farley and Raftery, 2002). As we wanted to optimize clustering performance in every method that we used in the simulation, we selected to use the *mclustBIC()* which determines the most optimal model characteristics (R Development Core Team, 2016, Fraley et al., 2012). This function has the flexibility to have the number of clusters specified or not specified.

Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a parts-based machine learning technique that uses a series of matrix manipulations to determine potential groups or likeness among objects (Devarajan, 2008, Lee and Seung, 1999). NMF has primarily been used to detect patterns in faces and text documents (Lee and Seung, 1999). However, recently NMF has been applied gene expression data from microarrays to discover molecular patterns (Brunet et al., 2004). The aim of NMF is to reduce the dimensionality of the data to find a small number of genes which are defined as a nonnegative linear combination of p genes (Devarajan, 2008). For our clustering analysis, we let our $N\ x\ G$ transformed matrix of counts be V which is decomposed into two matrices with non-negative counts (i.e., $V{\sim}WH$). W and H are matrices that are $p\ x\ k$ and $k\ x\ n$, respectively. The algorithm look for rank k of the factorization which represents the number of clusters (Lee and Seung, 1999). The rank is chosen to satisfy $(p+n)r < pn$ (Lee and Seung, 1999). It should be mentioned that the NMF algorithm may not always converge to the same solution for any given run (Brunet et al., 2004). To combat this challenge, Brunet et al. amended the initial NMF method proposed by Lee and Seung (1999) by adjusting the algorithm so to avoid numerical underflow (Brunet et al., 2004).

Some of the data transformations that were applied to the simulated datasets resulted in negative values which are unacceptable in NMF. Values for the NMF algorithm need to be nonnegative as the name suggests—a restriction that is unique to NMF. Hence, the minimum absolute value count in the dataset was added to all values of the dataset. NMF will not perform well if the dataset contains too much sparseness or values that are 0. Thus, for those dataset from the RE and RX categories, we added an additional count to all of the data values in the data set. Furthermore, for the NMF clustering method, the standard NMF method was used, “Brunets” method.

3.3.2.4.2 K Unknown Scenarios

The *K Unknown* scenarios would be most representative to that which would be faced in “real-life” when conducting a cluster analysis as the number of clusters is rarely known. The clustering analysis for *K Unknown* branch utilize all of the clustering methods that were used in the *K Known* clustering analysis with the addition of two different clustering methods. The two additional clustering methods include Recursive Partitioned Mixture Model (RPMM) clustering, and K-Means (KM) clustering. RPMM is a relatively new clustering methodology in comparison to KM which has been around for quite some time. The addition of these two clustering methods to the common exploratory analyses, *K Unknown*-type analyses, allows for a more comprehensive evaluation of available clustering methods. In this *K Unknown* scenarios, the number of clusters is unknown. Here we are ignoring how the data were simulated in clusters and use a data driven - algorithmic approach in combination with each of the five *K Unknown* clustering methods to estimate the number of clusters.

Recursively Partitioned Mixture Model Clustering

Recursively Partitioned Mixture Model (RPMM) clustering is another method that relies on mixture models to aid in the clustering of samples similarly to MC. Additionally, RPMM clustering also assumes that the data have some hierarchical structure as in HC. The combination of these two attributes make allow for model-based hierarchical clustering method for high-dimensional data; such as, RNA-Seq data (Houseman et al., 2008, Koestler, 2013). One caveat to this clustering method is that it will only cluster to a maximum of 2^r where r is the number of partitions algorithmically determined. In R, the *rpmm()* was used from the *RPMM* package (Houseman, 2014, R Development Core Team, 2016).

K-Means Clustering

K-Means (KM) clustering is one of the older clustering methods dating back to its first application nearly 40 years ago. KM groups objects into (k) fixed number of clusters so that the within-cluster sum of squares is minimized (Hartigan, 1979). The algorithm essentially shuffles all samples around searching for K clusters that have which have their respective within-cluster sum of squares minimized. For this type of clustering K must be known. Similarly to all other clustering methods mentioned above, the KM also has a function in R that will complete the procedures—*kmeans()* (R Development Core Team, 2016).

3.3.2.4.2.1 Selection of K Number of Clusters

Few formally defined algorithmic approaches have been developed to determine the number of clusters to be selected in clustering analyses. Most studies select K in a subjective manner as previously mentioned. Determination of the number of clusters for a specific analysis is a difficult task. Number of clusters are typically approximated based upon an experts advice or information from prior studies. Although, when no prior information is known about the data to be clustered, we turn to algorithmic approaches. In our study, we use two algorithmic approaches to estimate the number of clusters: (1) the Gap Statistic (Tibshirani, 2001), and (2) a model-based approach through using the best Bayesian Information Criterion (BIC) from the *mclust* package in R (Farley and Raftery, 2002, Fraley et al., 2012). When implementing the Gap Statistic, we developed code that would allow for the algorithm to work seamlessly with each of the five *K Unknown* clustering methods. The code modifications let the Gap Statistic select the number of appropriate clusters for the data while keeping the innate clustering method unaltered. However, execution of the model-based approach was conducted slightly different; rather than utilizing all of the five K Unknown clustering methods, only MC was used. In this instance, MC can determine the best BIC from the results of its EM (expectation-maximization) algorithm. The BIC that resembles the first decisive local maximum is indicative of the best model and in turn an estimation for the number of clusters that are present in the data (Fraley and Raftery, 1998).

Gap Statistic

Tibshirani et al. (2001) came up with a method that would select K such that the Gap Statistic would be optimized. The Gap Statistic is a measure that compares the within-cluster dispersion to that of the dispersion under the null (Tibshirani, 2001). According to Tibshirani, calculation of the Gap Statistic begins by clustering the data by varying the range of values for K; as well as, the within-dispersion measures (W_k) for $k = 1, 2, \dots, K$. From here, B reference datasets ($b = 1, \dots, \mathrm{B}$) are generated where the features are from a uniform distribution. Each dataset is then clustered to produce new within-dispersion measures (W_{kb}^{*}). The formula for the Gap Statistics becomes:

$$Gap\ Statistic = Gap(k) = \left({}^{1}/_{B}\right)\sum_{b} \log(W_{kb}^{*}) - \log(W_{k}).$$

For this particular study, we restricted the range K to $1 - 10$ clusters and the number of bootstraps, *B*, to 10. Once, the Gap Statistic is calculated for all variations of K, the estimated number of clusters can be calculated as:

$$\hat{k} = smallest\ k\ such\ that\ Gap(k) \geq Gap(k+1) - s_{k+1}$$

where $s_k = sd_k\sqrt{(1 + 1/B)}$ and $sd_k$is the standard deviation of the Gap Statistic for k. $\hat{k}$ is the new estimated number of clusters for a given dataset. For each of our transformed datasets from the four parent categories, we conducted a small simulation for each of the *D* datasets with $B =$ 10 bootstraps of the reference datasets for the calculation of the Gap Statistic for each of the clustering methods used in the *K Unknown* scenarios. This small simulation generated 100 values of $\hat{k}$ –one for each of the D datasets for any given combination of data transformation and clustering method. For each distinct combination mean $\hat{k}$ was calculated and used as the specified number of clusters. *K Unknown* clustering methods were then reran with the corresponding mean $\hat{k}$.

K Selection using Mclust

A model-based algorithm is another option to use to determine the number of clusters in a particular dataset. Recall from the MC section above, the f_c density from some multivariate normal distribution model with mean of $\boldsymbol{\mu_c}$ and covariance matrix $\sum_c$. Using different parameterizations for $\sum_c$, allows for different models to be passed through the EM algorithm as the number of clusters are varied (Fraley and Raftery, 1998). BICs are calculated for every possible combination of number of clusters and covariance matrix parameterization (Fraley and Raftery, 1998). The combination with the highest BIC yields the estimation of the number of clusters, say $\hat{k}_{MC}$, as previously mentioned above. Unlike the Gap Statistic implementation, we were only able to determine the number of clusters for each of the 100 simulated datasets that had been transformed. The estimated $\hat{k}_{MC}$ were then also re ran through all *K Unknown* clustering methods and evaluated.

3.3.2.5 Clustering Evaluation Methods

To summarize and compare the transformations and clustering methods, the following three evaluation criteria were used: Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), Clustering Error Rate (CER) (Witten et al.), and the Concordance Index (CI or C-Index) (Harrell et al., 1996). The ARI ranges in value between 0 and 1 and is computed as a measure of cluster similarity (Sibru et al., 2012, Hubert and Arabie, 1985). The decision to use the ARI instead of the Rand Index (Rand, 1971) was made because a constant value is not allowed for the expected value of two clustering (Rokach and Maimon, 2005). Values near 0 represent a lack of samples clustering to their "true" cluster; whereas, 1 indicates that samples cluster perfectly. ARI can be easily implemented using *adj.rand.index()* function in the fossil package in R (R Development Core Team, 2016, Rand, 1971, Hubert and Arabie, 1985). The CER is similar to the ARI; however, it is essentially the complimentary calculation without the adjustment. Lastly, the CI the probability that Sample 1 will cluster to c_1 if the sample was initially from c_1. Formally it is the probability between the predicted and the observed cluster assignments (Harrell et al., 1996). A CI value equal to 0.5 means that the probability of predicting the correct cluster assignment is no better than that of random chance or that there is no predictive ability. Values of CI that are closer to 1 indicate high predictive ability for objects to be clustered perfectly (Harrell et al., 1996). All of the distinct formulas or steps for each of these three evaluation criteria can be found in Appendix D.

3.4 Results

3.4.1 Simulated Data

For all simulation scenarios, data were simulated from a Negative Binomial distribution to represent four parent categories. As a reference, the four parent are: Top 100 MAD Genes with Equal Cluster Sized; 2) Random 100 MAD Genes with Equal Cluster Sizes; 3) Top 100 MAD Genes with Unequal Cluster Sizes; and 4) Random 100 MAD Genes with Unequal Cluster Sizes, abbreviated TE, RE, TX, and RX, respectively. Within each panel of Figure III-6, the x-axis contains the N samples and the y-axis the G genes. Each of the heatmaps are ordered by the correlation that is present between the different genes. Figure III-6 displays the variation between the parent scenarios; in addition, to the distinction of the clusters which are represented by the vertical dashed red lines. The varying number of clusters are a direct result of the effect size shifts that were applied to the mean and overdispersion parameters of the Negative Binomial distribution. As expected those parent categories that were selected randomly (RE and RX) from the Mayo Clinic data show more variation among the read counts that are present globally in the simulated datasets; whereas, less variation is present in the parent categories that were selected from the top 100 MAD genes (TE and TX).

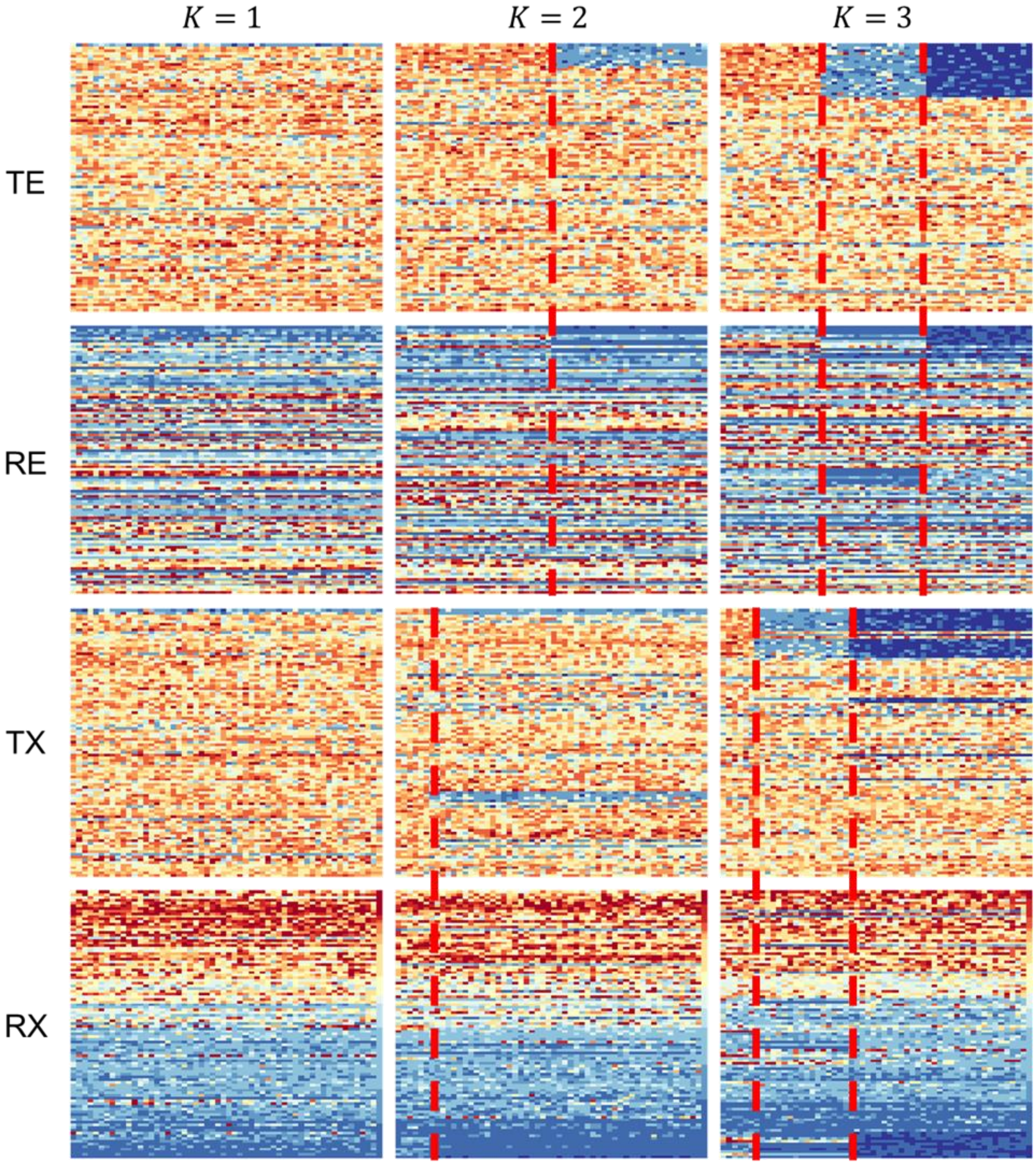
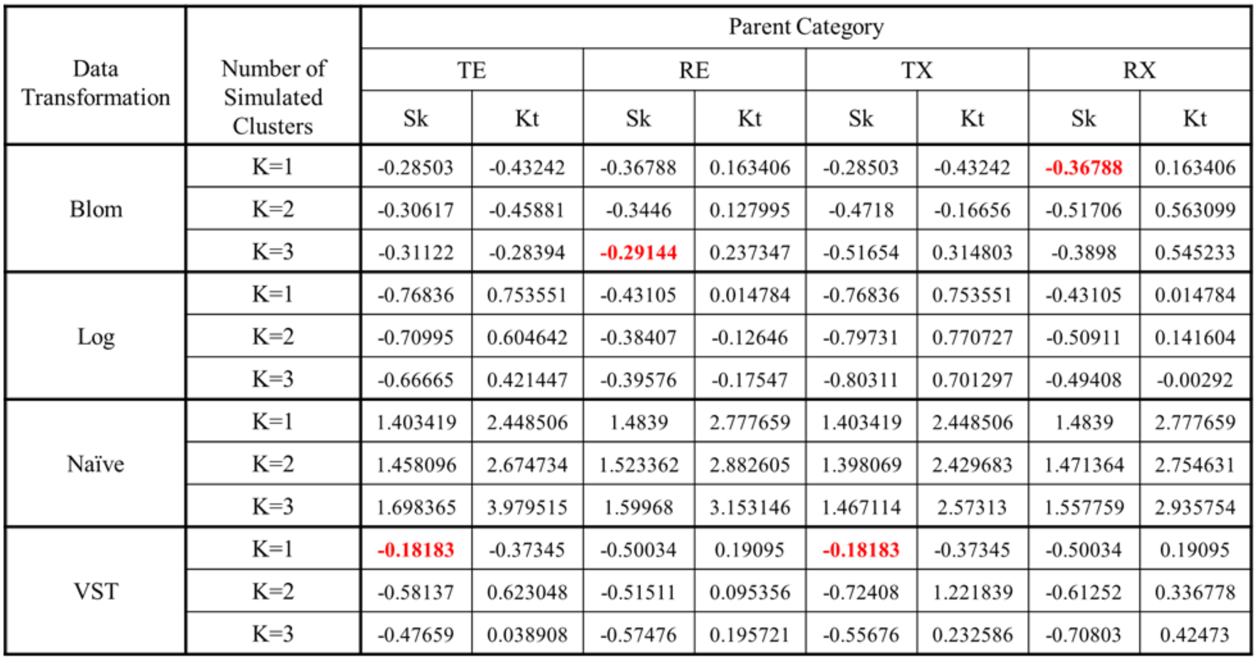


Figure III-6. Heatmaps for all four parent categories and for all K. Heatmaps from each of the four cateegories are plotted with each of the three different number of clusters (i.e., $K = 1$ as no clusters, $K = 2$ as two clusters, and $K = 3$ as three clusters). The dashed red lines help to display where clusters divide.

3.4.1.1 Comparison of Data Transformations

To compare the data transformations, we first looked at the measures of skewness and kurtosis, where data sampled from a Gaussian distribution would be around 0 and 3, respectively. All data transformation which numerically changed the data (i.e., Blom, Log, and VST) had on skewness values that more similar to that of a Gaussian distribution as compared to the naïve transformation (Table III-2). Though, the corresponding kurtosis values to those skewness values that are more normal all are platykurtic or have kurtosis values ≤ 3. Skewness values were closest to 0 for the RE and RX parent scenarios came from the Blom transformation and from the VST transformation for the TE and TX scenarios (Table III-2). This can also be visually seen in Figure III-7 where the best results of skewness are listed as the following combinations for the four parent categories: TE--VST transformation and $K = 1$; RE--Blom transformation and $K = 3$; TX--VST transformation and $K = 1$; and RX--Blom transformation and $K = 1$. We see that the Q-Q plots only minimal deviation from the theoretic quantile line which is an indication that data are approximately normal (Figure III-7). Figure III-7 also shows that the tails of the distribution are heavy, which correspond to the case when kurtosis value ≤ 3. Values for both skewness and kurtosis remained the same when $K = 1$ across all transformations selection-based parent scenarios of TE and TX, and RE and RX implying that method of data selection does not play a role in determining normality. For simulated clusters of $K = 2$ or $K = 3$, it is likely that the combination of varied cluster sizes and the effect shifts implemented in the NB distribution to form clusters play a role in the differences in normality between parent categories.



Data Transformation	Number of Simulated Clusters	Parent Category							
		TE		RE		TX		RX	
		Sk	Kt	Sk	Kt	Sk	Kt	Sk	Kt
Blom	K=1	-0.28503	-0.43242	-0.36788	0.163406	-0.28503	-0.43242	**-0.36788**	0.163406
	K=2	-0.30617	-0.45881	-0.3446	0.127995	-0.4718	-0.16656	-0.51706	0.563099
	K=3	-0.31122	-0.28394	**-0.29144**	0.237347	-0.51654	0.314803	-0.3898	0.545233
Log	K=1	-0.76836	0.753551	-0.43105	0.014784	-0.76836	0.753551	-0.43105	0.014784
	K=2	-0.70995	0.604642	-0.38407	-0.12646	-0.79731	0.770727	-0.50911	0.141604
	K=3	-0.66665	0.421447	-0.39576	-0.17547	-0.80311	0.701297	-0.49408	-0.00292
Naïve	K=1	1.403419	2.448506	1.4839	2.777659	1.403419	2.448506	1.4839	2.777659
	K=2	1.458096	2.674734	1.523362	2.882605	1.398069	2.429683	1.471364	2.754631
	K=3	1.698365	3.979515	1.59968	3.153146	1.467114	2.57313	1.557759	2.935754
VST	K=1	**-0.18183**	-0.37345	-0.50034	0.19095	**-0.18183**	-0.37345	-0.50034	0.19095
	K=2	-0.58137	0.623048	-0.51511	0.095356	-0.72408	1.221839	-0.61252	0.336778
	K=3	-0.47659	0.038908	-0.57476	0.195721	-0.55676	0.232586	-0.70803	0.42473

Table III-2. Normality summary of data transformations. Mean skewness (Sk) and mean kurtosis (Kt) values were calculated for each of the data transformations for each of K simulated clusters in the four parent categories. Those values that are bolded in red represent the closet value of skewness to the Normal distribution.

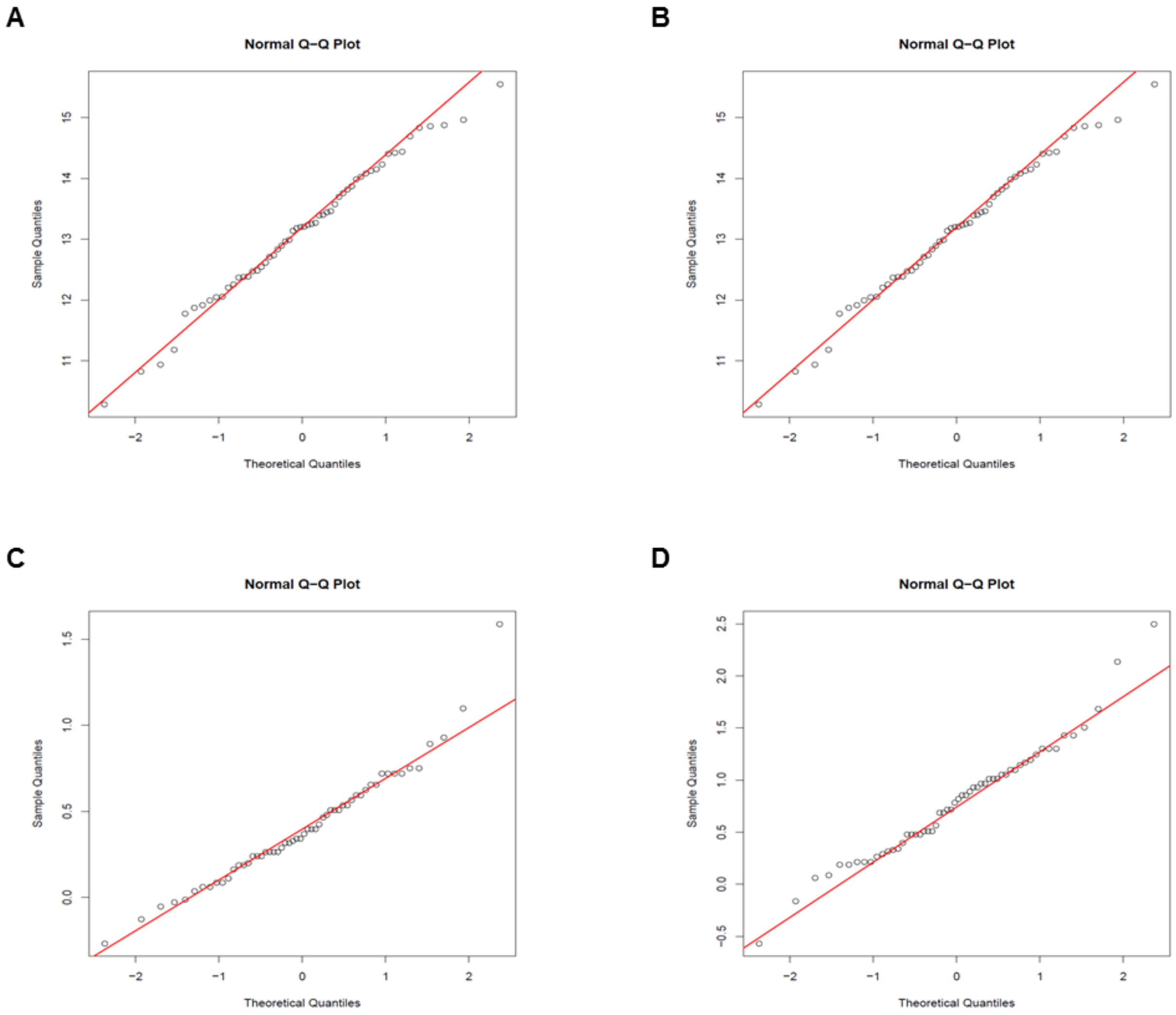


Figure III-7. Q-Q plots from data transformations whose skewness was most similar to Gaussian distribution. All Q-Q plots are from a sample of data from the scenario of those best skewness values highlighted in red in Table III-2. A) Q-Q plot for parent category TE, VST transformation, and, $K = 1$. B) Q-Q plot for parent category RE, Blom transformation, and, $K = 3$. C) Q-Q plot for parent category TX, VST transformation, and, $K = 1$. D) Q-Q plot for parent category RX, Blom transformation, and, $K = 1$. The red 45 degree line present in each panel provides a reference to where points should fall if from Gaussian distribution.

3.4.1.2 Comparison of Clustering Methods

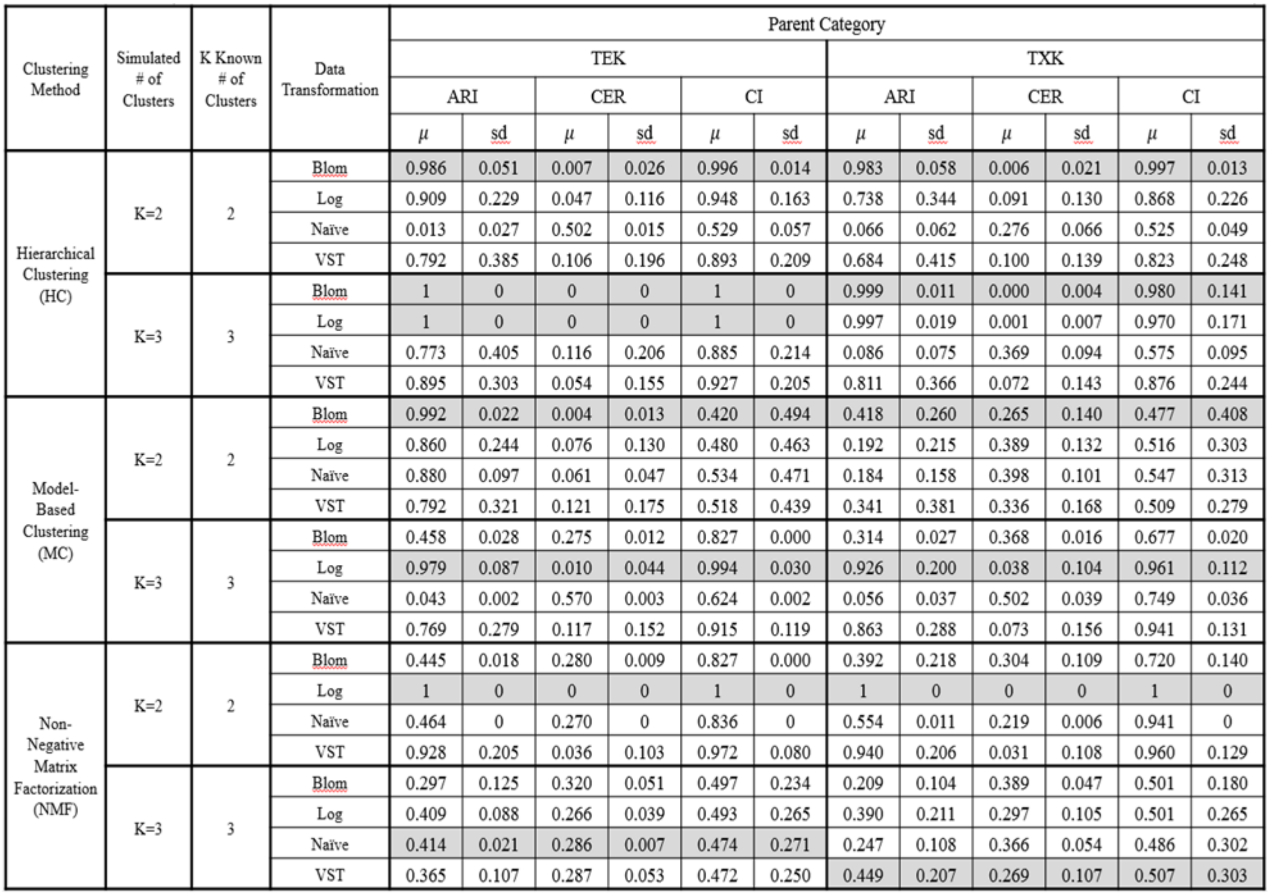
Clustering method performance was measured for both the *K Known* (Table III-3 and Table III-4) and *K Unknown* (Table III-5) simulations. Assessment of $K = 1$ scenario for *K Known* and

K Unknown scenarios are not presented in either of the summary tables as there were no clusters to compare sample assignment. This is due to the way in which $K = 1$ datasets were simulated to not reflect any clusters. However, we calculated the proportion of times in which the clustering algorithm selected the correct number of simulated clusters. This was completed for not only *K Unknown* for K=1, but for all other scenarios as well. Comparisons between the Gap Statistic's findings and the MC based selection for *K* were also made.

K Known Scenarios

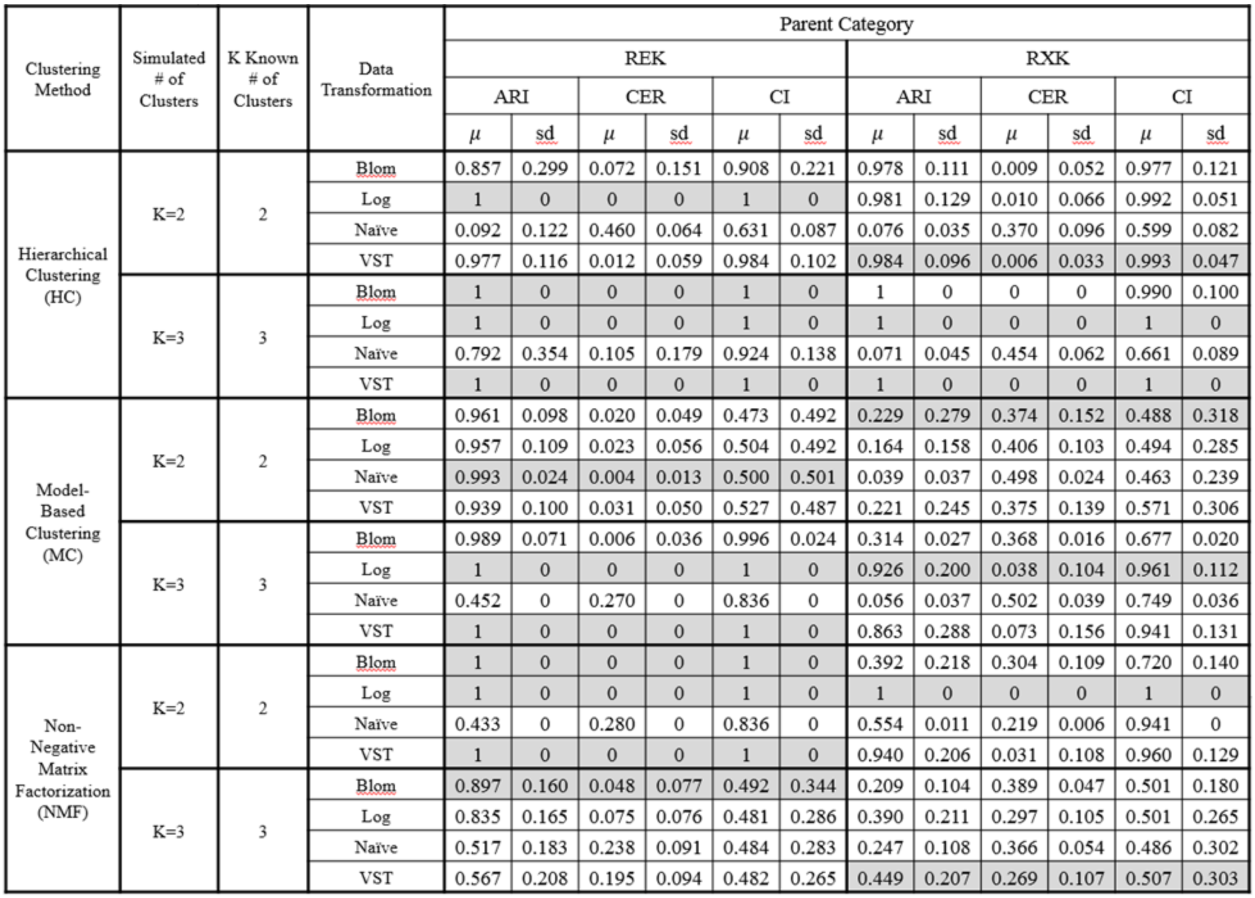
For the *K Known* simulation scenarios, combinations of all clustering methods for most data transformations and known K values had performance values that were better than random chance. That is, their mean CI values are greater than 0.5. Conversely, for the NMF clustering method, there are only minimal differences from 0.5. The small variations ranged between CI values or 0.48 to 0.55. When looking at the ARI and CER it is apparent that differences do exists between the pairing of data transformation and type of clustering method used. Notably, model-based clustering does not perform well in regards to selecting the correct clustering assignment when the Blom transformation is used with data that are highly variable prior to the transformation or rather for those data that represent the top 100 MAD genes (Table III-3 and Table III-4). The best overall performance was observed when model-based clustering was carried out on data that were simulated with three clusters and when a log transformation was applied (Table III-3, Figure III-8 and Figure III-9). For this combination of simulation parameters, parent category did not play a role.

Model-based clustering on average appears to have the greatest performance when it comes to correct cluster assignment evaluated by our clustering metrics. In Figure III-8 and Figure III-9, corresponding clustering evaluation metrics are present on the y-axis and data transformations on the x-axes. We can see that the model-based clustering (MC) line, in red, tends to be higher than the other clustering methods for ARI and CI and lowest for CER (Figure III-8 and Figure III-9). On the contrary, it appears that NMF performs the worst across data transformations. Though, there are a few instances where NMF works better when the data have not been transformed (i.e., the naïve transformation). This may also be due to the lack of assumptions required to carry out non-parametric methods--the assumption of normality is not needed. For all of the clustering performance metrics, there are drastic differences when comparing the Naïve transformation to the other data transformations. The Blom, Log, and VST transformations have similar results across the evaluation metrics for HC and MC clustering methods for data that were simulated from the selected random 100 MAD genes. Additionally, it appears that the performance of HC follows that of MC, and has only slightly poorer performance according to all metrics.



Clustering Method	Simulated # of Clusters	K Known # of Clusters	Data Transformation	Parent Category											
				TEK						TXK					
				ARI		CER		CI		ARI		CER		CI	
				μ	sd	μ	sd	μ	sd	μ	sd	μ	sd	μ	sd
Hierarchical Clustering (HC)	K=2	2	Blom	0.986	0.051	0.007	0.026	0.996	0.014	0.983	0.058	0.006	0.021	0.997	0.013
			Log	0.909	0.229	0.047	0.116	0.948	0.163	0.738	0.344	0.091	0.130	0.868	0.226
			Naïve	0.013	0.027	0.502	0.015	0.529	0.057	0.066	0.062	0.276	0.066	0.525	0.049
			VST	0.792	0.385	0.106	0.196	0.893	0.209	0.684	0.415	0.100	0.139	0.823	0.248
	K=3	3	Blom	1	0	0	0	1	0	0.999	0.011	0.000	0.004	0.980	0.141
			Log	1	0	0	0	1	0	0.997	0.019	0.001	0.007	0.970	0.171
			Naïve	0.773	0.405	0.116	0.206	0.885	0.214	0.086	0.075	0.369	0.094	0.575	0.095
			VST	0.895	0.303	0.054	0.155	0.927	0.205	0.811	0.366	0.072	0.143	0.876	0.244
Model-Based Clustering (MC)	K=2	2	Blom	0.992	0.022	0.004	0.013	0.420	0.494	0.418	0.260	0.265	0.140	0.477	0.408
			Log	0.860	0.244	0.076	0.130	0.480	0.463	0.192	0.215	0.389	0.132	0.516	0.303
			Naïve	0.880	0.097	0.061	0.047	0.534	0.471	0.184	0.158	0.398	0.101	0.547	0.313
			VST	0.792	0.321	0.121	0.175	0.518	0.439	0.341	0.381	0.336	0.168	0.509	0.279
	K=3	3	Blom	0.458	0.028	0.275	0.012	0.827	0.000	0.314	0.027	0.368	0.016	0.677	0.020
			Log	0.979	0.087	0.010	0.044	0.994	0.030	0.926	0.200	0.038	0.104	0.961	0.112
			Naïve	0.043	0.002	0.570	0.003	0.624	0.002	0.056	0.037	0.502	0.039	0.749	0.036
			VST	0.769	0.279	0.117	0.152	0.915	0.119	0.863	0.288	0.073	0.156	0.941	0.131
Non-Negative Matrix Factorization (NMF)	K=2	2	Blom	0.445	0.018	0.280	0.009	0.827	0.000	0.392	0.218	0.304	0.109	0.720	0.140
			Log	1	0	0	0	1	0	1	0	0	0	1	0
			Naïve	0.464	0	0.270	0	0.836	0	0.554	0.011	0.219	0.006	0.941	0
			VST	0.928	0.205	0.036	0.103	0.972	0.080	0.940	0.206	0.031	0.108	0.960	0.129
	K=3	3	Blom	0.297	0.125	0.320	0.051	0.497	0.234	0.209	0.104	0.389	0.047	0.501	0.180
			Log	0.409	0.088	0.266	0.039	0.493	0.265	0.390	0.211	0.297	0.105	0.501	0.265
			Naïve	0.414	0.021	0.286	0.007	0.474	0.271	0.247	0.108	0.366	0.054	0.486	0.302
			VST	0.365	0.107	0.287	0.053	0.472	0.250	0.449	0.207	0.269	0.107	0.507	0.303

Table III-3. Summary of evaluation metrics by clustering method and data transformation for *K Known* TEK and TXK parent categories. Three different evaluation metrics were used to evaluate the performance of the data transformation paired with clustering method.. The evaluation metrics include: Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI). Mean (μ) and standard deviation (sd) values were calculated for each metric. Fields with gray represent the highest performing transformation for *K* and clustering method by parent category. Note that some ties regarding performance exist.



Clustering Method	Simulated # of Clusters	K Known # of Clusters	Data Transformation	Parent Category											
				REK						RXK					
				ARI		CER		CI		ARI		CER		CI	
				μ	sd	μ	sd	μ	sd	μ	sd	μ	sd	μ	sd
Hierarchical Clustering (HC)	K=2	2	Blom	0.857	0.299	0.072	0.151	0.908	0.221	0.978	0.111	0.009	0.052	0.977	0.121
			Log	1	0	0	0	1	0	0.981	0.129	0.010	0.066	0.992	0.051
			Naïve	0.092	0.122	0.460	0.064	0.631	0.087	0.076	0.035	0.370	0.096	0.599	0.082
			VST	0.977	0.116	0.012	0.059	0.984	0.102	0.984	0.096	0.006	0.033	0.993	0.047
	K=3	3	Blom	1	0	0	0	1	0	1	0	0	0	0.990	0.100
			Log	1	0	0	0	1	0	1	0	0	0	1	0
			Naïve	0.792	0.354	0.105	0.179	0.924	0.138	0.071	0.045	0.454	0.062	0.661	0.089
			VST	1	0	0	0	1	0	1	0	0	0	1	0
Model-Based Clustering (MC)	K=2	2	Blom	0.961	0.098	0.020	0.049	0.473	0.492	0.229	0.279	0.374	0.152	0.488	0.318
			Log	0.957	0.109	0.023	0.056	0.504	0.492	0.164	0.158	0.406	0.103	0.494	0.285
			Naïve	0.993	0.024	0.004	0.013	0.500	0.501	0.039	0.037	0.498	0.024	0.463	0.239
			VST	0.939	0.100	0.031	0.050	0.527	0.487	0.221	0.245	0.375	0.139	0.571	0.306
	K=3	3	Blom	0.989	0.071	0.006	0.036	0.996	0.024	0.314	0.027	0.368	0.016	0.677	0.020
			Log	1	0	0	0	1	0	0.926	0.200	0.038	0.104	0.961	0.112
			Naïve	0.452	0	0.270	0	0.836	0	0.056	0.037	0.502	0.039	0.749	0.036
			VST	1	0	0	0	1	0	0.863	0.288	0.073	0.156	0.941	0.131
Non-Negative Matrix Factorization (NMF)	K=2	2	Blom	1	0	0	0	1	0	0.392	0.218	0.304	0.109	0.720	0.140
			Log	1	0	0	0	1	0	1	0	0	0	1	0
			Naïve	0.433	0	0.280	0	0.836	0	0.554	0.011	0.219	0.006	0.941	0
			VST	1	0	0	0	1	0	0.940	0.206	0.031	0.108	0.960	0.129
	K=3	3	Blom	0.897	0.160	0.048	0.077	0.492	0.344	0.209	0.104	0.389	0.047	0.501	0.180
			Log	0.835	0.165	0.075	0.076	0.481	0.286	0.390	0.211	0.297	0.105	0.501	0.265
			Naïve	0.517	0.183	0.238	0.091	0.484	0.283	0.247	0.108	0.366	0.054	0.486	0.302
			VST	0.567	0.208	0.195	0.094	0.482	0.265	0.449	0.207	0.269	0.107	0.507	0.303

Table III-4. Summary of evaluation metrics by clustering method and data transformation for *K Known* REK and RXK parent categories. Three different evaluation metrics were used to evaluate the performance of the data transformation paired with clustering method.. The evaluation metrics include: Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI). Mean (μ) and standard deviation (sd) values were calculated for each metric. Fields with gray represent the highest performing transformation for *K* and clustering method by parent category. Note that some ties regarding performance exist.

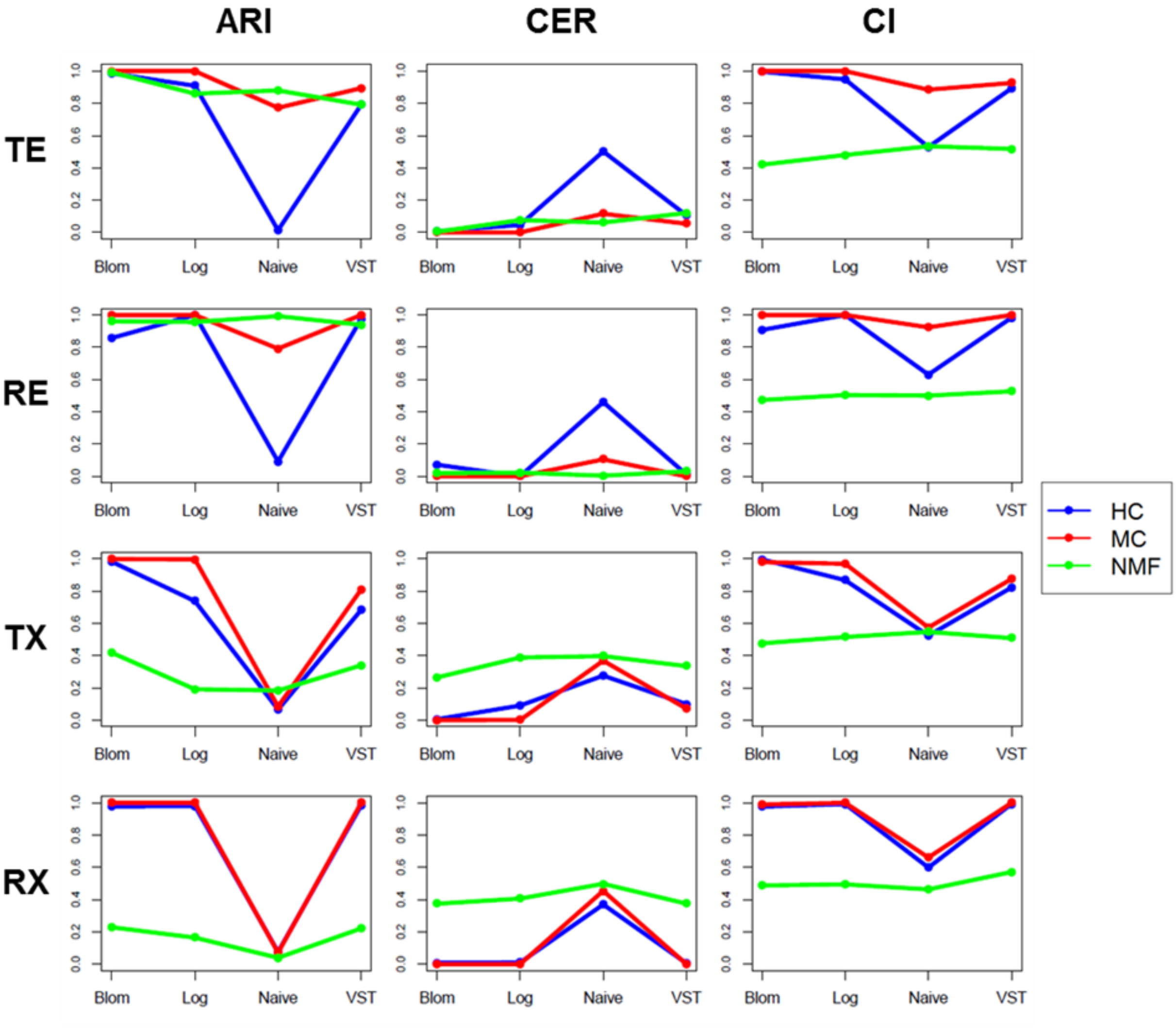


Figure III-8. Metric evaluation summary for all parent categories for $K = 2$. Mean Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI) are plotted for each of the four parent categories for $K = 2$. The Hierarchical Clustering (HC), Model-based Clustering (MC), and Non-Negative Matrix Factorization (NMF) are represented by the blue, red, and green lines, respectively.

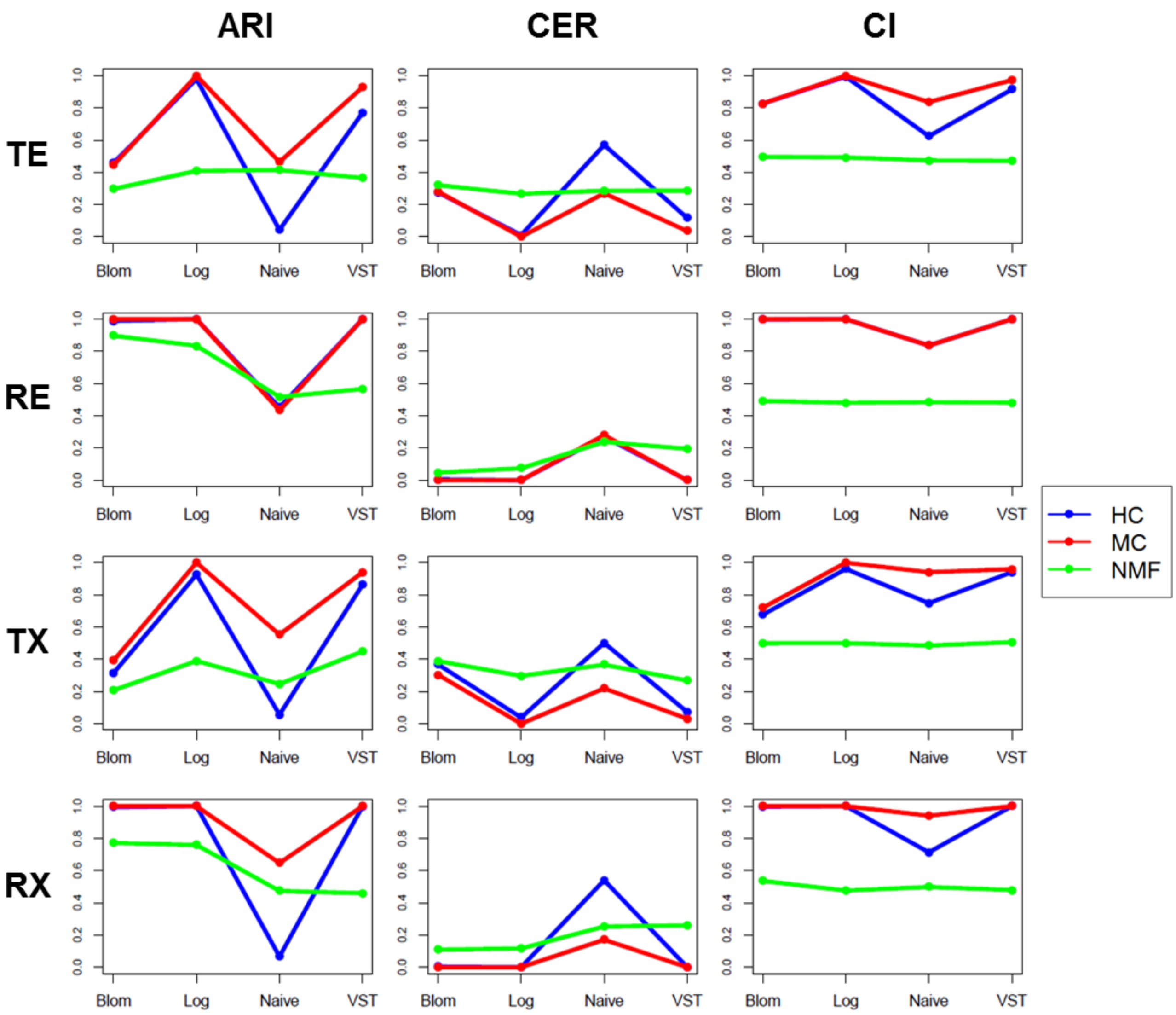
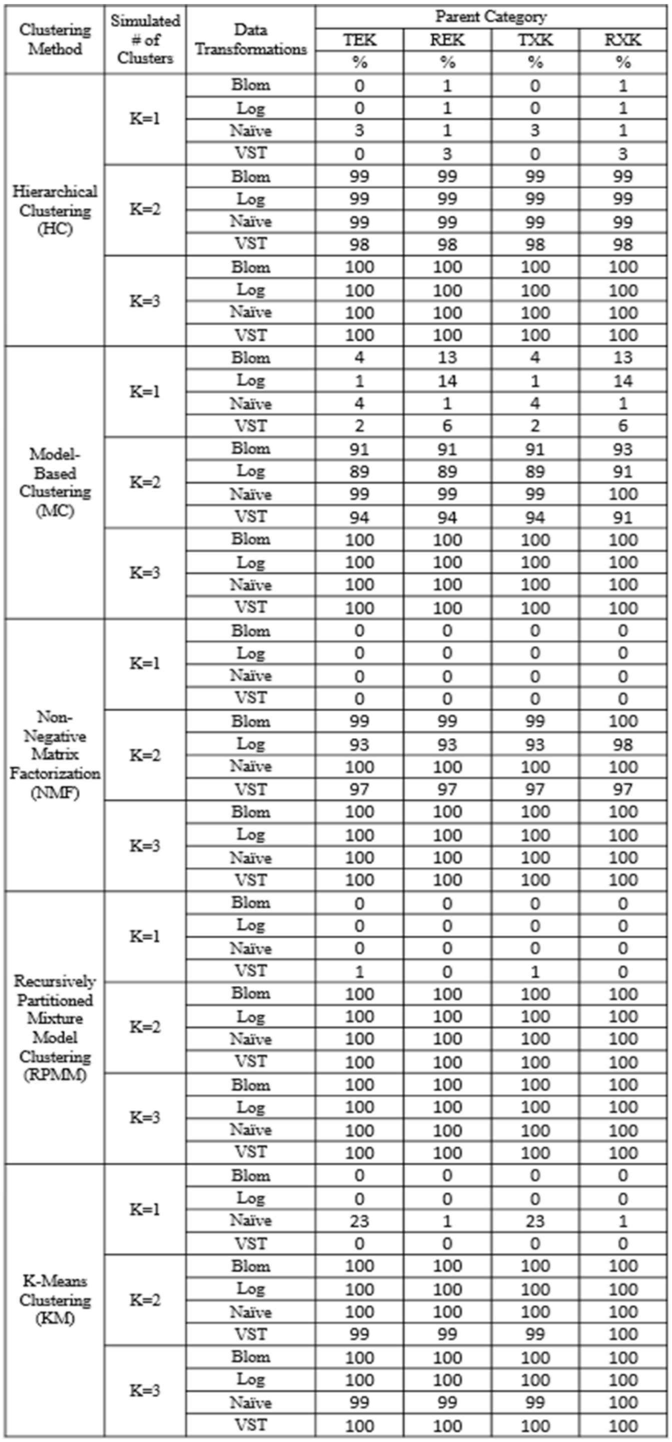


Figure III-9. Metric evaluation summary for all parent categories for $K = 3$. Mean Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI) are plotted for each of the four parent categories for $K = 3$. The Hierarchical Clustering (HC), Model-based Clustering (MC), and Non-Negative Matrix Factorization (NMF) are represented by the blue, red, and green lines, respectively.

K Unknown scenarios

For the *K Unknown* branch of the simulation, the same clustering performance metrics were used. Two additional clustering methods were also assessed— Recursively Partitioned Mixture Model (RPMM) and K-Means (KM). Prior to evaluation of the clustering methods, two algorithms were implemented. Both the Gap Statistic and model-base clustering via Bayesian Information Criteria (BIC) values were used to determine the optimal *K*. Results from the Gap Statistic and model-based approach can be found in Table III-5, and Table III-6, respectively. The Gap Statistic method often resulted with the choice of the data having no clusters $\hat{k} = 1$ regardless of the simulated number of clusters across all of *K Unknown* clustering methods the exception was for $K = 1$ the Gap Statistic tended to select $\hat{k} = 1$ correctly. At most, the Gap Statistic errored 23% of the time for the $K = 1$ scenarios. For $K = 2$ and $K = 3$, the Gap statistic failed in determining the correct number of simulated clusters nearly 100% regardless of parent category, clustering method, or data transformation. With the performance being so poor for the Gap Statistic in selecting the simulated *K*, we knew that our evaluation methods would not provide any useful additional information. This proved to be the case after brief evaluation. The mean CER and CI values for HC, MC, NMF, RPMM, and KM were very poor at values of 0.51 and 0.68 for $K = 2$ and $K = 3$, respectively. This was the case across all combinations of clustering methods, number of simulated clusters, and parent scenarios.



Clustering Method	Simulated # of Clusters	Data Transformations	Parent Category			
			TEK	REK	TXK	RXK
			%	%	%	%
Hierarchical Clustering (HC)	K=1	Blom	0	1	0	1
		Log	0	1	0	1
		Naïve	3	1	3	1
		VST	0	3	0	3
	K=2	Blom	99	99	99	99
		Log	99	99	99	99
		Naïve	99	99	99	99
		VST	98	98	98	98
	K=3	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	100	100	100	100
Model-Based Clustering (MC)	K=1	Blom	4	13	4	13
		Log	1	14	1	14
		Naïve	4	1	4	1
		VST	2	6	2	6
	K=2	Blom	91	91	91	93
		Log	89	89	89	91
		Naïve	99	99	99	100
		VST	94	94	94	91
	K=3	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	100	100	100	100
Non-Negative Matrix Factorization (NMF)	K=1	Blom	0	0	0	0
		Log	0	0	0	0
		Naïve	0	0	0	0
		VST	0	0	0	0
	K=2	Blom	99	99	99	100
		Log	93	93	93	98
		Naïve	100	100	100	100
		VST	97	97	97	97
	K=3	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	100	100	100	100
Recursively Partitioned Mixture Model Clustering (RPMM)	K=1	Blom	0	0	0	0
		Log	0	0	0	0
		Naïve	0	0	0	0
		VST	1	0	1	0
	K=2	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	100	100	100	100
	K=3	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	100	100	100	100
K-Means Clustering (KM)	K=1	Blom	0	0	0	0
		Log	0	0	0	0
		Naïve	23	1	23	1
		VST	0	0	0	0
	K=2	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	100	100	100	100
		VST	99	99	99	100
	K=3	Blom	100	100	100	100
		Log	100	100	100	100
		Naïve	99	99	99	100
		VST	100	100	100	100

Table III-5. Summary of the percentage (%) of times the Gap Statistic incorrectly selected the number of *K* simulated clusters. Percentages are given for each of the *K Unknown* clustering methods by *K* used in data simulation, data transformation, and parent category.

The poor performance of the Gap Statistic led us to explore other methods that are capable of algorithmically selecting the number of clusters present in a given dataset. We proceeded by implementing the model-based method for selecting *K* through the *mclust* R package. The results in terms of how often the algorithm selected the correctly the number of clusters were also discouraging. For all parent scenarios and all *K*, the algorithm incorrectly selected the number of clusters 100% of the time. We observe that the algorithm selects $\hat{k}_{MC}$ higher than the simulated number of clusters in the datasets. For $K = 1$, $K = 2$, $K = 3$, $\hat{k}_{MC}$ ranged from 2 clusters to 7 clusters, 3 clusters to 9 clusters, and 4 clusters to 9 clusters, respectively. While the algorithm does not directly pick up the exact number of clusters that were simulated in the datasets nor does the algorithm select $\hat{k}_{MC}$ to be equal to one nearly always, we can proceed with cluster evaluation as the as the greater value of $\hat{k}_{MC}$ still contains the number of simulated clusters. Rather, if $\hat{k}_{MC} = 4$ and our simulated datasets has $K = 3$, cluster assignments have the possibility to pick up all three clusters with potential that an outlying sample would be assigned to the fourth cluster.

Figure III-10 and Figure III-11 summarize the mean values of the clustering evaluation metrics similarly to Figures III-8 and III-9 for the *K Known* scenarios. K-Means, represented as the black line, clustering appears to perform the best when the number of optimal clusters is selected using the model-based clustering method’s built in approach for $K = 2\ and\ K = 3$ (Figure III-10 and Figure III-11). Conversely, Recursively Partitioned Mixture Model clustering performed worse than all other methods. We can also observe that HC, MC, and NMF clustering methods followed similar trends that were observed in the *K Known* scenarios. Specifically, for values of ARI and CER, we see drastic differences for the Naïve data transformation (Figure III-8, Figure III-9, Figure III-10, and Figure III-11). For the other three data transformations, it is difficult to select the one that might yield the most benefit to RNA-Seq in terms of clustering. According to the evaluation metrics that we used, different parent categories suggest that different data transformations are most beneficial. Another notable difference in these summaries is the lack of trends that exist across parent categories for all clustering methods; whereas, trends did exist in in the *K Known* clustering evaluation summary. We speculate that this is an artifact of carrying over $\hat{k}_{MC}$ from the MC selection algorithm applied to the transformed datasets and directly imputing it into each of the *K Unknown* clustering methods. No information was incorporated into the model-based algorithm regarding unique attributes present in each of the individual clustering method which may lead to the semi-irregular results. An interesting result of the $K = 2$ *Unknown* scenarios comes from comparing evaluation criteria across TX and RX (Figure III-10). KM, in black, has the highest performance when looking at ARI and CER. However, when looking at CI, we see that KM remains flat at 0.5 meaning the probability that the sample will cluster correctly is no greater than chance alone. This is a result that we wouldn't expect to have considering the mean values plotted for ARI and CER. KM evaluation criteria behavior in the *K Unknown* scenarios for $K = 3$ does not appear to have this drastic of a disagreement between any of the methods.

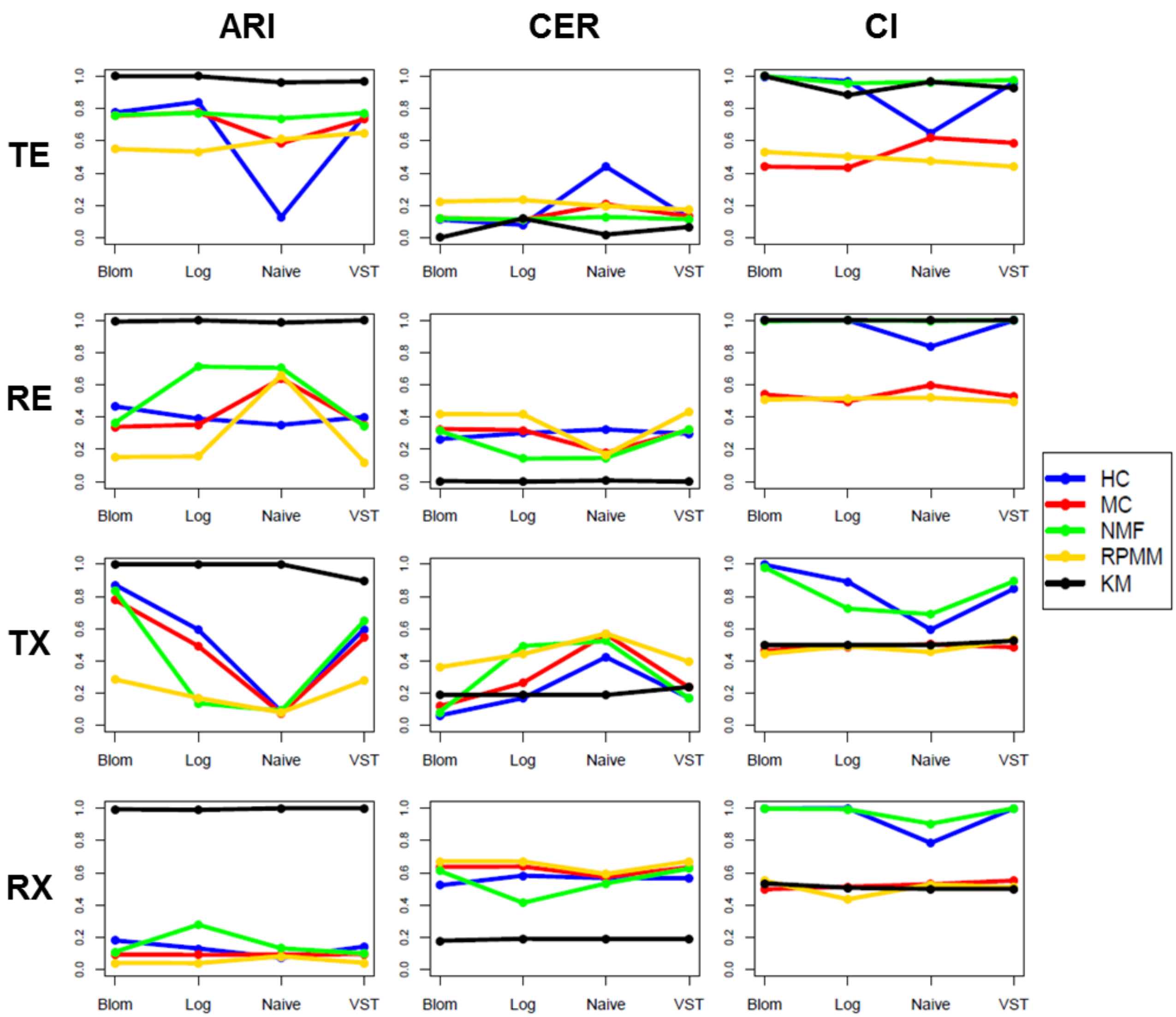


Figure III-10. Metric evaluation summary for all parent categories for $K = 2$ for model-based clustering selection of K. Mean Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI) are plotted for each of the four parent categories for $K = 2$. The Hierarchical Clustering (HC), Model-based Clustering (MC), Non-Negative Matrix Factorization (NMF), Recursively Partitioned Mixture Model (RPMM), and K-Means (KM) are represented by the blue, red, green, gold, and black lines, respectively.

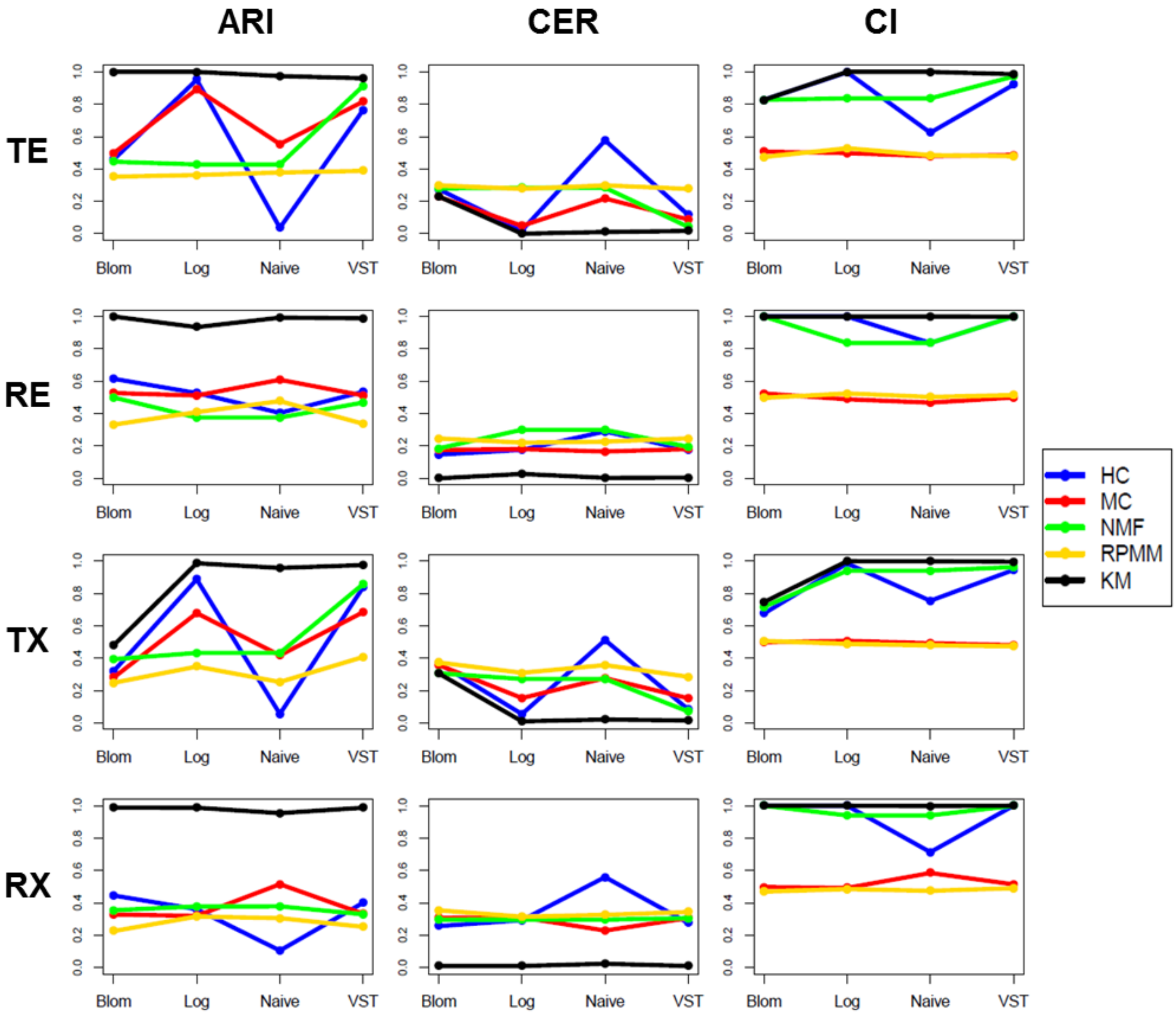


Figure III-11. Metric evaluation summary for all parent categories for K =3 for model-based clustering selection of K. Mean Adjusted Rand Index (ARI), Clustering Error Rate (CER), and Concordance Index (CI) are plotted for each of the four parent categories for $K = 3$. The Hierarchical Clustering (HC), Model-based Clustering (MC), Non-Negative Matrix Factorization (NMF), Recursively Partitioned Mixture Model (RPMM), and K-Means (KM) are represented by the blue, red, green, bold, and black lines, respectively.

3.4.1.3 Computational Resources

All *K Unknown* scenarios required much more computational time in comparison to *K Known* scenarios. The average time for the *K Known* scenarios was approximately 6,224.80 seconds or ~1.73 hours with standard deviation of .16 hours; whereas, the average time for the *K Unknown* scenarios was approximately 136,806.70 seconds or ~38 hours with standard deviation of ~4.3 hours. These average times were computed over all simulations in each of the given categories. In general, it was also true that as the simulated number of clusters in the simulation increased so did the amount of time it took to complete all of the calculations (Figure III-12). Additionally, there were no differences in computational time if the cluster sizes were equivalent or not. Computational time for simulations scenarios from both types of the data selection were similar to one another in both the *K Known* and *K Unknown* scenarios (Figure III-12). Rather, the amount of time for TE and TX scenarios and the scenarios from RE and RX were similar. RE and RX scenarios for the *K Unknown* had longer computational time than the TE and TX scenarios (Figure III-12). In the *K Known* scenarios the opposite is true—the TE and TX scenarios took longer to complete than did the RE and RX scenarios.

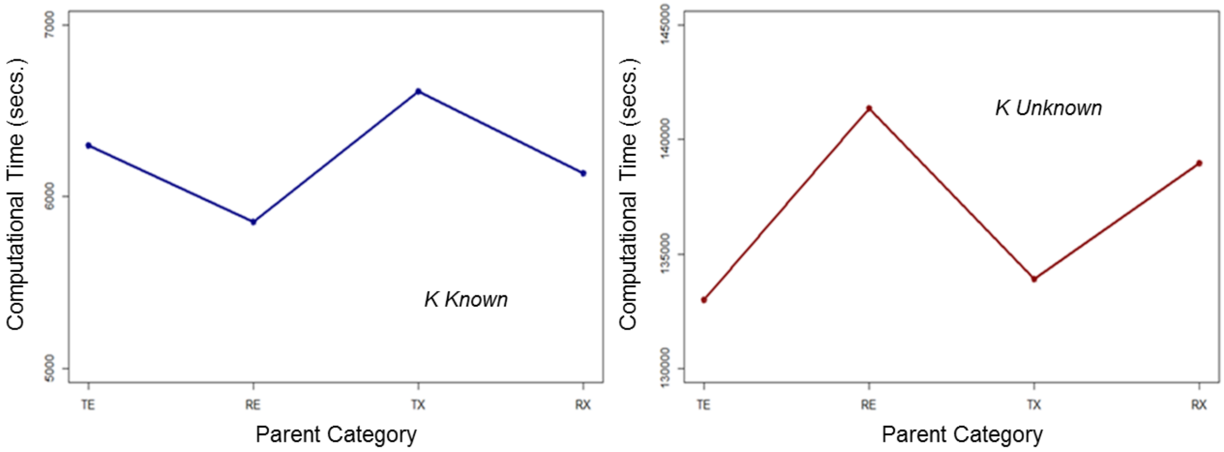


Figure III-12. Computational time for $K\ known$ and $K\ Unknown$ scenarios. Plots show the computational time that was taken to complete the simulations for each of the four parent scenarios. A) depicts the computational time for the $K\ Known$ scenarios; whereas, B) the $K\ Unknown$ scenarios.

3.5 Discussion

For RNA-Seq data there have not been many studies that have looked at the clustering performance of multiple clustering methods in combination with data transformations. Hence, there is little guidance for researchers as to which data transformations should be used for RNA-Seq data when conducting clustering analyses. Even if clustering analyses are exploratory in nature, they can provide valuable information regarding the relationship between genes or samples. In order to provide this information, it is important to have accurate and efficient statistical methods. In light of the minimal information and studies that are currently available, we conducted and compared the results of an extensive simulation study to assess clustering method performance when data selection, data transformations, number of simulated clusters, and clustering method were varied. Results from our simulation study provide insight to what could potentially be done to increase correct cluster assignment for samples; with or without prior information about how many clusters there might be. Additionally, the simulation study has revealed some of the challenges and difficulties that still remain for completing clustering analysis in RNA-Seq data. To combat biasing our performance results for our clustering methods, datasets were simulated to represent four parent categories that considered the way in which the data were selected and the size of the clusters.

We feel that the structure of the simulated data was improved upon from the Witten study from 2010 as parameters were obtained from “real-life” dataset. In Witten’s simulation study, the distributional parameters appeared to be arbitrarily selected. Values for the over-dispersion parameter, ϕ, appear to reflect values which are typically used to show over-dispersion within the data. Additionally, the mean of her model was composed of a multiplicative mixture of Uniform, Exponential, and Log-Normal distributions. While this way of approaching such simulation is typical, unique attributes from "real-life" data are missed. To better preserve the unique attributes that are present in actual RNA-Seq data samples, we obtained the NB MLEs to be used to simulate our datasets. Looking at Figure III-2 and Figure III-3 in Section 3.3, our simulated data appear to match the Mayo Clinic data well.

As RNA-Seq data do follow a NB distribution with high variability among gene read counts, implementation of data transformations are warranted. Without data transformations, the overdispersed nature of RNA-Seq create problems with many types of statistical methods. Moreover, many statistical methods, clustering methods included, are better adapted to handle data that follow more of a normal distribution. Hence, we applied the Blom, Log, and VST transformations to our simulated RNA-Seq data. In terms of skewness, we determined that all simulations made the data more normal. Specifically, the Blom transformation on average obtained the most normal data according to the mean skewness values (Table III-2). The data transformations; however, did not provide any benefit in handling tail behavior as denoted by mean kurtosis values presented in Table III-2.

Each data transformation was used for each of our clustering methods for the *K Known* and *K Unknown* simulation branches. In the assessment of clustering method performance in combination with the data transformations, we observed many expected results based off of previous literature from Witten, Yeung et al., and Thalamuthu et al.. While these authors' studies were completed in microarray data, model-based clustering (MC) was found to produce higher quality of clusters (Witten, 2011, Yeung, 2001, Thalamuthu et al., 2006). Notably, model-based clustering using *mclust* out-performed all other clustering methods in the *K Known* branch across all data transformations and evaluation metrics. Since the primary model used for MC is the Gaussian mixture model, it is reasonable that transforming data to look more normal would be highly beneficial for the performance of the MC method. Contrariwise, NMF lacks the level of performance in comparison to HC and MC. However, there does appear to be some benefit in using NMF when no transformation is made to the data. NMF does not use any model constraints to assign clusters as it is nonparametric, rather it seeks to minimize the generalized Kullback-Leibler divergence which is similar to conventional least squares. HC falls in the middle regarding its performance in comparison to MC and NMF. Though, performance could potentially be gained from using single linkage verses complete linkage especially in our parent scenario categories where the data selected were based on the top 100 MAD genes. Single linkage is better when outliers are present as the outliers are merged into clusters last rather than first (Chalise et al., 2014). The presence of potential outliers can be seen in Figure III-3.

Results from our *K Unknown* simulation branch revealed the difficulty in algorithmically selecting the number of clusters present in a given dataset when no expert advice is available. Implementation of the Gap Statistic purposed by Tibshirani et al. (2001) to determine the optimal number of clusters, $\hat{k}$, did not prove to provide any benefit in performance of any of the clustering methods assessed. In our Gap Statistic analysis to determine what the optimal $\hat{k}$ would be for any scenario, we nearly always ended up with the algorithm selecting our data to have no clusters, $\hat{k} = 1$, even when the data were purposefully simulated to have $K = 2$ and $K = 3$ clusters. After reviewing why this pneumonia was occurring, it was determined that the results are likely due to the fact that our data our data are not standardized nor may the effect shifts used to simulated the clusters in our datasets be large enough. This is likely the reason for the poor performance evaluation metrics in our clustering methods' analysis.

In addition to the Gap Statistic results for the *K Unknown* scenarios, the results from the second approach which used a model-based algorithm in combination with BIC, verified that any transformation to RNA-Seq data to make it more normal will be beneficial for downstream clustering analyses. Though, the utilization of an additional algorithm to select optimal *K*, did not provide much evidence regarding best practices when there is no information available regarding potential number of clusters. We can, however, say that our results suggest that using K-Means clustering is highly robust to missed assignment of clusters as it consistently performed better than all other clustering methods across our evaluation metrics. K-Means was also the only clustering method that had a relatively consistent performance trend across all parent categories. This is an interesting finding as K-Means clustering tends to perform best when clusters are of equal size. However, further investigation into the evaluation criteria of the $K = 2$ *Unknown* scenarios for TX and RX is needed to fully understand why the mean concordance index has such low performance for K-Means. Overall, the *K Unknown* performance results compared to the *K Known* performance results were worse for the mutual clustering methods used, HC, MC, and NMF, which is not unexpected. It would not be expected that by increasing uncertainty, which is present in our *K Unknown* branch of the study, that clustering performance would improve. Thus, not knowing the number of clusters in a given dataset is a challenge that researchers and statisticians will continually have to navigate.

This study does provide similar results to those which were found in microarray studies that looked at performance of clustering methods when the number of clusters is known. In particular, that model-based clustering demonstrates the highest level of performance when it comes to correctly assigning samples to their said clusters. Additionally, our results highly favor the use of the Log, base 2, transformation when it comes to conducting clustering analysis of RNA-Seq data. However, at this point, we feel that the results of the *K Unknown* simulation branch are inconclusive. The data and clustering performance would likely benefit from adding an additional step that would further standardize the data. Furthermore, there would be an opportunity to improve upon the Gap Statistic algorithm to better handle data that are not standardized. Likewise, there is room to advance those clustering methods used in this study, aside from MC, by developing algorithms to select the optimal number of clusters in a way that considers the method's theoretical background. There is also the potential to re-engineer MC's prominent Gaussian framework into a framework that is capable of handling discrete based distributions. More specifically, to design a model-based clustering algorithm that was built using a mixture of negative binomial distributions.

Other limitations of this study include that only a subset of data transformations and clustering methods are assessed. It is highly likely that there is not one best data transformation or clustering method for all scenarios. Additionally, this study is limited in that NB MLE parameters were only obtained from one type of cancer and histology type, ovarian cancer of the Serous histology, which many slightly bias the generalizability of the results across clustering methods for all types of genes or samples from different tissue or cancer types. Also, our study does not consider if selected features were linked or unlinked to producing inherit clusters. Future work to extend this study include: examining different effect sizes to apply to parameters of the NB distribution, varying N and G, normalizing the RNA-Seq data so that they are more standardized, and implementing any new clustering methods that have been adapted for count-type data. Furthermore, a natural extension of this study would be to look at clustering performance at the isoform level as it has the potential to further advance the knowledge behind different cancer signatures.

In conclusion, we found that RNA-Seq data requires caution when conducting clustering analyses. This is supported by our efforts to improve the performance of clustering methods through data transformations and common methods used to determine the number of clusters in a dataset. Our results suggest that a model-based clustering (MC) approach may be the best starting place for exploratory clustering analysis of RNA-Seq data types when the number of clusters is backed by prior knowledge. However, if no information is known about the number of clusters, one may want to investigate using K-Means clustering.

CHAPTER 4

Ethical Considerations for Precision Medicine— A Survey Protocol to Investigate Patient's Opinion Towards Genomic Sequencing

Janelle R. Noel-MacDonnell, Byron J. Gajewski, and Brooke L. Fridley

To be published in BMJ Quality and Safety

4.1 Abstract

The technologic advancements, specifically in high-throughput sequencing, that have recently occurred in the field of genomics have changed the way in which providers approach making treatment decisions. Phenotypes and clinical information are no longer are the primary determinant in the treatment decision process. By taking phenotypic and genotypic information together, more tailored treatments can be precisely selected. This approach to the treatment of cancer is becoming increasingly popular as in many cases the personalized treatments exhibit higher efficacy and provide patients with better quality of life. However, there is much ethical controversy that surrounds genomic sequencing in terms of patient privacy and security of the large amounts of data collected. Many survey studies have been completed to assess cancer patients' attitudes and perspective towards genomic sequencing and the ethics involved in the topic. Though, only few have been targeted solely towards cancer patients. As cancer patients are the largest population that has the potential to be affected by personalized medicine through genomic sequencing, it is critical that researchers and providers alike understand their opinions. We have developed a protocol to conduct a survey-based pilot study within the local University of Kansas network. We propose to conduct 8 one-on-one interviews and a focus group with 10 participants to aid in revision of a survey prototype. Upon revision of the pilot survey prototype, we hope to recruit approximately 32 participants, or more, who will attend a Saturday morning survey session. Participants will be given an approximately 22-item survey to complete which contains questions regarding their knowledge and opinion of genetic testing and its applications. Subsequently within the survey, participants will be asked demographic questions and be provided with a brief, educational overview of cancer and genomic sequencing. Results from this pilot study can later be used in defining and implementing a larger-scale, potentially national study.

Keywords: Precision Medicine, Genetic Sequencing, Medical Ethics, Sporadic Cancer, Hereditary Cancer, Patient Opinion Survey

4.2 Introduction

The big “C”, as cancer is often referred to, has become one of the hallmark diseases that has impacted nearly every individual in some way or another—a family member, oneself, or otherwise. There are likely many reasons this is the case; some of those reasons include risk factors such as tobacco, obesity, alcohol, infectious agents; as well as, potential environmental exposures (American Cancer Society, 2016). Although, potentially the number one reason for its high impact within the population is the sheer number of people have cancer or who were previously diagnosed with cancer and are in remission or disease free.

According to the World Health Organization (WHO), cancer has been recorded as one of the leading causes of death worldwide (World Health Organization, 2014). The trends of incidence and prevalence of cancer among the top three most populated countries in the world are very interesting. In China, the most populated country in the world, cancer is listed as the leading cause of death with an estimate of ~3 million cancer deaths having occurred in 2015 (Chen et al., 2016). However, in India, cancer does not even rank within the top 10 causes of death (Centers for Disease Control and Prevention[1], 2015). The estimated cancer mortality in India reported from GLOBOCAN 2012 was only 683,000 individuals (Ferlay et al., 2013). Continuing to the third highest populated country in the world, the United States, cancer again makes its way onto the list of leading causes of death. Several nationally recognized reports for the top leading causes of death in the United States rank cancer as the second most frequent causes of death (Centers for Disease Control and Prevention[2], 2015).

In 2013, in the United States there were 14,140,254 people living with cancer according to the latest published prevalence numbers (National Cancer Institute[1], 2016). While the national surveillance programs have not updated the prevalence numbers for 2016, they do provide estimates for the estimated number of new cases and deaths per 100,000 people. The estimated incidence, or new, cases of cancer and cancer deaths for 2016 are 448.7 per 100,000 and 168.5 per 100,000 people, respectively regardless of sex or type of cancer (National Cancer Institute[1], 2016). This translates into ~1,604,000 cancer incidence cases and ~617,000 cancer deaths (Ferlay et al., 2013). Furthermore, the estimated number of people in the United States that will have cancer or previously had cancer is approximately 19 million by the year 2024 (National Cancer Institute[2], 2016). The aforementioned collection of people have cancer types that are numerous and vast, and vary in terms of stage, grade, and histology. As there are so many cases associated with cancer, the costs that are associated with treatment of all of those patients is very high. The National Cancer Institute (NCI) estimated expenditure for cancer care in the United States increase by approximately $31 billion from 2010 to 2020 (National Cancer Institute[2], 2016). The estimated costs of cancer in the United States in 2010 was $124.57 billion (Mariotto et al., 2011). In addition to the dollar amount associated with cancer care, are both the difficult to quantify emotional and physical burden of the patient. Thus, there is a great need to research cancer from all aspects from cancer care treatment to performing good research, and to ensuring that patients, providers, and researchers have the best possible information available to them at any given cross section of time.

To promote collective efforts within the United States to reduce and prevent threats to the health in the general public, the government publishes an agenda of national topics and objectives for critical health related issues to be addressed (Office of Disease Prevention and Health Promotion, 2016). The overall goal of this agenda is to "attain high-quality, longer lives free of preventable disease, disability, injury, and premature death; achieve health equity, eliminate disparities, and improve the health of all groups; create social and physical environments that promote good health for all; and promote quality of life, healthy development, and healthy behaviors across all life stages" (Centers for Disease Control and Prevention: Division for Heart Disease and Stroke Prevention, 2014). The most recent agenda, *Healthy People 2020*, includes both "cancer" and "genomics" in its list of topics to be addressed (Office of Disease Prevention and Health Promotion, 2016). Specifically, for cancer the goal is to "reduce the number of new cancer cases; as well as, the illness, disability, and death caused by cancer" (Office of Disease Prevention and Health Promotion, 2016). Additionally, for genomics the goal is to "improve health and prevent harm through valid and useful genomic tools in clinical and public health practices" (Office of Disease Prevention and Health Promotion, 2016). The topic of genomics was never previously included in list of topics nor did it show up in any objectives to be addressed prior to the current rendition of the agenda. We believe this is directly related to the recent advancements of genomic technologies and understanding of the vast realm of genomics. While the two agenda topics items mentioned above can stand alone, the two have in recent years become fittingly married together as the field of cancer genomics. Though to better understand this marriage, we first look at each agenda item separately.

DNA is not a stagnant nor unchanging molecule. DNA is a forever changing molecule, in a person's lifetime their DNA can undergo a multitude of changes from repairs (i.e., DNA repair) to mutations (i.e., single nucleotide polymorphisms (SNP), insertions, deletions, etc.) and structural changes (i.e., chromosome duplication or altered chromosome structure) which results in variations of genetic material (Aguilera and Garcia-Muse, 2013). These variations taken together are often referred to as the hallmarks of cancer (Figure IV-1). Nearly all cancers found in humans can be categorized by these hallmarks (Negrini et al., 2010). The list of hallmarks proposed by Negrini et al. (2010) provide a fairly comprehensive overview of possible rationales which leads to the development of cancer.

Cancer is a very complex disease which utilizes a great amount of research hours and personnel. There are three groups which classify a majority of the cancers: familial cancers, hereditary cancers, and sporadic cancers. Familial cancers are caused by multiple variants that are often difficult to define. Those variants range from a unique combination of multiple genes, family history of cancer, and environmental stimuli (Coriell Personalized Medicine Collaborative, 2016, Sijmons, 2010). While the relationships between the hallmarks are not specifically depicted in Figure IV-I, we would expect there to be multiple hallmarks that had directional arrows pointing to other hallmarks as high variation of causes is characteristic of familial cancer. Families with familial cancer exhibit trends in cancer type greater than that expected by chance alone due to genetic syndromes and mutations in known cancer genes (Coriell Personalized Medicine Collaborative, 2016, National Cancer Institute[3], 2016).

Sporadic cancer are slightly easier to define as there are not as many components that are linked to its cause. Although, not all causes of sporadic cancers can be determined. Sporadic cancers (i.e., non-hereditary cancers) are classified by the lack of family history of given cancer and the absence of the individual having any type of genetic risk factor through an inherited gene mutation (National Cancer Institute[6], 2016). The genetic alterations that are found in these types of cancers are called somatic mutations (i.e., mutations which are observed in the tumor's genetic material but not in a person's inherited genetic material) and occur after conception (National Cancer Institute[7], 2016). The pathway in which these types of cancers develop is through an activation of the cell grow signaling which leads to DNA damage and DNA replication stress (Figure IV-1: Panel C) (Luo et al., 2009, Negrini et al., 2010). This causes downstream problems with genomic instability and reproduced cells evading cell death and senescence (Figure IV-1: Panel C) (Luo et al., 2009, Negrini et al., 2010). The initial activation of the cell grow signaling can be the result of a lifetime of genetic damage. Sporadic cancers account for approximately 60% of all cancers and tend to occur later in life (Coriell Personalized Medicine Collaborative, 2016, Anderson, 1992).

Hereditary cancers are different than sporadic cancers in that they are linked to genetic inheritance from parents (i.e., genomic instability, mutations in key cancer causing genes) (Figure IV-I: Panels A and B). Rather, hereditary cancers are associated with a mutation in determined susceptible germline gene that cause an individual to have an increased risk to develop cancer (Coriell Personalized Medicine Collaborative, 2016). If an individual develops cancer that has been determined to be linked to such mutation of a gene, likely through genetic sequencing, is termed a hereditary cancer. Cases of hereditary cancers are generally found in younger individuals (Anderson, 1992). Unfortunately, some of the susceptible germline gene mutations have been tracked and shown to have a lifetime risk up to an 85% chance of developing cancer (Coriell Personalized Medicine Collaborative, 2016). The most published hereditary cancers occur from mutations in BRCA1 and/or BRAC2 resulting in breast and ovarian cancer (National Cancer Institute[4], 2015, Walsh et al., 2010, King et al., 2003). A summarized list of mutated genes and the related cancer types of the identified 50 hereditary cancer syndromes can be found on the National Cancer Institute’s website (National Cancer Institute[5], 2013). While the actual percentage of hereditary cancers is small, the syndromes which arise from them is consistently growing (Strahm and Malkin, 2006). It should be noted that it is not always the case that if an individual has a germline mutation that if they get cancer that it will be a direct result of the germline mutation. A different mutation may be linked to the cause of the cancer. Moreover, on occasion genes that are inherited and mutated through the germline, genes which predispose an individual to cancer, also play a role in pathogenesis in mutated counterparts in sporadic cancers (Strahm and Malkin, 2006).

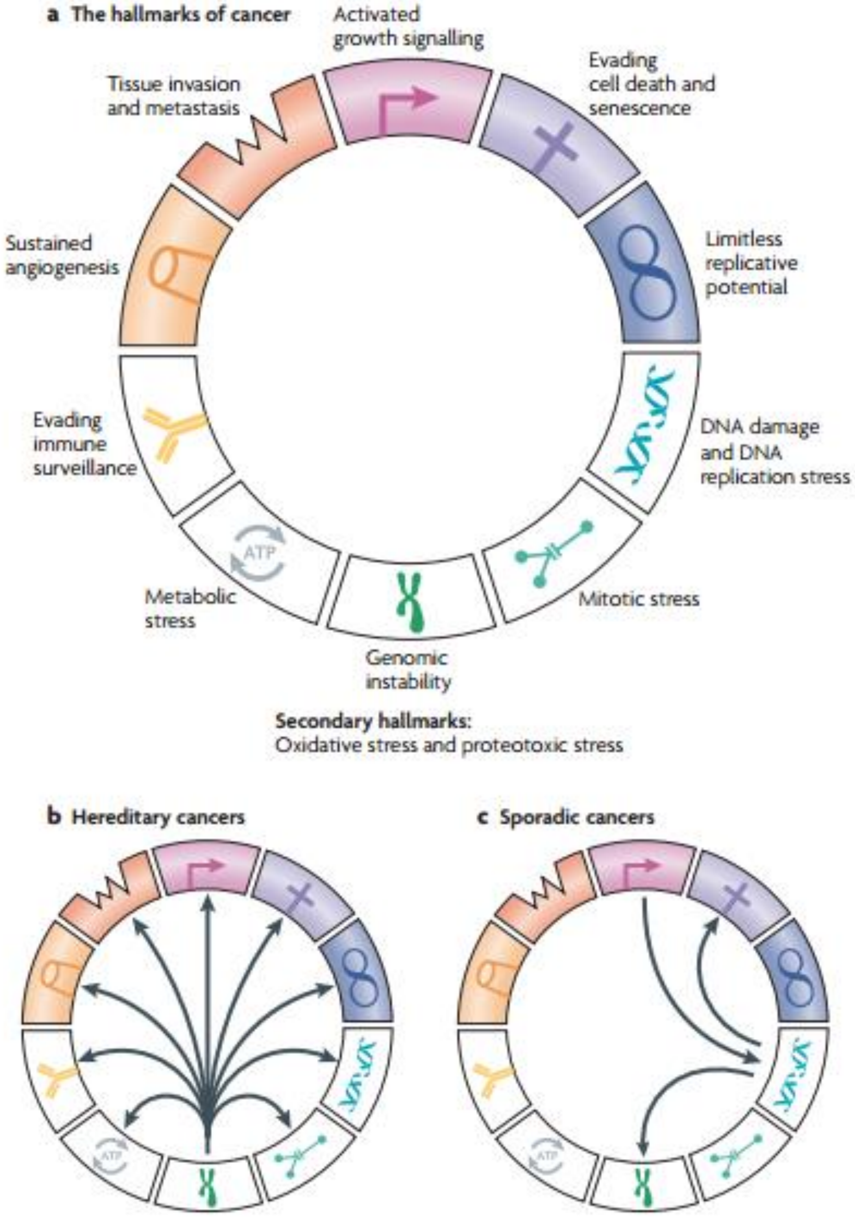

a The hallmarks of cancer
Activated growth signalling
Evading cell death and senescence
Limitless replicative potential
DNA damage and DNA replication stress
Mitotic stress
Genomic instability
Metabolic stress
ATP
Evading immune surveillance
Sustained angiogenesis
Tissue invasion and metastasis
Secondary hallmarks:
Oxidative stress and proteotoxic stress
b Hereditary cancers
c Sporadic cancers

Figure IV-1. Overview of the hallmarks of cancer (Luo et al., 2009, Negrini et al., 2010)**.** Found within the sections of the circle are many of the identified hallmarks that play a role in the development of cancer (Panel A). Those hallmarks that lack shaded backgrounds are those hallmarks that were added in 2010. The arrows in the inner circle of panel B) and Panel C) depict the relationships hallmarks have with each other in hereditary and sporadic cancers, respectively (Luo et al., 2009, Negrini et al., 2010).

Cancer is fueled by changes in an individual's genetic material that occur irregularly or changes that are prompted by the environment (i.e., radiation exposure, smoking, sun exposure, etc.). Hence, it is fitting that a large component to understanding and treating cancer would be to understand the variations present in genetic material. The field of study which studies genes, gene function, and their technologies is genomics (World Health Organization[2], 2002). To gain insight into genetic material, genetic sequencing, can be performed which determines the combination of the nucleic acids in DNA and/or RNA. Together, these combinations make up the genetic sequence, which provides genes that are present and information about the structure and function of those genes (National Human Genome Research Institute[2], 2015). However, it should be noted that genetic testing is not exclusively limited to DNA and RNA, analyses can be performed on chromosomes, proteins, and metabolites. Also, there are also three different types of DNA sequencing that are utilized—whole genome and targeted (i.e., sequencing of specific areas and exome).

Sequencing technologies have improved greatly over the years. Prior to 2004, microarray technology was used to determine genetic sequence, but it was drastically limited in the amount of sequencing information that could be output. Recent advancements in technology have led to the popularity of using Next-Generation Sequencing (NGS) for genetic sequencing. The advent of NGS technologies has not been around for that long of a time frame, but it has already enabled researchers to study genetic material at a level that surpassed early expectations (Van Dijk et al., 2014). It wasn't until 2004 that NGS became commercially available (Mardis, 2008). Sequencing technology has improved greatly over the years in terms of increased speed, efficiency, lower costs, and higher accuracy to provide "exquisite sensitivity and resolution" (Walsh et al., 2010, Mardis and Wilson, 2009). Currently, cost for whole genome sequencing

(WGS) is ≤ ~$1,000 which make it available for smaller labs, research centers, clinics, and population-wide (Van Dijk et al., 2014, Shen et al., 2015, Ku et al., 2013). Addressing the objective listed in Healthy People 2020 for genomics, it is fitting that in the near future providers and patients would be encouraged to utilize the application of genomics in the treatment of common diseases.

The fields of genomics and cancer are projected to weave together more so in coming years than they are currently. Today our understanding of the molecular nature of cancer is due largely in part to next-generation sequencing (NGS) techniques (Yang et al., 2012). Additionally, some of the largest advancements in the field of genomics have been in the area of cancer biology (Balmain et al., 2003). More specifically, genomic sequencing and molecular profiling gained popularity as their results led to better understanding of the complexities of cancer (Balmain et al., 2003). As put by Catenacci et al. (2014), we are at a “critical point” in modern-day medicine where cancer treatment and care decisions are being driven by the plethora of data produced by NGS. With the costs of sequencing becoming more reasonable, more providers are regularly using it for purposes of cancer diagnostics in both a discovery and confirmatory context (Shen et al., 2015, Ku et al., 2013). In familial cancer, researchers are able to conduct WGS within the family to determine if offspring have germline mutations of genes which predisposed them to cancer (Shen et al., 2015, Ku et al., 2013). Furthermore, sequencing can determine single changes in a nucleic acid base of a portion of DNA which can disrupt proteins responsible for normal cell function (The Cancer Genome Atlas, 2010). However, the primary goal of genomic sequencing is to identify those somatic or germline mutations that might be candidates for targeted drug therapies (Gingras et al., 2016).

This is a form of precision medicine. As its name suggests, precision medicine, or also synonymously termed personalized or individualized medicine, is the tailoring of disease treatments and/or interventions to the unique characteristics, both genotypic and phenotypic, that an individual has (Ciardiello et al., 2014). Next-Generation Sequencing is becoming more commonly used to determine "best" drug therapy through a data-informed decision when it comes to placing a patient on a treatment that they will likely benefit from based upon given molecular biomarkers (Yang et al., 2012, Gingras et al., 2016). The goals of personalized medicine are to "increase the probability of efficacy and/or decreasing the probability of serious adverse events"(Vicini et al., 2016). Cancer is at the "frontline" of personalized medicine as additional considerations are now being given to molecular biomarkers and not solely phenotypes when developing eligibility criteria in clinical trials (Ciardiello et al., 2014). Many clinical therapies in development are largely associated with defined biomarkers (Ciardiello et al., 2014). As more cancer therapeutics are being developed specifically for certain biomarkers, it is essential that researchers document their findings to allow other researchers to further advance personalized medicine in the cancer patient population. In the last 10 years, few databases (i.e., The Genomics of Drug Sensitivity in Cancer (GDSC) database, Mutations and Drugs Portal (MDP) database, and canSAR) have been created to keep track of which molecular features influence a drug response in cancer cells through a combination of cell line drug sensitivity data, genomic data, and data from the analyses of genomic features (Yang et al., 2012, Taccioli et al., 2015, Bulusu et al., 2014).

While great progress has been made in the area of precision medicine for cancer, it has not been without challenges—scientific challenges do to cancer's complexity, and otherwise due to ethical considerations. There is much controversy that surrounds genetic testing. At the center of the controversy is that genomic sequencing has the ability to identify an individual, reveal if the individual is at an increased risk of developing certain types of cancer, expose other diseases that the individual may have which leads to valid concerns of patient privacy, discrimination, and security of the large amounts of data (Ciardiello et al., 2014). More specifically, whole genome and exome sequencing produces large-scale data of which has the potential to have both medical and social influence as levels of result's uncertainty can still be present (Fiore and Goodman, 2015). While much genetic data has been released and stored in publicly available databases without any direct link to individuals, there are still some medical-Sequencing databases that contain patient identifying variables such as: demographic information, clinical information, etc. (Foster and Sharp, 2006). That being said, data that is obtained from sequencing should be stored using the same degree of security that is used to protect other entities that house personal health information (i.e., electronic medical records and some databases).

Fortunately and unfortunately with genetic sequencing there is no way to tune the results. Rather, intermittently sequencing reveals a medical finding that is different from that which a researcher was looking for—referred to as "an incidental finding". Incidental findings can expose information about paternity, risks of certain diseases or syndromes, etc. which may have drastic implications mental health, well-being, or even how an individual proceeds care. The question that often arises with these incidental findings is, "Do you tell your patient of them?" This has created much controversy in recent years with genomic sequencing becoming more readily used in diagnosing patients. In general, there are two sides to this argument. Side one, should providers only present results and findings for the current medical issue that a patient has; or side two, should a provider lay out any findings from medical tests. Previously proposed guidelines in general seem to advise patient providers to communicate such findings when they are found, especially if results are analytically valid and clinically significant (Fabsitz et al., 2010, Wolf et al., 2012, Zawati and Knoppers, 2012).

In general, current practice for reporting and communicating incidental findings to patients is dependent on the following subjective criteria: "variant frequency, the potential for medical intervention to mitigate disease, the strength of association between specific gene abnormalities and the condition, and penetrance (i.e., proportion of individuals carrying a particular variant of a gene) of those genes" (McGuire et al., 2013, Green et al., 2013, Middlelton et al., 2016). Following this recommendation, only approximately 1% of patients would have a qualifying incidental finding (Green et al., 2013). However, much debate is still had about those incidental findings that are of uncertain significance (Hofman, 2016). When patients were asked whether or not they would want to know about any incidental findings, the consensus was to allow the individual patient to decide based upon their moral, political, and religious values (Townsend et al., 2012, Freedman, 1987, Foster and Sharp, 2006).

Several other ethical concerns that arise from genetic testing are: adequacy of patient consent; familial genetic testing; additional germline testing for individuals with early onset sporadic cancers (i.e., if patient was diagnosed prior to age 55); targeted vs. whole genome sequencing; unnecessary treatment due to false positive results of testing; stereotyping/stigmatization; disparities in access to additional testing, counseling, and the way in which insurance companies reimburse; and additional privacy and discriminatory concerns (Fiore and Goodman, 2015, Foster and Sharp, 2006). Although, governmental bodies are helping address some of the ethical concerns that have developed surround genetic sequencing. In 2008, the Genetic Non-Discrimination Act (GINA) was passed which forbids employers and health insurance companies to discriminate against individuals based on genetic test or family history (Fiore and Goodman, 2015). While GINA was a step in the right direction, there is still work to be done to protect patients and address ethics associated with genomic sequencing. It is highly likely that as the general community becomes more educated about genomic sequencing as a whole, acceptance of its implementation will increase.

4.3 Motivation and Objectives for the Survey

Understanding cancer patients’ opinions regarding precision medicine in terms of genomic sequencing is critical, especially as there are many ethical concerns that arise. By obtaining patients’ opinions, providers can potentially deliver better, more efficacious treatment to their patients. Additionally, it provides insight to surveillance groups that develop guidelines which address approaches to care of patients. Current literature is saturated with surveys that have been given to providers regarding their opinion to the use of genomic sequencing in patients with varying types of cancer. This is likely due to the fact that providers have acquired some education regarding genomic testing. Other surveys that seek individuals’ opinions towards genomic sequencing in cancer care utilize participants that are from some type of health care profession or science researcher area, or participants that from the general public (Middlelton et al., 2016, Henneman et al., 2013). The few published studies that use patients’ opinion towards genomic sequencing, are not enough to fully gain a consensus of the general cancer patient population. Hence, we propose a protocol for implementing a local patient opinion survey targeting cancer patients, which could later be evaluated for content validity, reliability, and duplicity. Upon this evaluation, the survey could be revised and executed in a larger population at a later date. Prior to the development of the survey, we plan to conduct a series of one-on-one interviews and a host a focus group with various types of cancer patients. The primary objective of the survey would be to determine basic cancer patients' opinions towards genomic sequencing in a pilot study with the hopes conduct a larger-scale study. Additionally, we would like to gain some insight to patient opinions regarding targeted drug therapies, and briefly germline sequencing.

4.4 Proposed Pilot Study Design

In developing a protocol to implement our survey, there are many topics that needed to be addressed; such as, what are the concerns that patients have, who the study will be given to, how many participants will take the survey, participant recruitment, questionnaire development, and potential statistical analyses. For this pilot survey study, we plan to use a convenience sample of cancer patients being treated within the University of Kansas network of providers. Using a small sample of varied providers, we will aim to schedule up to 8 one-on-one interviews or a focus group which contains up to 10 people as recommended. Once data is summarized from those participants in either the one-on-one interviews, we will recruit participants to take our pilot survey study. Those participants that agree to participate in a morning survey session group will be given a survey questionnaire to complete. After completion of the survey questionnaire, participants will also be given the opportunity to voice opinions and discuss with other survey participants about their questions and concerns of having genomic testing completed, and in turn completed to determine treatment options. In the following section, we describe each of these topics.

4.4.1 Pilot Study Approval

Prior to moving forward with the conduction of the one-on-one interviews, focus group, and future pilot survey study, approval needs to be attained from the Institutional Review Board (IRB). This protocol, interview/focus group questions, supplementary documents including the consent form, recruitment “Save the Date”, and potential survey questionnaire will be submitted to the electronic IRB system to be reviewed by the IRB committee. We assume that since this study would involve “no more than minimal risk” that it would receive an expedited review by the committee. Once approval is given from the IRB to proceed with the study, we would begin reaching out to providers to help recruit participants for our study. This pilot survey study will be conducted in compliance with this protocol and Good Clinical Practice (GCP) guidelines. Any changes to this protocol or study documents will be submitted to the IRB as an amendment for review and approval. All copies of completed consent forms and survey questionnaires will be securely retained for five years after the completion of the study the Principal Investigator (PI). Furthermore, the master electronic data will be encrypted with password protection. Only necessary members of the research team will be given data files. No patient identifying information will be collected on the survey. Lastly, all survey administrators and research team members will undergo brief training and detailed instruction regarding to insure that everyone is competent in their respective roles.

4.4.2 One-on-one Interviews and Focus Group

In order to develop a survey that accurately obtains patients’ opinions towards the previously mentioned topics, we propose to conduct multiple one-on-one interviews and/or a small focus group to gain deep insight into areas where cancer patients have concerns in terms of the ethics behind precision medicine and genetic sequencing. The use of one-on-one interviews and a focus group, we hope, will provide a comfortable environment for participants to thoughtfully respond to our questions or ask for clarification; and provide awareness to areas or domains that we may need to consider to be addressed in our questionnaire. Recruitment of participants for either the one-on-one interviews or focus groups will be recruited from other principal investigators that we have worked with in the past. These providers will discuss with their patients about the opportunity to be a part of our study. Depending on the patient's interest level, the provider will attempt to schedule a 30 minute interview with a member from our research team at their next follow-up appointment or they will be given a "Save the Date" (Appendix E) regarding when the focus group session would be help. The exact date of the focus group will be determined by research team upon approval from the IRB.

4.4.3 Sample Frame and Sample Size Justification

Participants for these one-on-one interviews, focus group, and future survey will be obtained from a convenience sample made up from cancer patients within the University of Kansas network. Providers within the University of Kansas network will be asked to aid in our recruitment effort of patients by discussing our on-going study and handing their eligible patients one of the "Save the Date" cards to attend our focus group or survey session depending on where we are at in our study timeline (Appendix E). In the early stages of our study, providers will also be responsible for scheduling one-on-one interviews at an agreeing patient's next follow-up visit with a designated research team member interviewer. As we are targeting recruitment of cancer patients, we plan to reach out to previous PIs that our research team has worked with for filling out one-on-one interviews and focus groups. For our pilot survey portion of the study, we plan to recruit from providers from The University of Kansas Cancer Center (KUCC), the Medical Oncology Division in the Department of Internal Medicine at the University of Kansas Medical

Center (KUMC), and the Hematology/Oncology Division at KU's Westwood campus. We assume that we will have adequate success in recruiting from previous PIs that we have worked with to fill our one-on-one interviews.

Our justification for recruiting providers and in turn participants from three different locations is to reach the widest range of cancers amongst pilot survey participants. Within each of the three recruitment locations that we will utilize to recruit patients from, there are approximately 80 listed providers with credentials of M.D. (i.e., medical doctor), O.D. (i.e., osteopathic doctor), or P.A. (i.e. physician assistant). Reaching out to these providers to see if they would be willing to assist us in completing this pilot survey study will be fairly feasible as e-mail, phone number, and/or office location are available through either the KUMC website (www.kumc.edu), KUMC email directory through Microsoft Outlook, or through "Find a Doctor" on the health grades website (www.healthgrades.com). However, we believe that not all internet-listed providers see patients in the clinic or if the websites includes the most up-to-date list of providers which would reduce our sample of providers, say to 40 providers if we are being conservative. Furthermore, due to the busy schedule of providers, we assume that only half of those providers that see patients in the clinic will be able and willing to take on the additional responsibility in helping us recruit participants. Assuming during a one week period a provider sees at least two patients that have been diagnosed within the last year, our potential sample size of individuals over the course of four weeks would be 160 individuals. Our hopes are that up to 20% of those patients whom received the information from their provider and "Save the Date" card will actually attend the scheduled pilot survey session (Appendix E). This equates to ~32 participants, or more. Ideally, both the focus group and pilot survey session would be held on a Saturday morning around 10:00 a.m. at a University of Kansas facility that had easy-assessable parking near the facility. The exact date of the survey session would be determined by research team following the study approval from the IRB.

4.4.4 Eligibility Criteria and Compensation

For a patient to be eligible to participate in any portion of our study, they must meet two criteria. First, the patient must have been diagnosed with any stage of cancer within the last year from the date of the study event they are attending. Secondly, the patient must be at least 20 years of age. Any participant in either a one-on-one interview, the focus group, or in the pilot study will receive $50 for their time.

4.4.5 Questionnaire Development for Interviews, Focus Group, and Pilot Study

The study questionnaires for this research study includes questions asked to the patient regarding the following: demographics, understanding of genomics and genetic testing, and opinions related genomic sequencing. Questions used in the initial one-on-one interviews and in the focus group will follow closely to those found in Appendix F. However, the both the research team member interviewer and the focus group moderator will be trained to inquire further to facilitate more discussion on topics that participants fill strongly about. Specifically the survey questionnaire will be divided into three sections to better facilitate the flow of the survey. Section one contains only basic demographic questions that address the background of the individual filling out the survey. Section two will briefly educate the individual about genetics, cancer, and genetic testing. Lastly, section three will contain questions to obtain patient's opinion which will be broken into appropriate domains in our pilot survey study (i.e., education, precision medicine ethics, etc.). Responses to each of the survey questions were designed to be highly inclusive and provide insight. To minimize non-response to questions, most questions are given the response option "Prefer Not to Answer". The proposed prototype of the survey questionnaire can be found in (Appendix F). Revisions to this prototype would be made to reflect the findings from the one-on-one interview and focus group to better address our objectives. Additional questions that were asked by either the interviewer or moderator will be considered by research team to determine best way to include them in the pilot survey study.

4.4.6 Data Collection

4.4.6.1 One-on-one Interview and Focus Group Study Portion

Participants that take part in either a one-on-one interview or the focus group will be asked to sign a consent form prior to being asked any information pertaining to the study. In the one-on-one interview the research team member interviewer will ask the participant to fill our questions similar to those found in Section one of the survey prototype found in Appendix F. Once these responses were recorded, interviewer would start a digital voice recorder to capture all dialect. Interviewer would take notes as interview took place and would use digital voice playback to fill in any missed information after completion of the interview. Collection of focus group data would be collected very similarly to that of the one-on-one interviews. The only difference that would occur is that the prototype pilot survey in its entirety would be given to all participants. Once participants complete the prototype pilot survey, moderator will facilitate open session for questions, comments, and clarification. A digital voice recording of this open session will be taken and turned into a transcript for research team to review. Relevant responses and notes as determined by research team member would be summarized for review. Changes to the pilot survey prototype will then be made reflect participants, moderator, and interview's feedback and comments.

4.4.6.2 Pilot Survey Study Portion

Participants that show up to the survey session will also be asked to sign a consent form prior to receiving the survey questionnaire to complete. The entire survey session will be led by two independent survey administrators each of whom were educated about the survey topics and instructed on how to run the survey session. We propose to collect data from a 22-item, paper-based survey consisting of primarily check box responses and few free text questions. This 22-item questionnaire would be revised to reflect those findings from the one-on-one interviews and focus groups. Additionally, study participants are given a space at the end of the survey to provide feedback and comments about the survey and/or session. Study participants will be given a paper copy of the survey to record their responses and be directed to turn in their completed survey to one of the survey session administrators.

Data from paper survey are then entered manually into a digital database (i.e., Microsoft Excel) by two research team members. Data will then be matched to determine agreement of entered responses to aid with Quality Control. Any disagreeing results between the two datasets entered by the research team members will be adjusted by a different research team member through review of paper survey responses. Each of the survey questionnaires will be assigned a questionnaire ID which will allow for rectification of disagree responses. Free text fields, comments, and feedback will be summarized in a list format. Once Quality Control is conducted by the research team members, data will be sent to statisticians for analysis.

4.4.7 Data Analysis

After the data are given to the statistician, descriptive analyses will be completed for all survey questions. For all of the check box questions, frequency and percentage of response will be calculated. However, for those free text questions, inclusive lists of all responses will be developed. If it is applicable for frequencies and percentages to be reported, the statistician would create such list. Shell tables for these analyses can be found in Appendix G. It would be left up to the statistician's digression for what type of statistical software (i.e., R Statistical Software or SAS) they wanted to use to complete the analysis. Data will be examined for ceiling and/or flooring effects for ordered responses to see if responses would need to be adjusted in the larger study. Additionally, missing values would be tracked for each question. Depending on the percentage of missing values, data imputation may be considered. Furthermore, for Questions 14, 20, and 21 mean response would be calculated. Questions 14 and 20 will be assigned the following values for given responses: 0 = "Unknown or Prefer Not to Answer", 1 = "Disagree", 2 = "Neither Disagree or Agree", and 3 = "Agree". Similarly, for Question 21, 0 = "Unknown or Prefer Not to Answer", 1 = "I would decline additional genetic testing", 2 = "I would accept additional genetic testing".

Under this current pilot study, no subsequent statistical analysis would be complete. Although, in the larger-scale study it might be useful to compare differences among responses based on age group. For instance look at responses for patients <50 years of age vs. 50 years or older through a chi-square test. With a larger-scale study, more elaborate statistical analyses can be conducted as larger sample size would although for many statistical methods to be adequately powered.

4.5 Discussion

The need for additional survey studies that target cancer patients' opinions towards their care is evident from our review of the literature. As genomic sequencing is become more regularly used in diagnosing and decisions about treatment options, it makes sense to survey the opinions of cancer patients regarding the topic. To expand the current literature, we developed a protocol for a patients' opinion study including one-on-one interviews, a focus group, and a pilot survey study which would be implemented within the University of Kansas network. The objectives of the study would be to patients' opinions towards genomic sequencing; and to gain some insight to their opinions in using genomic sequencing to determine targeted drug therapy options. Our proposed protocol for implementing our pilot survey study is fairly basic in comparison to many pilot survey studies. However, we believed that it is necessary to take this step-wise approach in its development to have a meaningful and influential survey study. Our sampling population is already considered to be heavily burdened and could highly benefit from such a study. Our hope would that information from this pilot study would drive conduction of a larger-scale, potentially nationwide, grant-funded survey.

When conducting survey studies, there are many logistical issues that must be considered for the study to be successful. If the proposed pilot survey study were to be approved by the IRB, we believe that we would be able to successfully reach our targeted sample size of ~32 participants in our pilot survey study. Once the one-on-one interviews and focus group portions of the study are completed, we have confidence in the patient providers will be willing to take part in providing their patients will information about our study as the University of Kansas has a mission statement prioritizes research productivity. Additionally, we trust that our developed study material (i.e., "Save the Dated", initial questions, and survey questionnaire prototype) will be easily understood by study participants. All study material wording has been designed at an appropriate reading level (approximately at an eighth grade reading level). Concurrently, the focus group and survey session would be scheduled to take place outside of the normal work week for most individuals, and will only require them to spend up to an hour or hour and a half at the determined KU facility location. Furthermore, the burden of the one-on-one interview to the participant is also reduced by scheduling it together with the patient's next follow-up.

Although, as with any research study, our proposed pilot survey study contains limitations. Our study is limited in that we were unable to find any published information of any type regarding a patient survey implemented at KUMC. We are also limited in that those participants that complete the study are from a convenience sample and likely are interested in voicing their opinions. Hence, the results may be lacking opinions of those that do not participate can lead to bias in our results. Additionally, our questionnaire prototype is not a validated. Though, we are optimistic that the dialect between patients and study portion administrators will provide valuable feedback and comments to improve our survey to be conducted on a national level. Despite these limitations, this pilot survey study has the potential to provide valuable information regarding patients' opinions towards genomic sequencing while keeping the patient burden very low.

CHAPTER 5

Discussion

The RNA- Sequencing (RNA-Seq), a Next-Generation Sequencing (NGS) technology, has greatly changed the landscape of the field of genomics with its accuracy, breadth of data, and sensitivity verses previously used technologies such as microarrays. Throughout this dissertation, many of the RNA-Seq analyses are compared to similar types of analyses that were completed in microarrays. This is done as many researchers and scientists alike seek to translate analyses methods once used in microarrays to be used in RNA-Seq. However, in doing so, there are many considerations that need to be given and challenges to be addressed. Unlike microarray data which is continuous in nature and most often normally distributed, RNA-Seq data consists of discrete data, or count data. The properties mentioned for microarray data make its analyses more straightforward as numerous method's assumptions can be readily met.

While the literature is saturated with studies that compare the biological differences between microarray and RNA-Seq technologies, there are only few that extend the differences between the two technologies in terms of the statistical analyses. In this dissertation, we sought to expand the knowledge of some of the different statistical challenges that arise from RNA-Seq data; as well as, address some of the ethical concerns involved in genomic sequencing. The principle findings from each chapter are as follows. In Chapter II's differential expression study a lack of precision amongst selection of similar genes that are differentially expressed when comparing differential expression analysis methods from our empirical analysis. Results from our simulation study, which first examined empirical Type I error rate and later empirical power, were as expected. In general, models (i.e, LM, LMM, GLM, GLMM) that were fit according to the data (i.e., if the data were paired ($\rho = 0.3$ and 0.5) or unpaired ($\rho = 0$) between measurements) had control of the empirical Type I error rate at a level of 0.05 less regardless of the distribution (i.e., Bivariate Normal, Bivariate Poisson, or Bivariate Negative Binomial) the

RNA-Seq data were simulated from. In those simulation scenarios where control of the Type I error rate was established, we observed that the empirically calculated power increase with increases in mean shift and sample size which is a typical relationship. Those scenarios with conservative Type I error rate values significantly lower than 0.05 had extremely low empirical power. Achievement of adequate power depended heavily on the sample size for scenarios where data were simulated from the Bivariate Poisson and Bivariate Negative Binomial distributions. Additionally, in Chapter III's data transformation and clustering method assessment, we found that it is highly challenging to transform data to "look" more normal. Despite all of the data transformations that were applied, no transformation equated to the exact skewness and kurtosis values found in normally distributed data. Moreover, our results suggest a model-based clustering is the most robust approach to clustering analysis when some knowledge is previously known about the number of clusters in the data. Conversely, if the number of clusters in unknown, K-Means clustering would perform best in determining most likely clusters. Lastly, in Chapter IV no specific findings were observed as the chapter focuses on the setup of a pilot survey study. However, we assume that highly valuable findings would be obtained to allow for an extension of the survey to a larger scale.

Each of the studies presented within this dissertation can be extended in multiple ways in the future. Findings in Chapter II, provokes implementing either a sandwich estimator or utilization of the method of moments to better handle the paired structure and limitation of current models. To do so it Kauerman and Carroll (2000) suggest implementation of a robust covariance matrix estimator. In the study in Chapter III, one may want to consider applying the data transformations and clustering methods to an actual dataset to see how the performance of the methods compare to those results found in our simulation study. This study also motivates a development of a new algorithm to determine the number of unknown clusters within a dataset. This new algorithm would need to consider that range of gene expression values and be tuned to handle small effect changes between clusters. Both of the studies that utilize RNA-Seq data can be extended to smaller sample sizes and compared with the current findings. This is a critical expansion of these studies as often times researchers do not have the resources to have larger sample sizes in their experiments. Another general need that is lacking from the realm of RNA-Seq research in terms of statistical methodology is the application of non-parametric approaches to analyses. A natural extension of the proposed protocol for the pilot survey study would be to actually conduct the study and begin the process of obtaining IRB approval, reach out to providers from the three mentioned departments within the University of Kansas network, and conduct the study portions.

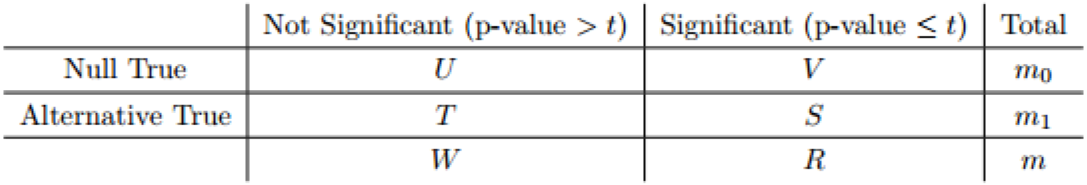
Throughout the process of completing this dissertation, many lessons have been learned. Aside from the findings of the studies in Chapters II and III, this dissertation has greatly improved my statistical programming skills, my approach to designing figures and tables, and my overall take on conducting meaningful research. Both of the studies in Chapter II and Chapter III, gave me the sense that sequencing technology is outpacing the types of statistical analyses to be conducted on its data. We are able to acquire massive amounts of complex data, but when attempting to answer research questions there is often opposition between findings which are statistically relevant and/or clinically relevant. Hence, researchers, statisticians, bioinformaticians, and other bio-related scientists are tasked to work together to address those unanswered research questions, develop new hypotheses, and overcome the challenges that are associated with the “omic” big data.

APPENDICES

Appendix A: Local False Discovery Rate (FDR) calculation details

The following calculation is used in the *qvaule* package in Bioconductor from Storey et. al. (2015) to obtain the local False Discovery Rate (lFDR).

Consider m hypotheses (i.e., $H_1, H_2, \ldots, H_m$) are conducted where each hypothesis results in a p-value, p_g, for $g = 1, \ldots m$ (i.e., p_1, p_2,…, p_m). These null hypothesis is that some g^th gene is not differentially expressed. These p-values corresponding to the tested null hypotheses are considered to be statistically significant if $p_i \leq 0.05$. Results from the testing can be placed into the 2 x 2 contingency table below,



	Not Significant (p-value $> t$)	Significant (p-value $\leq t$)	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
	W	R	m

(Storey, 2010, Benjamini and Hochberg, 1995). Here, V is the number of false positives, or rather the number of Type I errors; and R is the total count of all significant null hypothesis. Using the information from the 2 x 2 contingency table we can determine the FDR

$$FDR = \boldsymbol{E}\left[\frac{V}{R}\middle| R > 0\right]\mathbf{Pr}(R > 0)$$

(Benjamini and Hochberg, 1995). Though, others have proposed additional extensions of the above FDR. One extension is the lFDR which is used to quantify the probability of $H_g = true|p - value_g$ (Efron and Tibshirani, 2002, Efron et al., 2001). It follows as such from Liao et. al. (2004)(Liao et al., 2004):

Let

$$z_g = \begin{cases} 1 \; if\ the\ gth\ gene\ is\ differentially\ expressed \\ 0 \; if\ the\ gth\ gene\ is\ not\ differentially\ expressed \end{cases}$$

be modeled as a Bernoulli trial with probability $1 - \pi_0$, where $\pi_0 = \frac{m_o}{m_1}$. Also, let f_0 and f_1 be the density of $p_g|z_g = 0$ and $p_g|z_g = 0$, respectively. Here, $f_0 \sim Uniform$ (0,1). That said, we have $f(y) = \pi_0 + 1 - \pi_0 f_1(y)$ which is the two component mixture model that all p_g come from. Hence, the

$$lFDR(t) \equiv \Pr\left(H_g = true\middle|p_g = t\right) = \frac{\pi_0}{\pi_0 + 1 - \pi_0 f_1(t)}.$$

Appendix B: Comparison of overlapping Differential Expression (DE) genes after using estimations for False Discovery Rate (FDR) as re-evaluation criteria

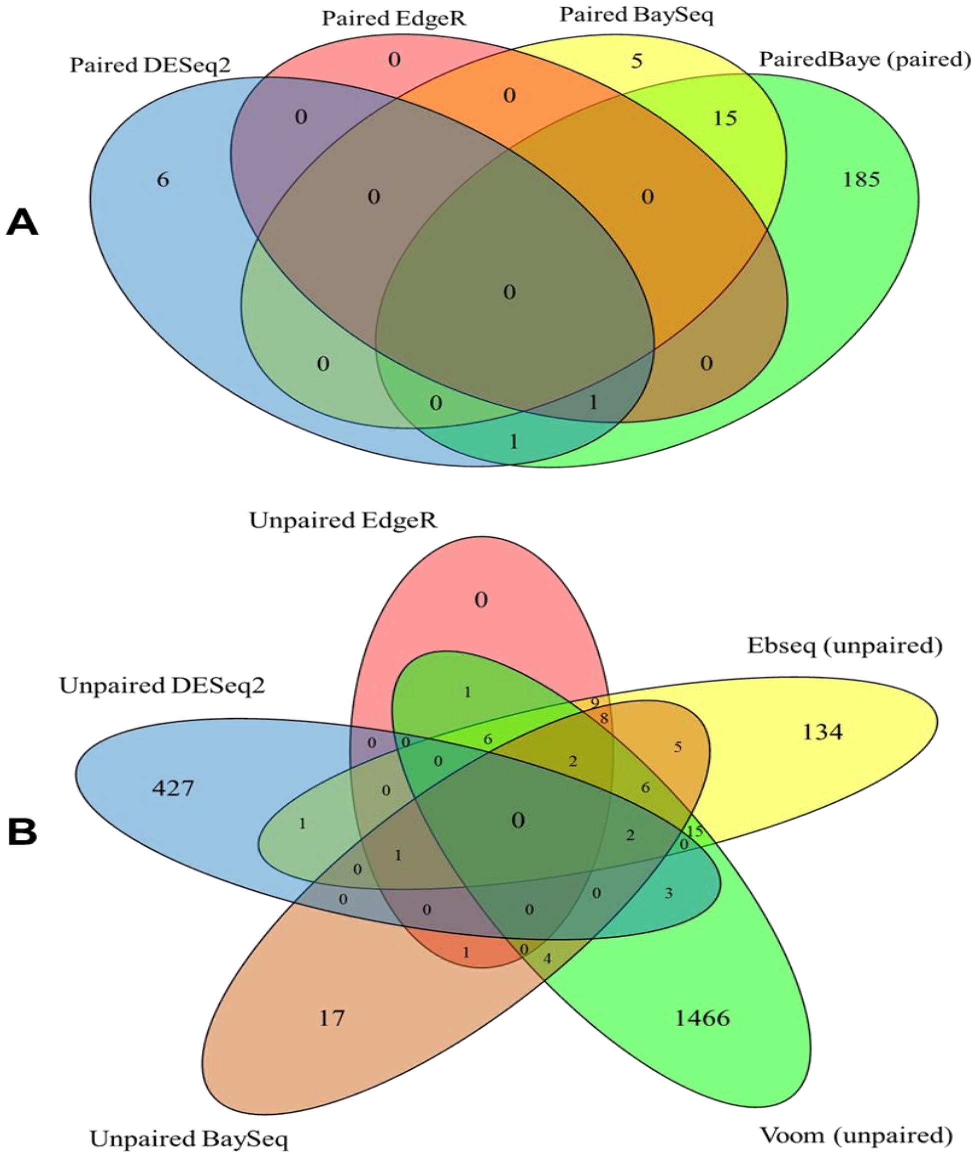


Figure II-A1. Comparison of Differentially Expressed (DE) genes found in unpaired and paired methods after re-evaluation. The Venn diagrams above contain the number of Differentially Expressed (DE) genes that were determined by each method. The overlapping portions of the Venn diagrams represent the number of DE genes selected to be the same between the compared DE methods. A) contains comparisons of DE gens found using paired designs; and B) contains comparisons of DE genes found using unpaired designs minus results from unpaired BaySeq. DE genes were determined using re-evaluation criteria (estimated FDR < 0.2).

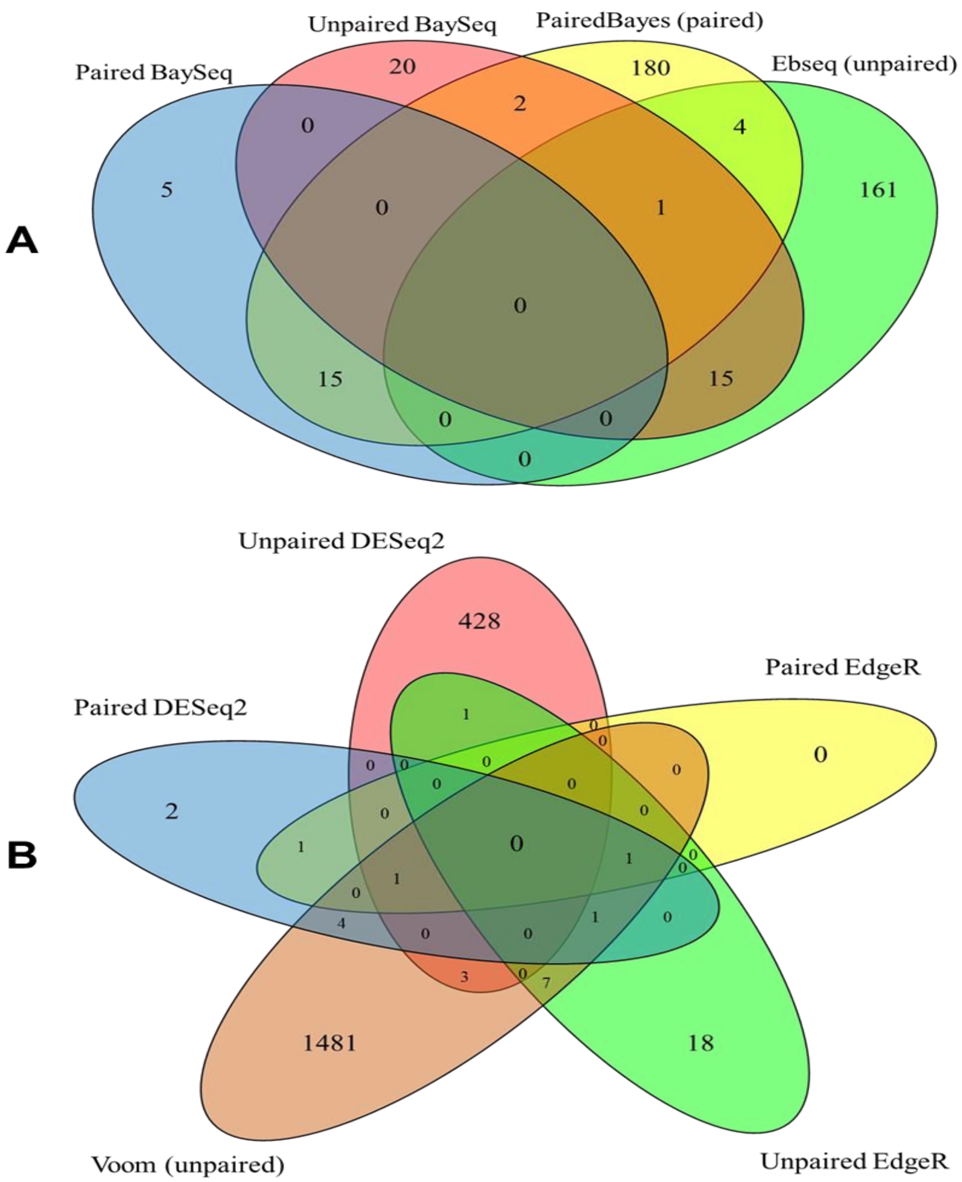


Figure II-A2. Comparison of Differentially Expressed (DE) genes found using Bayesian and Frequentist theoretical backgrounds after re-evaluation. The Venn diagrams above contain the number of Differentially Expressed (DE) genes that were determined by each method. The overlapping portions of the Venn diagrams represent the number of DE genes selected to be the same between the compared DE methods. A) contains comparisons of DE gens found using Bayesian methods; and B) contains comparisons of DE genes found using Frequentist methods with our results from CuffDiff.. DE genes were determined using re-evaluation criteria (estimated FDR < 0.2)

Appendix C: Additional correlation summaries from simulation study

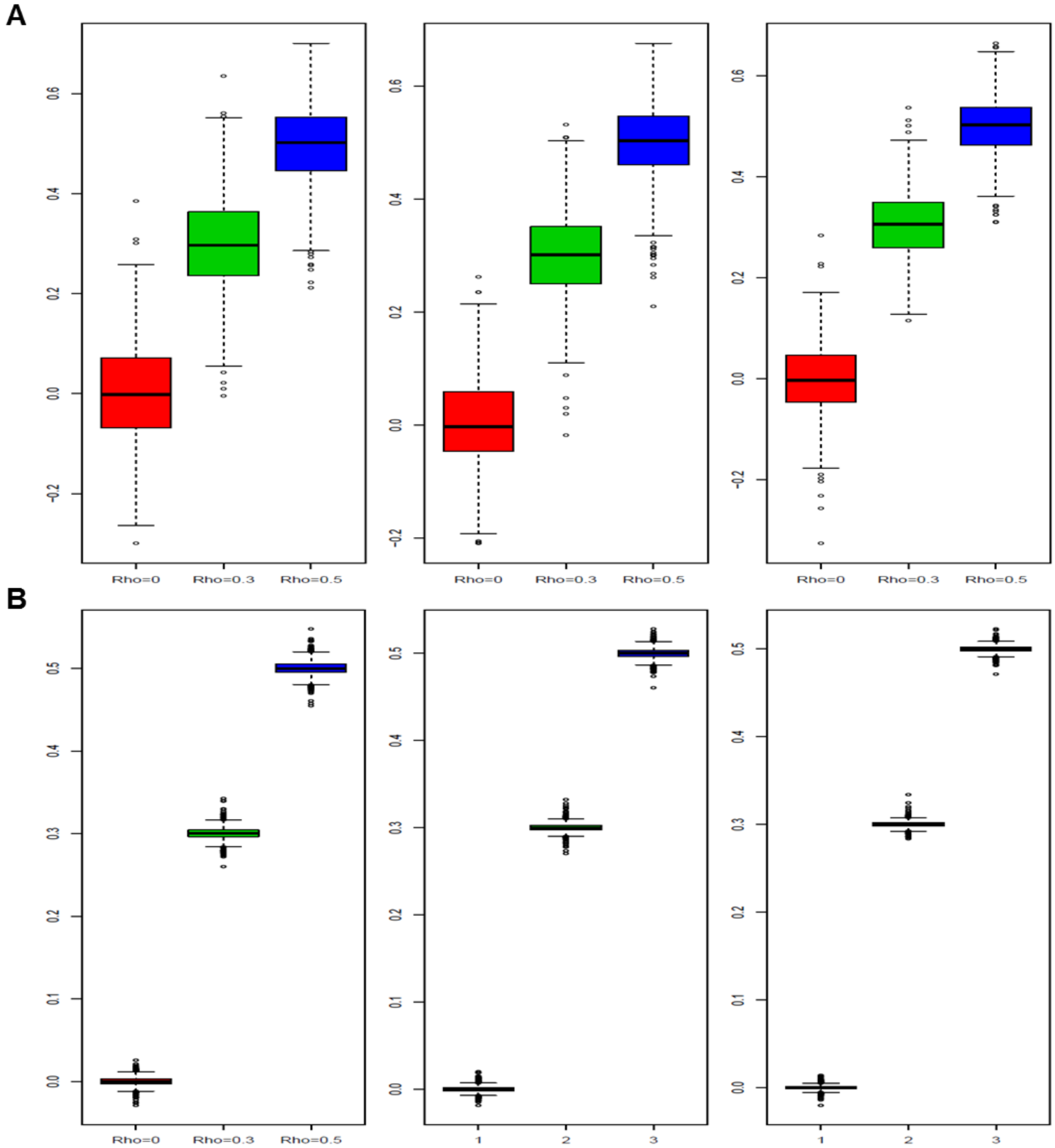
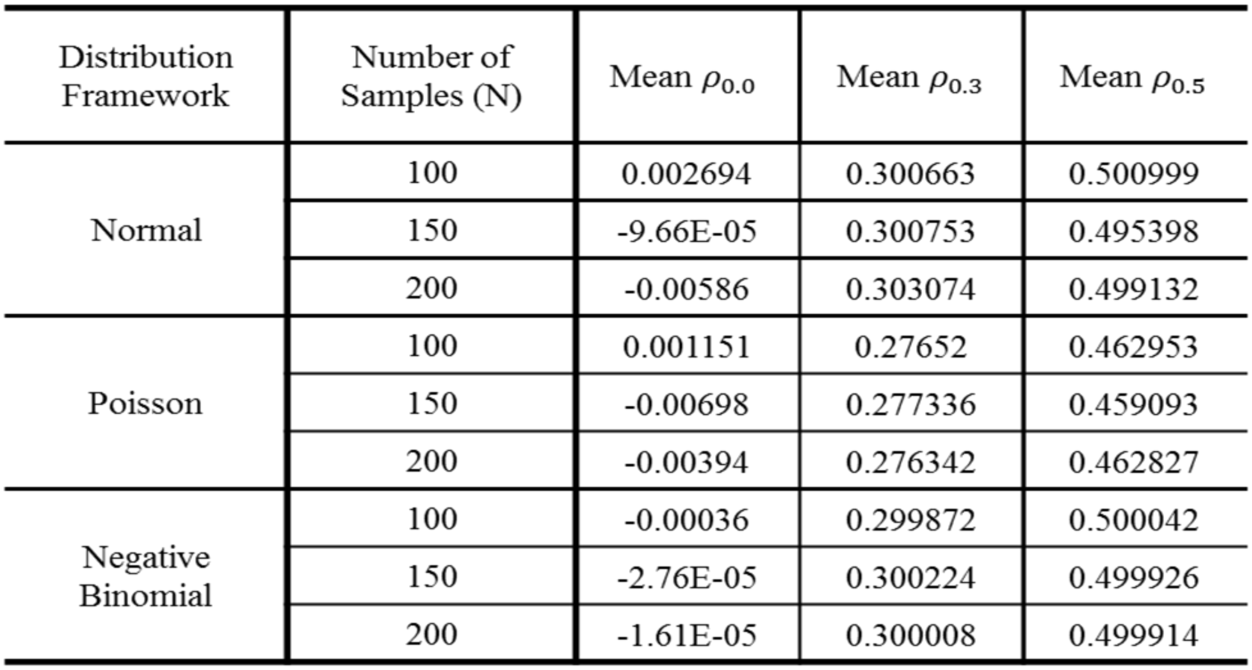


Figure II-A3. Correlation variation summary for simulated data from the Bivariate Poisson and Bivariate Negative Binomial distributions. A) depicts the variability of correlations in simulated data for $\rho = 0, 0.3,$ and 0.5 from the Bivariate Poisson distribution. B) depicts the variability of correlations in simulated data for $\rho = 0, 0.3,$ and 0.5 from the Bivariate Negative Binomial distribution.



Distribution Framework	Number of Samples (N)	Mean $\rho_{0.0}$	Mean $\rho_{0.3}$	Mean $\rho_{0.5}$
Normal	100	0.002694	0.300663	0.500999
	150	-9.66E-05	0.300753	0.495398
	200	-0.00586	0.303074	0.499132
Poisson	100	0.001151	0.27652	0.462953
	150	-0.00698	0.277336	0.459093
	200	-0.00394	0.276342	0.462827
Negative Binomial	100	-0.00036	0.299872	0.500042
	150	-2.76E-05	0.300224	0.499926
	200	-1.61E-05	0.300008	0.499914

Table II-1A. Correlation summary for simulated data from the Bivariate Normal, Bivariate Poisson, and Bivariate Negative Binomial distributions under the alternative with shift of 0.5. Table contains a summary of average correlations from the simulated data for N = 100, 150, and 200 for correlations $\rho = 0, 0.3,$ and 0.5. Data were simulated for unequal means with mean shift of 0.5 added.

Appendix D: Formulas for Evaluation Criteria for Clustering Methods

For each of our D datasets with $N = 56$ samples (i.e., $D = \{d_1, d_2, \dots, d_n\}$), two types of partitions are made U and V each of which divides D into k or r mutually disjoint subsets.

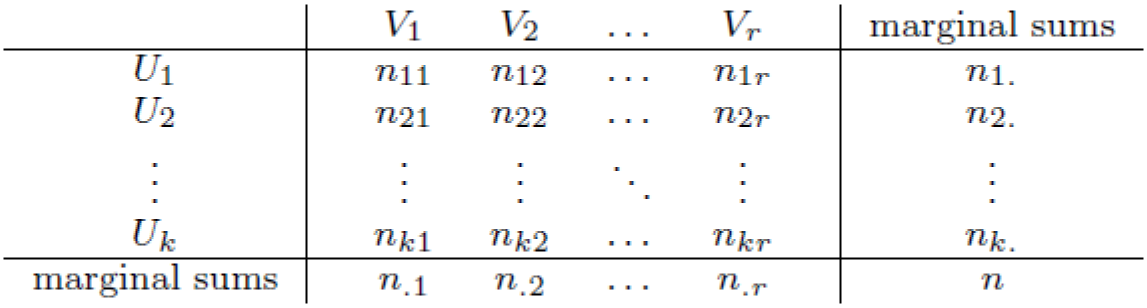
Partition U:

$$D = \bigcup_{i=1}^{k} U_i \text{ and } U_i \cap U_j = \emptyset \quad \forall\, i \neq j \quad \rightarrow \quad U = \{U_1, U_2, \dots, U_k\}$$

Partition V:

$$D = \bigcup_{i=1}^{k} V_i \text{ and } V_i \cap V_j = \emptyset \quad \forall\, i \neq j \quad \rightarrow \quad V = \{V_1, V_2, \dots, V_r\}$$

Evaluation for clustering agreement is based off identifying pairs (d_i, d_j) of data that are the from the same or different partition(Rabbany and Zaiane, 2015). Counts of these pairs are taken from the following table:



	V_1	V_2	$\dots$	V_r	marginal sums
U_1	n_{11}	n_{12}	$\dots$	n_{1r}	$n_{1.}$
U_2	n_{21}	n_{22}	$\dots$	n_{2r}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_k	n_{k1}	n_{k2}	$\dots$	n_{kr}	$n_{k.}$
marginal sums	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.r}$	n

The ij^{th} element in the table is the intersection of the two partitions (i.e., $n_{ij} = |U_i \cap V_j|$), and the marginal sums are $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$. To determine whether a pair belong to the same or different partition, identification of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are computed using sums from the table. Both the Adjusted Rand Index (ARI) and the Clustering Error Rate (CER) utilize this table's information.

Adjusted Rand Index (ARI) (Hubert and Arabie, 1985):

Assumes that the above table is randomly constructed with fixed marginal sums. Rather the size of the clusters in the partition are fixed. With these two assumptions, we get the Adjusted Rand Index

$$Adjusted\ Rand\ Index\ (ARI) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r}\binom{n_{ij}}{2} - {\sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{i=1}^{r}\binom{n_{.j}}{2}}\Big/{\binom{n}{2}}}{\frac{1}{2}\left[\sum_{i=1}^{k}\binom{n_{i.}}{2} + \sum_{i=1}^{r}\binom{n_{.j}}{2}\right] - {\sum_{i=1}^{k}\binom{n_{i.}}{2}\sum_{i=1}^{r}\binom{n_{.j}}{2}}\Big/{\binom{n}{2}}}$$

Clustering Error Rate (CER) (Witten, 2011):

The Clustering Error Rate that is used in Witten's manuscript (2001) is 1-Rand Index. The formula for the Rand Index (Rand, 1971) is:

$$Rand\ Index\ (RI) = 1 + \frac{1}{n^2 - n}\left(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - \left(\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2\right)\right)$$

Hence,

$$Clustering\ Error\ Rate\ (CER) = 1 - 1 + \frac{1}{n^2 - n}\left(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - \left(\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2\right)\right)$$

$$= \frac{1}{n^2 - n}\left(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - \left(\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2\right)\right)$$

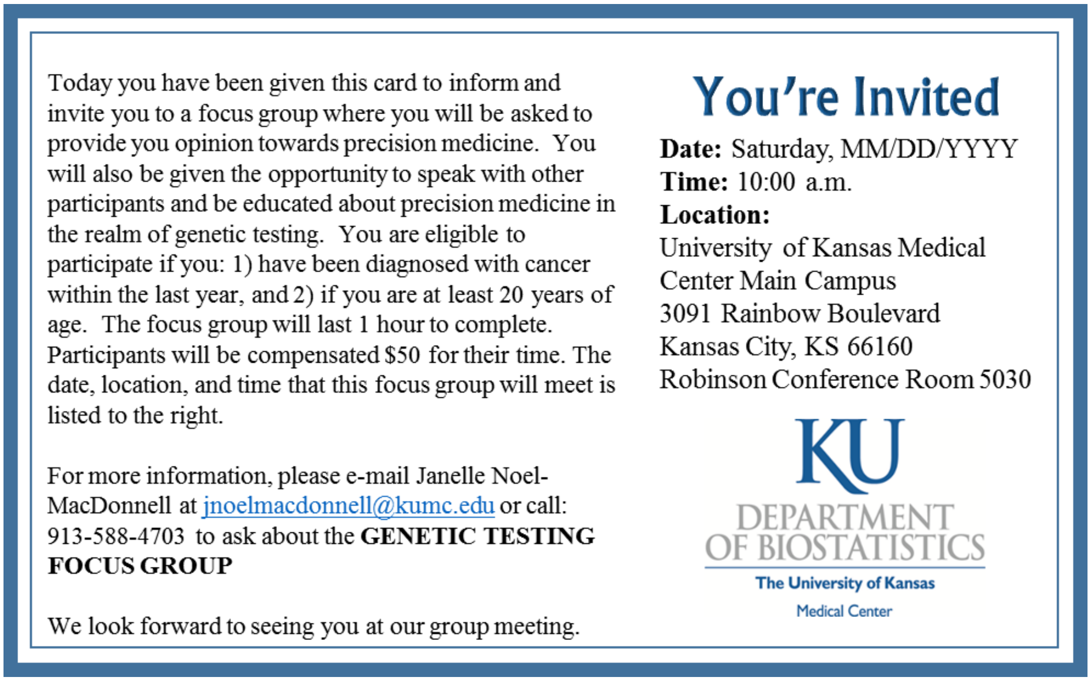
Concordance Index (CI or C-Index) (Harrell et al., 1996):

To calculate the Concordance Index (CI), we make comparisons between the cluster assignments of samples determined by the clustering methods to our simulated cluster assignments. The steps to achieve this are:

1) Organize simulated cluster assignments into pairs (d_i, d_j) where $i \neq j$. Using notation from our study (k_i, k_j)
2) If $k_i > k_j$, determine if $\hat{k}_i > \hat{k}_j$ or $\hat{k}_i = \hat{k}_j$ which determines if the predicted cluster assignments are concordant or discordant. If $\hat{k}_i > \hat{k}_j$ add 1 to running sum as prediction is concordant to observed simulated cluster assignment. If $\hat{k}_i = \hat{k}_j$ add 0.5 to running sum. Let the running sum be *s*.
3) Tally number of times $k_i > k_j$, say *n*.
4) Calculate CI as

$$Concordance\ Index\ (CI) = n/s$$

Appendix E: Patient Recruiting Cards



Today you have been given this card to inform and invite you to a focus group where you will be asked to provide you opinion towards precision medicine. You will also be given the opportunity to speak with other participants and be educated about precision medicine in the realm of genetic testing. You are eligible to participate if you: 1) have been diagnosed with cancer within the last year, and 2) if you are at least 20 years of age. The focus group will last 1 hour to complete. Participants will be compensated $50 for their time. The date, location, and time that this focus group will meet is listed to the right.

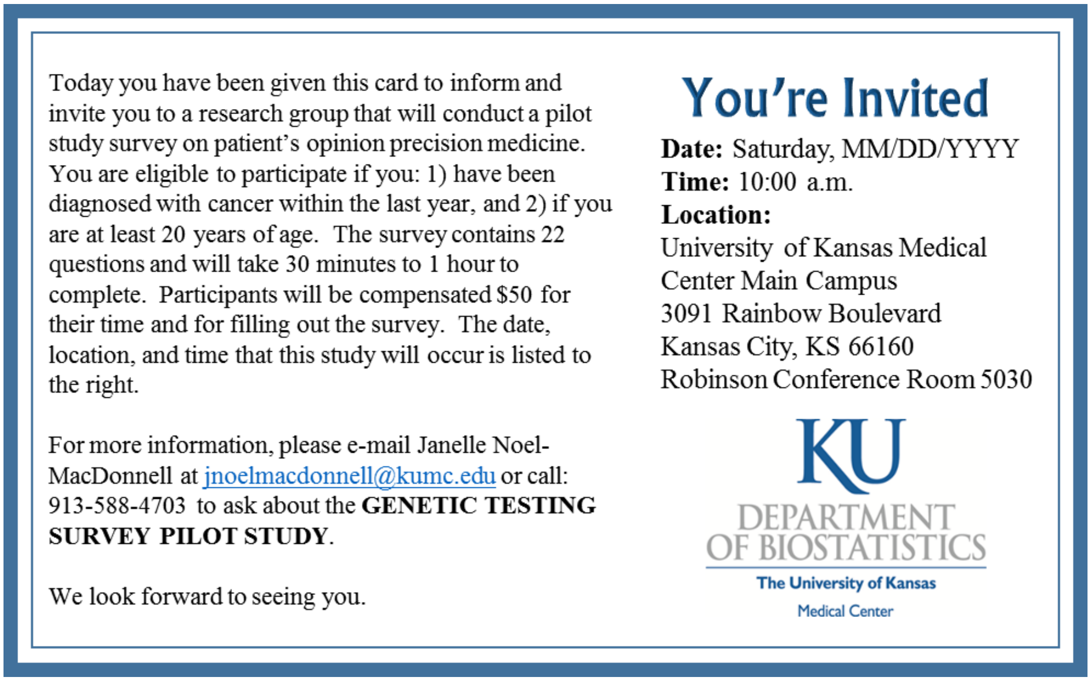
For more information, please e-mail Janelle Noel-MacDonnell at jnoelmacdonnell@kumc.edu or call: 913-588-4703 to ask about the **GENETIC TESTING FOCUS GROUP**

We look forward to seeing you at our group meeting.

You're Invited

Date: Saturday, MM/DD/YYYY
Time: 10:00 a.m.
Location:
University of Kansas Medical
Center Main Campus
3091 Rainbow Boulevard
Kansas City, KS 66160
Robinson Conference Room 5030

KU
DEPARTMENT
OF BIOSTATISTICS
The University of Kansas
Medical Center



Today you have been given this card to inform and invite you to a research group that will conduct a pilot study survey on patient's opinion precision medicine. You are eligible to participate if you: 1) have been diagnosed with cancer within the last year, and 2) if you are at least 20 years of age. The survey contains 22 questions and will take 30 minutes to 1 hour to complete. Participants will be compensated $50 for their time and for filling out the survey. The date, location, and time that this study will occur is listed to the right.

For more information, please e-mail Janelle Noel-MacDonnell at jnoelmacdonnell@kumc.edu or call: 913-588-4703 to ask about the **GENETIC TESTING SURVEY PILOT STUDY**.

We look forward to seeing you.

You're Invited

Date: Saturday, MM/DD/YYYY
Time: 10:00 a.m.
Location:
University of Kansas Medical
Center Main Campus
3091 Rainbow Boulevard
Kansas City, KS 66160
Robinson Conference Room 5030

KU
DEPARTMENT
OF BIOSTATISTICS
The University of Kansas
Medical Center

Appendix F: Survey Questionnaire

KU
DEPARTMENT
OF BIOSTATISTICS
The University of Kansas
Medical Center

University of Kansas Medical Center
Department of Biostatistics
5028 Robinson (5th Floor)
3901 Rainbow Boulevard
Kansas City, KS 66160

Pilot Study Questionnaire

Questionnaire ID: #####

Pilot Study Title: Cancer Patient Opinions Towards Genetic Testing in Cancers

Thank you for participating in this focus group and consenting to take part in this survey which measures cancer patient's opinion towards genetic sequencing in cancers. The survey below contains three separate sections. Section 1 contains general questions that will provide information about your background. Section 2 will briefly educate or refresh your knowledge of the concepts that you will asked questions about in Section 3. Lastly, Section 3 is comprised of questions to help us obtain patient's opinions regarding the use of genetic sequencing to aid in cancer treatment. Below Section 3 is space to leave comments and feedback regarding this focus group and/or survey. Be assured that all answers to this survey will be kept confidential. Once you have completed all questions, please turn in survey to administer to receive your compensation for you time and participation.

---

*Please be sure to select a single response for each question.

Section 1: Background Questions

1. **What is your age?**

□ 20 – 29 Years □ 30 – 39 Years □ 40 – 49 Years

□ 50 – 59 Years □ 60 – 69 Years □ 70+ Years

□ Prefer Not to Answer

2. What is your gender?

□ Female □ Male □ Other

□ Prefer Not to Answer

3. What is your race/ethnicity?

□ White □ Black or African American □ American Indian or Alaska Native

□ Hispanic or Latino □ Asian □ Other ____________________________

□ Unknown

□ Prefer Not to Answer

4. What is your marital status?

□ Single □ Married □ Divorced □ Separated

□ Prefer Not to Answer

5. What is your highest level of education?

□ Less Than High School □ High School / GED □ Some College □ 4 – Year College Degree

□ Graduate or Professional Degree

□ Prefer Not to Answer

6. What is your household income level?

□ $0 - $15,999 □ $16,000 - $24,999 □ $25,000 - $49,999

□ $50,000 - $99,999 □ $100,000+

□ Prefer Not to Answer

7. How many children do you have?

□ No Children □ 1 Child □ 2 – 3 Children □ 4 – 5 Children

□ 6+ Children

□ Prefer Not to Answer

8. What type of cancer do you currently have?

____

9. Do you have a family history of cancer?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

9a. If you answered “Yes” to question 9 above, what type of cancer(s) are part of your family history of cancer?

____

Section 2: Genetic Testing Education

Genetic information can be viewed as the blue-print for every individual. This blue-print contains specific information that makes individuals have specific characteristic traits (i.e., hair color, presence of dimples, hairline, etc.). The major component of this blue-print is DNA which is made up of four base pairs (i.e., Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)) which code for genes (Figure IV-S1). DNA is found in form of chromosomes within the nucleus of nearly every cell in the human body. The DNA within each of the 23 pairs of chromosomes guide the cells in the body to grow and develop. Though when a gene becomes mutated in a specific area, it can cause cell to misbehave. Depending on the mutation, the human body's immune system may respond by killing off the cell. However, when the immune system does not recognize the mutation, the cell divides and copies at rapid rate leading to cancer. With the advancement of technology, researchers are able to look directly at an individual's DNA and/or those mutations that are found in cancerous cells through genetic testing (Figure IV-S2). Often times the information found from genetic testing of cancer can lead to more targeted treatment therapies.

(Designed figure showing relationship between genes, cells, tissues, and the body)

Figure IV-S1. Overview of DNA within the human body

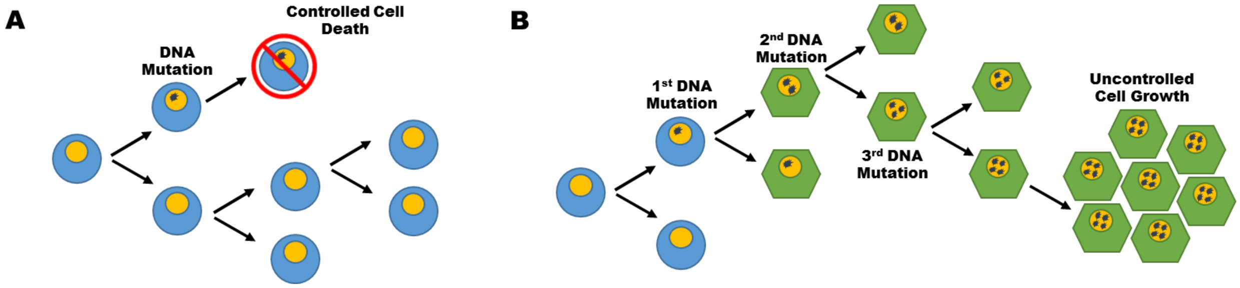


Figure IV-S2. Types of cell copying in the human body. New cells are continually being made in the human body by coping information from older cells. Occasionally, a mutation of the DNA occurs which causes various reactions to the cell. Panel A displays the immune system's response to a DNA mutation. Panel B displays the uncontrollable grow (i.e., cancer) that occurs when the immune system does not recognize the mutations that occur.

Section 3: Patient Opinion Questions

10. Please rate your current level of understanding of genetic testing?

□ None □ Very Poor □ Poor

□ Fair □ Good □ Very Good

□ Unknown

□ Prefer Not to Answer

11. Have you ever had genetic testing completed for your current cancer?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

12. Have you ever had any genetic testing completed in your past for any reason?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

13. Have any members of your immediate family (i.e., parents and/or grandparents) that have had cancer, had genetic testing completed?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

14. Generally speaking, do you think that conducting genetic testing is ethical?

□ Disagree □ Neither Disagree or Agree □ Agree

□ Unknown

□ Prefer Not to Answer

15. Has your medical provider talked to you about the possibility of having genetic testing completed to determine targeted options for cancer treatment?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

16. Has your medical provider scheduled an appointment for you to have genetic testing completed on a sample of your cancer (i.e., your tumor)?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

17. If you answered "No" or "Unknown" to question 15 would you want to have genetic testing completed if it were an option to determine targeted cancer treatment?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

□ Answered "Yes" for Question 15

18. If the results from genetic testing revealed an incidental finding would you want to know about it? (An *incidental finding* is a potential medically relevant finding that was found unintentionally or that is unrelated to tested medical condition.)

□ Yes □ No □ Unknown

□ Prefer Not to Answer

19. If genetic testing revealed that there was a clinical trial testing a new drug that would be a treatment option for your cancer, would you consider enrolling?

□ Yes □ No □ Unknown

□ Prefer Not to Answer

20. Do you think patient providers should promote genetic testing to cancer patients?

□ Disagree □ Neither Disagree or Agree □ Agree

□ Unknown

□ Prefer Not to Answer

21. How do you feel about having additional genetic testing of your germline to determine if you have a mutation or many mutations in cancer-predisposing genes?

□ I would decline additional genetic testing

□ I would agree to additional genetic testing

□ Unknown

□ Prefer Not to Answer

22. Which of the following, if any, do you feel influenced your responses to any of the questions in Section 3? (Select all that apply)

□ Lack of Knowledge/Information Regarding Survey Topic

□ Religious Beliefs

□ Costs Associated with Genetic Testing

□ Insurance Concerns

□ Previous Experience

□ Other ___________________________

□ Prefer Not to Answer

Please provide us with any feedback or comments pertaining to this survey or focus group session.

____

____

____

____

Thank you for your time and responses. Please see survey administrator to receive your compensation for participation.

Appendix G: Shell Tables and Feedback Space to Complete after Quality Control is Completed on Survey Data

Table IV-A1. Demographic summary from Section 1 of survey questionnaire

Table contains summary of responses from Section 1 of survey questionnaire. Responses are given as a frequency and percentage.

Questions	Question Responses	Response Frequency (%)
1. What is your age?	20-29 Years 30-39 Years 40-49 Years 50-59 Years 60-69 Years 70+ Years Prefer Not to Answer	
2. What is your gender?	Female Male Other Prefer Not to Answer	
3. What is your race/ethnicity?	White Black or African American American Indian of Alaska Native Hispanic or Latino Asian Other Unknown Prefer Not to Answer	
4. What is your marital status?	Single Married Divorced Separated Prefer Not to Answer	
5. What is your highest level of education?	Less Than High School High School/ GED Some College 4 – Year College Degree Graduate or Professional Degree Prefer Not to Answer	

6. What is your household income level?	$0-$15,999 $16,000-$24,999 $25,000-$49,999 $50,000-$99,999 $100,000+ Prefer Not to Answer	
7. How many children do you have?	No Children 1 Child 2 – 3 Children 4 – 5 Children 6+ Children Prefer Not to Answer	
8. What type of cancer do you have currently?	**Responses will be listed with their frequency	
9. Do you have a family history of cancer?	Yes No Unknown Prefer Not to Answer	
8a. If you answered “Yes” to question 8 above, what type of cancer(s) are part of your family history of cancer?	**Top responses will be listed with their frequency. Complete list would be presented in list format below.	

Question 8 unique responses with frequency provided in “( )”:

Question 9a unique responses with frequency provided in “( )”:

Table IV-A2. Summary of patient's opinion responses from Section 3 of survey questionnaire

Table contains summary of responses from Section 1 of survey questionnaire. Responses are given as a frequency and percentage.

Questions	Question Responses	Response Frequency (%)
10. Please rate your current level of understanding of genetic testing?	None Very Poor Poor Fair Good Very Good Unknown Prefer Not to Answer	
11. Have you ever had genetic testing completed for your current cancer?	Yes No Unknown Prefer Not to Answer	
12. Have you ever had any genetic testing completed in your past for any reason?	Yes No Unknown Prefer Not to Answer	
13. Have any members of your immediate family (i.e., parents and/or grandparents) that have had cancer, had genetic testing completed?	Yes No Unknown Prefer Not to Answer	
14. Generally speaking, do you think that conducting genetic testing is ethical?	Disagree Neither Disagree or Agree Agree Unknown Prefer Not to Answer	
15. Has your medical provider talked to you about the possibility of having genetic testing completed to determine targeted options for cancer treatment?	Yes No Unknown Prefer Not to Answer	

16. Has your medical provider scheduled an appointment for you to have genetic testing completed on a sample of your cancer (i.e., your tumor)?	Yes No Unknown Prefer Not to Answer	
17. If you answered “No” or “Unknown” to question 15 would you want to have genetic testing completed if it were an option to determine targeted cancer treatment?	Yes No Unknown Prefer Not to Answer Answered “Yes” for Question 15	
18. If the results from genetic testing revealed an incidental finding would you want to know about it?	Yes No Unknown Prefer Not to Answer	
19. If genetic testing revealed that there was a clinical trial testing a new drug that would be a treatment option for your cancer, would you consider enrolling?	Yes No Unknown Prefer Not to Answer	
20. Do you think patient providers should promote genetic testing to cancer patients?	Disagree Neither Disagree or Agree Agree Unknown Prefer Not to Answer	
21. How do you feel about having additional genetic testing of your germline to determine if you have a mutation or many mutations in cancer-predisposing genes?	I would decline additional genetic testing I would agree to additional genetic testing Unknown Prefer Not to Answer	

22. Which of the following, if any, do you feel influenced your responses to any of the questions in Section 3? (Select all that apply)	Lack of Knowledge/Information Regarding Survey Topic Religious Beliefs Costs Associated with Genetic Testing Insurance Concerns Previous Experience Other Prefer Not to Answer	

Question 22 written-in responses for “Other” response options:

Table IV-A3. Mean response of patient’s opinion for select questions

Questions	Question Responses	Mean Response
14. Generally speaking, do you think that conducting genetic testing is ethical?	Disagree Neither Disagree or Agree Agree Unknown Prefer Not to Answer	
20. Do you think patient providers should promote genetic testing to cancer patients?	Disagree Neither Disagree or Agree Agree Unknown Prefer Not to Answer	
21. How do you feel about having additional genetic testing of your germline to determine if you have a mutation or many mutations in cancer-predisposing genes?	I would decline additional genetic testing I would agree to additional genetic testing Unknown Prefer Not to Answer	

List IV-A1. Comments and feedback provide by participants at the end of study

REFERENCES

Aguilera, A. & Garcia-Muse, T. 2013. Causes of Genome Instability. *Annual Reviews Genetics,* 32**,** 1-32.

Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. 2006. Microarray Data Analysis: From Disarray to Consolidation and Consensus. *Nature Reviews: Genetics,* 7**,** 55-65.

American Cancer Society 2016. Cancer Facts & Figures.

Anders, S. & Huber, W. 2010. Differential Expression Analysis for Sequence Count Data. *Genome Biology,* 11.

Anders, S., Mccarthy, D. J., Chen, Y., Okoniekshie, M., Smyth, G. K., Huber, W. & Robinson, M. D. 2013. Count-based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor. *Nature Protocol,* 8**,** 1765-1786.

Anders, S. a. H., W. 2010. Differential Expression Analysis for Sequence Count Data. *Genome Biology,* 11.

Anderson, D. E. 1992. Familial Versus Sporadic Breast Cancer. *Cancer Syndromes,* 70**,** 1740-1746.

Auer, P. & Doerge, R. 2010. Statistical Design and Analysis for RNA Sequencing Data. *Genetic Socitey of America,* 185**,** 405-416.

Balmain, A., Gray, J. & Ponder, B. 2003. The Genetics and Genomics of Cancer. *Nature Genetics,* 33**,** 238-244.

Barbiero, A. & Ferrari, P. A. 2014. Simulation of Correlated Poisson Variables. *Applied Stochastic Models in Business and Industry,* 31.

Bateson, W. & Gregor, M. 1913. *Mendel's Principles of Heredity,* University Press.

Beasley, T. M., Erickson, S. & Allison, D. B. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet,* 39**,** 580-95.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society,* 57**,** 289-300.

Brown, P., And Botstein, D. 1999. Exploring the New World of the Genome with DNA microarrays. *Nature Genetics,* 21.

Brunet, J., Tamayo, P., Golub, T. R. & Mesirov, J. P. 2004. Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *Proceedings of the NAtional Academy of Sciences,* 101**,** 4164-4169.

Bulusu, K. C., Tym, J. E., Coker, E. A., Schierz, A. C. & Al-Lazikani, B. 2014. CanSAR: Updated Cancer REsearch and Drug Discovery Knowledgebase. *Nucleic Acids Research,* 42**,** D1040-D1047.

Casella, G. & Berger, R. L. 2002. *Statistical Inference*, Duxbury Press.

Catenacci, D. V. T., Amico, A. L., Nielsen, S. M., Geynisman, D. M., Rambo, B., Carey, G. B., Gulden, C., Fackenthal, J., Marsh, R. D., Kindler, H. L. & Olopade, O. I. 2014. Tumor Genome Analysis Includes Germline Genome: Are We Ready for Surprises? *International Journal of Cancer,* 136**,** 1559-1567.

Centers for Disease Control and Prevention^1 2015. CDC in India.

Centers for Disease Control and Prevention^2 2015. Leading Causes of Death.

Centers for Disease Control and Prevention: Division for Heart Disease and Stroke Prevention 2014. Healthy People 2020.

Chalise, P., Koestler, D. C., Bimali, M., Yu, Q. & Fridley, B. L. 2014. Integrative clustering methods for high-dimensional molecular data. *Transl Cancer Res,* 3**,** 202-216.

Chen, G., Jaradar, S., Banerjee, N., Tanaka, T., Ko, M. S. H. & Zhang, M. Q. 2002. Evaluation and Comparison of Clustering Algoritms in Analysing ES Cell Gene Expression Data. *Statistica Sinica*, 12.

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X. Q. & He, J. 2016. Cancer statistics in China, 2015. *CA: A Cancer Journal for Clinicians,* 66**,** 115-132.

Chong, E. Y., Huang, Y., Wu, H., Ghasemzadeh, N., Uppal, K., Quyyumi, A. A., Jones, D. P. & Yu, T. 2015. Local False Discovery Rate Estimation Using Feature Reliability in LC/MS Metabolomics Data. *Scientific Reports,* 5.

Chung, L. M., Ferguson, J. P., Zheng, W., Qian, F., Bruno, V., Montogomery, R. R. & Zhao, H. 2013. Differential Expression Analysis for Paired RNA-Seq Data. *BMC Bioinformatics,* 14:110.

Ciardiello, F., Arnole, D., Casali, P. G., Cervantes, A., Douillard, J. Y., Eggermont, A., Eniu, A., Mcgregor, K., Peters, S., Piccart, M., Popescu, R., Van Cutsem, E., Zielinski, C. & Stahel, R. 2014. Delievering Precision Medicin in Oncology Today and in Future – The Promise and Challenges of Personalized Cancer Medicine: A Position paper by European Society for Medical Oncology (ESMO). *Annals of Oncology* 25**,** 1673-1678.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., Mcpherson, A., Szczesniak, M., Gaffney, D., Elo, L., Zhang, X. & Mortazavi, A. 2016. A survey of best practices for RNA-Seq Data. *Genome Biology,* 17.

Coriell Personalized Medicine Collaborative 2016. Is Cancer Genetic?

Crick, F. H. C. 1970. Central Dogma of Molecular Biology. *Nature,* 227.

Dahm, R. 2005. Fredrich Miescher and The Discovery of DNA. *Developmental Biology,* 278**,** 274-288.

Darwin, C. 1872. *On the Origins of Species by Means of Natural Selection,* Akron, Ohio, The Werner Company.

Derisi, J., Iyer, V., and Brown, P., 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science,* 278.

Devarajan, K. 2008. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. *Plos One Computational Biology,* 4.

Dillies Et. Al. 2012. A Comprehensive Evaluation of Normalization methods for Illumiuna High-Throughput RNA Sequencing Data Analysis. *Bioinformatics*.

Durbin, B. P., Hardin, J. S., Hawkins, D. M. & Rocke, D. M. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics,* 18 Suppl 1**,** S105-10.

Efron, B. & Tibshirani, R. 2002. Empirical Bayes Methods and False Discovery Rates for Microarrays. *Genetic Epidemiology,* 23**,** 70-86.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. 2001. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96**,** 1151-1160.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A,* 95**,** 14863-8.

Erhardt, V. 2009. corcounts: Generate Correlated Count Random Variables.

Erhardt, V. & Czado, C. 2008. A Method for Approximately Sampling High-Dimensional Count Variables with Prespecified Pearson Correlation.

Fabsitz, R. R., Mcguire, A., Sharp, R. R., Puggal, M., Beskow, L. M., Biesecker, L. G., Bookman, E., Burke, W., Burchard, E. G., Church, G., Clayton, E. W., Eckfeldt, J. H., Fernandez, C. V., Fisher, R., Fullerton, S. M., Gabriel, S., Gachupin, F., James, C., Jarvik, G. P., Kittles, R., Leib, J. R., O’donnell, C., O'rourke, P. P., Rodriguez, L. L., Schully, S. D., Shuldiner, A. R., Sze, R. K. F., Thakuria, J. V., Wolf, S. M. & Burke, G. L. 2010. Ethical and Practical Guidelines for Reporting Genetic Research Results To Study Participants: Updated Guidelines from an NHLBI Working Group. *Circulation. Cardiovascular genetics,* 3**,** 574-580.

Farewell, V. T. & Sprott, D. A. 1988. The Use of a Mixture Model in the Analysis of Count Data. *Biometrics***,** 1191-1194.

Farley, C. & Raftery, A. E. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association,* 97**,** 611-631.

Ferlay, J., Soerjomataram, I., Ervik, M. & Al., E. 2013. GLOBOCAN 2012: Estimated Cancer Incidence. *World Health Organization: International Agency for Research on Cancer,* 10.

Fiore, R. N. & Goodman, K. W. 2015. Precision Medicine Ethics: Selected Issues and Developments in Next-Generation Sequencing, Clinical Oncology, and Ethics. *Current Opinion Oncology,* 28.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection: A Complete Variorum Edition. *Oxford University Press*.

Foster, M. W. & Sharp, R. R. 2006. Ethical Issues in Medical-Sequencing Research: Implications of Genotype-Phenotype Studies for Individuals and Populations. *Human Molecular Genetics,* 15**,** R45-R49.

Fraley, C. & Raftery, A. E. 1998. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *In:* WASHINGTON, U. O. (ed.).

Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clutering, Classification, and Density Estimation Technical Report No. 597. Department of Statistics, University of Washington.

Freedman, B. 1987. Equipoise and The Ethics of Clinical Research. *New England Journal of Medicine,* 317**,** 141-145.

Gelman, A. 2013. Understanding Posterior P-values. *Electronic Journal of Statistics*.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. & Zhang, J. 2004. Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology,* 5.

Gingras, I., Sonnenblick, A., De Azambuja, E., Paesmans, M., Delaloge, S., Aftimos, P., Piccart, M. J., Sotiriou, C., Ignatiadis, M. & Azim, H. A. 2016. The current use and attitudes towards tumor genome sequencing in breast cancer. *Scientific Reports,* 6**,** 22517.

Green, P. J. 1984. Iteratively REweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society,* 46**,** 149-192.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Krof, B. R., Martin, C. L., Mcguire, A. L., Nussbaum, R. L., O'daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S. & Biesecker, L. G. 2013. ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. *In:* GENOMICS, A. C. O. M. G. A. (ed.).

Guo, Y. L., C.; Ye, F.L Shyr, Y. 2013. Evaluation of Read Count Based RNAseq Analysis Methods. *BMC Genomics,* 14.

Hardcastle, T. J. 2016. baySeq: Empirical Bayesian Analysis of Patterns of Differential Expression in Count Data. Bioconductor.

Hardcastle, T. J. & Kelly, K. A. 2010. baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics,* 11.

Harrell, F. E., Lee, K. L. & Mark, D. B. 1996. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, And Measuring And Reducing Errors. *Statistics in Medicine,* 15**,** 361-387.

Hartigan, J. A., And Wong, M.A. 1979. Algorithm AS 136: A K-means Clustering Algorithm. *Journal of the Royal Statistical Society,* 28**,** 100-108.

Heather, J. M. & Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

Henneman, L., Verneulen, E., Van El, C. G., Claassen, L., Timmermans, D. R. M. & Cornel, M. C. 2013. Public Attitudes Towards Genetic Testing Revisited: Comparing Opinions Between 2002 and 2010. *European Journal of Human Genetics*, 21**,** 793-799.

Hofman, B. 2016. Incidental Findings of Uncertain Significance: To Know or Not to Know – That is Not the Question. *BMC Medical Ethics*, 17.

Holley, R. W. 1965. Structure of a Ribonucleic Acid. *Science*, 147**,** 1462-1465.

Houseman, E. A., And Koestler, D.C. 2014. RPMM: Recursively Partitioned Mixture Model. R package version 1.20 ed.

Houseman, E. A., Christensen, B. C., Yeh, R., Marsit, C. J., Karagas, M. R., Wrensch, M., Nelson, H. H., Wiemels, J., Zheng, S., Wiencke, J. K. & Kelsey, K. T. 2008. Model-based Clustering of DNA Methlylation Array Data: A Recursive-partitioning algorithm for high-dimensional Data Arising as a Micture of Beta Distributions. *BMC Bioinformatics,* 9.

Huang, A. 2008. Similarity Measures for Text Document Clustering. *In:* THE UNIVERSITY OF WAIKATO, D. O. C. S. (ed.). Hamilton, New Zealand.

Hubert, L. & Arabie, P. 1985. Comparing Partitions. *Journal of Classification,* 2**,** 196-218.

Jiang, D., Tang, C. & Zhang, A. 2004. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 16.

Johnson, N. L., Kotz, S. & Balakrishnan, N. 1997. *Discrete Multivariate Distributions,* New York, John Wiley and Sons, Inc.

Karlis, D. & Ntzoufras, I. 2006. Bayesian Analysis of the Differences of Count Data. *Statistics in Medicine,* 25**,** 1885-1905.

Kauermann, G. & Carroll, R. J. 2000. The Sandwich Variance EstimatorL Efficiency Properties and Coverage Probability of Confidence Intervals. *Stata Journal*.

Khafri, S., Kazemnejad, A. & Eskandari, F. 2008. Hierarchical Bayesian Analysis of Bivariate Poisson Regression Model. *World Applied Sciences Journal,* 4.

King, M., Mars, J. H. & Mandell, J. B. 2003. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science* 302**,** 643-646.

Koestler, D. C., Christensen, B.C., Marsit, C.J., and Kelsey, K.T., 2013. Recursively Partitioned Mixture Model Clustering of DNA Methlyation Data Using Biologically Informed Correlation Structures. *Statistical Applications in Genetics and Molecular Biology,* 12**,** 225-240.

Ku, C. S., Cooper, D. N. & Roukos, D. H. 2013. Clinical relevance of cancer genome sequencing. *World Journal of Gastroenterology : WJG,* 19**,** 2011-2018.

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.

Lee, D. D. & Seung, H. S. 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Letter to Nature,* 401.

Lee, H. S. 1996. Analysis of Overdispersed Paired Count Data. *Canadian Journal of Statistis,* 24**,** 319-326.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M. & Kendziorski, C. 2013. EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments. *Oxford University Press*.

Liao, J. G., Lin, Y., Selvanyagam, Z. E. & Shih, W. J. 2004. A Mixture Model for Estimating the Local False Discovery Rate in DNA Microarray Analysis. *Bioinformatics,* 20.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Kevin, M. B., Investigators;, A., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G. & Macgregor, S. 2010. A Versatile Gene-Based Test for Genome-wise Association Studies. *The American Journal of Human Genetics,* 87**,** 139-145.

Liu, P. & Si, Y. 2014. Cluster Analysis of RNA-Sequencing Data. *In:* DATTA, S. & NETTLETON, D. (eds.) *Statistical Analysis of Next Generation Sequencing Data.* Springer.

Love, M. I., Anders, S. & Huber, W. 2014. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biology,* 15.

Love, M. I., Anders, S. & Huber, W. 2016. Differential Analysis of Count Data--the DESeq2 Package.

Luo, J., Solimini, N. L. & Elledge, S. J. 2009. Prinicples of Cancer Therapy: Oncogene and Non-Oncogene Addiction. *Cell,* 136**,** 823-837.

Makretsov, N., Huntsman, D. G., Nielson, T. O., Yorida, D. G., Peacock, M., Cheang, M. C. U., Dunn, S. E., Hayes, M., Van De Rijn, M., Bajdik, C. & Gilks, C. B. 2004. Hierarchical Clustering Analysis of Tissue Microarray Immunostaining Data Identifies Prognostically Signigicant Groups of Breast Carcinoma. *Clinical Cancer Research,* 10**,** 6143-6151.

Mardia, K. V. 1970. *Families of Bivariate Distributions,* London, Griffin.

Mardis, E. R. 2008. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics,* 9**,** 387-402.

Mardis, E. R. & Wilson, R. K. 2009. Cancer Genome Sequencing: A Review. *Human Molecular Genetics,* 18**,** R163-R168.

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**,** 1509-1517.

Mariotto, A. B., Yabroff, K. R., Shao, Y., Feuer, E. J. & Brown, M. L. 2011. Projections of the Cost of Cancer Care in the U.S. : 2010-2020. *Journal of the National Cancer Institute*.

Maxam, A. M. & Gilbert, W. 1977. A New Method for Sequencing DNA. *Proceedings of the National Academy of Sciences,* 74**,** 560-564.

Mccullagh, P. & Nelder, J. A. 1989. *Generalized Linear Models,* Boca Raton, Chapman & Hall/CRC.

Mcguire, A. L., Joffe, S., Koenig, B. A., Biesecker, B. B., Mccullough, L. B., Blumenthal-Barby, J. S., Cualfield, T., Terry, S. F. & Green, R. C. 2013. Ethics and Genomic Incidental Findings. *Science,* 340**,** 1047-1048.

Mclachlan, G. J., Bean, R. W. & Peel, D. 2002. A Mixture Model-Based Approach to The Clustering of Microarray Expression Data. *Bioinformatics,* 18**,** 413-422.

Medvedovic, M. & Sivaganesan, S. 2002. Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles. *Bioinformatics,* 18**,** 1194-1206.

Medvedovic, M., Yeung, K. Y. & Bumgarner, R. E. 2004. Bayesian Mixture Model Based Clustering of Replicated Microarray Data. *Bioinformatics,* 20**,** 1222-1232.

Middlelton, A., Morley, K. I., Bragin, E., Firth, H. V., Hurles, M. E., Wright, C. F. & Parker, M. 2016. Attitudes of nearly 7000 Health Professionals, Genomic Researchers, and Publics Toward the Return of Incidental Results from Sequencing Research. *European Journal of Human Genetics,* 24**,** 21-29.

Monti, W., Savage, K.J., Kutok, J.L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R.C.T., Cin, P.D., Lass, C., Pinkus, G.S., Salles, G., Harris, N.L., Dalla-Favera, R., Habermann, T. M., Aster, J.C., Golub, T.R., and Shipp, M.A. 2005. Molecular Profiling of Diffuse Large B-cell Lymphoma Identifies Robust Subtypes Including One Characterized by Host Inflammatory response. *Blood*, 105.

Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L., and Wold, B. 2008. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nature Methods,* 5**,** 621-628.

National Cancer Institute^{1}. 2016. *SEER Stat Fact* [Online]. [Accessed].

National Cancer Institute^{2}. 2016. *Cancer Statistics* [Online]. [Accessed].

National Cancer Institute^{3} 2016. *Familial Cancer*. *NCI Dictionary of Cancer Terms.* National Cancer Institute.

National Cancer Institute^{4} 2015. BRCA1 and BRCA2: Cancer Risk and Genetic Testing.

National Cancer Institute^{5} 2013. Genetic Testing for Hereditary Cancer Synfromes.

National Cancer Institute^{6} 2016. Sporadic Cancer. *NCI Dictionary of Cancer Terms.* National Cancer Institute.

National Cancer Institute^{7} 2016. Somatic Mutation. National Cancer Institute.

National Human Genome Research Institute^{2} 2015. DNA Sequencing.

National Human Genome Research Institute, National Institutes of Health Department of Health and Human Services & Office of Science U.S. Department of Energy. 2010. *The Human genome Project Completion: Frequently Asked Questions* [Online]. [Accessed].

Negrini, S., Gorgoulis, V. G. & Halazonetid, T. D. 2010. Genomic Instability--An Evolving Hallmark of Cancer. *Nature Reviews,* 11.

Nielsen, T., West, R., Linn, S., Alter, O., Knowling, M., O'connell, J. S. Z., Fero, M., Sherlock, G., Pollack, J., Brown, P., Botstein, D. And Van De Rijin, M., 2002. Molecular Characterisation of Soft Tissue Tumours: A Gene Expression Study. *The Lancet,* 359.

Office of Disease Prevention and Health Promotion 2016. Healthy People.

Qu, Y. & Xu, S. 2004. Supervised Cluster Analysis for Microarray Data Based on Multivariate Gaussian Mixture. *Bioinformatics,* 20**,** 1905-1913.

Quackenbush, J. 2001. Computational Analysis for Microarray Data. *Nature Reviews: Genetics,* 2**,** 418-427.

R Development Core Team 2016. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rabbany, R. & Zaiane, O. R. 2015. Generalization of Clustering Agreements and Distances for Overlapping Clusters and Network Communities. *Data Mining and Knowledge Discovery,* 29**,** 1458-1485.

Rand, W. M. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association,* 66.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. & Betel, D. 2013. Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data. *Genome Biology,* 14.

Reeb, P. D., And Steibel, J.P. 2013. Evaluating Statistical Analysis Models for RNA Sequencing Experiments. *Frontiers in Genetics*.

Rencher, A. C. & Christensen, W. F. 2012. *Methods of Multivariate Analysis,* Hoboken, New Jersey, John Wiley & Sons, Inc.

Robinson, M. D., Mccarthy, D. J. & Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

Robinson, M. D. a. S., G. K. 2007. Moderated Statistical Tests for Assessing Differences in Tag Abundance. *Bioinformatics,* 23**,** 2881-2887.

Robinson, M. D. a. S., G. K. 2008. Small-sample Estimation of NEgative Binomial Dispersion, with Application to SAGE data. *Oxford Journals: Science and Mathematics,* 9.

Robles, J. a. Q., S. E., Stephen, S.J.; Willson, S.R.; Burden, C.J.; Talyor, J.M. 2012. Efficient Experiemental Design and Analysis Strategies for the Detection of Differential Expression Using RNA-Sequencing. *BMC Genomics,* 13.

Rokach, L. & Maimon, O. 2005. Clustering Methods. *Data Mining and Knowledge Discovery Handbook.* US: Springer.

Rozencweig, M., Nicaise, C., Beer, M., Crespeigne, N., Van Rijmenant, M., Lenaz, L. & Kenis, Y. 1983. Phase I Study of Carboplatin Given on a Five-day Intravenous Schedule. *Journal of Clinical Oncology,* 1**,** 621-626.

Sanger, F. 1980. Fredrerick Sanger--Biographical.

Schena, M., Shalon, D., Davis, R., Brown, P., 1995. Quatitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science,* 270**,** 467-470.

Seyednasrollah, F., Laiho, A., and Elo, L.L. 2013. Comparison of Software Packages for Detecting Differential Expression in RNA-Seq Studies. *Briefings in Bioinformatics*.

Shannon, W., Culverhouse, R., and Duncan, J., 2003. Analyzing Microarray Data Using Cluster Analysis. *Pharmacogenomics,* 4**,** 41-52.

Shen, T., Pajaro-Van De Stadt, S. H., Yeat, N. C. & Lin, J. C. H. 2015. Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Frontiers in Genetics,* 6**,** 215.

Si, Y., Liu, P., Li, P., and Brutnell, T.P.L 2013. Model-Based Clustering for RNA-Seq Data. *Bioinformatics*

Sibru, A., Kerr, G., Crane, M. & Ruskin, J. 2012. RNA-Seq vs. Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering. *PLOS ONE*.

Sijmons, R. H. 2010. Identifying Patients with Familial Cancer Syndromes. *Cancer Syndromes*.

Smyth, G. K. 2005. Limma: Linear Models for Microarray Data. *In:* GENTLEMAN, R. C., V.J.; HUBER, W.; IRIZARY, R.A.; DUDOIT, S. (ed.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* New York: Springer.

Soneson, C. & Delorenzi, M. 2013. A Comparison of Methods for Differential Expression Analysis of RNA-Seq data. *BMC Bioinformatics,* 14.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffery, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E. & Borresen-Dale, A. 2001. Gene Expression Patterns of Breast Carcinomas Distinguish Tumore Subclasses with Clinical Implications. *PNAS,* 98**,** 10869-10874.

Storey, J. D. 2010. False Discovery Rates. *In:* UNIVERSITY, P. (ed.). Princeton.

Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. 2015. qvalue: Q-value estimation for false discovery rate control. R package version 2.1.1.

Strahm, B. & Malkin, D. 2006. Hereditary Cancer Predisposition in Children: Genetic Basis and Clinical Implications. *International Journal of Cancer,* 119**,** 2001-2006.

Taccioli, C., Sorrentino, G., Zannini, A., Caroli, J., Beneventano, D., Anderlucci, L., Lolli, M., Bicciato, S. & Del Sal, G. 2015. MDP, a database linking drug response data to genomic information, identifies dasatinib and statins as a combinatorial strategy to inhibit YAP/TAZ in cancer cells. *Oncotarget,* 6**,** 38854-38865.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X. & Tseng, G. C. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics,* 22**,** 2405-12.

The Cancer Genome Atlas 2010. Cancer Genomics: What Does if Mean for You? : National Institutes of Health.

Tibshirani, R., Walther, G., and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society,* 63.

Townsend, A., Adam, S., Birch, P. H., Lohn, Z., Rousseau, F. & Friedman, J. M. 2012. "I Want to Known What's in Pandora's Box": Comparing Stakeholder Prespectives on Incidental Findings in Clinical Whole Genomic Sequencing. *American Journal of Medical Genetics Part A,* 158A**,** 2519-2525.

Trabzuni, D., (Ukbec);, T. U. K. B. E. C. & Thomson, P. C. 2014. Analysis of Gene Expression Data Using a Linear Mixed Model/Finite Mixture Model Approach: Application to Regional Differences in the Human Brain. *Bioinformatics,* 30**,** 1555-1561.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. & Pachter, L. 2013. Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq. *Nature Biotechnology,* 31.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelly, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. 2012. Differential Gene and Transcript Expression Analysis of Rna-Seq Experiments with TopHat and Cufflinks. *Nature Protocol,* 7**,** 562-578.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. 2014. Ten Years of Next-Generation Sequencing Technology. *Trends in Genetics,* 30**,** 418-426.

Vicini, P., Fields, O., Lai, E., Litwack, E. D., Martin, A.-M., Morgan, T. M., M.A.;, P., Papaluca, M., Perez, O. D., Ringel, M. S., Robson, M., Sakul, H., Vockley, J., Zaks, T., Dolsten, M. & Sogaard, M. 2016. Precision Medicine in the Age of Big Data: The Present and Future Role of Large-Scale Unbiased Sequencing in Drug Discovery and Development. *Clinical Pharmacology and Theraputeics,* 99**,** 198-207.

Walsh, T., Lee, M. K., Casadei, S., Thornton, A. M., Stray, S. M., Pennil, C., Nord, A. S., Mandell, J. B., Swisher, E. M. & King, M. 2010. Detection of Inherited Mutations for Breast and Ovarian Cancer using Genomic Capture and Massively Parallel Sequencing. . *PNAS.*

Wang, Z., Gerstein, M. & Snyder, M. 2009. RNA-Seq: A Revoluntionary Tool for Transcriptomics. *National Review of Genetics,* 10**,** 57-63.

Watson, J. D. & Crick, F. H. C. 1953. Molecular Structure of Nucleic Acids. *Nature,* 171**,** 737-738.

Weinstien, J. N., Et. Al. 1997. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science,* 275.

Wilkins, M. H. F. 1963. Molecular Configuration of Nucleic Acids. *Science,* 140**,** 941-950.

Witten, D. M. 2011. Classification and Clustering of Sequencing Data Using a Poisson Model. *The Annals of Applied Statistics,* 5**,** 2493-2518.

Wolf, S. M., Crock, B. N., Van Ness, B., Lawrenz, F., Kahn, J. P., Beskow, L. M., Cho, M. K., Christman, M. F., Green, R. C., Hall, R., Illes, J., Keane, M., Knoppers, B. M., Koenig, B. A., Kohane, I. S., Leroy, B., Maschke, K. J., Mcgeveran, W., Ossorio, P., Parker, L. S., Petersen, G. M., Richardson, H. S., Scott, J. A., Terry, S. F., Wilfond, B. S. & Wolf,

W. A. 2012. Managing Incidental Findings and Research Results in Genomic Research Involving Biobanks & Archived Datasets. *Genetics in medicine : official journal of the American College of Medical Genetics,* 14**,** 361-384.

World Health Organization^2 2002. Genomics and World Health: Report of the Advisory Committee of Health Research. Geneva.

World Health Organization 2014. World Cancer Report 2014. *In:* STEWART, B. W. & WILD, C. P. (eds.).

Yahav, I. & Shmueli, G. 2011. On Generating Multivariate Poisson Data in Management Science Applications. *Applied Stochastic Models in Business and Industry,* 28**,** 91-102.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoor, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., ;, Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., Mcdermott, U. & Garnett, M. J. 2012. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic *Nucleic Acids Research,* 41**,** D955-D961.

Yee, T., And Hastie, T.J. 2003. Reduced-rank Vectore Generalized Linear Models. *Statistical Modeling*, 3**,** 15-41.

Yee, T. W., And Wild, C. J., 1996. Vector Generalized Additive Models. *Journal of the Royal Statistical Society*, B**,** 481-493

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L., 2001. Model-Based Clustering and Data Transformations for Gene Expression Data. *Bioinformatics,* 17**,** 977-987.

Your Genome. 2016. *What is the 'Central Dogma'?* [Online]. Available: http://www.yourgenome.org/facts/what-is-the-central-dogma [Accessed 7/24/2016].

Zawati, M. N. H. & Knoppers, B. M. 2012. International normative perspectives on the return of individual research results and incidental findings in genomic biobanks. *Genet Med,* 14**,** 484-489.

Zhao, S., Fund-Leung, W., Bittner, A., Ngo, K. & Liu, X. 2014. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE,* 9.

Zwiener, I., Frisch, B. & Binder, H. 2014. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One,* 9**,** e85150.