INTEGRATIVE DATA ANALYSIS OF MICROARRAY AND RNA-SEQ

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Qi Wang

In Partial Fulfillment of the Requirements
of the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Statistics

June 2018

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Integrative Data Analysis of Microarray and RNA-seq

By

Qi Wang

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Seung Won Hyun
Chair

Dr. Ke (Kurt) Zhang

Dr. Rhonda Magel

Dr. Shaobin Zhong

Approved:

6/29/2018
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Background: Microarray and RNA sequencing (RNA-seq) are two commonly used high-throughput technologies for gene expression profiling for the past decades. For global gene expression studies, both techniques are expensive, and each has its unique advantages and limitations. Integrative analysis of these two types of data would provide increased statistical power, reduced cost, and complementary technical advantages. However, the complete different mechanisms of the high-throughput techniques make the two types of data highly incompatible.

Methods: Based on the degrees of compatibility, the genes are grouped into different clusters using a novel clustering algorithm, called Boundary Shift Partition (BSP). For each cluster, a linear model is fitted to the data and the number of differentially expressed genes (DEGs) is calculated by running two-sample t-test on the residuals. The optimal number of cluster can be determined using the selection criteria that is penalized on the number of parameters for model fitting. The method was evaluated using the data simulated from various distributions and it was compared with the conventional K-means clustering method, Hartigan-Wong's algorithm. The BSP algorithm was applied to the microarray and RNA-seq data obtained from the embryonic heart tissues from wild type mice and *Tbx5* mice. The raw data went through multiple preprocessing steps including data transformation, quantile normalization, linear model, principal component analysis and probe alignments. The differentially expressed genes between wild type and *Tbx5* are identified using the BSP algorithm.

Results: The accuracies of the BSP algorithm for the simulation data are higher than those of Hartigan-Wong's algorithm for the cases with smaller standard deviations across the five different underlying distributions. The BSP algorithm can find the correct number of the clusters using the selection criteria. The BSP method identifies 584 differentially expressed genes between the wild type and *Tbx5* mice. A core gene network developed from the differentially expressed genes showed a set of key genes that were known to be important for heart development.

Conclusion: The BSP algorithm is an efficient and robust classification method to integrate the data obtained from microarray and RNA-seq.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF APPENDIX TABLES

CHAPTER 1. INTRODUCTION

1.1. Global Gene Expression Profiling

Gene expression profiling is a way to measure the activity of multiple genes simultaneously, sometimes even the entire genome. In most cases, gene expression profiling is used to distinguish the expression levels of the genes between the control group and the treatment group. Microarray and RNA sequencing (RNA-seq) are two commonly used technologies for gene expression profiling.

In the last two decades, microarrays have been widely used and have become a standard tool for biological research. The cost of microarray is relatively cheap due to the technology development and the availability of commercial platforms. The microarray technology uses the hybridization between specific DNA sequences (known as probes) and target cDNA samples to measure the expression levels of target genes. In the National Center for Biotechnology Information (NCBI) PubMed website, there are 102,310 related articles as a result of searching the keyword “microarray” (June 2018). The oldest article was published in 1992. The number of publications increased dramatically from 296 in the year 2000 to 6,418 in the year 2017 (PubMed, 2018).

Microarray technology provides an effective way to study gene expression levels of a large number of genes simultaneously. However, there are some limitations while using this technology. First, the probe design on the microarray highly relies on the existing knowledge of the genome sequence. Second, the background noise of the signals is usually high. Lastly, since there are different platforms available from many commercial suppliers, gene expression levels obtained from different platforms can not be directly compared.

In the last decade, a sequence-based approach (RNA sequencing) has been developed and became popular. There are 14,668 published articles at PubMed as a result of searching the keywords “RNAseq” or “RNA-seq” in June 2018, with the oldest publications in 2008. While in 2017 there were already 3,644 published papers related to RNA-seq (PubMed, 2018). RNA-seq technology uses next-generation sequencing (NGS) technology to sequence short cDNA fragments (called reads) transcribed from mRNA. It can estimate the gene expression by counting the number of reads mapped to the gene. One advantage of this technology is that it is very useful for non-model organisms because it does not require prior knowledge of the genome sequence. Also, it has low background signals and is highly accurate for estimating gene expression levels. Furthermore, it has the ability to distinguish isoforms and allelic expression (Wang, Gerstein, & Snyder, 2009). Some of the disadvantages of RNA-seq includes relatively high cost compared to microarray, complicated data analysis, lack standard protocol, and etc.

Since both microarray and RNA-seq have been widely used in recent years, the gene expression data of many organisms obtained from both technologies are available. The goals of this study were to find an effective way to integrate data from both microarray and RNA-seq and increase the power of statistical testing. The data integration between microarray and RNA-seq will be helpful to find differentially expressed genes in the genome.

1.2. Microarray Data

The microarray technology is based on a simple property of DNA: the two strands of DNA can be separated in heat and restored to form the double helical structure in low temperature. This property is the foundation for DNA hybridization. Since the microarray technology was developed in the late 1980s, it has been used in large-scale gene studies of gene expression profiling (Liu, et al., 2013), single nucleotide polymorphism (SNP) detection (Jacob, et al., 2015), alternative splicing detection (Kamtchueng, et al., 2014), fusion gene detection (Løvf, et al., 2013), and so on.

The microarray technology can be broadly divided into two categories based on fabrication: spotted (cDNA) microarray (Schena, Shalon, Davis, & Brown, 1995) (DeRisi, et al., 1996) and oligonucleotide in situ (such as Affymetrix) microarray (Fodor, et al., 1991). In spotted microarray, the pre-designed and synthesized probes are spotted onto glass and will hybridize to their complementary cDNA targets. Fluorescent signals will be generated during the hybridization. While in oligonucleotide in situ microarray, photolithographic synthesis is used to generate probes one nucleotide at a time. Due to the high reproducibility and ease of construction, Affymetrix microarray has become widely used by more and more researchers.

A microarray experiment should start with the experimental design. There are two major elements that need to be considered while designing a microarray experiment: replicates and sample size. Replicates, especially biological replicates, are essential for making conclusions of treatment effects (Churchill, 2002). Technical replicates are also necessary for some cases, such as quality-control studies (Allison, Cui, Page, & Sabripour, 2006). Sample size also plays an important role in microarray design. A larger sample size would provide more power for statistical analysis, which is always recommended (Zien, Fluck, Zimmer, & Lengauer, 2004; Pawitan, Michiels, Koscielny, Gusnanto, & Ploner, 2005; Zehetmayer, Graf, & Posch, 2015).

Batch effects should also be considered while constructing microarray experiments. Luo, et al (2010) define batch effects as the systematic biases between batches (samples) in microarray analysis. There are various causes for batch effects: platform differences between samples; experimental procedure differences between laboratories; differences based on the equipment used in the experiments; sample collection/storage conditions and so on (Luo, et al., 2010). Batch effects would introduce unwanted variability into data. Thus, minimizing or even eliminating such batch effects is crucial for microarray experiments.

After obtaining the raw data, which is image signals from microarray experiments, a preprocessing stage is often required before making any statistical inference. The preprocessing stage of the microarray data analysis refers to normalization, data filtering, and transformation (Allison, Cui, Page, & Sabripour, 2006).

Normalization is usually the first thing to do after the image signals of microarray are converted to expression values. The gene expression values of the control genes by experiment design are treated as constant since their expressions should not change among treatment conditions. One example of the control genes is housekeeping genes. Then a global normalization can be performed among different samples and sometimes even platforms (Bilban, Buehler, Head, Desoye, & Quaranta, 2002). Bolstad, et al use three normalization methods, cyclic loess, contrast-based method, and quantile normalization, to reduce the expression variation across arrays (Bolstad, Irizarry, Åstrand, & Speed, 2003).

Data filtering is the process of removing the probes with a low expression percentage. For instance, the m/n filter removes the genes whose number of expression in the samples was less than m among a total of n microarray chips (Pounds & Cheng, 2005). This is a way to control the random or technical errors and to make sure that the differentially expressed genes are due to treatment effects (Gusnanto, Calza, & Pawitan, 2007).

Transformation of the expression values is usually necessary since normalization is one of the assumptions for many traditional statistical methods such as linear model, two-sample t-test, and the analysis of variance (ANOVA). A common transformation used in the microarray study is the logarithm transformation (Rocke & Durbin, 2003), which uses the log value of the expressions as the dependent variable in statistical analysis. With a constant variance after transformation, many statistical analyses can be conducted.

The property of microarray data can be described in a few areas: sensitivity, specificity, reproducibility, and accuracy (Draghici, Khatri, Eklund, & Szallasi, 2006). Sensitivity refers to the ability of the probes hybridizing with the targets at a low concentration rate. The minimum detection limit for microarray is approximately two to ten copies per cell (Holland, 2002; Kane, et al., 2000). Specificity defines the ability of a probe to hybridize to a specific target and discern between similar sequences. The length of a probe is one of the factors that influences specificity (Jayaraman, Hall, & Genzer, 2006). Shorter probes (~25 bp) have higher specificity than the longer probes (60 bp), but lower sensitivity (Relógio, Schwager, Richter, Ansorge, & Valcárcel, 2002). The length of the microarray probes is suggested to be about 150 base pairs (Chou, Chen, Lee, & Peck, 2004). Reproducibility measures the ability of the technology to achieve the same or similar results under repeated measurements. The reproducibility rate for Affymetrix microarray is shown to be around 80-90% concordance for experiments conducted within one facility (MAQC Consortium, et al., 2006). The correlation across different microarray platforms is between 0.7 and 0.8 (David, et al., 2005). Accuracy describes how the measured quantity agrees with the true value. Traditionally, the two-channel cDNA microarray has a higher accuracy than the single-channel oligonucleotide microarray since cDNA microarray measures expression ratio while oligonucleotide microarray measures absolute transcript concentrations (Czechowski, Bari, Stitt, Scheible, & Udvardi, 2004). The cost for microarray is at least $100 per sample (Mcloughlin, 2011), normally in the range of $100 to $300.

Microarray was developed in the 1990s and soon became a hot topic for gene expression profiling due to its low cost and high throughput. There are many well-established software packages available for microarray, which makes the microarray data analysis easy to perform. Microarray also have some limitations. Since the probes are pre-designed, it has limited use on the detection of the alternative splicing. The probe design requires sequence information of the genome. Microarray can not be used if the sequence of the genome is unknown. Also, researchers find out that there exists a widespread spatial bias in the probes and targets hybridization, which means that the probe spot position in the microarray actually affects the probe-targets hybridization (Steger, et al., 2011).

1.3. RNA-seq Data

The RNA-seq technology utilizes the NGS technology which was developed from the first-generation Sanger sequencing (Sanger & Coulson, 1975; Sanger, Nicklen, & Coulson, 1977). Sanger sequencing is sometimes called chain-termination sequencing with a ‘plus and minus’ system. The principle used in Sanger sequencing is that the modified di-deoxynucleotidetriphosphates (ddNTPs) can not form phosphodiester bonds with other nucleotides like the normal deoxynucleosidetriphosphates (dNTPs), thus the DNA strand that ends with ddNTPs can not be extended by a DNA polymerase. The first DNA genome, bacteriophage phi X174 (or ΦX174), was obtained using Sanger sequencing (Sanger, et al., 1977).

The NGS is superior to Sanger sequencing in the aspects of high speed, high throughput, dynamic range, and reduced cost. The first NGS instrument was launched by 454 Life Sciences in 2005 (Margulies, et al., 2005). The sequencing mechanism of 454 is pyrosequencing, which is based on the sequencing by synthesis principle. Three major commercial NGS systems are available: Roche/454, Illumina, and SOLiD. Since Illumina is widely used with a better coverage and depth than other sequencing technologies at a fixed cost, let's use Illumina sequencing as an example to introduce the NGS technology.

The process of Illumina NGS technology includes four steps: library preparation, cluster generation, sequencing, and data analysis (Illumina, 2017). In the library preparation step, the sample DNA is randomly fragmented, and adapters are ligated to both ends of the fragments. Then these single-stranded fragments randomly bind to their complementary adapters that are immobilized on the surface of the flow cell. Bridge amplification is used to form a clonal cluster for each bound fragment. The sequence by synthesis technology used by Illumina was developed from Sanger sequencing. During each sequencing cycle in Illumina, one of the four fluorescently-labeled dNTPs is added to the nucleic acid chain. The fluorescence emission is captured after laser excitation. The nucleic bases are determined by the fluorescent wave length and intensity. The RNA-seq data analysis process, including Illumina sequencing data, is introduced as follows.

The raw data obtained from next-generation sequencing, including RNA-seq, is the sequence of the DNA fragments, called read. Raw reads in RNA-seq refer to the short sequences obtained directly from the experiments. Most of the time, raw reads are stored in the FASTQ format, which records both the nucleotide sequences and the corresponding quality scores (Cock, Fields, Goto, Heuer, & Rice, 2010). The raw reads normally go through a quality control process to ensure that the reads are of decent quality from the experiments for downstream analysis. Quality control methods include the base/sequence quality, GC content, sequence duplication levels, and Kmer content (Andrews, 2016).

Once the raw reads have been verified to have good quality, they can be mapped to the reference genome or transcriptome. This process is called sequence alignment or mapping. This is a crucial step in the preprocessing stage because the accuracy of the alignment affects the gene expression levels. The read alignment to the reference is a process of similarity search (Wang, 2013).

The mapped reads need to be quantified after read alignment to estimate the gene expression levels. One straightforward way to quantify the reads is using the number of alignments (counts) to represent the expression level of each gene. Thus, if some reads have multiple alignments in the genome, expression accuracy for these genes is relatively low compared to others that have unique alignments.

After getting the count for each gene, statistical inference of the data can be made by assuming the count data follows Poisson distribution or Negative Binomial distribution. Several methods can also be used to normalize gene expression levels: Reads Per Kilobase per Million reads (RPKM), Fragments Per Kilobase of exon per Million fragments mapped (FPKM), or Transcripts Per Million (TPM). RPKM (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) only considers the reads that can be mapped to either known exons or candidate exons based on the NCBI gene models. TPM is the fraction of transcripts for an isoform (Li, Ruotti, Stewart, Thomson, & Dewey, 2010). FPKM is very similar to RPKM and can be used for calculating the gene expression levels (Trapnell, et al., 2010). The only difference is that FPKM is used for the paired-end reads instead of the single-end reads in RPKM. FPKM counts the two paired-end reads as one fragment. Because the TPM value is a measure of fraction and is not highly related with the number of reads in the library, it is believed to be more comparable than RPKM or FPKM between samples from different experiments (Conesa, et al., 2016).

In general, the quality of RNA-seq data is better than that of microarray data. The performances of Illumina sequencing and Affymetrix microarray have been compared by several studies. Marioni et al showed that the reproducibility of RNA-seq is higher than microarray with less technical variation based on the Poisson model (Marioni, Mason, Mane, Stephens, & Gilad, 2008). Fu et al compared the absolute transcript measurements between microarray and RNA-seq and evaluated their accuracy by shotgun mass spectroscopy quantification (Xing, et al., 2009). RNA-seq is more accurate than microarray with regards to measuring the absolute transcript level. The inter-site reproducibility for RNA-seq is 95%, which is higher than microarray (SEQC/MAQC-III Consortium, 2014). RNA-seq has a wider dynamic range than microarray, which provides better sensitivity (Sîrbu, Kerr, Crane, & Ruskin, 2012; Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014).

RNA-seq utilizes NGS technology to measure the gene expression levels. Since it has RNA sequence as a result, it can be used for splicing variant detection. Unlike microarray, RNA-seq does not require genome information as input. RNA-seq can be used to create a transcriptome through *De novo* assembly. Even though RNA-seq has a wider dynamic range and better sensitivity than microarray, it still has some disadvantages. One of them is its relatively high cost than microarray. Currently, the cost of RNA-seq per sample is about ten times higher than that of microarray. The RNA-seq technology has sequencing errors, which means that the sequence of the reads is not always 100% accurate. The sequencing errors can be divided into three categories depending on their cause: position-specific errors, sequence-specific errors, and systematic errors (Meacham, et al., 2011). These sequencing errors would cause problems for accurate SNP detection. Since most of the software designed for RNA-seq are Linux-based with scripts written in different programming languages, researchers normally need to have special bioinformatic training for RNA-seq data analysis. Due to the huge amount of raw data obtained from RNA-seq (> 5 GB), it requires a larger computer resources to process and store the data.

Even though there are plenty of software and packages available for RNA-seq data analysis, there is not yet one standard protocol.

1.4. Motivation and Goals of Study

Since invention of the microarray technology in the 1990s, it has been widely used by researchers in large-scale studies for gene expression profiling. The relatively low cost made microarray affordable by many laboratories. There are well-established methods for microarray data analysis. But there are some disadvantages with microarrays. For instance, microarray requires pre-designed probes, thus only the expression level of those genes is measured. Also, unspecific hybridization reduces the measurement accuracy in microarray and causes a high background noise level.

The RNA-seq technology was developed about a decade ago and it soon became a preferred method for gene expression profiling by many researchers. Even though the expression levels measured by RNA-seq are more reliable than those measured by microarray, the relative high cost of RNA-seq limits the number of samples used in each RNA-seq experiment. With a smaller sample size, the power of the downstream statistical tests would be lower and false conclusions might be made because of it. Moreover, by removing the technology effects in the process of microarray and RNA-seq data integration, the artifacts by individual labs would also be removed.

There is an enormous amount of publications on gene expression data sets using either the microarray or RNA-seq technology available online. NCBI's Gene Expression Omnibus (GEO) website contains expression data from over 70,000 experiments (Edgar, Domrachev, & Lash, 2002). There are 423 series related to the heart tissue of *Mus musculus* in GEO (June 2018). Most of the series published prior to 2014 were conducted using microarray. RNA-seq has been more frequently used in recent years. For some laboratories, it is possible that their research was conducted using microarray, later switching to RNA-seq because of better performance. Thus, data integration of microarray and RNA-seq would be useful to better utilize the existing information without excess cost. Besides, the increase of sample size by data integration would increase the reliability of the statistical inference. However, some challenges remain for data integration of microarray and RNA-seq.

First, the exist of the batch effects makes the data integration harder. Here is an example to better understand the question. Suppose some researchers want to study the development of the embryonic heart in mice to find the genes that play important roles in this process. There are two data sets that are available in GEO: GSE73544 (Nie, et al., 2015) and GSE66965 (Wei, 2015). GSE73544 has microarray gene expression data from embryonic day (E) 12.5 wild type (WT) mouse heart. GSE66965 contains the RNA-seq gene expression data from E13.5 WT mouse heart. Since these two data sets were generated from different technologies, they can not be simply combined to conduct statistical analysis. These two experiments were completed by different labs in different locations at different times. These factors can all affect gene expression measurements and introduce unwanted variation into the data.

Secondly, the expression values measured by microarray and RNA-seq have different distributions. There is a magnitude problem which means that the range of expression values from one technology is not consistent with the range of the values from the other technology. Therefore, direct comparisons between the two technologies will not work since the expression value of 2 in microarray does not equal to the same expression value of 2 in RNA-seq. In microarray, the expression values of the genes rely on the binding between the cDNA/cRNA targets and the pre-designed probes. Following hybridization between targets and probes, the detectable fluorescent signal gets transformed to digital values, which represents the gene expression levels. Thus, the expression levels in microarray are believed to follow a normal distribution. But the raw data in RNA-seq are counts, which are non-negative integers. Poisson distribution and negative binomial distribution are commonly used to describe the distribution of count data. Sometimes, the count data are further normalized using RPKM, FPKM, or TPM.

Thirdly, the microarray and RNA-seq have different biases. A systematic spatial bias exists in microarray probe-target hybridization, which is caused by the lateral diffusion (Steger, et al., 2011). Research also shows that the amplification bias caused by long probe – ploy(A) – tail distance largely influences the number of differentially expressed genes (DEGs) detection (Wim, et al., 2007). The biases in RNA-seq includes GC bias and sequencing errors. GC bias is caused by the over-representation of the GC-rich sequences over AT-rich sequences in Illumina (Benjamini & Speed, 2012). The sequencing errors might be caused by the location of the reads, sequence of the reads, or genomic position (Meacham, et al., 2011).

In this dissertation, a novel method for data integration of microarray and RNA-seq, Boundary Shift Partition (BSP) algorithm, is proposed and applied to E9.5 embryonic heart expression data collected from wild type and *Tbx5* mice.

1.5. Data Integration Methods

Because of technology development, various types of genomic data from different sources have become available and have been applied in the fields of functional genomics (Evangelistella, et al., 2017; Chudasama, et al., 2018), epigenomics (Laird, 2010; Bien, et al., 2017; Zhang, et al., 2017), metagenomics (Tringe, et al., 2005; Rodriguez-Brito, Rohwer, & Edwards, 2006; Abayasekara, et al., 2017), and so on. The genomic studies include differentially expressed gene analysis, single nucleotide polymorphism (SNP), copy number variation (CNV), gene co-expression networks, etc. Each of these studies provides the scientists with a different angle for understanding the mystery of the whole genome. With the increasing amount of genomic data, data integration becomes a helpful tool for genomic analysis as it could increase the statistical power by using a larger sample size and could be used for cross validation.

The data analysis using different platforms or results is difficult without universal standards or controls. This phenomenon in microarray or RNA-seq data analysis is sometimes called the “Tower of Babel.” In other words, the comparison of the gene expression results obtained using different platforms or technologies is not an easy task. The expression values might represent different expression levels of the genes even though the values are the same from different analysis.

This section focuses on the data integration of genomic data from a statistical perspective, mainly the gene expression data obtained from microarray and RNA-seq technologies. The purpose of this section is to review some of the algorithms and methods for combining gene expression data. Currently, most of the algorithms for data integration are within a technology, such as combining two microarray data sets. For the analysis including both microarray and RNA-seq, most researchers analyzed the expression data separately to find the DEGs and then compared results. For instance, a *Drosophila melanogaster* embryo development study (Sîrbu, Kerr, Crane, & Ruskin, 2012) involved three expression data sets: RNA-seq, single-channel, and dual-channel microarrays. The Limma (Ritchie, et al., 2015) package in R was used to perform the differential expression analysis for the two microarray data sets and the DESeq (Anders & Huber, 2010) package was used to find DEGs in the RNA-seq data set. A similar DEGs comparison study between technologies was conducted using rats with exposed chemicals (Wang, et al., 2014). The DEGs were identified in each data set before they were compared and analyzed among the data sets. Few methods have been designed for analyzing combined data sets from microarray and RNA-seq.

Various methods of data integration for both within technologies and between technologies are discussed in this section. This review could be useful for improving methodologies of data integration for genomic data and help researchers identify their research directions.

The methods for data integration can be broadly divided into four categories: data transformation, Location-Scale (LS) methods, Matrix Factorization (MF) methods, and model-based integration (Hamid, et al., 2009; Lazar, et al., 2012; Ritchie, et al., 2015).

1.5.1. Data Transformation

Data integration based on transformation refers to transforming part or all of the expression data obtained from different sources so that all data will follow the same distribution after transformation.

The print-tip loess normalization is used to normalize the log-ratios of the gene expression levels from the two-color cDNA microarray (Smyth & Speed, 2003). The print-tip loess normalization can adjust both the spatial and intensity trends in the data. Composite loess normalization can be used when control spots are available.

Wang et al use several transformation methods including normalization transformation and global median transformation to integrate 2,968 expression profiles of 131 microarray studies obtained from NCBI GEO website (Wang, Srivastava, & Schwartz, 2010). Tissue-sensitive genes are identified using the integrated data.

A Training Distribution Matching (TDM) approach is developed to normalize RNA-seq data so that the transformed RNA-seq data would have a similar distribution to the microarray data (Thompson, Tan, & Greene, 2016). TDM normalizes RNA-seq data using the quantile information of the data set and ensures the transformed RNA-seq data will fall in the same range as the microarray log2 transformed data. The TDM method is applied to breast cancer data sets for unsupervised and supervised classification using Partitioning Around Medoids (PAM) (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2015) and LASSO multinomial logistic regression (Friedman, Hastie, & Tibshirani, 2010), respectively. As a result, TDM performs well compared to the quantile normalization, nonparanormal transformation, and log2 transformation on a range of data.

An expression visualization and integration platform (expVIP) is developed to combine RNA-seq data sets of a crop species for DEG analysis (Borrill, Ramirez-Gonzalez, & Uauy, 2016). The study integrates the reads of 418 wheat samples from 16 RNA-seq data sets. expVIP requires the raw reads, the reference genome, and metadata of the experiments as input. The reads are quality controlled using fastQC (Andrews, 2016) and quantified using kallisto (Bray, Pimentel, Melsted, & Pachter, 2016). sleuth (Pemental, Bray, Puente, Melsted, & Pachter, 2017) is used for differential gene expression analysis, which utilizes the kallisto quantifications and bootstraps.

One of the disadvantages for data transformation method is that the batch effects are not considered. Therefore, the normalized data would still have the batch effects confounded with the expression values.

1.5.2. Location-Scale Methods

LS methods use a collection of techniques to transform the original expression data from different batches in a similar range in terms of equal mean (location) and/or variance (scale).

After transformation, the expression values of the genes from different batches should be comparable and can be used in the downstream analysis.

One of LS methods is Batch Mean Centering (BMC). It transforms the original data by simply subtracting the mean value of a given gene from each sample to remove the batch effects (Sims, et al., 2008). Shen, Ghosh and Chinnaiya (2004) develop a two-stage Bayesian mixture modeling strategy to convert the original data into [-1, 1]. Over the past decade, some more complex LS methods, such as the Empirical Bayes (EB) method (also known as Combat) (Johnson, Li, & Rabinovic, 2007), the Cross-Platform Normalization (XPN) method (Shabalin, Tjelmeland, Fan, Perou, & Nobel, 2008), and Distance Weighted Discrimination (DWD) (Marron, Todd, & Ahn, 2007; Huang, Lu, Liu, & Marron, 2012), have been widely used in batch effect removal methods comparison. The details of these three LS methods will be introduced in the following paragraphs.

The EB method, developed by Johnson, et al (2007), first standardizes the genes using the least-squares approach so that the expression data would have a similar mean and variance. The standardized data is assumed to have a normal distribution, and the parameters can be estimated using Bayesian approach. In the end, the EB batch-adjusted data is calculated. The EB method is more robust and does not require as many samples as DWD.

The first step in the XPN algorithm (Shabalin, Tjelmeland, Fan, Perou, & Nobel, 2008) is median centering and standardizing the gene expression values in each sample to remove batch effects. Then a block linear model can be applied to the combined data using K-means clustering. Model parameters can be estimated using maximum likelihood estimations. Finally, the batch effect adjusted values can be calculated using estimations of the model parameters. One limitation of XPN is that it can only be used to analyze microarray expression data from two batches.

The DWD method is a margin-based classification, which is a modification of the Support Vector Machine (SVM) (Vapnik V. N., 1995; Vapnik, Golowich, & Smola, 1997). The main idea of the DWD method is to find the optimal hyperplane that maximizes the projected distance of all the data on this hyperplane (margin) (Marron, Todd, & Ahn, 2007; Huang, Lu, Liu, & Marron, 2012). Like XPN method, DWD can only analyze data from two batches at a time. A stepwise DWD has been developed to compare data in three batches (Benito, et al., 2004).

Granatum is a software for analyzing Single-cell RNA sequencing (scRNA-Seq) data (Zhu, et al., 2017). Granatum uses ComBat (Johnson, Li, & Rabinovic, 2007) and median alignment to remove batch effects from data sets of the normalized expression values before differential expression analysis.

Even though the transformed expression values would have similar mean (location) and standard deviation (scale) using LS method, the normalized distributions are not guaranteed.

1.5.3. Matrix Factorization Methods

The MF based methods remove the most important variation from the data set under the assumption that differences across batches bring more variation on the expression data than differences on biological groups. The Singular Value Decomposition (SVD) (Alter, Brown, & Botstein, 2000) and Principal Component Analysis (PCA) are the two commonly used methods for matrix factorization. Normally, the vector/principal component that contains the highest variation of the data is removed and the result is the batch effect adjusted expression data. Some MF methods will be introduced in the following paragraphs.

The Surrogate Variable Analysis (SVA) is one of the widely used MF method for batch effect removal (Leek & Storey, 2007). The first step of SVA is to detect the batch effects using SVD iteratively on the residual matrix to remove any structure. A weight is given to every gene which represents the signature significance of the expression heterogeneity. Then the surrogate variables can be constructed using the probability weights and SVD on the reduced expressed matrix.

The Frozen Surrogate Variable Analysis (fSVA) is an adjustment of SVA to improve prediction accuracy in microarray analysis (Parker, Bravo, & Leek, 2014). The fSVA first uses SVA for batch effects correction. Then batch effects in new samples would be removed using the results from SVA. Sample prediction can be applied using the classifier that was obtained within these batch effects removal samples.

The RUV-2 (Remove Unwanted Variation, 2-step) algorithm (Gagnon-Bartsch & Speed, 2012) uses the same linear model of SVA to remove the batch effects. RUV-2 applies the factor analysis on the negative control genes since they are believed to be unassociated with the interested genes. Thus, it makes sure that the biological effects of interest would not be removed along with the batch effects from the original data. The performance of the RUV-2 method is showed to be comparable to Combat and SVA when applied to several data sets.

An updated version of RUV-2 (Jacob, Gagnon-Bartsch, & Speed, 2016) can be used to remove batch effects while the genes of interest are unobserved. Since the expression values of the replicate samples should only be affected by the unwanted batch effects, the new RUV-2 method uses replicate samples to estimate and remove the batch effects.

The thresholding singular value decomposition (T-SVD) regression method is to be used for the prediction of microRNA (miRNA)-gene regulation and long noncoding RNA (lncRNA)-gene regulation (Ma, Xiao, & Wong, 2014). First, a sparse but not orthogonal matrix is calculated using a thresholding-based regularized multivariate regression. Then the Sparse Orthogonal Decomposition Algorithm (SODA) is applied to the matrix to make it orthogonal while maintaining its sparsity.

The MF method works better for the data obtained from similar sources. In another word, the expression values from different samples have the same distribution. If the samples have completely different distributions, like microarray log-fold change data and RNA-seq count data, the MF method is not applicable.

1.5.4. Model-based Integration

In the model-based integration, only the final statistical results obtained from different data sets are merged. Meta-analysis is an example for the model-based integration, in which typically the effect sizes or p-values are combined before meta-analysis model fitting.

Rau, et al. compare the performance of p-value combination methods and a global negative binomial generalized linear model (GLM) with fixed effect method on two RNA-seq data sets of human melanoma cell lines (Rau, Marot, & Jaffrézic, 2014). In the individual p-value combination meta-analysis, the raw p-value of each gene is calculated by fitting a negative binomial GLM using the gene count in DEseq (Anders & Huber, 2010) package. The p-values of individual analysis are combined using the inverse normal approach (Marot & Mayer, 2009) and the Fisher combination approach (Fisher, 1970). In the global differential analysis, a negative binomial GLM with fixed study effect is used to calculate the p-value for each gene. The results show that global GLM with fixed study effect work well for small numbers of studies and low inter-study variability.

BayesMetaSeq identifies DEGs by fitting a Bayesian hierarchical model that assumes gene counts follow a negative binomial distribution with hyperparameters baseline, effect size and dispersion vectors (Ma, Liang, & Tseng, 2017). Model parameters are estimated using Markov chain Monte Carlo (MCMC) sampling. The DEGs classification is done by using a Dirichlet process Gaussian mixture model. The BayesMetaSeq is applied to a RNA-seq integrated data set of three brain samples related to the human immunodeficiency virus (HIV) transgenic rat. Compared to edgeR-Fisher and DEseq-Fisher, BayesMetaSeq detected more DEGs using the same significant levels.

Lyu and Li propose a rank-based semi-parametric model for the DEGs detection with microarray and RNA-seq combined data set (Lyu & Li, 2016). The genes are assumed to belong to three categories: non-DEGs, up-regulated DEGs and down-regulated DEGs, which are located in the middle, top, and bottom of the rank list, respectively. The log-fold changed expression values are classified into three components using an extended copula mixture model (Li, Brown, Huang, & Bickel, 2011). The method is applied to a data set from Microarray Quality Control (MAQC) and Sequencing Quality Control (SEQC) projects and the results are compared to several methods including DEseq (Anders & Huber, 2010) and eBays (Smyth, 2004). The Lyu and Li’s method has the lowest average ranks of the fold change by the top DEGs and better enrichment compared to other methods presented in the paper.

The model-based integration often lacks the robustness since there are underlying assumptions for each model. If one of the assumption is invalid, the performance of the model-based integration would probably become poor.

1.5.5. Other Methods

Besides the batch effects removal methods, some other batch effects related studies have been done. For instance, Reese, et al (2013) use PCA and guided principal component analysis (gPCA) to perform a statistical test for detecting batch effects. The test statistic is defined as the ratio of the variance of the first principal in gPCA to the variance of the first principal in PCA. The p-value of the test statistic is estimated using simulation from a permutation distribution.

A supervised classification analysis by using the median rank scores of the expression values and quantile discretization has been developed for significant gene prediction (Warnat, Eils, & Brors, 2005). Median Rank Scores (MRS) is a LS method that can transform the original expression values in different platforms into a similar numerical range (Tödling & Spang, 2003). The equal frequency binning is applied to further transform the data set (Liu, Hussain, Tan, & Dash, 2002). In the end, the Support Vector Machine (SVM) is used for the supervised classification analysis.

In addition to the expression values of the genes, some other continuous or categorical variables can also be used for classification. A GLM with elastic net (Zou & Hastie, 2005) penalty has been used to build a multinomial classifier for disease subtypes (Hughey & Butte, 2015). The leave-one-study-out cross-validation for the elastic net classifier can then be used for analyzing significant genes for the subtypes of disease.

1.6. Clustering Approaches

Clustering analysis refers to the algorithms that group objects (genes or samples in gene expression profiling) based on their similarities. Clustering analysis is one of the unsupervised learning approach, which means that the original data doesn't include some label/level information of the observations. Two widely used clustering algorithms in gene expression profiling field are hierarchical clustering (Eisen, Spellman, Brown, & Botstein, 1998; Sirinukunwattana, Savage, Bari, Snead, & Rajpoot, 2013) and K-means clustering (Li, et al., 2010; Iam-On & Boongoen, 2012; Nazeer, Sebastian, & Kumar, 2013). This section will mainly focus on the K-means clustering as it usually served as a comparison method in clustering analysis.

K-means clustering is a well-known clustering method. In brief, K-means clustering is trying to divide *N* objects into *K* clusters based on the distance between objects and cluster centers. K-means clustering selects the best partition by minimizing the within-cluster sum of squares. It requires the number of clusters, *K,* as input.

There are four basic steps for K-means algorithm. First, the K-means algorithm randomly selects *K* objects as the centers of the *K* clusters. In each iteration, the Euclidean distance between each object and each cluster center is calculated. And the object will be assigned to its nearest cluster, i.e. the cluster with the smallest Euclidean distance. Thirdly, the new cluster centers are calculated after each iteration. Last, the second and third steps will be repeated until convergence. The Euclidean distance of two vectors, **X** and **Y** in *n* dimensions, is calculated as follows:

$$D(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1.1}$$

There are mainly two assumptions for K-means clustering. K-means clustering performs well on spherical or ball-shaped data (Jain, 2010). Also, the number of objects in each cluster should approximately even with equal variance.

This dissertation focuses on how to combine and analyze data from the microarray and RNA-seq technologies. Chapter 2 introduces the embryonic heart data set of microarray and

RNA-seq, as well as the preprocessing analysis for the data sets. The Boundary Shift Partition (BSP) algorithm is proposed in Chapter 3. In Chapter 4, the performance of BSP algorithm is evaluated and compared with the Hartigan-Wong's (Hartigan & Wong, 1979) K-means algorithm using simulated data. The results of applying the BSP algorithm to the preprocessed embryonic heart data sets are presented in Chapter 5. Chapter 6 includes the conclusion and discussion.

CHAPTER 2. EMBRYONIC HEART DATA AND DATA PROCESSING

2.1. Data Sets

The microarray and RNA-seq data used in this study were obtained from the heart tissue of wild type mice. Thus, all the samples in this study were biological replicates. The microarray data had 45,220 probes for four samples and the RNA-seq data had 36,594 genes for two samples. An annotation file for the microarray data was also available. Table 2.1 and Table 2.2 show part of the data for microarray and RNA-seq, respectively. The gene expression values for microarray were pre-normalized, as were the expression values for RNA-seq.

Table 2.1. Part of Microarray Gene Expression Data

Gene Symbol	MA1	MA2	MA3	MA4
Akt1	28256.95	31073.26	28898.22	24935.76
H2-Q7	1662.139	1885.987	1745.079	1469.927
Kdr	2828.991	2184.211	3196.065	1801.478
Tyrp1	6.717917	9.32741	5.445747	27.12331
Gpi1	163500.7	163215.3	168793.5	151467
Hmbs	21400.02	19857.29	19153.49	16631.1
Ntrk2	946.7777	1253.803	1199.267	882.634
Olfr1307	9.887807	15.01386	16.75826	16.17594
Olfr166	37.82956	41.1092	46.04584	31.2603
Rps7	132366.4	140177	138795.1	119004.4
Rb1	2046.421	1179.788	1492.828	1456.258
Rps18	135911.4	142896.8	141766.1	135984.5
Ppp1r2	1170.884	1246.634	908.9469	939.2718
Egfr	632.85	320.6998	492.6504	513.2288
Hist1h2af	119479.4	109577.6	126294.6	99295.84
Hmgb1	45355.12	43946.57	43733.9	39502.78
Ldha	115796.2	151865.7	129909.3	109036.9
Ndufa1	11520.16	13568.99	9853.245	8809.788
Morf4l1	42213.25	52661.69	44065.17	47247.77

Table 2.2. Part of RNA-seq Gene Expression Data

Gene Symbol	RNA3	RNA11
Gnai3	88.1996	87.6813
Pbsn	0	0
Cdc45	28.3955	31.8773
H19	1935	1991.3
Scml2	1.98629	1.7267
Apoh	0	0
Narf	16.7858	16.1682
Cav2	1.73598	1.37976
Klf6	12.2728	13.0533
Scmh1	13.3895	14.5688
Cox5a	159.566	161.959
Tbx2	6.74803	4.63926
Tbx4	1.03696	0.370719
Zfy2	0	0
Ngfr	15.3549	13.6166
Wnt3	0.626531	0.284542
Wnt9a	0.825546	0.699762
Fer	9.22548	10.1839
Xpo6	35.2881	35.1514
Tfe3	13.3063	11.8945

2.2. Data Sets Combination

To analyze the potential relationship between microarray and RNA-seq, the two separate data sets needed to be combined together. The genes that were only tested in either one of the data sets had to be removed for the final combined data set. In the end, the total number of genes in the combined data set was 14,857. For each gene, there were four samples from microarray technology and two samples from RNA-seq technology. Table 2.3 shows part of the combined data set. The following study of this research was based on this combined data.

Table 2.3. Part of Combined Gene Expression Data

ID	Gene	MA1	MA2	MA3	MA4	RNA3	RNA11
1	0610005C13Rik	9.170	62.082	9.016	13.971	0.343	0.605
2	0610007P14Rik	30743.800	41798.900	30191.950	25841.170	42.795	48.368
3	0610009B22Rik	2689.288	2867.673	2208.953	1974.990	21.416	24.036
4	0610009O20Rik	509.854	545.401	434.529	409.202	22.130	22.761
5	0610010F05Rik	379.490	532.325	350.831	358.410	9.278	9.826
6	0610010K14Rik	2474.878	2673.511	2546.653	1853.097	53.963	46.830
7	0610011F06Rik	6572.091	5595.217	6144.279	5095.386	20.136	17.472
8	0610012G03Rik	3273.217	4917.466	3205.454	2633.260	13.095	14.455
9	0610025J13Rik	19.970	14.762	8.790	17.921	0.208	0.000
10	0610030E20Rik	928.720	1417.446	928.723	1211.190	9.844	9.969
11	0610037L13Rik	2541.094	2094.012	2203.223	1513.617	44.595	42.082
12	0610039K10Rik	5.463	9.310	4.100	13.537	0.188	0.181
13	0610040B10Rik	26.726	18.682	28.430	25.734	1.993	2.285
14	0610040J01Rik	5873.463	10218.140	6296.872	5715.141	1.991	2.855
15	1110001J03Rik	13413.240	19702.430	14653.450	16726.380	30.242	28.063
16	1110002L01Rik	5.590	30.719	4.224	13.489	8.597	8.743
17	1110004E09Rik	17528.280	19552.970	17579.300	15785.640	19.316	19.656
18	1110004F10Rik	30966.340	30948.160	29484.190	24864.760	97.337	96.518
19	1110007C09Rik	1864.882	1994.757	1955.108	1819.455	5.822	6.616
20	1110008E08Rik	5.617	9.480	4.251	13.553	0.000	0.000

2.3. Magnitude Problem Solving Using Data Transformation

After a basic analysis of the combined data set, there was a magnitude problem in the gene expression values between microarray technology and RNA-seq technology. Table 2.4 shows the minimums and maximums for each sample. It is clear that the average minimum difference of the gene expression values between microarray and RNA-seq technology was 7.72, while the average maximum difference was around 306,984.43. The maximum gene expression value of microarray was about 100 times higher than that of RNA-seq. Therefore, direct comparison between microarray and RNA-seq would be a problem. Besides the magnitude problem, all six samples were highly skewed to the right, no matter if the samples were from microarray or RNA-seq. To solve the magnitude problem and better analyze the data, transformation of the original data was necessary. Box-Cox transformation, logarithm and cube root transformation were applied to this combined data set.

Table 2.4. Minimums and Maximums of the Six Samples

Samples	Minimum	Maximum	Standard deviation
MA1	5.27	293,570.9	15,549.546
MA2	9.08	310,030.5	16,293.554
MA3	3.94	323,838.0	15,260.399
MA4	12.59	312,757.1	13,984.681
RNA3	0	2,757.4	73.005
RNA11	0	3,371.99	77.567

The one parameter Box-Cox transformation is a type of power transformation (Box & Cox, 1964), which is defined as:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0 \\ \ln(y_i) & if\ \lambda = 0 \end{cases}, \tag{2.1}$$

Since the minimum expression value of RNA-seq was 0, 10^{-5} was added to all the data points to avoid problems during transformations. Logarithm and cubic root transformations are commonly used transformations in research. Based on the results of these two transformations, the logarithm and cubic root transformation was used to conduct the following research.

After applying the Box-Cox transformation to the combined data, the ranges of the gene expression values of the two technologies were close to each other. The Box-Cox transformation summary is shown in Table 2.5. The magnitude problem of the expression values no longer existed after the Box-Cox transformation. However, the minimum of the RNA-seq values was

-7.14, which would cause problems for explanation since it was difficult to describe the gene expression levels with negative values.

Table 2.5. Box-Cox Transformation Summary

Sample	λ	Minimum	Maximum
MA1	-0.02	1.65	11.13
MA2	-0.02	2.16	11.17
MA3	-0.02	1.35	11.21
MA4	-0.02	2.47	11.18
RNA3	0.14	-7.14	14.51
RNA11	0.14	-7.14	15.13

Table 2.6. Logarithm and Cubic Root Transformation Summary

	Minimum	Maximum	Standard deviation	Skewness
Log(MA1)	1.68	12.59	2.834	0.024
Log(MA2)	2.21	12.64	2.741	0.126
Log(MA3)	1.37	12.68	2.883	0.0004
Log(MA4)	2.53	12.69	2.514	0.259
$\sqrt[3]{RNA3}$	0	14.02	1.487	1.201
$\sqrt[3]{RNA11}$	0	14.99	1.493	1.229

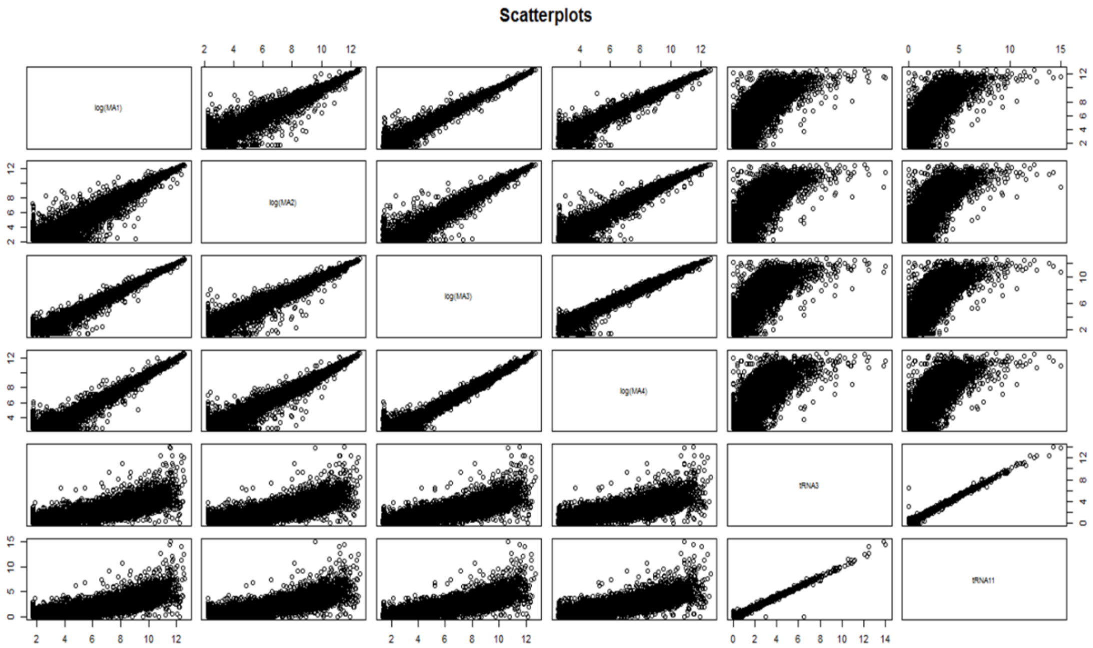


Figure 2.1. The Scatterplots for the Transformed Data.

Table 2.6 shows the results from the logarithm and cubic root transformation. This transformation took the log of the microarray values and cubic root of the RNA-seq values. Therefore, the original data was transformed to have a similar range. After the transformation, all the values remained positive. The following study was conducted based on this logarithm and cubic root transformation. Figure 2.1 shows the scatterplot of the transformed data set.

2.4. Linear Relationship Analysis

To further understand the relationship between the microarray values and RNA-seq values without considering the gene effect, some regression models was fitted. The average gene expression values from the microarray were treated as the dependent variable and the average gene expression levels from the RNA-seq were treated as the independent variable. The 1st order, 2nd order, 3rd order regression models and the logarithmic model were used to fit the transformed data. The models used are as follows:

$$tMA = \alpha\, tRNA + \varepsilon \tag{2.2}$$

$$tMA = \alpha\, tRNA + \beta\, tRNA^2 + \varepsilon \quad (2.3)$$

$$tMA = \alpha\, tRNA + \beta\, tRNA^2 + \gamma\, tRNA^3 + \varepsilon \quad (2.4)$$

$$tMA = \alpha \log(tRNA) + \varepsilon \quad (2.5)$$

For each model fitted in the regression analysis, the p-value was less than 0.05 with 14,857 observations. Table 2.7 shows the adjusted R-squares for each model. Based on the adjusted R^2, the 3rd order model was best among these four models. The adjusted R^2 for the 3rd order model was 0.7702. Figure 2.2 shows several plots for checking 3rd order model assumptions. Figure 2.3 is the scatterplot with fitted 3rd order model. Since the 3rd order model had the highest adjusted R^2 value to fitting the average microarray and RNA-seq transformed data, this model was also used to fit each sample in microarray and RNA-seq. The adjusted R^2 values are reported in Table 2.7 and 2.8. The four samples of the transformed microarray and the mean expression value of transformed microarray served as the dependent variable in each model. The two samples of transformed RNA-seq and the mean expression value of RNA-seq were used as the independent variable. The adjust R^2 values of the 3rd order model among each sample were close to each other. Therefore, the 3rd order regression model was the best among the four models to describe the relationship between microarray and RNA-seq gene expression levels without considering the individual gene effect.

Table 2.7. Adjusted R^2 for the Models

Model	Adjusted R^2
1st order model	0.6918
2nd order model	0.7623
3rd order model	0.7702
Logarithmic model	0.3994

Table 2.8. Adjusted R^2 for the 3rd Order Models

	tRNA3	tRNA11	tRNA
tMA1	0.768	0.765	0.768
tMA2	0.771	0.775	0.775
tMA3	0.771	0.769	0.772
tMA4	0.739	0.739	0.741
tMA	0.768	0.768	0.770

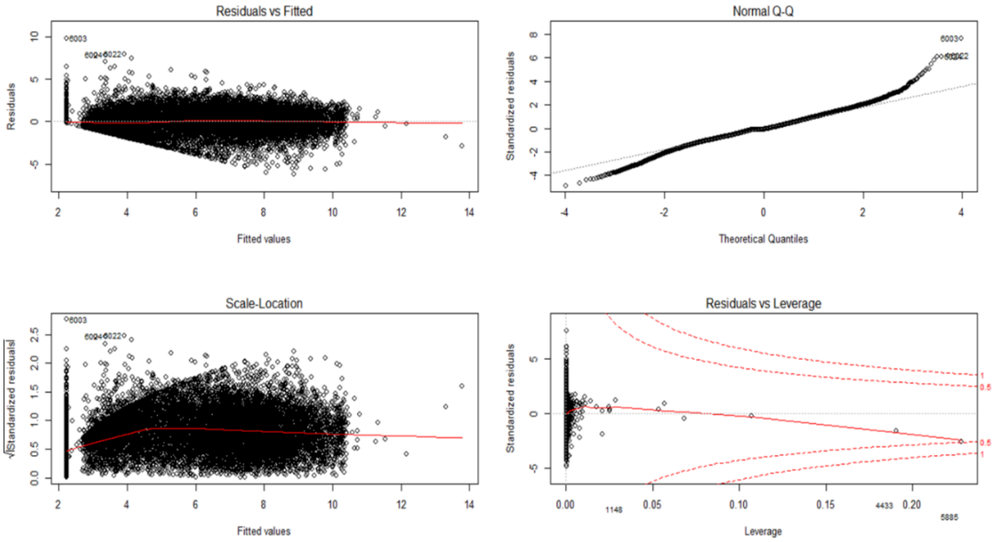


Figure 2.2. The 3rd Model Assumption Checking Plots

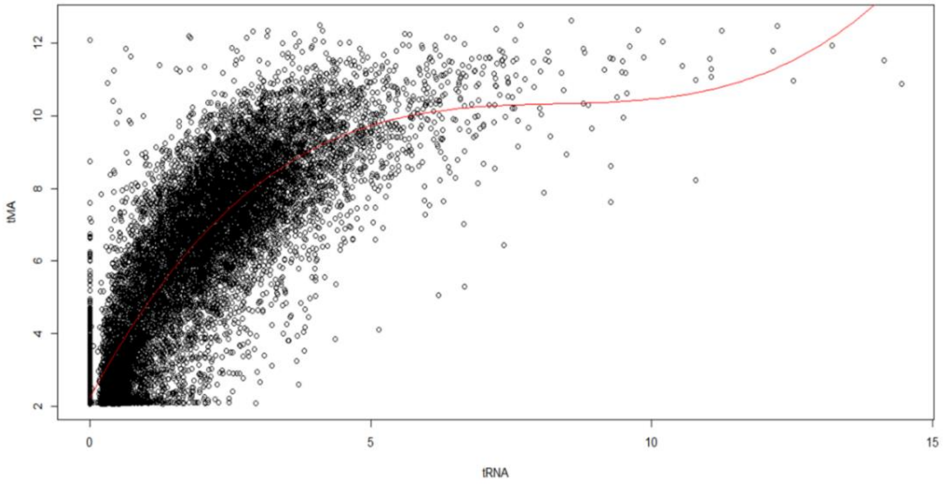


Figure 2.3. The Scatterplot with Fitted 3rd Order Model

2.5. Noise Minimization and Quantile Normalization

To minimize the background noise from the microarray and RNA-seq technologies, the mismatch values within the 5th percentile of each technology were removed. For instance, if the average expression value of one gene from microarray was within the lowest 5% of all microarray gene expression values, but its mean RNA-seq expression value was greater than 5% of the non-zero RNA-seq values, this gene was labeled as a mismatch gene and vice versa.

Quantile normalization was applied to this removed mismatch data set to further understand the relationship between microarray and RNA-seq technology. The logarithm and cubic root transformed data were used to conduct this quantile normalization. For each sample, the transformed gene expression values were ranked from smallest to largest. The genes in the same rank among the six samples were forced to stay the same. In this case, the mean values of these genes were applied. Therefore, it was guaranteed that all six samples had exactly the same distribution.

For the mismatch gene detection, there were 156 genes that have an expression value within the lowest 5% of the microarray expression levels but higher than 5% of non-zero RNA-seq values. And 1,889 genes had an expression value within the lowest non-zero 5% of the RNA-seq expression leves (genes with RNA-seq expression value of 0 included) but higher than 5% of microarray values. A total of 2,045 mismatch genes had been removed from the combined data set.

Table 2.9 shows the sample statistic summaries before and after quantile normalization which include minimum, median, maximum, mean and standard deviation. The data after quantile normalization had been used to fit another 3rd order regression model. This model had a significant p-value that was less than 0.05 with all terms that were significant in the model. The adjusted R^2 was 0.7207, which is actually about 5% lower than the previous results. The model assumption plots and fitted curve plot are shown in Figure 2.4 and 2.5.

Table 2.9. Sample Summaries Before and After Quantile Normalization

	Minimum	Median	Maximum	Mean	SD.
Log(MA1)	1.68089	6.521271	12.58987	6.139841	2.64616
Log(MA2)	2.207173	6.566698	12.64443	6.22337	2.601999
Log(MA3)	1.373153	6.456845	12.688	6.05967	2.688006
Log(MA4)	2.533276	6.279776	12.65318	6.063591	2.423351
$\sqrt[3]{RNA3}$	0	1.865335	14.02275	1.975917	1.429577
$\sqrt[3]{RNA11}$	0	1.8528	14.99554	1.963013	1.439169
Log(MA)	2.067095	6.495878	12.62841	6.176559	2.538741
$\sqrt[3]{RNA}$	0	1.862883	14.45526	1.978734	1.423216
QN	1.33144	4.923787	13.26563	4.736925	2.183008
QN-MA	1.305396	4.925272	13.12782	4.737567	2.168986
QN-RNA	1.329872	4.932203	13.16276	4.7373	2.178561

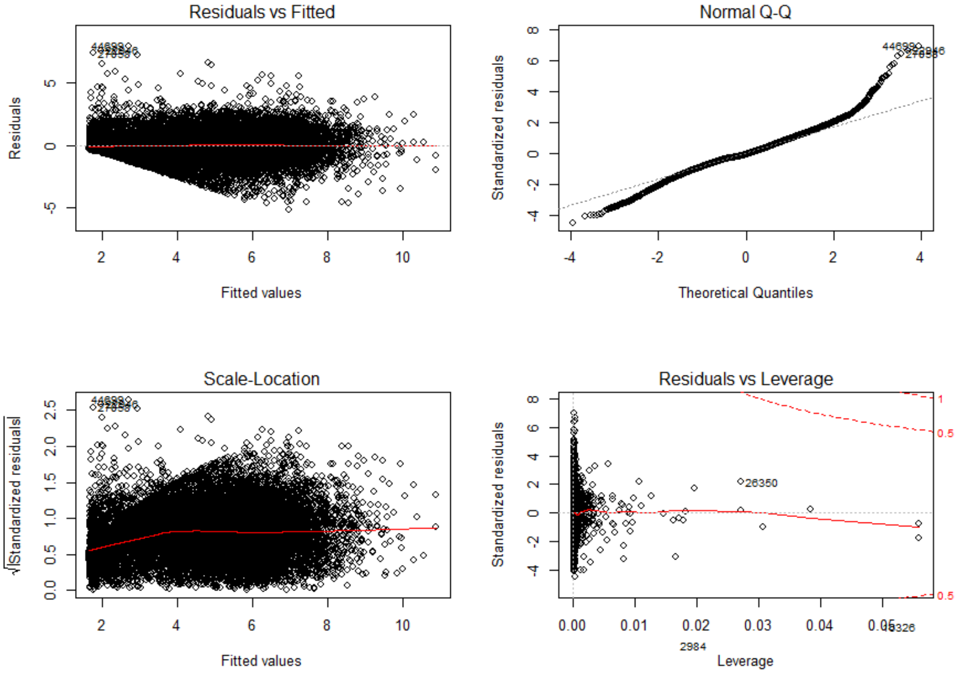


Figure 2.4. The 3rd Model Assumption Checking Plots for Quantile Normalized Data

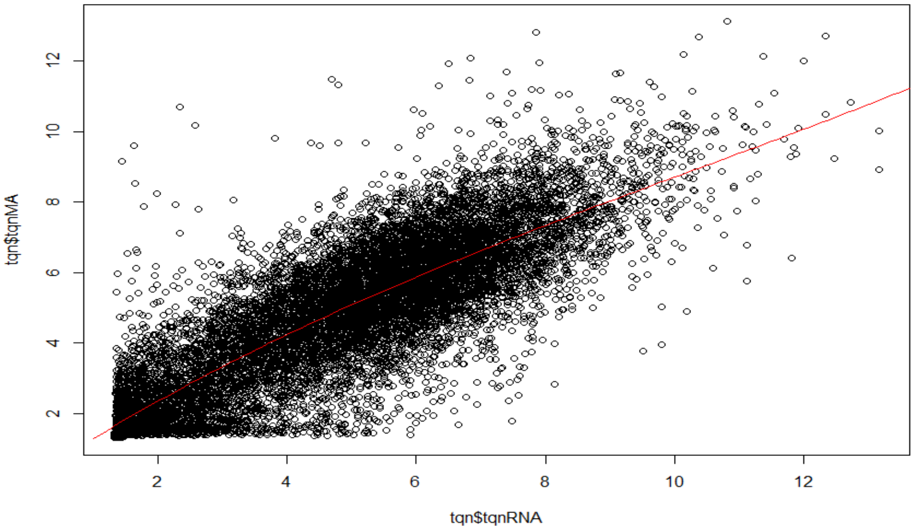


Figure 2.5. The Scatterplot with Fitted 3rd Order Model for Quantile Normalized Data

2.6. Inconsistency Between Microarray and RNA-seq

To study the gene expression difference measured by different technologies, the quantile normalized data were used. Ideally, the quantile normalized expression levels of the same gene among the six samples should stay the same. Therefore, the differences between the mean quantile normalized microarray expression values and the mean quantile normalized RNA-seq expression values should be close to 0. Thus, the study of the difference can help illuminate the inconsistencies of the two technologies.

The difference was calculated by using the mean quantile normalized microarray expression values to subtract the mean quantile normalized RNA-seq expression values within each gene. Table 2.10 shows the basic statistics about the difference. Heat maps were used to have a better visual understanding of the difference. Figure 2.6 and 2.7 are the heat maps for the difference of transformed data before and after quantile normalization, respectively.

Table 2.10. Statistic Summary for the Gene Expression Difference Between Two Technologies

	Minimum	Median	Maximum	Mean	SD.
Difference	-5.83923	0.02498	8.35044	0.0002676	1.20344

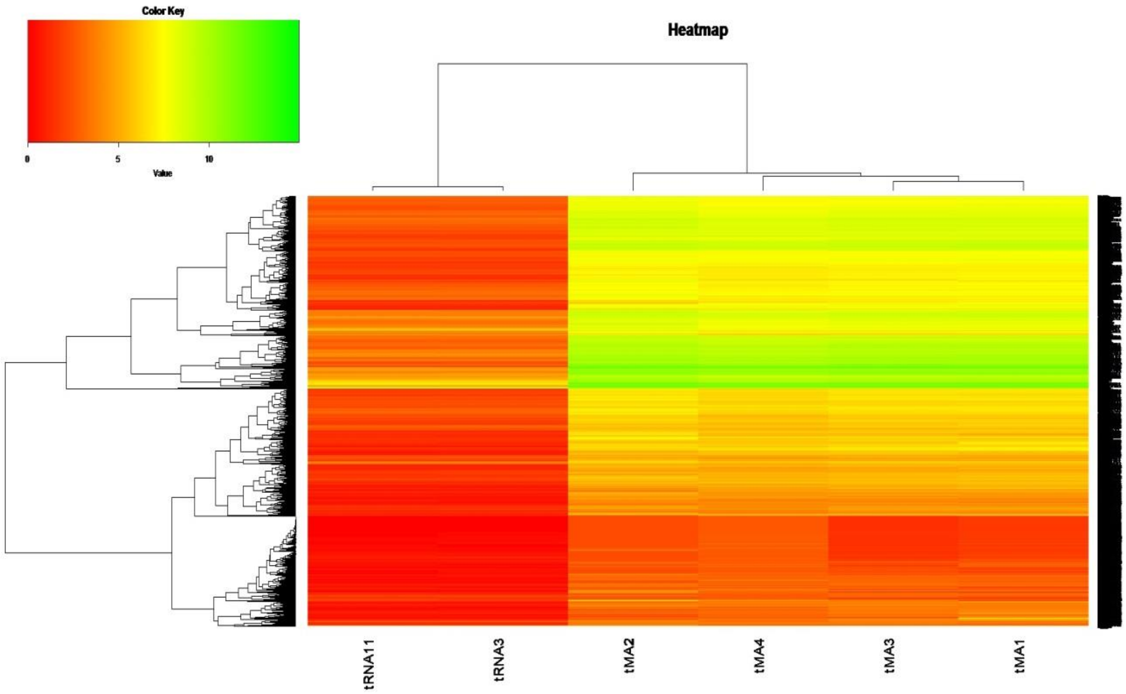


Figure 2.6. Heat Map for the Difference of the Logarithm and Cubic Root Transformed Data

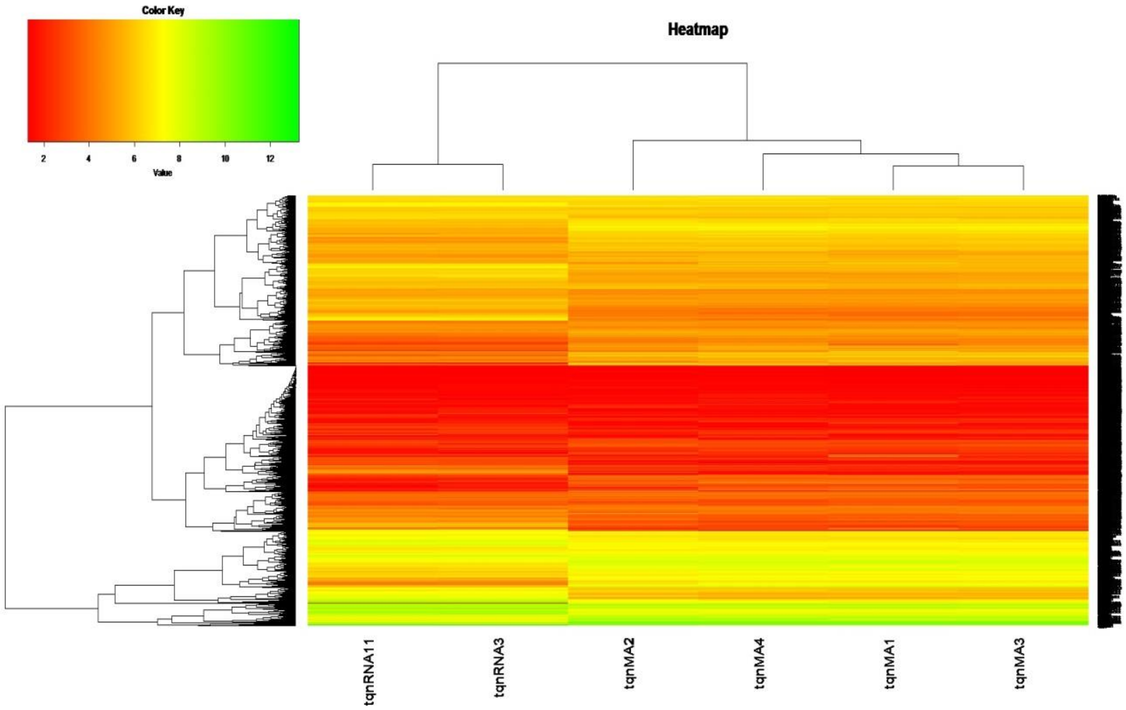
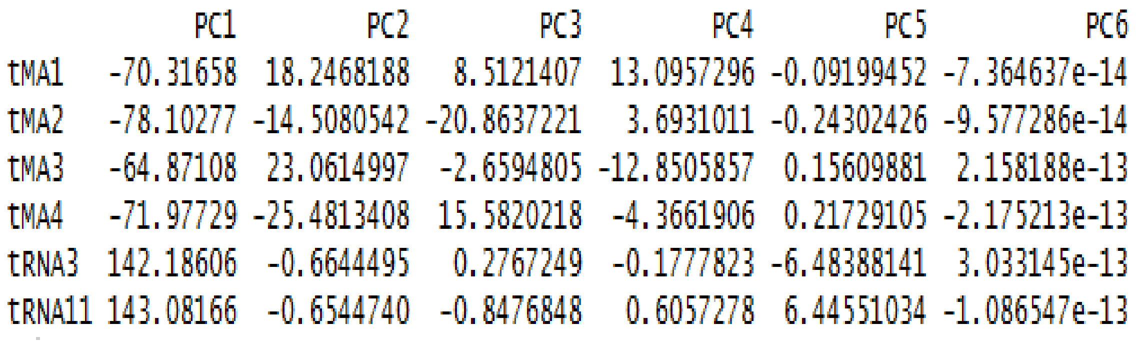


Figure 2.7. Heat Map for the Difference of the Quantile Normalized Transformed Data

2.7. Minimizing Differences in Microarray and RNA-seq

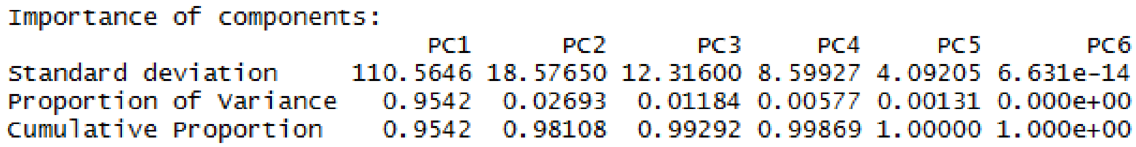
To remove the technology effect from the logarithm and cubic root transformed data, principal component analysis was applied to the data set. First, the data was centered and scaled. Then, the components were calculated by a singular value decomposition of the data matrix. The component(s) that distinguish between microarray and RNA-seq were removed.

The loading matrix and importance of components from principal component analysis are shown in Figure 2.8 and 2.9. Thus, the 1st principal component contained over 95% of the overall variability in the data set. In order to remove the component which distinguishes between the two technologies, the p-values from the t-test was used. Basically, for each component, the two-sample t-test was conducted between different technologies. Table 2.11 shows the p-values for each component. It is clear that the 1st principal component had a p-value that was less than 0.05. Therefore, the 1st principal component would be removed.



```
              PC1         PC2         PC3         PC4         PC5           PC6
tMA1    -70.31658  18.2468188   8.5121407  13.0957296 -0.09199452 -7.364637e-14
tMA2    -78.10277 -14.5080542 -20.8637221   3.6931011 -0.24302426 -9.577286e-14
tMA3    -64.87108  23.0614997  -2.6594805 -12.8505857  0.15609881  2.158188e-13
tMA4    -71.97729 -25.4813408  15.5820218  -4.3661906  0.21729105 -2.175213e-13
tRNA3   142.18606  -0.6644495   0.2767249  -0.1777823 -6.48388141  3.033145e-13
tRNA11  143.08166  -0.6544740  -0.8476848   0.6057278  6.44551034 -1.086547e-13
```

Figure 2.8. Loading Matrix of Principal Components for Original Data



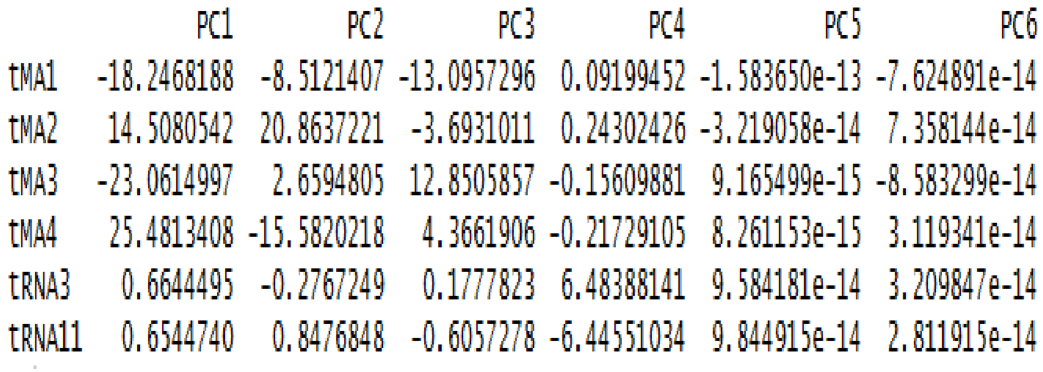
```
Importance of components:
                            PC1      PC2      PC3     PC4     PC5       PC6
Standard deviation     110.5646 18.57650 12.31600 8.59927 4.09205 6.631e-14
Proportion of Variance   0.9542  0.02693  0.01184 0.00577 0.00131 0.000e+00
Cumulative Proportion    0.9542  0.98108  0.99292 0.99869 1.00000 1.000e+00
```

Figure 2.9. Importance of Components for Original Data

Table 2.11. P-values of Student's t-tests for Original Data

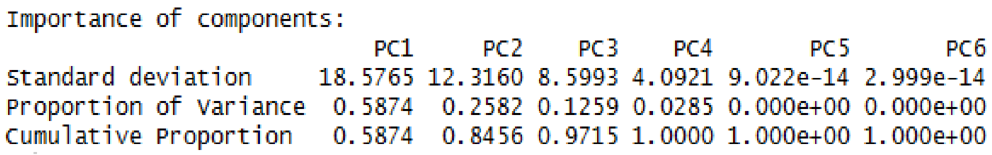
	PC1	PC2	PC3	PC4	PC5	PC6
P-value	2.76e-06	0.9394	0.9605	0.9576	0.9972	0.6193

Figure 2.10 and 2.11 show the loading matrix and importance of components from principal component analysis after the removal of the 1st PC, respectively. Based on the results of the t-tests (Table 2.12), none of the p-values is greater than 0.05. Thus, the technology effect has been removed from the original data.



```
               PC1         PC2         PC3         PC4           PC5           PC6
tMA1   -18.2468188  -8.5121407 -13.0957296  0.09199452 -1.583650e-13 -7.624891e-14
tMA2    14.5080542  20.8637221  -3.6931011  0.24302426 -3.219058e-14  7.358144e-14
tMA3   -23.0614997   2.6594805  12.8505857 -0.15609881  9.165499e-15 -8.583299e-14
tMA4    25.4813408 -15.5820218   4.3661906 -0.21729105  8.261153e-15  3.119341e-14
tRNA3    0.6644495  -0.2767249   0.1777823  6.48388141  9.584181e-14  3.209847e-14
tRNA11   0.6544740   0.8476848  -0.6057278 -6.44551034  9.844915e-14  2.811915e-14
```

Figure 2.10. Loading Matrix of Principal Components for the 1st PC Removed Data



```
Importance of components:
                           PC1     PC2    PC3    PC4       PC5       PC6
Standard deviation     18.5765 12.3160 8.5993 4.0921 9.022e-14 2.999e-14
Proportion of Variance  0.5874  0.2582 0.1259 0.0285 0.000e+00 0.000e+00
Cumulative Proportion   0.5874  0.8456 0.9715 1.0000 1.000e+00 1.000e+00
```

Figure 2.11. Importance of Components for the 1st PC Removed Data

Table 2.12. P-values of Student's t-tests for the 1st PC Removed Data

	PC1	PC2	PC3	PC4	PC5	PC6
P-value	0.9394	0.9605	0.9576	0.9972	0.6193	0.3429

Table 2.13 shows the descriptive statistics of the six samples after removing the 1st principle component. The six samples follow similar distributions after the 1st principle component removal. The mean values of Microarray samples and the mean values of RNA-seq samples were calculated. Figure 2.12 shows the relationship between Microarray and RNA-seq after the 1st principle component removal. They are almost perfect correlated based on the figure. This is because over 95% of the variation in the original data was removed when removing the 1st principle component. Even though the technology difference was successfully removed, it also removed too much variation from the data.

Table 2.13. Data Summary after 1st Principle Component Removal

	tMA1	tMA2	tMA3	tMA4	tRNA3	tRNA11
Minimum	0.5223	0.5217	0.5322	1.3758	1.0134	0.5570
1st Quantile	3.0829	2.8443	3.0507	2.8482	2.9153	2.9047
Median	5.0543	4.9268	5.1016	4.7793	4.9523	4.9644
3rd Quantile	6.3749	6.2904	6.4585	6.1033	6.2833	6.2917
Maximum	12.785	12.525	12.346	12.495	12.280	12.995
Mean	4.7760	4.7085	4.8015	4.6675	4.7337	4.7382
Standard deviation	2.1980	2.1205	2.2721	1.9676	2.1059	2.1203

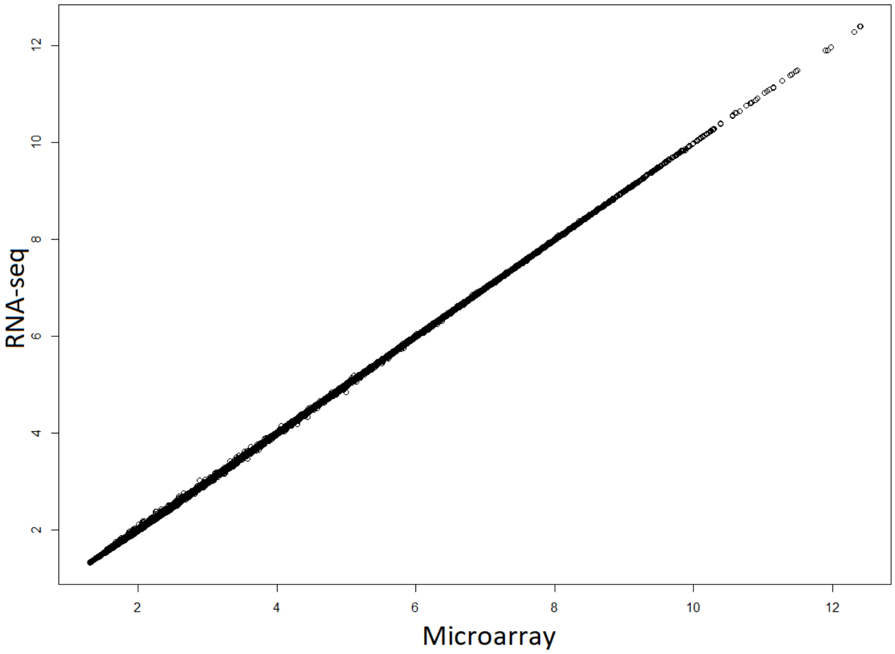


Figure 2.12. The Relationship between Microarray and RNA-seq after the Removal of the 1st Principle Component

2.8. Probe Alignments

The probe alignments to the reference genome were applied to the data set to have a better understanding of the expression values of microarray and RNA-seq because the existence of multiple alignments implied that the real gene expression levels were lower than the values in the experiment. In microarray, the gene is represented by a pre-designed short sequence called probes. In RNA-seq, the expression level of a gene is represented by the number of copies of the sequence fragments named as reads. Ideally, the probes in microarray and the reads in RNA-seq should be gene specified. In other words, the probes and reads should be unique to the genes that they are representing. However in reality, due to the small length of probes (25 – 150 bp in general (Chou, Chen, Lee, & Peck, 2004)) and reads (30 – 200 bp in general (Chang, Wang, & Li, 2014)) comparing to the large genome size from 0.49 Mbp of *Mycoplasma genitalium* (Huber, et al., 2002) up to 670 Gbp in *Polychaos dubium* (McGrath & Katz, 2004), the sequence of probes and reads might not be unique to the specific gene. Therefore, the probe alignment was necessary to analyze the expression accuracy.

The process of probe alignments was mapping the probes of microarray to the reference genome, the mouse genome in this case. The number of alignments for each sequence was recorded. If a sequence has multiple alignments in the reference genome, it means that some sequences other than the specific gene could also contribute to the gene expression levels, which create a false positive result. Sometimes, a probe might not find an alignment to the reference genome at all. There are several reasons to cause this problem: for instance, one might make mistakes while doing the alignment, or the sample has a unique sequence in this region that differs from the reference genome. Thus, the probe alignments would help to understand the statistical trends of the expression values in microarray.

The probe alignments were applied to three data sets: the original combined data set with 14,857 genes, the data set of 12,812 genes after removal of the mismatch genes and the data set with the mismatch genes containing 2,045 genes. A brief summary of the probe alignments is shown in Table 2.14. Thus, the number of genes with multiple alignments is 1,027 in the 14,857 genes data set, 826 in the 12,812 genes data set and 201 in the 2,045 mismatch genes data set. The percentage of the multiple alignment probes in the mismatch data set (9.8289%) is the highest among these three data sets. But the percentage of probes with no alignment to the reference genome in the mismatch data set (9.8778%) is lower than the other two data sets. The total percentage of the genes including multiple alignments and no alignment is about the same in these three data sets.

Table 2.14. Summary for Probe Alignments of the Three Data Sets

	# of Alignments	Data sets		
Multiple alignment	[2,5]	917	734	183
	(5,10]	70	59	11
	(10,∞)	40	33	7
NAs	0	1,840	1,638	202
Total		2,867	2,464	403
Total number of genes		14,857	12,812	2,045

Since mismatch data set showed a different performance in comparison to the other two data sets, detailed probe alignments were applied to the mismatch data set. Table 2.15 shows the summary for the probe alignments applied to the mismatch data set. Based on the definition of the mismatch genes, there were two cases: low microarray but high RNA-seq values and high microarray low RNA-seq values. The low microarray but high RNA-seq gene represented that the average expression value of one gene from microarray was within the lowest 5% of all microarray gene expression values, but its mean RNA-seq expression value was greater than 5th quantile of the non-zero RNA-seq values. If the average expression value of one gene from RNA-seq was within the lowest 5% of all non-zero RNA-seq gene expression values, but its mean microarray expression value was greater than 5th quantile of the microarray values, this gene was a high microarray low RNA-seq gene. There were 156 genes in the low microarray high RNA-seq category and 1,889 genes in the high microarray low RNA-seq category. As the results show, the percentage of probes with multiple alignments in the high microarray low RNA-seq category (10.4288%) is much higher than that in the low microarray high RNA-seq category (2.5641%). Part of the reasons to have a such results is that microarray normally has a higher background noise than RNA-seq, therefore, it is expected to see some genes with low RNA-seq expressions values but high microarray expression values.

Table 2.15. Summary for Probe Alignments of the Mismatch Data Set

	# of Alignments	Low microarray high RNA-seq	High microarray low RNA-seq
Multiple Alignments	[2,5]	4	179
	(5,10]	0	11
	(10,∞)	0	7
NAs	0	22	180
Total		26	377
Total Number of genes		156	1,889

CHAPTER 3. METHODOLOGY DEVELOPMENT

Suppose the gene expression data set contains the expression values for *N* genes which are collected from *P* samples. Among the *P* samples, p_1 samples are analyzed using microarray technology, p_2 samples are analyzed using RNA-seq technology, $p_1 + p_2 = P$. All these *P* samples are assumed to be collected from the same conditions. In another word, these *P* samples are biological replicates to each other.

To measure the technology difference among the *P* samples, the mean expression value difference between microarray and RNA-seq technology for each gene can be calculated as follows:

$$Diff_i = MA_i - RNA_i \tag{3.1}$$

where MA_i denotes the mean expression value of gene *i* for the samples of microarray technology and RNA_i denotes the mean expression value of gene *i* for the samples of RNA-seq technology.

Let's assume that $Diff_i$ follows a mixture normal distribution with *K* clusters. The variances of these *K* clusters are equal. Thus, each expression value can be fitted in a linear model with the form:

$$Expression_{ijp} = \alpha_i + \beta_j Tech + \varepsilon_{ijp} \tag{3.2}$$

where α_i represents the mean expression value of gene *i* for the RNA-seq technology samples, $i = 1, 2, 3, \dots, N$. β_j represents the technology difference between microarray and RNA-seq for cluster j, $j = 1, 2, 3, \dots, K$. ε_{ijp} represents the random error for gene *i* in cluster *j* of sample p, $p = 1, 2, 3, \dots, P$. ε_{ijp} follows a normal distribution with mean 0 and variance σ^2, whereas $\varepsilon_{ijp} \sim N(0, \sigma^2)$.

The Boundary Shift Partition (BSP) algorithm involves 3 steps: partition initialization, assignment of genes on the edges, and stop of the algorithm under selected criteria.

3.1. Partition Initialization

Since gene expression data normally includes expression values from thousands of genes, the $Diff_i$ values are used to reduce the calculation complexity. Thus, the expression value of each gene is represented by a single value, $Diff_i$.

The $Diff_i$ values are first sorted from smallest to largest, denoted as $SDiff_i$. Then $K - 1$ partition points can be randomly chosen among the $N - 1$ possible partition points for the N $SDiff_i$ values to have K clusters. An example of the random partition for 3 (K) clusters among 10 (N) genes is shown in Figure 3.1. Each circle represents a $SDiff_i$ value. There are 9 (10-1) possible partition points among these 10 genes. The 2 (3-1) random partition points between the third gene and the fourth gene, as well as the fifth gene and the sixth gene divide the 10 genes into 3 clusters.

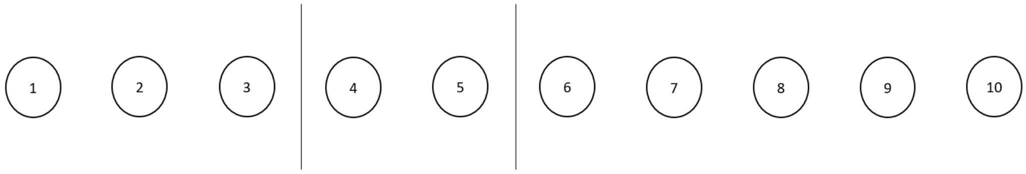


Figure 3.1. Example of the Random Partition

3.2. Assignment of Genes on the Edges

Since the $Diff_i$ values are sorted, only the genes on the edge of the clusters can possibly be moved to its nearby cluster. Let σ_j denote the standard deviation of the $SDiff_i$ values for cluster j.

$$\sigma_j = \sqrt{\frac{1}{n_j - 1}\sum_{j=1}^{n_j}\left(SDiff_i - C_j\right)^2} \qquad (3.3)$$

where n_j is the number of genes in cluster j. C_j is the mean expression value for cluster j, $C_j = \frac{1}{n_j}\sum_{i=1}^{n_j} SDiff_i$. Let G_e and G_{e*} denote the edge genes for partition divider D $(D = 1, 2, 3, \ldots, K-1)$ for cluster j and $j+1$, respectively. Thus, it's possible that G_e can belong to its nearby cluster ($j+1$). Let's assign G_e to its nearby cluster and calculate the standard deviation for cluster j and the standard deviation for the nearby cluster after the assignment, denoted as $\sigma_j^{\prime}$ and $\sigma_{j-nearby}^{\prime}$, respectively. G_e can be assigned to its nearby cluster if $\sigma_j^{\prime} + \sigma_{j-nearby}^{\prime} < \sigma_j + \sigma_{j-nearby}$. Keep moving the edge gene until it reaches the local minimum. Repeat this process for gene G_{e*}. Compare the local minimum of G_e, G_{e*}, and the original partition and choose the cutting point that provides the smallest value of the sum of the standard deviations. Repeat these steps until all the edges for the *K* clusters can't be assigned to their nearby clusters.

3.3. Stop of the Algorithm Under Selected Criteria

If none of the genes on the edges can be assigned to its nearby cluster, calculate the sum of the standard deviation for the *K* clusters, $\Sigma = \sum_{j=1}^{K} \sigma_j$. Each run of the algorithm can find a local minimum for Σ. To make sure the algorithm reaches the global minimum, the algorithm can be stopped if the Σ has not changed in the last *M* iterations.

3.4. Selection Criteria for Optimal Clusters

The optimal number of clusters is selected based on the number of differential measured genes (DMGs) in each cluster. After the BSP algorithm, the data set is divided into *K* clusters. Then the mean values of the RNA-seq technology for each gene (α_i) and the technology difference between microarray and RNA-seq for each cluster (β_j) can be estimated using the mean estimation. The residuals (ε_{ijp}) can be calculated for each expression value for the *K* clusters using the formula: $\varepsilon_{ijp} = Expression_{ijp} - \alpha_i - \beta_j Tech$. The two-sample t-test will be applied on the residuals for gene *i* between microarray and RNA-seq samples. The gene is defined to be differentially expressed if the p-value of the t-test is less than the significant level α. Normally α is chosen to be 0.05.

The selection criteria for the optimal number of clusters is defined as follows:

$$log_2\left(\frac{\#DMGs + 1}{N}\right) + k \tag{3.4}$$

The optimal number of cluster is the one that minimizes the selection criteria.

3.5. Algorithm Summary

The differentially expressed genes are detected using the BSP algorithm. The difference between the mean value of microarray and RNA-seq for each gene is used for gene clustering. A linear model is fitted for the genes in each cluster. The two-sample t-test uses the residuals of the linear model between microarray and RNA-seq technologies for DMGs detection. This algorithm clusters the genes into K partitions by minimizing the standard deviations of the difference values between microarray and RNA-seq.

Suppose the data set contains *N* genes, and the optimal number of clusters is *K*.

Step 1. Randomly select *K-1* partition points to form *K* clusters for the sorted *N* objects.

Step 2. For the two genes of each partition divider D $(D = 1, 2, 3, \dots, K - 1)$ in cluster *j* and *j*+*1*, assign each gene to its nearby cluster and calculate the standard deviations of the two clusters before and after assignment, respectively. Keep moving the edge until the local minimum is reached. Choose the cutoff point of these two clusters that minimize the sum of standard deviations for the two clusters.

Step 3. Go back to Step 2 for the next edge or go to Step 4 if all the edges have been analyzed.

Step 4. Record the minimum value of the sum of the standard deviations for all the clusters. Go back to Step 1.

Step 5. Stop the algorithm if the minimum value of the sum of the standard deviations has not changed in the last *M* iterations.

A flowchart of the algorithm is shown in Figure 3.2.

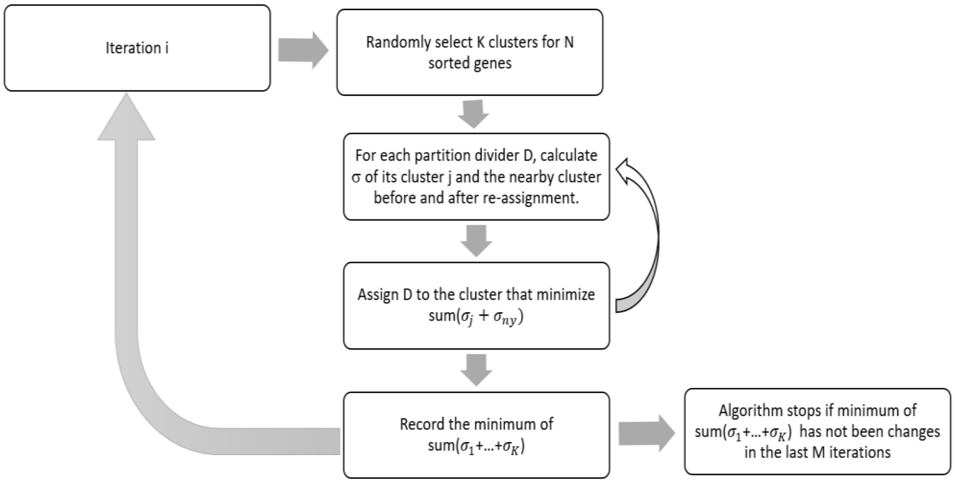

Iteration i
Randomly select K clusters for N sorted genes
For each partition divider D, calculate σ of its cluster j and the nearby cluster before and after re-assignment.
Assign D to the cluster that minimize sum($\sigma_j + \sigma_{ny}$)
Record the minimum of sum(σ_1+...+σ_K)
Algorithm stops if minimum of sum(σ_1+...+σ_K) has not been changes in the last M iterations

Figure 3.2. A Flowchart of the BSP Algorithm

CHAPTER 4. SIMULATION

To evaluate the performance of the BSP algorithm, six simulation data sets were generated in order to resemble situations with different distributions and a magnitude of variations. Also, the Hartigan-Wong's (Hartigan & Wong, 1979) K-means algorithm was used to compare the performance of the BSP algorithm.

4.1. Data Generation

For the simulation data, 2,000 genes were generated into four clusters. There were six replicates for each gene, four replicates were from microarray technology and two replicates were from RNA-seq technology. The simulation parameters on the linear model: $Expression_{ijp} = \alpha_i + \beta_j Tech + \varepsilon_{ijp}$, were estimated using the transformed embryonic heart data. In the model, α_i represents the mean expression value of gene *i* for the RNA-seq technology samples. β_j represents the technology difference between microarray and RNA-seq for cluster j. ε_{ijp} represents the random error for gene *i* in cluster j of sample p.

Based on the range of the differences between microarray and RNA-seq in the embryonic data, the parameters of β_j were chosen to be -0.5, 2, 4.5, and 7 for the four clusters, respectively. Table 4.1 shows the number of genes and the true β_j value for each cluster.

Table 4.1. Simulation Data Summary

Cluster	1	2	3	4
β_j	-0.5	2	4.5	7
# of genes	532	499	470	499

The mean expression values for RNA-seq (α_i) were generated from uniform distribution *U(0, 14)*, based on the range of the gene expression values for the RNA-seq samples. The error terms of all the four clusters were assumed to have equal variance σ^2 with mean 0.

Normality was one of the assumptions for the BSP algorithm of fitting linear model. The model assumed that the genes in each cluster follow a normal distribution with equal variances among all the clusters. Therefore, the residuals of the linear model followed a normal distribution, $\varepsilon_{ijp} \sim N(0, \sigma^2)$. To test the robustness of the BSP algorithm, some other distributions, besides normal distribution, were used to generate the error terms of the simulation data: Exponential distribution, Gamma distribution, Gaussian mixture distribution, and Beta distribution. The error terms were generated from these four distributions with mean 0 and standard deviation ranging from 0.1 to 1.5. Due to parameters' boundaries, the standard deviations for beta distribution were in the range of 0.1 to 0.5. The error terms of these four distributions were first generated with shifted means. Then the error terms were subtracted by their means to shift to 0. Given mean μ and standard deviation σ, the calculation of the parameters for each distribution is shown in Table 4.2. The error terms were generated using corresponding functions in R software.

Table 4.2. The Formulas and Parameter Estimates for the Four Distributions

Distribution	$Parameter\ estimates$	
Exponential	$\lambda = \frac{1}{\sigma}$	
Gamma	$shape: \alpha = \frac{\mu^2}{\sigma^2}$	$scale: \beta = \frac{\sigma^2}{\mu}$
Mixture	0.5*N(0, σ)+0.5*N(μ, σ)	
Beta	$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)$	$\beta = \alpha\left(\frac{1}{\mu} - 1\right)$

Therefore, the RNA-seq samples for each gene were generated using the formula: $\alpha_i + \varepsilon_{ijp}$. The microarray samples for each gene were generated using the formula: $\alpha_i + \beta_j + \varepsilon_{ijp}$.

4.2. Simulation Method

The BSP method was applied to the simulation data to test the accuracy of this method for 2 to up to 20 clusters. The stop criteria for the BSP algorithm was that the sum of the standard deviations did not change in the last 10 iterations. Then the residuals were calculated based on the fitted linear model: $Expression_{ijp} = \alpha_i + \beta_j Tech + \varepsilon_{ijp}$. Finally, the optimal number of clusters was chosen related to the minimum of the selection criteria values, which calculated using the number of DMGs for each cluster.

Hartigan-Wong's (Hartigan & Wong, 1979) K-means method was chosen as a comparison method for the BSP algorithm. Hartigan-Wong's K-means method was applied to the simulated data by using the *kmeans()* function in R software (R Core Team, 2016). The default parameter setting was used. After the K-means classification, a similar process can be used to calculate the residuals and the optimal number of clusters. Figure 4.1 shows a briefly flowchart for the process of simulation analysis.

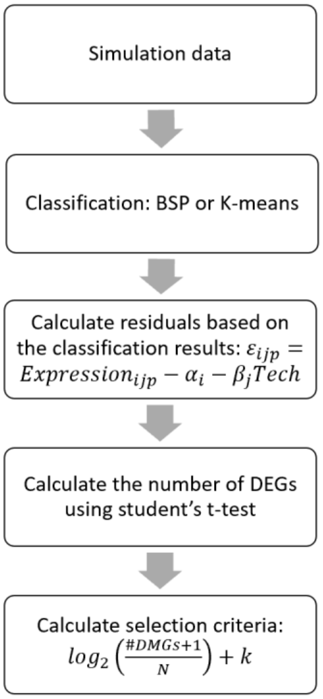

Simulation data
Classification: BSP or K-means
Calculate residuals based on the classification results: $\varepsilon_{ijp} = Expression_{ijp} - \alpha_i - \beta_j Tech$
Calculate the number of DEGs using student's t-test
Calculate selection criteria: $log_2\left(\frac{\#DMGs+1}{N}\right) + k$

Figure 4.1. The Flowchart of Simulation Methods. BSP: Boundary Shrift Partition; K-means: Hartigan-Wong's K-means method; DMGs: Differentially Measured Genes.

Two evaluation parameters were used to compare the performance of the BSP and K-means methods: misclassified rate (MR) and the accurate number of clusters. The misclassified rate of the algorithm was defined as the percentage of the genes that were misclassified to their own clusters for the case with true number of clusters (4). For instance, there were n out of N genes that were not successfully classified to their correct clusters for the four-clusters scenario. The MR was calculated using the formula: $MR = \frac{n}{N} \times 100\%$. The accurate number of clusters for the algorithm was defined as the number of times that the selection criteria correctly identified the true number of clusters as the optimal number of clusters using the residuals of the fitted linear model.

4.3. Normal Distribution

4.3.1. Misclassified Rate Comparison

Simulation data set 1 was generated with different values of the standard deviation for ε_{ijp} to check the performance of the BSP algorithm. The error terms were assumed to follow a normal distribution, $\varepsilon_{ijp} \sim N(0, \sigma^2)$. The values of σ ranged from 0.1 to 1.5 increasing by 0.1. For each case of σ, BSP and K-means were run 20 times to check the consistency of their results. As mentioned in the previous section, the 2,000 genes were generated in each case which belong to four clusters. Figure 4.2 shows the MR comparison between BSP and K-means. Detailed MR of BSP and K-means are showed in Appendix A.1 and A.2, respectively.

Based on the figure, it's clear that the results of BSP were more stable with small values of σ compared to those of K-means. For the first four cases, BSP almost successfully identified the true cluster of all genes for the 20 runs with the MRs less than 5%. The MR curve for BSP increased as standard deviation increased. As $\sigma =$ 1.5, the MR was approximately 30%. The

MR curve for K-means started at 17%, decreasing to near 0% and then increased as variance increased. For K-means, when $\sigma > 1.3$, MR increased to over 20%.

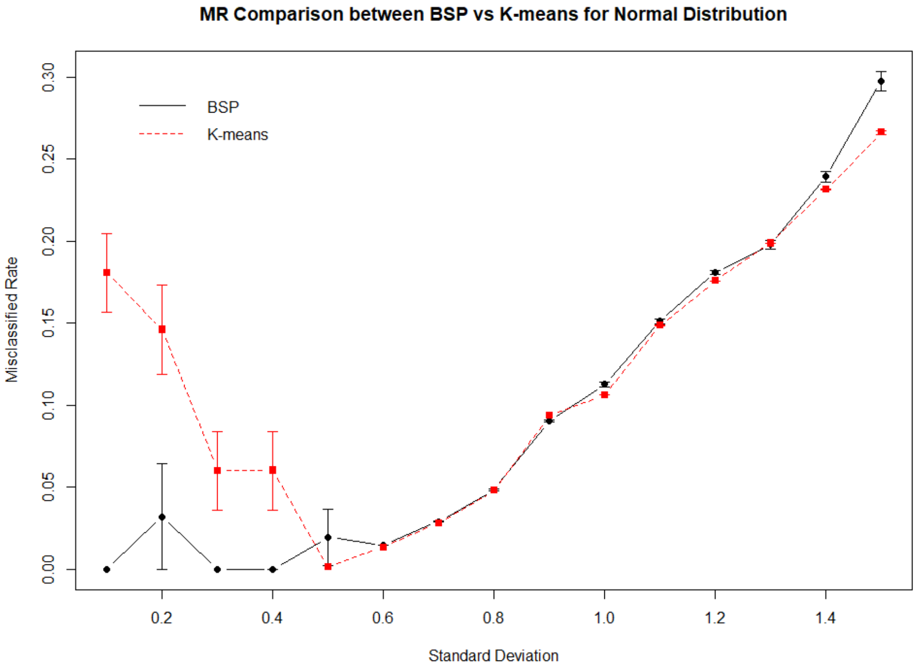


Figure 4.2. The Misclassified Rate Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Normal Distribution. Data shown as mean MR ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents MR. BSP: black solid line; K-means: red dashed line.

The bars on the plot represent the standard error for each case of σ. The standard errors of the K-means algorithm for the first four cases were very large, probably due to the limited number of starting point for the K-means function in R. The K-means algorithm didn't find the global minimum in these cases. For the σ values between 0.6 to 1.4, the performance of BSP and K-means were almost identical. As σ increased, the accuracy of BSP was a little bit higher than K-means. Overall, the BSP algorithm was more reliable to identify the true clusters of the genes than the K-means method, especially in the cases with smaller values of σ.

After obtaining the classification results of BSP and K-means for simulation data set 1, the residuals of the fitted linear model were calculated for both methods. The number of DMGs for each cluster was identified based on the t-test of the residuals. The level of significance was chosen to be 0.5. The results for the number of DMGs are shown in Appendix A.3 and A.4 for BSP and K-means, respectively.

Due to the high MRs of the K-means in the first four cases, the numbers of DMGs for the high MR runs were much higher than the runs with lower MRs. For example, as MR increased from 0% to 23.5% and 24.9%, the number of DMGs increased from 14 to 972 and 1033 for the first case of K-means with $\sigma = 0.1$ (Details shown in Appendix A.2 and A.4). After the first four cases, BSP and K-means had similar numbers of DMGs for different σ cases. Figure 4.3 shows the plot comparison between BSP and K-means.

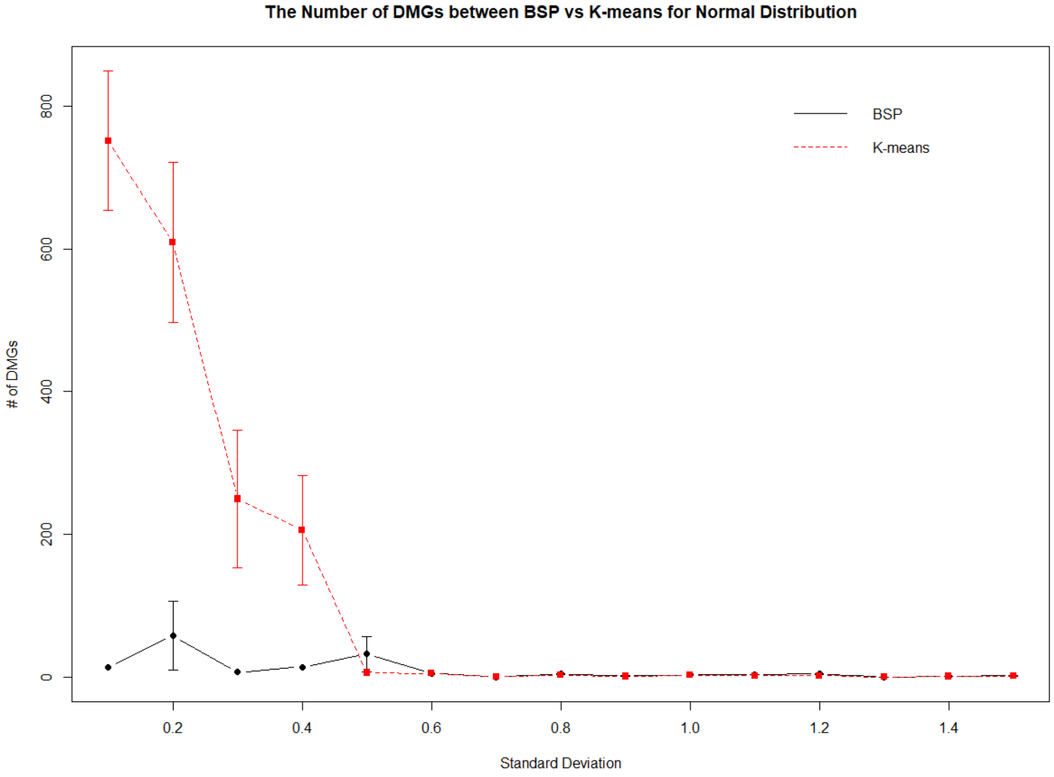


Figure 4.3. The Number of Differentially Measured Genes Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Normal Distribution. Data shown as mean number of DMGs ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the number of DMGs. BSP: black solid line; K-means: red dashed line.

Similar to the results of MR, the standard errors of the K-means method for the first four σ values were very high compared to the rest of the results. The number of DMGs for K-means started around 700, and as σ increased, the number of DMGs decreased to close to 0. The number of DMGs for the BSP algorithm was very stable, which was approximately 0.

4.3.2. The Accurate Number of Clusters Comparison

The selection criteria was applied to the 15 cases of σ (simulation data set 1) to find out the optimal number of clusters in each case. Appendix A.5 and A.6 show the selection criteria accuracy levels of BSP and K-means, respectively, in which 1 means the true number of clusters were correctly identified in that run and 0 means four-clusters was mis-identified as the optimal. Figure 4.4 shows selection criteria accuracy for BSP and K-means.

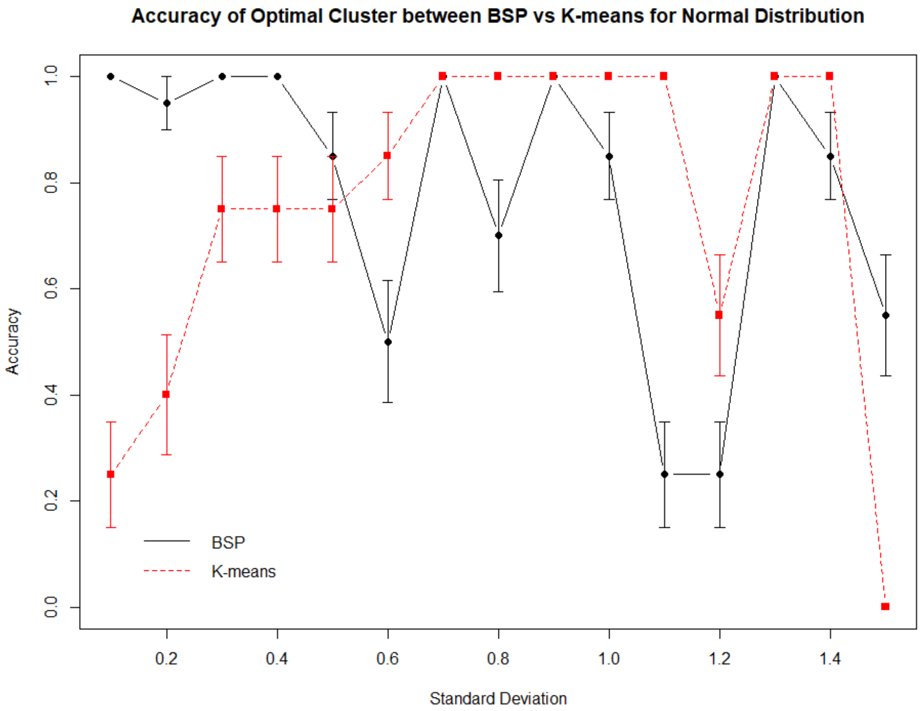


Figure 4.4. The Selection Criteria Accuracy Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Normal Distribution. Data shown as mean accuracy ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the accuracy. BSP: black solid line; K-means: red dashed line.

A high level of variance of the optimal number of clusters accuracy for both BSP and K-means was found. For BSP, the selection criteria identified the true number of clusters for the first four cases with over 90% accuracy. In comparison, the accuracies for the same first four case results of K-means were much lower, which might be due to the high MRs of K-means. For the remaining cases, the selection criteria correctly identified the real number of clusters for cases with $\sigma = 0.7, 1.0, \text{and } 1.3$ of the BSP results, as well as $\sigma = 0.7, 0.8, 0.9, 1.0, 1.1, 1.3, \text{and } 1.4$ of the K-means results. The selection criteria did not work well for the remaining cases. For instance, the selection criteria could not find four clusters as optimal in most of the 20 runs for case $\sigma = 1.1$ and 1.2 of the BSP results. An indistinguishable situation also happened for K-means case $\sigma = 1.5$, which the selection criteria accuracy was 0% since the minimum of the selection criteria appeared at three-clusters case.

For both the BSP and K-means algorithms, the selection criteria correctly identified the optimal number of clusters in about half of the 15 different σ cases. Both algorithms had cases that the selection criteria could not find the true number of clusters, at least in most of the 20 runs for these cases.

4.4. Exponential Distribution

4.4.1. Misclassified Rate Comparison

Simulation data set 2 used exponential distribution to generate the standard deviations of ε_{ijp}, $\varepsilon_{ijp} \sim Exp(\lambda)$. As shown in Table 4.2, the values of λ were calculated using the formula: $\lambda = \frac{1}{\sigma}$. Since in exponential distribution, the mean and standard deviation are the same, the error terms were subtracted by σs to have a mean of 0. Similar to simulation data set 1, the values of σ ranged from 0.1 to 1.5. Two-thousand genes were generated in each case which belong to four clusters. Figure 4.5 shows the MR comparison between BSP and K-means for exponential distribution. More details on the MRs of BSP and K-means are shown in Appendix A.7 and A.8, respectively.

As shown in Figure 4.5, the MR curve for BSP increased as standard deviation increased, which started at 0% for the first three cases. The MR curve for K-means started at 17%, decreasing to close to 2.5%, which was almost identical to the MR of BSP for $\sigma = 0.6$. Then both MR curves shared a similar trend and increased as variance increased.

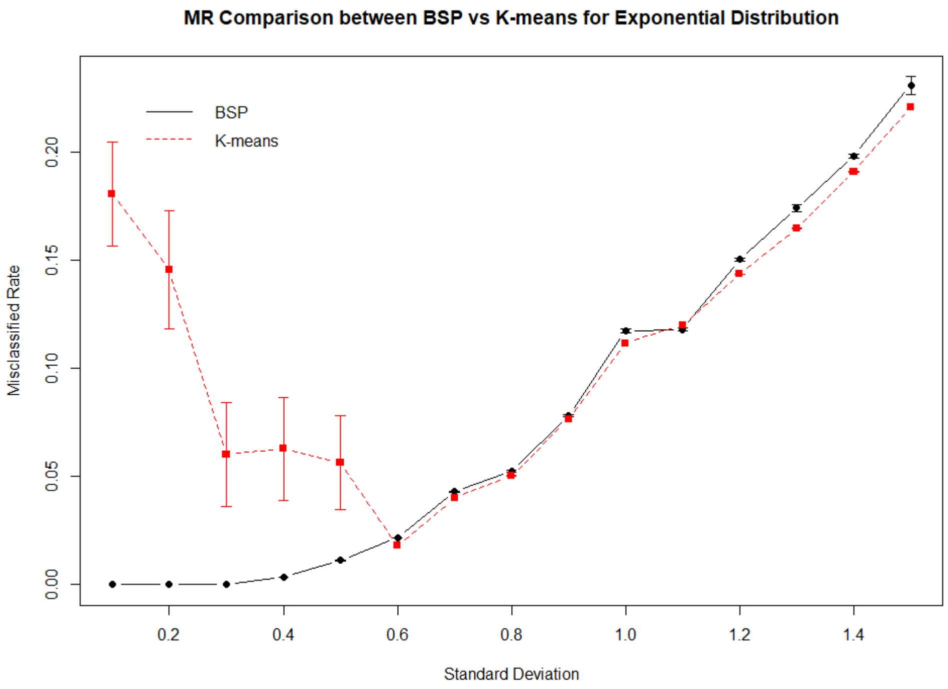


Figure 4.5. The Misclassified Rate Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Exponential Distribution. Data shown as mean MR ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents MR. BSP: black solid line; K-means: red dashed line.

The number of DMGs was identified using the t-test for the residuals of the of BSP and K-means methods. The level of significance was chosen to be 0.5. The results for the number of DMGs are shown in Appendix A.9 and A.10 for BSP and K-means for exponential distribution, respectively. Figure 4.6 shows the plot comparison between BSP and K-means.

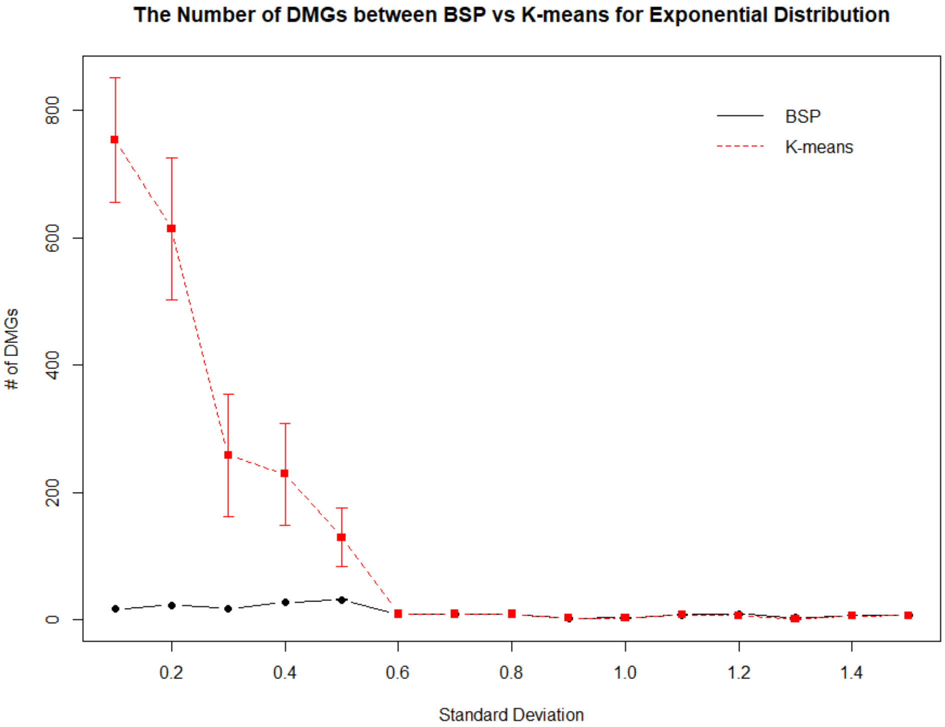


Figure 4.6. The Number of Differentially Measured Genes Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Exponential Distribution. Data shown as mean number of DMGs ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the number of DMGs. BSP: black solid line; K-means: red dashed line.

Based on Figure 4.6, the average numbers of DMGs of K-means for the first five cases were higher than those of BSP. The number of DMGs started around 700, and when σ increased the number of DMGs decreased to close to 0 for $\sigma > 0.5$. The number of DMGs for the BSP algorithm was very stable and close to 0.

4.4.2. The Accurate Number of Clusters Comparison

The selection criteria values were calculated for the 15 cases of σ (simulation data set 2) to verify the optimal number of clusters. The selection criteria accuracy of BSP and K-means are shown in Appendix A.11 and A.12, respectively. Figure 4.7 shows selection criteria accuracy of BSP and K-means.

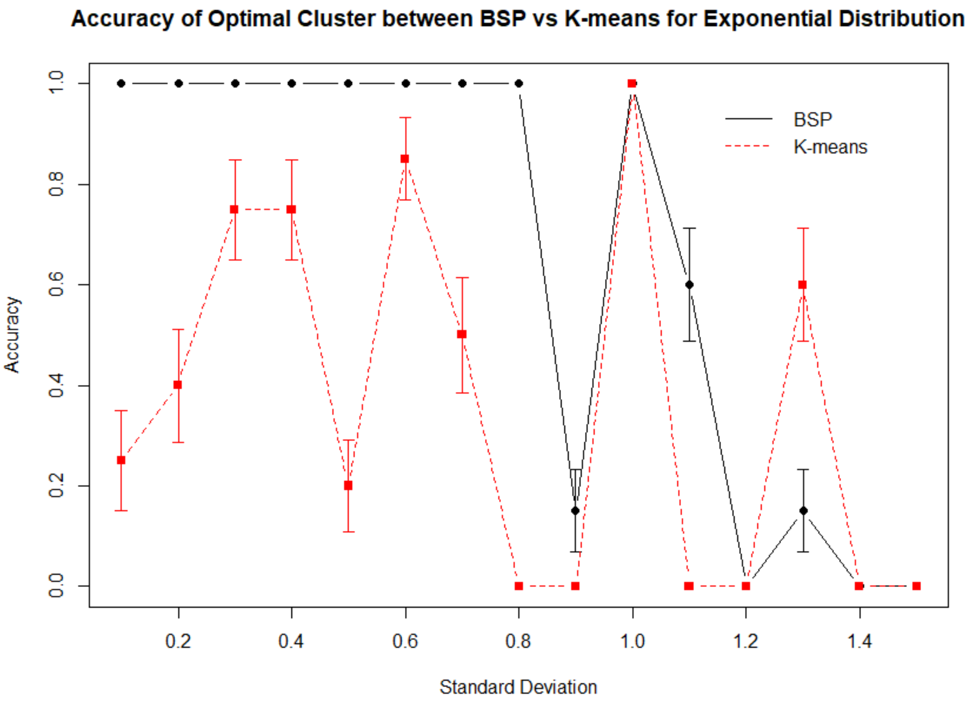


Figure 4.7. The Selection Criteria Accuracy Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Exponential Distribution. Data shown as mean accuracy ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the accuracy. BSP: black solid line; K-means: red dashed line.

In Figure 4.7, the selection criteria for BSP identified the true number of clusters for the first eight cases with 100% accuracy. While the accuracy for K-means ranged between 20% and 90% for these cases. For the rest cases, the selection criteria correctly identified the real number of clusters for $\sigma = 1.0$ of both BSP and K-means results. Similarly, both BSP and K-means had 0% accuracy for cases with $\sigma = 1.2, 1.4,$ and $1.5.$ Moreover, K-means had three more cases of 0% accuracy, $\sigma = 0.8, 0.9,$ and $1.1.$ In summary, BSP had higher accuracy using the selection criteria compared to K-means, even though they both had cases that the selection criteria could not find the true number of clusters for the 20 runs.

4.5. Gamma Distribution

4.5.1. Misclassified Rate Comparison

Simulation data set 3 generated the error terms (ε_{ijp}) of the linear model with different values of the standard deviation, same as previous distributions. The error terms were assumed to follow a gamma distribution, $\varepsilon_{ijp}{\sim}Gamma(\alpha, \beta)$. The shape and scale parameters, α and β, were calculated using the formulas: $\alpha = \frac{\mu^2}{\sigma^2}$ and $\beta = \frac{\sigma^2}{\mu}$, respectively. The values of σ ranged from 0.1 to 1.5 increasing by 0.1. For each case of σ, BSP and K-means were run 20 times for the 2,000 genes. Figure 4.8 shows the MR comparison between BSP and K-means for gamma distribution. More details on the MR values of BSP and K-means are shown in Appendix A.13 and A.14, respectively.

As shown in Figure 4.8, BSP successfully identified the true cluster of all genes for the first three cases with 0% MR. After the first three cases, the MR curve for BSP increased as standard deviation increased. When $\sigma = 1.5$, the MR of BSP was approximately 25%. The MR curve for K-means started around 17%, it decreased for the first six cases to close to 0% and increased in the remaining cases as variance increased. The standard errors of MR for the K-means algorithm in the first five cases were approximately 2.5%, which were much larger than the rest of the cases.

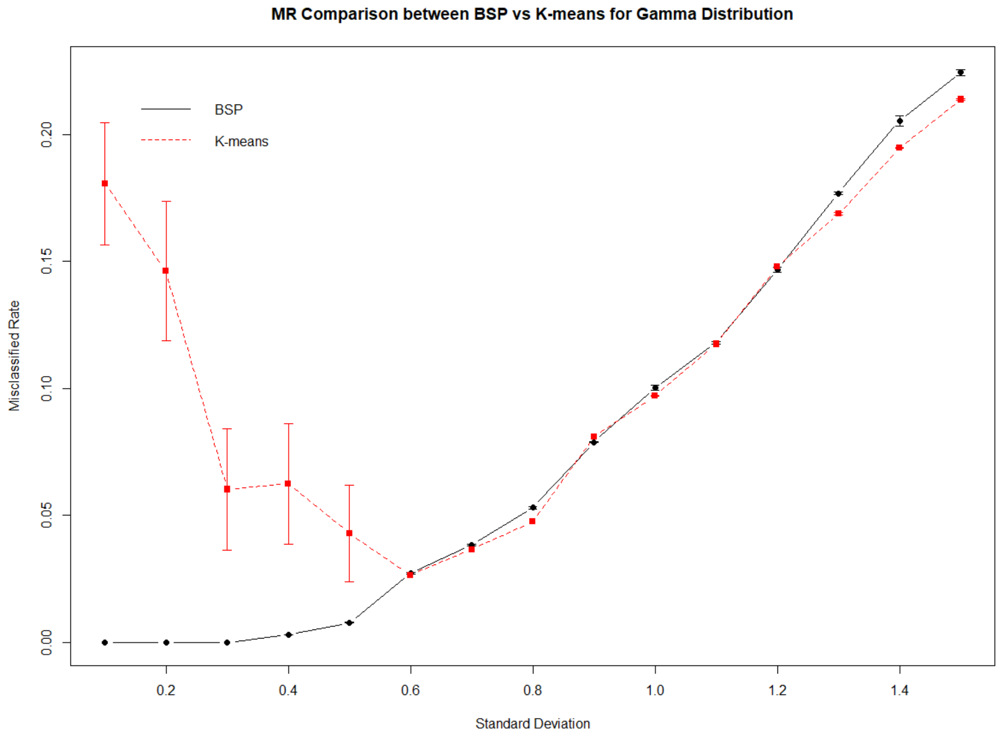


Figure 4.8. The Misclassified Rate Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gamma Distribution. Data shown as mean MR ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents MR. BSP: black solid line; K-means: red dashed line.

The number of DMGs were calculated for both the BSP and K-means classifications, which are shown in Appendix A.15 and A.16 for BSP and K-means, respectively. The plot comparison between BSP and K-means for gamma distribution is shown in Figure 4.9.

The number of DMGs for the BSP algorithm was close to 0 for all the 15 cases. While the number of DMGs for K-means started around 750, and as σ increased the number of DMGs decreased to close to 0 at $\sigma = 0.6$. Then for the remaining cases, the number of DMGs curves for both BSP and K-means had a similar trend.

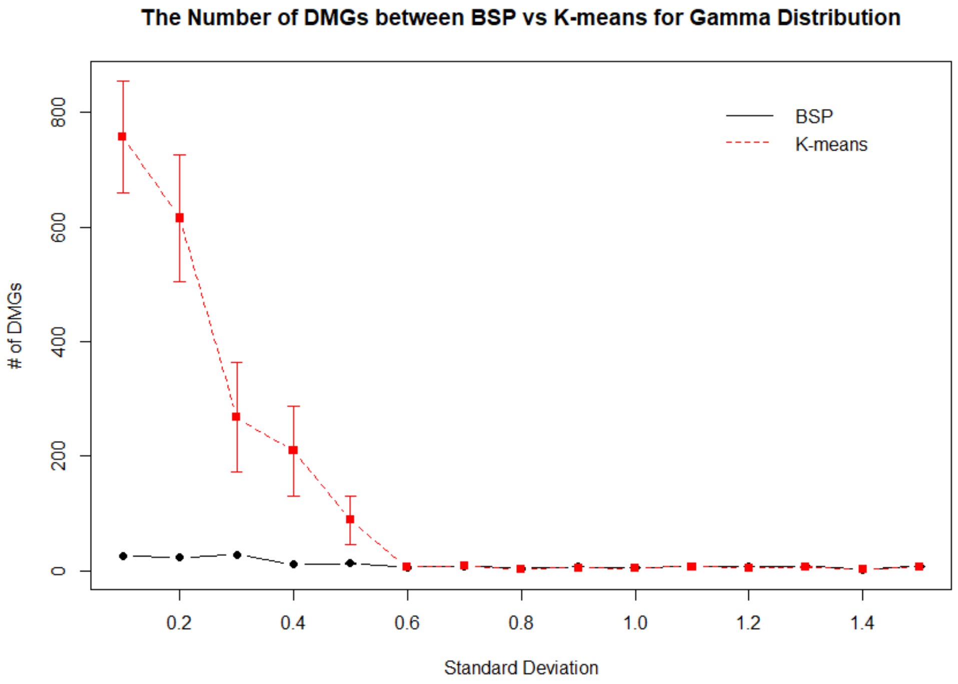


Figure 4.9. The Number of Differentially Measured Genes Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gamma Distribution. Data shown as mean number of DMGs ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the number of DMGs. BSP: black solid line; K-means: red dashed line.

4.5.2. The Accurate Number of Clusters Comparison

The accurate number of clusters were also calculated for gamma distribution to verify the selection criteria accuracy for each case. Appendix A.17 and A.18 show the selection criteria accuracy for BSP and K-means, respectively. The plot of selection criteria accuracy for BSP and K-means of gamma distribution is in Figure 4.10.

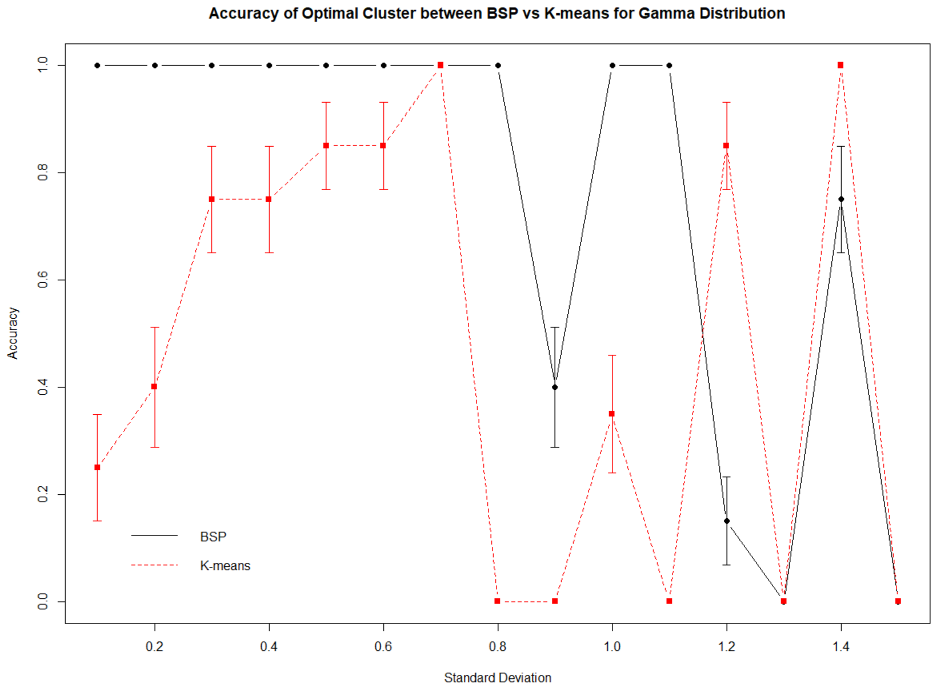


Figure 4.10. The Selection Criteria Accuracy Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gamma Distribution. Data shown as mean accuracy ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the accuracy. BSP: black solid line; K-means: red dashed line.

In Figure 4.10, the selection criteria accuracy varied greatly for both the BSP and K-means algorithms, ranging between 0% to 100%. For BSP, the selection criteria identified the true number of clusters for the first eight cases, as well as the cases of $\sigma = 1.0$ and 1.1, with 100% accuracy. On the other hand, the selection criteria only correctly identified the true number of clusters for two cases ($\sigma = 0.7$ and 1.4) for K-means results. For the remaining cases, the selection criteria accuracies of K-means were lower than those of BSP except for $\sigma = 1.2$. Overall, the accurate number of clusters was correctly identified in more cases for the BSP classification than that for the K-means classification. And the selection criteria accuracy of BSP had relatively smaller standard errors compared to those of K-means.

4.6. Gaussian Mixture Distribution

4.6.1. Misclassified Rate Comparison

Simulation data set 4 assumed the error terms follow a gaussian mixture distribution: $\varepsilon_{ijp} \sim N(\mu_z, {\sigma_z}^2)$, where $\mu_z = (0, \mu)$, $\sigma_z = (\sigma, \sigma)$, and $w_z = (0.5, 0.5)$. The values of σ were in the range between 0.1 and 1.5. The error terms were subtracted by $\frac{\mu}{2}$ to make the means shifted to 0. Both BSP and K-means run 20 times for each case of σ. As mentioned in the previous section, the true number of clusters for the simulation data was four. Figure 4.11 shows the MR comparison between BSP and K-means for gaussian mixture distribution. Individual MR values of the 15 cases for BSP and K-means are showed in Appendix A.19 and A.120, respectively.

In Figure 4.11, the MR results of BSP were more stable compared to those of K-means, especially for small values of σ. The MR for BSP started at 0% for the first three cases and gradually increased to more than 30% at $\sigma = 1.5$. The MR curve for K-means decreased for the first five cases from about 17% to 0% and then increased as variance increased.

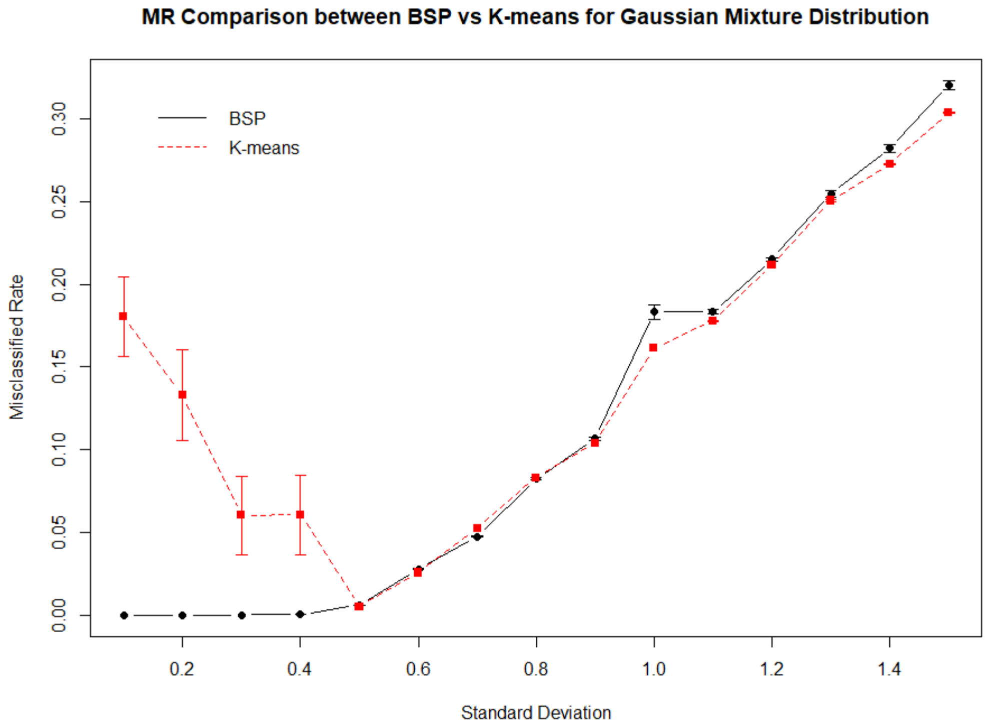


Figure 4.11. The Misclassified Rate Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gaussian Mixture Distribution. Data shown as mean MR ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents MR. BSP: black solid line; K-means: red dashed line.

After receiving the classification results of BSP and K-means for simulation data set 4, the residuals of the fitted linear model were calculated for both methods. The number of DMGs was identified based on the t-test of the residuals with the level of significance as 0.5. The results for the number of DMGs are shown in Appendix A.21 and A.22 for BSP and K-means, respectively. Figure 4.12 shows the plot comparison of the number of DMGs between BSP and K-means for gaussian mixture distribution.

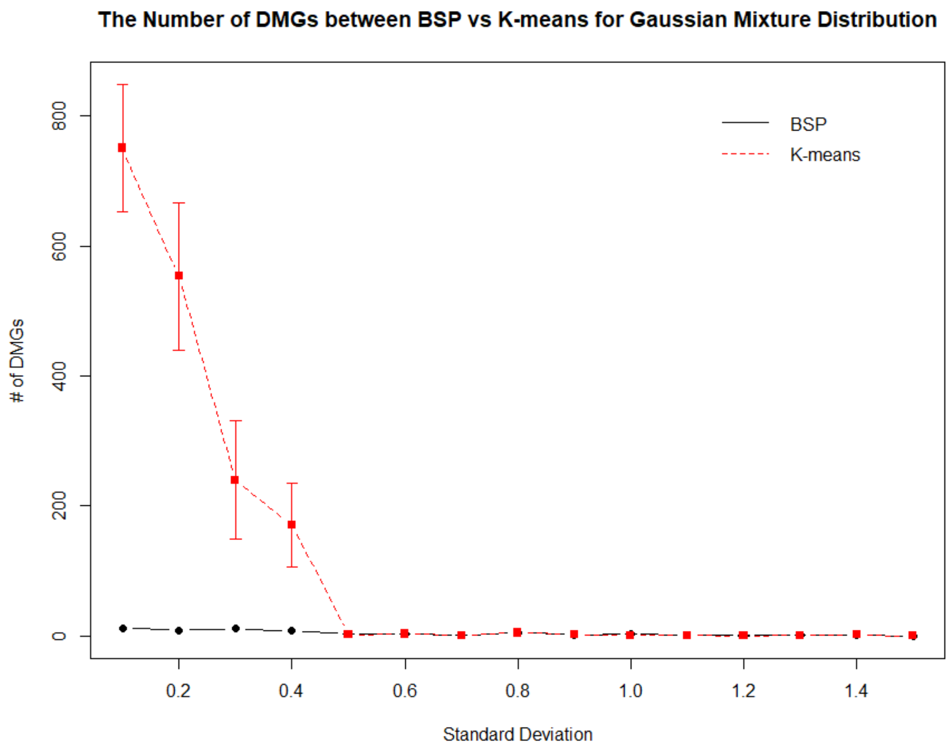


Figure 4.12. The Number of Differentially Measured Genes Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gaussian Mixture Distribution. Data shown as Mean number of DMGs ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the number of DMGs. BSP: black solid line; K-means: red dashed line.

Like the results of MR, the standard errors of the number of DMGs for the K-means method were very high in the first four cases compared to the rest of the cases. The number of DEGs for K-means were over 700 in the first case, and as σ increased the number of DMGs decreased to close to 0 after the fifth case. The number of DMGs for BSP algorithm was much stable and varied around 0.

4.6.2. The Accurate Number of Clusters Comparison

After obtaining the number of DMGs for the 15 cases of σ, the selection criteria were used to verify the optimal number of clusters. The selection criteria accuracy of gaussian mixture distribution are shown in Appendix A.23 and A.24 for BSP and K-means, respectively. The plot comparison of the accuracy for BSP and K-means is shown in Figure 4.13.

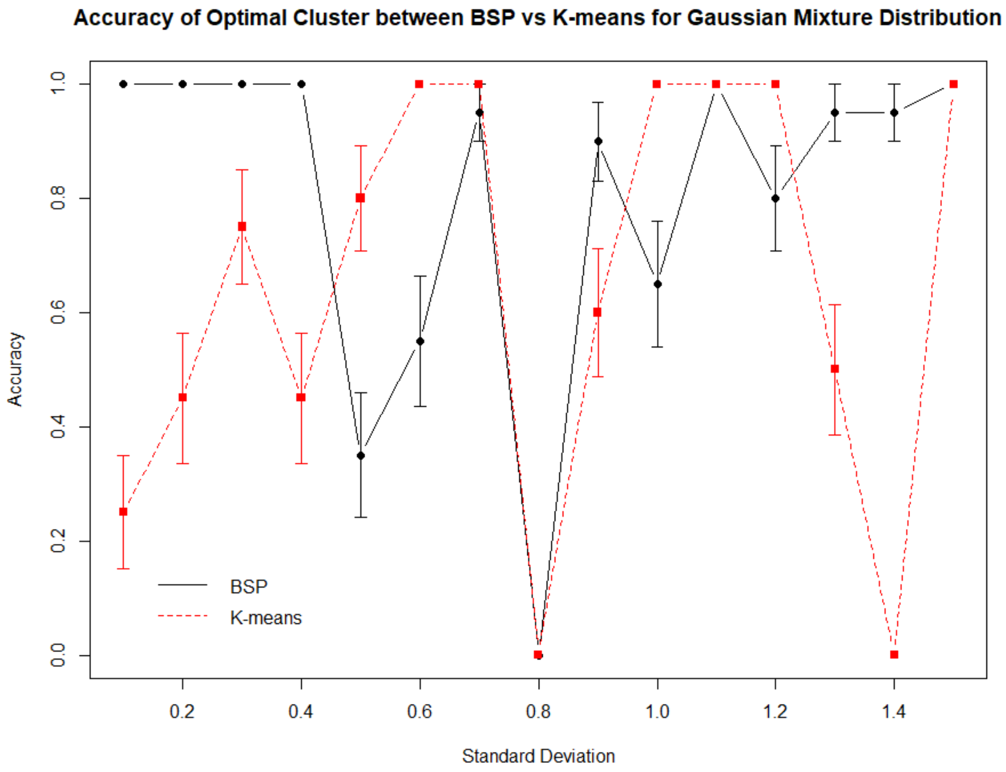


Figure 4.13. The Selection Criteria Accuracy Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Gaussian Mixture Distribution. Data shown as mean accuracy ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the accuracy. BSP: black solid line; K-means: red dashed line.

For BSP, the selection criteria had an accuracy over 60% in 12 out of 15 cases. While for K-means, only 9 out of 15 cases had accuracy more than 60%. The selection criteria correctly identified the real number of clusters five times and six times for BSP and K-means, respectively. K-means had one more case (2) with 0% accuracy than the cases of BSP (1).

4.7. Beta Distribution

4.7.1. Misclassified Rate Comparison

The beta distribution was used to generate the simulation data set 5. The values of the standard deviation for ε_{ijp} ranged from 0.1 to 0.5 due to the parameters' boundaries of beta distribution. The error terms of the expression values for the 2,000 genes were assumed to follow a beta distribution: $\varepsilon_{ijp} \sim beta(\alpha, \beta)$, where $\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)$ and $\beta = \alpha\left(\frac{1}{\mu} - 1\right)$. Then the mean values of the errors were subtracted from the error terms. For each case of σ, BSP and K-means were run 20 times to check the consistency of their results. Figure 4.14 shows the MR comparison between BSP and K-means for beta distribution. Detailed MR of BSP and K-means are showed in Appendix A.25 and A.26, respectively.

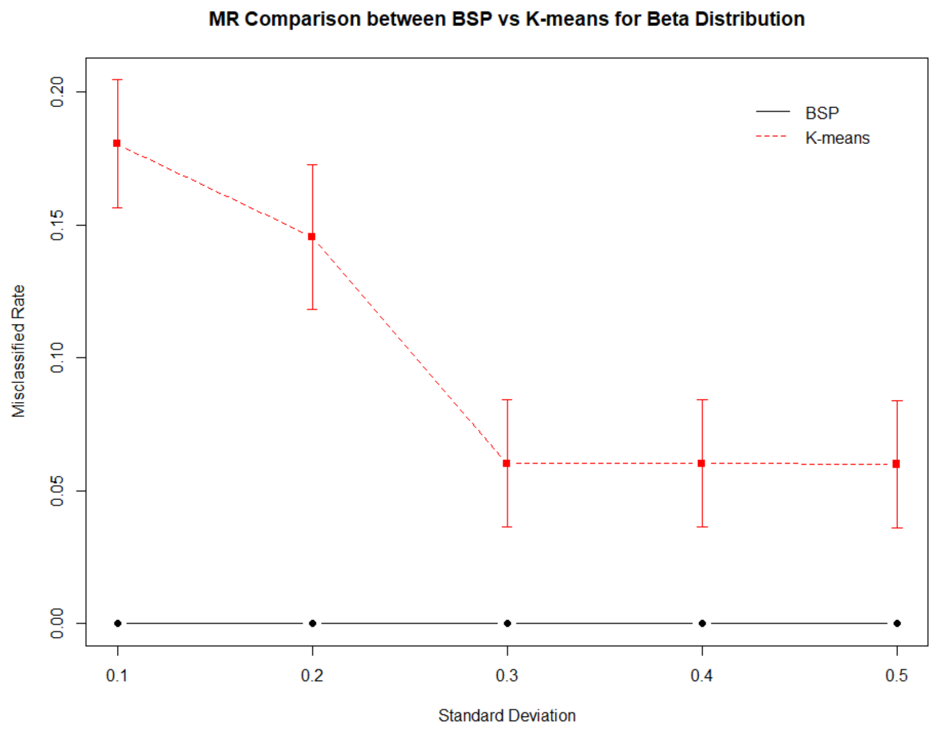


Figure 4.14. The Misclassified Rate Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Beta Distribution. Data shown as mean MR ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents MR. BSP: black solid line; K-means: red dashed line.

The residuals of fitted linear model were used to calculate the number of DMGs, which are shown in Appendix A.27 and A.28 for BSP and K-means, respectively. Figure 4.15 shows the plot comparison between BSP and K-means for beta distribution.

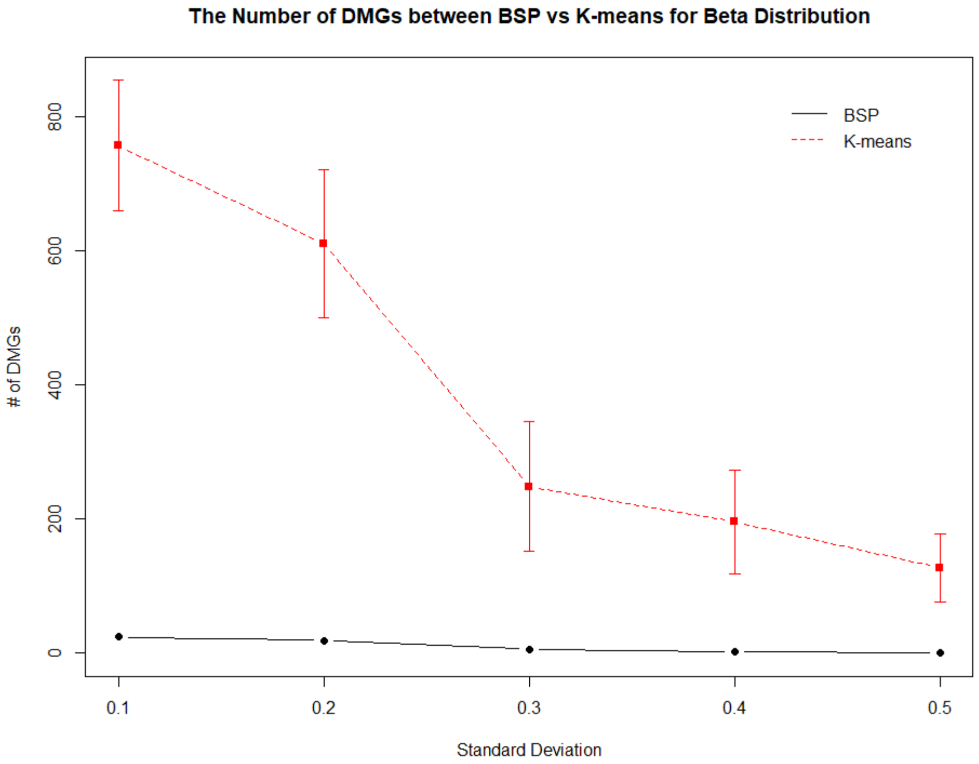


Figure 4.15. The Number of Differentially Measured Genes Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Beta Distribution. Data shown as mean number of DMGs ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the number of DMGs. BSP: black solid line; K-means: red dashed line.

Similar to the results of MR, the number of DMGs for BSP was varied around 0. And the number of DMGs for K-means had a decreasing trend started at approximately 700 to about 100.

4.7.2. The Accurate Number of Clusters Comparison

The selection criteria accuracies were also calculated for the 15 cases of σ, which are shown in Appendix A.29 and A.30 for BSP and K-means, respectively. Figure 4.16 shows selection criteria accuracy of BSP and K-means for the five cases of beta distribution.

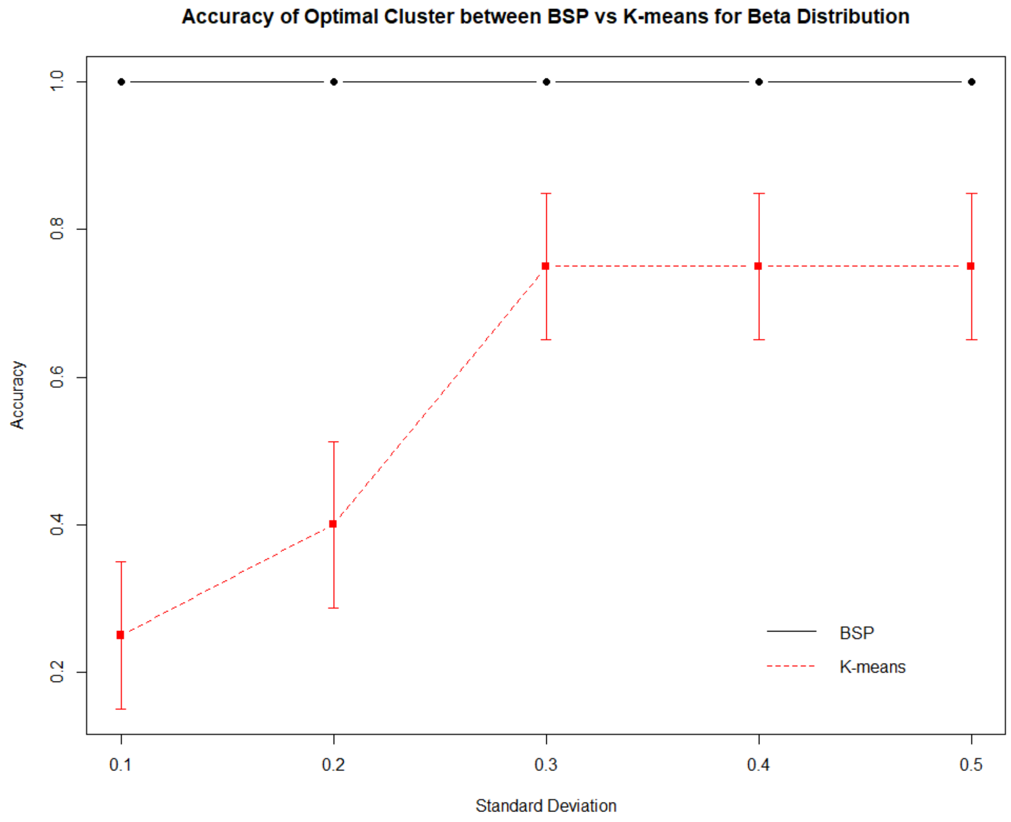


Figure 4.16. The Selection Criteria Accuracy Comparison between Boundary Shift Partition and Hartigan-Wong's K-means for Beta Distribution. Data shown as mean accuracy ± SE of 20 runs with different standard deviations. X-axis depicts standard deviation and y-axis represents the accuracy. BSP: black solid line; K-means: red dashed line.

Based on Figure 4.16, the selection criteria identified the true number of clusters for BSP with 100% accuracy in all five cases. In comparison, the selection criteria accuracies were below 80% for the K-means algorithm. The results indicated that the BSP classification was more stable than K-means on identifying the true number of clusters using the selection criteria, at least with small values of the standard deviation for the error terms.

4.8. Algorithm Time Complexity

Regarding the computational complexity for the BSP algorithm, the running time is $O(NK - K^2) \approx O(NK)$ for the worst scenario, where K is the number of clusters and N is the total number of genes in the data set. Since in reality, K is also unknown. The BSP algorithm considers a range of possible K values, which is between 2 and 20 in this research. Let $Ķ$ denotes the maximum number of clusters tested in BSP. In another word, the BSP algorithm would find the partitions for 2 to up to $Ķ$ clusters. The running time for this process would be $O(NĶ^2 - Ķ^3) \approx O(NĶ^2)$.

Another simulation data set 6 was generated to study the computational time of the algorithm. The number of genes in each case was 1,000, 2,500, 5,000, 7,500, 10,000, 12,500, 15,000, 20,000, 25,000, and 30,000. Similar with the previous simulation data, there were four clusters for each scenario and the standard deviation of the residuals (σ) was chosen to be 0.5. The algorithm searched the optimal number of cluster in the range of 2 clusters to 20 clusters. The stop criteria of the algorithm was that the sum of the standard deviations did not change in the last 10 iterations. Figure 4.17 shows the relationship between computational time and the gene size.

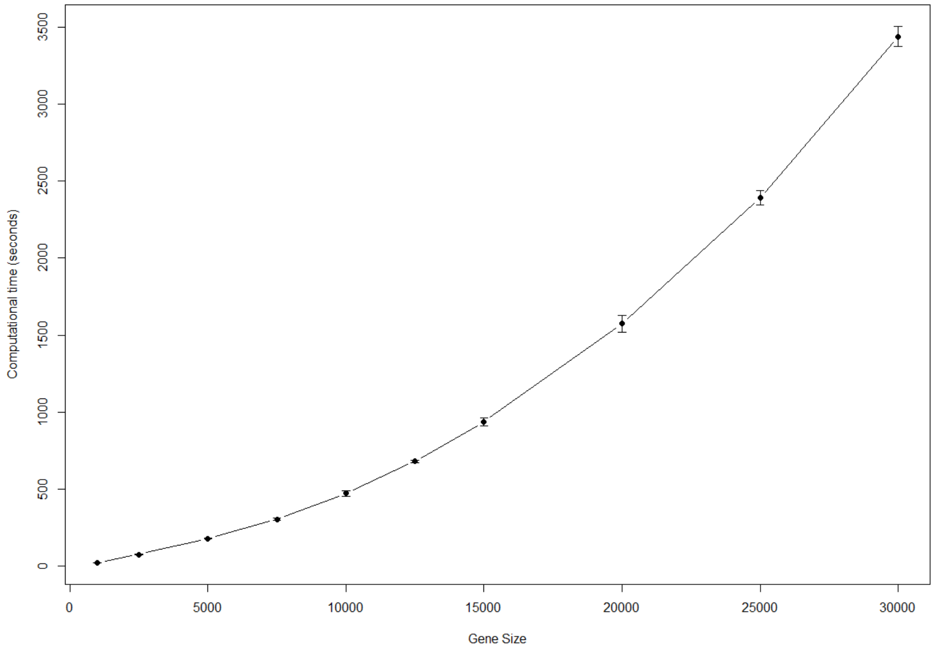


Figure 4.17. The Relationship Between Computational Time and the Gene Size. X-axis depicts the number of genes and y-axis represents the computational time in seconds.

As shown is Figure 4.17, the computation time increased when the number of genes in the data set increased. On average, the algorithm took 3437.44 seconds to classify 30,000 genes into 2 clusters to 20 clusters. One thing to be noticed, the computational time would vary for each run since the stop criteria was that the minimum of the sum of the standard deviations does not change in last 10 iterations. Figure 4.18 shows the computational time of one run for different number of clusters with the simulation data of 15,000 genes.

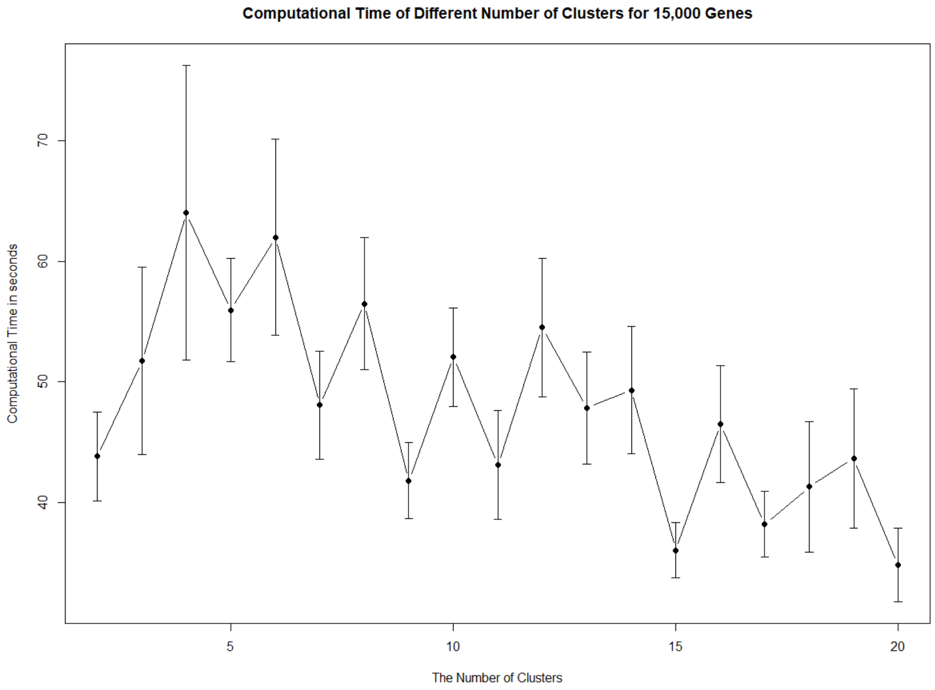


Figure 4.18. The Computational Time of Different Number of Clusters for 15,000 Genes. X-axis depicts the number of clusters and y-axis represents the computational time in seconds.

In Figure 4.18, the computational time changed between approximately 30 seconds and 100 seconds. There was a slightly decreasing trend for the computational time with the number of cluster increased, which indicated that the computational time had a negative relationship with the number of clusters.

Since normality is one of the assumptions for the residuals of the linear model, simulation data set 1 was generated using normal distribution. For simulation data set 1, the BSP algorithm successfully classified the 2,000 genes into their true clusters with less than 5% mean MRs where small standard deviations ($\sigma \leq 0.8$) were used, shown in Figure 4.2. There was clearly an increasing trend of the MR curve for BSP as the standard deviation increased. On the other hand, the MR curve for K-means started at a very high average MR (approximately 17%), and it decreased to close to 0% at $\sigma = 0.5$ then increased as σ increased.

The number of DMGs for the BSP algorithm was very stable with all the DMGs near 0 for the 15 cases of normal distribution (Figure 4.3). The DMGs curve for K-means started at about 700, decreased in the first five cases and then shared a similar trend for the DMGs curve of BSP.

The selection criteria accuracy for normal distribution varied greatly for both BSP and K-means, shown in Figure 4.4. The accurate number of clusters identified using the selection criteria for BSP in the first four cases was very high (over 95%) compared to those of K-means. As standard deviation increased, the selection criteria accuracy varied in the range of 0% to 100% for both BSP and K-means. Thus, the selection criteria worked well for the BSP classification with small values of standard deviation ($\sigma \leq 0.4$).

Even though normality is one of the assumptions for the linear model, the residuals are not always guaranteed to follow a normal distribution, which means the assumption might be invalid. Four other distributions, Exponential distribution, Gamma distribution, Gaussian Mixture distribution, and Beta distribution, were chosen to check the robustness of the BSP algorithm when normality was not valid. These four distributions shared similar patterns of the normal distribution for both the BSP and K-means classification methods. It is safe to say that performance of BSP is very stable with different underlying distributions.

CHAPTER 5. APPLICATION TO THE EMBRYONIC HEART DATA

5.1. Wild Type Embryonic Heart Data

The BSP algorithm was applied to the transformed embryonic heart data set with six samples consisting of 12,812 genes each. The four microarray samples used the logarithm transformation and the two RNA-seq samples used the cubic root transformations. Figure 5.1 shows the scatterplots of microarray and RNA-seq before and after transformation.

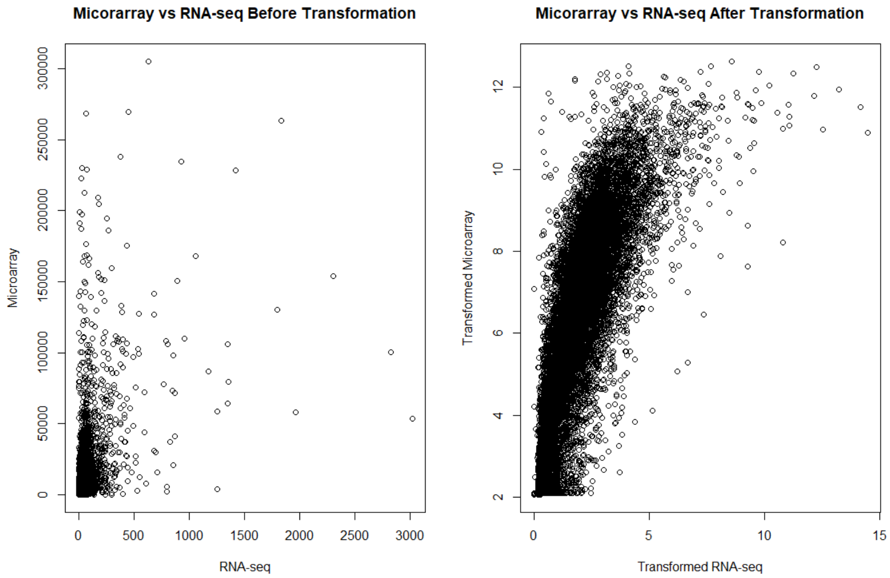


Figure 5.1. The Scatterplots of Microarray and RNA-seq Before and After Transformation. The plot on the left was the scatterplot of microarray and RNA-seq before transformation. The plot on the right was the scatterplot after transformation.

The BSP algorithm stop criteria was that the minimum of the sum of the standard deviations did not change in last 1,000 iterations. The increase on the number of stop iterations compared to the simulation analysis would help the BSP algorithm to reach the global minimum. The regression model was fitted using the algorithm results with the technology effects and gene effects.

To fit the linear model, the expression values were treated as the dependent variable and the gene effect and technology effect were treated as the independent variables. Because the technology had two levels, only 1 indicator variable was used for the technology effect. The RNA-seq technology was used as the baseline.

$$\text{Technology} = \begin{cases} 0, & if\ Tech = RNAseq \\ 1, & if\ Tech = Microarray \end{cases} \tag{5.1}$$

Thus, the simple linear model fitted was in this form:

$$Expression_{ijp} = \alpha_i + \beta_j Tech + \varepsilon_{ijp} \tag{5.2}$$

The residual for each observation was calculated to identify the number of the DMGs for each cluster using the t-test. The optimal number of clusters was chosen as the one that minimized the selection criteria:

$$log_2\left(\frac{\#DMGs + 1}{N}\right) + k \tag{5.3}$$

The algorithm detected five clusters to be the optimal number of clusters for the embryonic heart data set, shown in Table 5.1. The corresponding number of the DMGs was 471 using 0.05 as the significant level, which was about 3.676% (471/12,812) of the genes in the data set. All the DMGs are listed in Appendix B. Since the percentage of the DMGs was less than 5%, the type I error rate, which is usually chosen as 0.5, was well controlled.

Table 5.1. The Number of DMGs and Criteria for the Embryonic Heart Data Set

Cluster	# of DMGs	Criteria
2	5052	0.657716
3	2211	0.465928
4	956	0.257167
5	471	0.237435
6	391	0.969502
7	372	1.897824
8	299	2.583611
9	293	3.554464
10	297	4.57396
11	302	5.597966
12	288	6.529718
13	278	7.478913
14	290	8.539667
15	283	9.504539
16	249	10.32058
17	263	11.39919
18	240	12.26768
19	270	13.43694
20	255	14.35479

The summary of the probe alignments for the DMGs with their clusters is shown in Table 5.2. 30.79% (145/471) of the genes among all the DMGs had either multiple alignments or no alignment. Compared to the values in Table 2.14, the total percentage of the genes with multiple alignments and NAs in the DMGs was much higher than that of genes for the whole data set. For genes with multiple alignments, the percentage increased from 6.45% (826/12,812) in the whole data set (shown in Table 2.14) to 21.66% (102/471) in DMGs. But the percentage of genes with NAs decreased from 12.78% (1,638/12,812, shown in Table 2.14) to 9.13% (43/471) in the

DMGs. Therefore, multiple alignments (> 30%) is one of the reasons for the genes to be identified as differentially expressed.

Table 5.2. Probe Alignments Summary for DMGs

	# of Alignments	DMGs
	[2,5]	80
Multiple Alignments	(5,10]	14
	(10,∞)	8
NAs	0	43
Total		145

Table 5.3 shows the DMGs probe alignments based on the five clusters. All the DMGs were either in cluster 1 or in cluster 5, which means that these DMGs were most likely to have extreme values. The majority of the DMGs (77.27%) in cluster 1 did not have alignments (NAs). While 92.09% (92/101) of the DMGs in cluster 5 had at least two alignments.

Table 5.3. Probe Alignments Summary for DMGs by Clusters

	# of Alignments	Cluster 1	Cluster 5
	[2,5]	8	72
Multiple Alignments	(5,10]	2	12
	(10,∞)	0	8
NAs	0	34	9
Total		44	101

5.2. $\boldsymbol{Tbx5^{+/-}}$ Embryonic Heart Data

Tbx5 belongs to the T-box transcription factor family and involves in the regulation of cardiac and forelimb developmental processes. Mutations of *Tbx5* cause Holt–Oram syndrome (HOS; OMIM142900) in humans (Li, et al., 1997; Bruneau, et al., 2001). HOS is a rare penetrant disorder (affects ∼1 in 100,000 livebirths) which may result in congenital heart disease including abnormal heart rates and arrhythmias (Basson, et al., 1999; Petal, Silcock, McMullan, Brueton, & Cox, 2012).

Four samples of the *Tbx5* embryonic heart data were obtained from *Tbx5* heterozygous mouse line. The Affymetrix microarray technology was used to measure the expression values of *Tbx5* samples. Same as the four wild type microarray samples, the expression values of the *Tbx5* samples were logarithm transformed to prevent the magnitude problem between the microarray and RNA-seq technology. Then the expression values of the 12,812 selected genes from the six wild type samples were combined with those from the four *Tbx5* microarray samples to study the DEGs between the two genotypes.

According to the BSP results of the wild type embryonic heart data, shown in Table 41, the optimal number of cluster was five. The same parameter estimations of the linear model were used to calculate the residuals of the genes for the *Tbx5* samples. The DEGs were defined to be the genes that had the Benjamini and Yekutieli's (BY) adjusted (Benjamini & Yekutieli, 2001) p-value less than 0.05 in the two-sample t-test between *Tbx5* samples and wild type samples.

There were 584 DEGs found between *Tbx5* samples and wild type samples. The probe alignments for these 584 genes by their clusters are shown in Table 5.4.

Table 5.4. Probe Alignments Summary for DEGs between Genotypes by Clusters

	# of Alignments	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Multiple Alignments	[2,5]	5	8	4	5	8
	(5,10]	1	1	0	1	1
	(10,∞)	0	0	0	0	0
NAs	0	18	24	27	15	9
Total		24	33	31	21	18

The number of the DEGs for the five clusters was 68, 116, 164, 136, and 100, respectively. The percentage of the NAs in cluster 1 to cluster 5 were ranging from 50% to about

90% (75.00%, 72.73%, 87.10%, 71.43%, 50.23%, respectively). The percentage of the multiple alignments and NAs varied from 15% to 35% for these five clusters.

A gene co-expression network analysis was conducted using the GeneNet (Schaefer, Opgen-Rhein, & Strimmer, 2015) package in R software. The residual matrix for the 584 DEGs was used as input to calculate the partial correlation in the Graphical Gaussian Models. Among these DEGs, *Tbx5* co-expressed with four genes, *Osr1*, *Adamts1*, *Wnt4*, and *Lhx1*. A gene co-expression network among these four genes was generated using GeneMANIA (Warde-Farley, et al., 2010) website, shown in Figure 5.2.

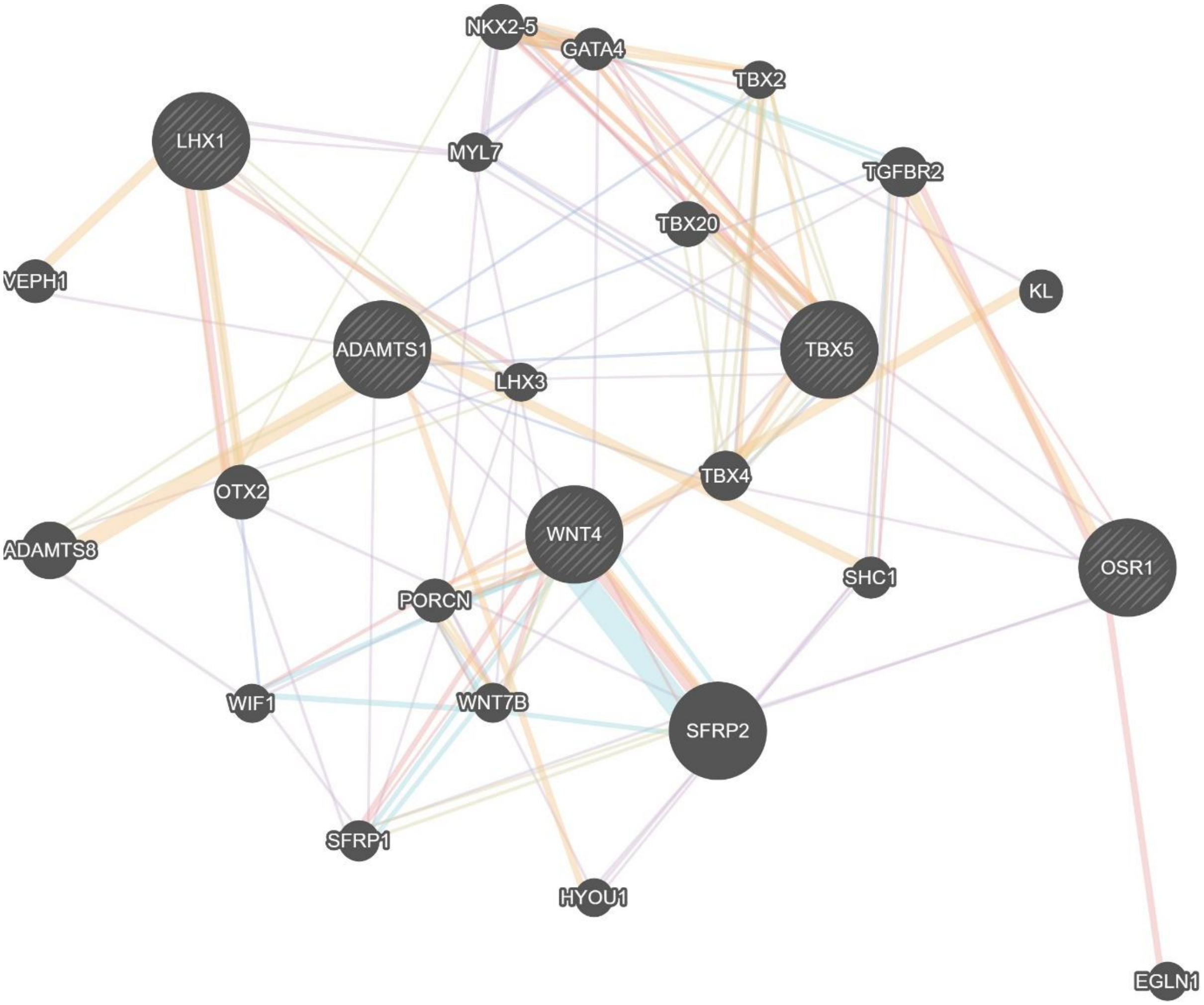


Figure 5.2. Gene Co-expression Network of *Tbx5*, *Osr1, Adamts1, Wnt4,* and *Lhx1*.

In conclusion, the BSP algorithm was applied to the transformed embryonic heart data set with 12,812 genes of the six wild type samples. 3.676% (471/12,812) of the genes in the mice heart data set were identified to be differentially measured between the microarray and RNA-seq technologies by this method. Since all the samples in this data set were biological replicates, it assumed that the expression levels for these 12,812 genes among the 6 samples were the same. The percentage of the DMGs (3.676%) was the type I error rate for the BSP algorithm, which was lower than the commonly used cutoff 0.5. In another word, the type I error rate was well controlled by the BSP classification method.

The results of BSP algorithm were also used to identify the DEGs between the four *Tbx5* samples and the six wild type samples. 4.56% (584/12,812) of the genes was identified as differentially expressed between these two genotypes. As expected, *Tbx5* was one of the DEGs among these 584 genes. *Gata4* and *Tbx5* are the two transcription factors that interact with each other and play important roles in heart development (Bruneau, et al., 2001; Moskowitz, et al., 2007; McCulley & Black, 2012). The BSP algorithm successfully listed *Gata4* as the one of the DEGs using 0.05 as the significant levels for t-test between *Tbx5* samples and wild type samples. *Gata4* was down-regulated in the *Tbx5* samples due to the down-expression of *Tbx5* gene.

CHAPTER 6. DISCUSSION

Microarray and RNA-seq are two commonly used high-throughput technologies for transcriptome analysis. Microarray is relatively inexpensive, and the data analysis is relatively easy with the availability of many user-friendly software tools. Nevertheless, microarray technology has several limitations, such as high background noise level and inaccuracy of probe detection. On the other hand, RNA-seq is a newer technology with a higher level of sensitivity and specificity. However, due to the large amount of raw sequencing data obtained from RNA-seq, the preprocessing and analysis of RNA-seq data is time consuming and requires researchers proficient with have bioinformatics knowledge and programming languages. Most of the software designed for RNA-seq are Linux-based with scripts written in different programming languages. Even though there are plenty of software packages available for RNA-seq data analysis, there is not yet one standard protocol or pipeline. With all the advantages and disadvantages for microarray and RNA-seq, data integration would help to reduce the experimental cost and increase the statistical power by increasing the sample size. The goal of this research is to find an efficient way to combine the data sets from these two technologies and identify the DEGs using the combined data.

The data integration between microarray and RNA-seq is more challenging than integration within technologies. There are two methods, which we are aware of, that were proposed to solve this kind of problem: Training Distribution Matching (TDM) approach (Thompson, Tan, & Greene, 2016) and a rank-based semi-parametric model (Lyu & Li, 2016). The TDM normalizes the RNA-seq data using quantile information to ensure a same distribution between RNA-seq and microarray data. The TDM shows a similar performance as quantile normalization in both simulated and real data. The TDM has a higher accuracy than quantile normalization in the cases with higher noise level. One of the downsides of TDM is that it assumes the discrepancy between microarray and RNA-seq is the same for all the genes. The rank-based semi-parametric model classifies the gene expression levels into three categories, non-DEGs, up-regulated DEGs, and down-regulated DEGs, using an extended copula mixture model. The rank-based semi-parametric model has a better performance on DEGs detection compared to other methods including DEseq and eBays. But the rank-based semi-parametric model has less statistical power compared with parametric models. Both the TMD and rank-based semi-parametric model fail to consider the variable discrepancy among the subsets of genes, which is confounding within the data set.

By comparison, the BSP removed the batch effects from the integrated data set by removing the technology effects in the linear model. Other than removing batch effect, there are more advantages for BSP algorithm. The performance of the BSP algorithm was consistent in the simulation data of the five different underlying distributions. The BSP algorithm correctly identified the true clusters of greater than 95% genes in almost all the cases with standard deviations ≤ 0.6 among the five distributions, which indicates that the BSP algorithm is robust for various distributions.

Additionally, BSP results had a higher accuracy than the results of K-means in the cases of small standard deviations. Moreover, the number of DMGs identified by BSP were close to type I error rate 5% in all cases with different values of σ. Since in the simulation data, all the genes were generated under the same condition among samples, there were no DMGs theoretically. Therefore, any DMGs in simulation data are false positive, which is a component of the type I error. This result indicated that the type I error rate was well controlled by our method.

Since the BSP algorithm used the sorted difference values of the genes as input, it highly reduced the computational complexity. The computational time for BSP was $O(NK - K^2) \approx O(NK)$ for the worst scenario, which was linear to the number of genes in the data set. Moreover, the computational time had a slightly decreasing trend as the number of cluster increasing. Our simulation study validated the theoretical results of the time complexity. Therefore, BSP is an efficient algorithm for data integration between microarray and RNA-seq.

However, there are some limitations to the BSP algorithm. First of all, the BSP might be trapped in a local minimum in some of the runs, as shown in some of the simulation results. One possible solution is to increase the number of iterations in the BSP stop criteria, and it is always recommended to use a large number of iterations in the data sets with large numbers of genes. But the computational time would be increased accordingly. Another solution is to generate more optimal initial partitions with other conventional clustering algorithms.

Secondly, we have only applied the BSP algorithm to the complete randomized design, which is the simplest experimental design. We expect the BSP algorithm can be applied to more complicated design when the corresponding linear model is fitted to the data. A related question is, can our method be used for the study where there are no matched samples between microarray and RNA-seq? Theoretically it is impossible to integrate non-matching data because the technology effect is confounded with the treatment effects. Nevertheless, it is feasible to use matched samples from a different study that uses the same microarray and RNA-seq platforms. Future studies will be devoted to implement the method for such non-matching scenario and to test the performance.

Thirdly, the selection criteria for the optimal number of clusters was chosen to detect the “elbow point”. It worked well for BSP in the cases with small standard deviations. More extensive simulation study should be conducted for validating the selection criteria. Future research need to be done for providing the theoretical proof of the selection criteria.

When our method was used for comparing the gene expression profiles between wildtype and *Tbx5*$^{+/-}$ in embryonic mouse hearts, 584 genes have been found to be exponentially expressed. A number of the DEGs were known to be involved in heart development and their mutations induce congenital heart diseases. For instance, *Osr1* gene was on the top DEGs that showed high partial correlation with *Tbx5* in our analysis. The previous studies by our lab and other research groups have found that *Osr1* is the downstream gene of *Tbx5* and the loss of function of *Osr1* induces atrial septal defects and out flow track abnormality (Xie, et al., 2012; Zhou, et al., 2015; Zhang, et al., 2016).

In this study, we have developed the BSP algorithm for robustly, efficiently and accurately removing the technology effect between microarray and RNA-seq and thus provided integrative analysis of microarray and RNA-seq data. The simulation study and real data application showed that the proposed method achieved well controlled type I error rate and better performance than the conventional clustering method, K-means. This study provides researchers a novel and rigorous approach to combine microarray and RNA-seq data, for an increased power for the downstream statistical analysis.

CHAPTER 7. REFERENCES

Abayasekara, L. M., Perera, J., Chandrasekharan, V., Gnanam, V. S., Udunuwara, N. A., Liyanage, D. S., . . . Tharmakulasi. (2017). Detection of bacterial pathogens from clinical specimens using conventional microbial culture and 16S metagenomics: a comparative study. *BMC Infectious Diseases, 17*(1), 631.

Abdi, H., & William, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(40), 433-459.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014). *Molecular Biology of the Cell* (6 ed.). Garland Science.

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.

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America, 97*(18), 10101-10106. doi:10.1073/pnas.97.18.10101

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Andrews, S. (2016, August 3). *FastQC*. Retrieved from Babraham Bioinformatics: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baldi, P., & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics, 17*, 509-519.

Basson, C. T., Huang, T., Lin, R. C., Bachinsky, D. R., Weremowicz, S., Vaglio, A., . . . Seidman, C. E. (1999). Different TBX5 interactions in heart and limb defined by Holt–Oram syndrome mutations. *Proceedings of the National Academy of Sciences of the United States of America, 96*(6), 2919-2924.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., & Marron, J. (2004). Adjustment of systematic microarray data biases. *Bioinformatics, 20*(1), 105-114. doi:https://doi.org/10.1093/bioinformatics/btg385

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research, 40*(10), e72. doi:10.1093/nar/gks001

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165-1188. doi:10.1214/aos/1013699998

Bien, S., Auer, P., Harrison, T., Qu, C., Connolly, C., Greenside, P., . . . Banbury. (2017). Enrichment of colorectal cancer associations in functional regions: Insight for using epigenomics data in the analysis of whole genome sequence-imputed GWAS data. *PLoS One, 12*(11), e0186518. doi:10.1371/journal.pone.0186518

Bilban, M., Buehler, L. K., Head, S., Desoye, G., & Quaranta, V. (2002). Normalizing DNA microarray data. *Current Issues in Molecular Biology, 4*, 57-64.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry, 72*(1), 291-336. doi:10.1146/annurev.biochem.72.121801.161720

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185-193.

Borrill, P., Ramirez-Gonzalez, R., & Uauy, C. (2016). expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant physiology, 170*(4), 2172-2786. doi:10.1104/pp.15.01667

Box, G. E., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211-252.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology, 34*(5), 525-527. doi:10.1038/nbt.3519

Bruneau, B. G., Nemer, G., Schmitt, J. P., Charron, F., Robitaille, L., Caron, S., . . . Seidman, J. G. (2001). A Murine Model of Holt-Oram Syndrome Defines Roles of the T-Box Transcription Factor Tbx5 in Cardiogenesis and Disease. *Cell, 106*(6), 709-721. doi:https://doi.org/10.1016/S0092-8674(01)00493-7

Burrows, M., & Wheeler, D. (1994). *A block-sorting lossless data compression algorithm.* Palo Alto, Calfornia: Digital Equipment Corporation.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., & Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research, 10*, 2022-2029.

Chang, Z., Wang, Z., & Li, G. (2014). The impacts of read length and transcriptome complexity for De Novo assembly: A simulation study. *PLoS ONE, 9*(4), e94825.

Chou, C.-C., Chen, C.-H., Lee, T.-T., & Peck, K. (2004). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research, 31*(12), e99.

Chudasama, P., Mughal, S., Sanders, M., Hübschmann, D., Chung, I., Deeg, K., . . . Sch. (2018). Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nature Communications, 9*, 144. doi:10.1038/s41467-017-02602-0

Churchill, G. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics, 32*, Suppl:490-495. doi:10.1038/ng1031

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767-1771. doi:10.1093/nar/gkp1137

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13. doi:10.1186/s13059-016-0881-8

Czechowski, T., Bari, R. P., Stitt, M., Scheible, W.-R., & Udvardi, M. K. (2004). Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *The Plant Journal : for cell and molecular biology, 38*(2), 366-379. doi:10.1111/j.1365-313X.2004.02051.x

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., . . . Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics, 11*, S7. doi:10.1186/1471-2164-11-S4-S7

David, P., Gvr, C., Joel, G., Joanne, H., Jonathon, P., Chang, K., . . . S, K. E. (2005). Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics, 6*(1), 63. doi:10.1186/1471-2164-6-63

de Boer, B., van den Berge, G., de Boer, P. A., Moorman, A. F., & Ruijter, J. M. (2012). Growth of the developing mouse heart: An interactive qualitative and quantitative 3D atlas. *Developmental Biology, 368*(2), 203-213.

Deluca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., . . . Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics, 28*(11), 1530-1532. doi:10.1093/bioinformatics/bts196

DeRisi, J., Penland, L., Brown, P. O., Meltzer, P. S., Ray, M., Chen, Y., . . . Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics, 14*(4), 457–460. doi:10.1038/ng1296-457

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics, 22*(2), 101-109. doi:10.1016/j.tig.2005.12.005

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30*(1), 207-210.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 14863-14868.

Evangelistella, C., Valentini, A., Ludovisi, R., Firrincieli, A., Fabbrini, F., Scalabrin, S., . . . Harfouche, A. (2017). De novo assembly, functional annotation, and analysis of the giant reed (Arundo donax L.) leaf transcriptome provide tools for the development of a biofuel feedstock. *Biotechnology for Biofuels*, 138. doi:10.1186/s13068-017-0828-7

Fisher, R. A. (1970). *Statistical Methods for Research Workers* (14 ed.). Edinburgh: Oliver and Boyd.

Fodor, S. P., Read, J., Pirrung, M. C., Stryer, L., Lu, A. T., & Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *science, 251*(4995), 767.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*(21), 768–769.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Satistical Software, 33*(1), 1-22.

Gagnon-Bartsch, J. A., & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics, 13*(3), 539-552. doi:doi:10.1093/biostatistics/kxr034

Gong, Y., Huang, H.-T., Liang, Y., Trimarchi, T., Aifantis, I., & Tsirigos, A. (2017). lncRNA-screen: an interactive platform for computationally screening long non-coding RNAs in large genomics datasets. *BMC Genomics, 18*(1), 434. doi:10.1186/s12864-017-3817-0

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., . . . Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-seq unified mapper (RUM). *Bioinformatics, 27*(18), 2518-2528. doi:10.1093/bioinformatics/btr427

Gusnanto, A., Calza, S., & Pawitan, Y. (2007). Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology, 18*(2), 182-193.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., & Beyene, J. (2009). Data Integration in Genetics and Genomics: methods and challenges. *Human Genomics and Proteomics: HGP, 2009*, 869093. doi:10.4061/2009/869093

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108. doi:10.2307/2346830

Holland, M. J. (2002). Transcript abundance in yeast varies over six orders of magnitude. *The Journal of Biological Chemistry, 277*(17), 14363-14366. doi:10.1074/jbc.C200101200

Huang, H., Lu, X., Liu, Y., & Marron, J. (2012). R/DWD: Distance Weighted Discrimination for Classification, Visualization and Batch Adjustment. *Bioinformatics, 28*(8), 1182-1183.

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., & Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature, 417*, 63-67.

Hughey, J. J., & Butte, A. J. (2015). Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Research, 43*(12), e79. doi: https://doi.org/10.1093/nar/gkv229

Iam-On, N., & Boongoen, T. (2012). A New Locally Weighted K-Means for Cancer-Aided Microarray Data Analysis. *Journal of Medical Systems, 36*(Suppl 1), 43-49. doi:10.1007/s10916-012-9889-0

Illumina, I. (2017). *An introduction to Next-Generation Sequencing Technology.* Retrieved from Illumina: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Jacob, C., Tan, J., Miller, B., Tan, A., Takala-Harrison, S., Ferdig, M., & Plowe, C. (2015). A microarray platform and novel SNP calling algorithm to evaluate Plasmodium falciparum field samples of low DNA quantity. *BMC Genomics, 15*(1), 719. doi:10.1186/1471-2164-15-719

Jacob, L., Gagnon-Bartsch, J. A., & Speed, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics, 17*(1), 16-28. doi:doi:10.1093/biostatistics/kxv026

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666. doi:doi.org/10.1016/j.patrec.2009.09.011

Jayaraman, A., Hall, C. K., & Genzer, J. (2006). Computer Simulation Study of Molecular Recognition in Model DNA Microarrays. *Biophysical Journal, 91*(6), 2227-2236. doi:10.1529/biophysj.106.086173

Johnson, W., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127.

Kamtchueng, C., Delannoy, A., Wilhelm, E., Léger, H., Benecke, A., & Bell, B. (2014). Alternative Splicing of TAF6: Downstream Transcriptome Impacts and Upstream RNA Splice Control Elements. *PLoS One, 9*(7), e102399. doi:10.1371/journal.pone.0102399

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., & Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research, 28*(22), 4552-4557. doi:https://doi.org/10.1093/nar/28.22.4552

Kawasaki, E. S. (2006). The end of the microarray tower of babel: Will universal standards lead the way? *Journal of Bimolecular Techniques, 17*(3), 200-206.

Kerr, M., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., & Churchill, G. A. (2000). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica, 12*, 203-217.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*, R36. doi:10.1186/gb-2013-14-4-r36

Laird, P. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews. Genetics., 11*(3), 1991-203. doi:10.1038/nrg2732

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10*(3), R25.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., . . . Nowe, A. (2012). Batch effect removalmethods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics, 14*(4), 469-490.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics, 26*(4), 493-500. doi:https://doi.org/10.1093/bioinformatics/btp692

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research, 18*(11), 1851–1858. doi:10.1101/gr.078212.108

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., . . . Nelson, T. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature Genetics, 42*(12), 1060-1067. doi:10.1038/ng.703

Li, Q. Y., Newbury-Ecob, R. A., Terrett, J. A., Wilson, D. I., Curtis, A. R., Yi, C. H., . . . Brook, J. D. (1997). Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nature Genetics, 15*, 21-29. doi:doi:10.1038/ng0197-21

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics, 5*(3), 1752-1779. doi:10.1214/11-AOAS466

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics, 25*(5), 713-714. doi: https://doi.org/10.1093/bioinformatics/btn025

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery, 6*, 393-423.

Liu, M., Hou, X., Zhang, P., Hao, Y., Yang, Y., Wu, X., . . . Guan, Y. (2013). Microarray gene expression profiling analysis combined with bioinformatics in multiple sclerosis. *Molecular Biology Reports, 40*(5), 3731-3737. doi:10.1007/s11033-012-2449-3

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEET RANSACTIONSO N INFORMATIONT HEORY, 28*(2), 129-137.

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

Løvf, M., Thomassen, G. O., Mertens, F., Cerveira, N., Teixeira, M. R., Lothe, R. A., & Skotheim, R. I. (2013). Assessment of Fusion Gene Status in Sarcomas Using a Custom Made Fusion Gene Microarray. *PLoS ONE, 8*(8), e70649. doi:10.1371/journal.pone.0070649

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., . . . Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The Pharmacogenomics Journal, 10*, 278-291. doi:10.1038/tpj.2010.57

Lyu, Y., & Li, Q. (2016). A semi-parametric statistical model for integrating gene expression profiles across different platforms. *BMC Bioinformatics, 17*(Suppl 1), 5. doi:10.1186/s12859-015-0847-y

Ma, T., Liang, F., & Tseng, G. C. (2017). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using Bayesian hierarchical models. *Journal of the Royal Statistical Society. Series C, Applied Statistics, 66*(4), 847-867. doi:10.1111/rssc.12199

Ma, X., Xiao, L., & Wong, W. H. (2014). Learning regulatory programs by threshold SVD regression. *Proceedings of the National Academy of Sciences of the United States of America, 111*(44), 15675-15680. doi:10.1073/pnas.1417808111

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). Cluster: Cluster Analysis Basics and Extensions.

MAQC Consortium, Reid, L., Jones, W., Shippy, R., Warrington, J., Baker, S., . . . Setterquist, R. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology, 24*(9), 1151-1161. doi:10.1038/nbt1239

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Godwin, B. C. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.

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

Marot, G., & Mayer, C.-D. (2009). Sequential analysis for microarray data based on sensitivity and meta-analysis. *Statistical Applications in Genetics and Molecular Biology, 8*(1), Articale 3. doi:10.2202/1544-6115.1368

Marron, J., Todd, M. J., & Ahn, J. (2007). Distance-Weighted Discrimination. *Journal of the American Statistical Association, 102*(480), 1267-1271.

McCulley, D. J., & Black, B. L. (2012). Transcription Factor Pathways and Congenital Heart Disease. *Current Topics in Developmental Biology, 100*, 253-277. doi:doi: 10.1016/B978-0-12-387786-4.00008-7

McGrath, C. L., & Katz, L. A. (2004). Genome diversity in microbial eukaryotes. *Trends in Ecology and Evolution, 19*(1), 32-38.

Mcloughlin, K. S. (2011). Microarrays for Pathogen Detection and Analysis. *Briefings in Functional Genomics, 10*(6), 342-353. doi:10.1093/bfgp/elr027

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D., Singer, M., & Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics, 12*, 451. doi:10.1186/1471-2105-12-451

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*, 621-628.

Moskowitz, I. P., Kim, J. B., Moore, M. L., Wolf, C. M., Peterson, M. A., Shendure, J., . . . Seidman, C. E. (2007). A Molecular Pathway Including Id2, Tbx5, and Nkx2-5 Required for Cardiac Conduction System Development. *Cell, 129*(7), 1365–1376. doi:https://doi.org/10.1016/j.cell.2007.04.036

Nazeer, K. A., Sebastian, M., & Kumar, S. M. (2013). A novel harmony search-K means hybrid algorithm for clustering gene expression data. *Bioinformation, 9*(2), 84-88. doi:10.6026/97320630009084

Nie, J., Jiang, M., Zhang, X., Tang, H., Jin, H., Huang, X., . . . Yang, Z. (2015). Post-transcriptional Regulation of Nkx2-5 by RHAU in Heart Development. *Cell Reports, 13*(4), 723-732.

Parker, H. S., Bravo, H. C., & Leek, J. T. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ, 2*(e561). doi:10.7717/peerj.561

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics, 21*(13), 3017-3024. doi:https://doi-org.ezproxy.lib.ndsu.nodak.edu/10.1093/bioinformatics/bti448

Pemental, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods, 14*(7), 687–690. doi:10.1038/nmeth.4324

Petal, C., Silcock, L., McMullan, D., Brueton, L., & Cox, H. (2012). TBX5 intragenic duplication: a family with an atypical Holt–Oram syndrome phenotype. *European Journal of Human Genetics, 20*, 863-869. doi:doi:10.1038/ejhg.2012.16

Pounds, S., & Cheng, C. (2005). Statistical development and evaluation of microarray gene expression data filters. *Journal of Computational Biology, 12*(4), 482-495.

PubMed. (2018, October). (U.S. National Library of Medicine) Retrieved from PubMed: https://www.ncbi.nlm.nih.gov/pubmed/

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rau, A., Marot, G., & Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics, 15*, 91. doi:10.1186/1471-2105-15-91

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., Andrade, M. d., . . . Eckel-Passow, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics, 29*(22), 2877-2883. doi:10.1093/bioinformatics/btt480

Relógio, A., Schwager, C., Richter, A., Ansorge, W., & Valcárcel, J. (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research, 30*(11), e51.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics., 16*(2), 85-97. doi:10.1038/nrg3868

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), e47. doi:10.1093/nar/gkv007

Rocke, D. M., & Durbin, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics, 19*, 966-972.

Rodriguez-Brito, B., Rohwer, F., & Edwards, R. A. (2006). An application of statistics to comparative metagenomics. *BMC Bioinformatics, 7*, 162. doi:10.1186/1471-2105-7-162

Sanger, F., & Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology, 94*(3), 441-448. doi:10.1016/0022-2836(75)90213-2

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., . . . Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature, 265*(5596), 687–695. doi:10.1038/265687a0

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74*(12), 5463-5467. doi:10.1073/pnas.74.12.5463

Schaefer, J., Opgen-Rhein, R., & Strimmer, K. (2015, 08). GeneNet: Modeling and Inferring Gene Networks. Retrieved from http://strimmerlab.org/software/genenet/

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467-470.

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control consortium. *Nature Biotechnology, 32*(9), 903-914. doi:10.1038/nbt.2957

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., & Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics, 24*(9), 1154-1160.

Shen, R., Ghosh, D., & Chinnaiyan, A. M. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics, 5*, 94.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., . . . Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Medical Genomics, 1*, 42.

Sîrbu, A., Kerr, G., Crane, M., & Ruskin, H. J. (2012). RNA-Seq vs Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering. *PLoS ONE, 7*(12), e50986. doi:10.1371/journal.pone.0050986

Sirinukunwattana, K., Savage, R., Bari, M., Snead, D., & Rajpoot, N. (2013). Bayesian Hierarchical Clustering for Studying Cancer Gene Expression Data with Unknown Statistics. *PLoS One, 8*(10), e75748. doi:10.1371/journal.pone.0075748

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*(1), Article 3. Retrieved from http://www.statsci.org/smyth/pubs/ebayes.pdf

Smyth, G. K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods, 31*(4), 265-273. doi:doi.org/10.1016/S1046-2023(03)00155-5

Steger, D., Berry, D., Haider, S., Horn, M., Wagner, M., Stocker, R., & Loy, A. (2011). Systematic Spatial Bias in DNA Microarray Hybridization Is Caused by Probe Spot Position-Dependent Variability in Lateral Diffusion. *PLoS ONE, 6*(8), e23727. doi:doi.org/10.1371/journal.pone.0023727

Thompson, J. A., Tan, J., & Greene, C. S. (2016). Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ, 4*, e1621. doi:10.7717/peerj.1621

Tödling, J., & Spang, R. (2003). Assessment of Five Microarray Experiments on gene expression profiling of breast cancer. *Proceedings of the seventh annual international conference on Research in computational molecular biology.* Berlin, Germany: ACM.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics, 25*(9), 1105-1111. doi:https://doi.org/10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511-515. doi:10.1038/nbt.1621

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511-515. doi:10.1038/nbt.1621

Tringe, S., Von Mering, C., Kobayashi, A., Salamov, A., Chen, K., Chang, H., . . . Rubin, E. (2005). Comparative Metagenomics of Microbial Communities. *Science, 308*(5721), 554-557.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 5116-5121.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer.

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. *Advances in Neural Information Processing Systems. 9*, pp. 281-287. Cambridge: MIT Press.

Wang, C., Gong, B., Bushel, P., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., . . . Megherbi, D. (2014). A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data. *Nature Biotechnology, 32*(9), 926-932. doi:10.1038/nbt.3001

Wang, L., Srivastava, A., & Schwartz, C. (2010). Microarray data integration for genome-wide analysis of human tissue-selective gene expression. *BMC Genomics, 11*(Suppl 2), S15. doi:10.1186/1471-2164-11-S2-S15

Wang, M. D. (2013). In the Spotlight: Bioinformatics. *IEEE Reviews in Biomedical Engineering, 6*, 3-8. doi:10.1109/RBME.2012.2228311

Wang, Z., Gerstein, M., & Snyder, M. (2009, 1). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, pp. 57-63.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57-63. doi:10.1038/nrg2484

Warde-Farley, D., Donaldson, S., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Wright, G. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research, 38*, W214-W220. doi:10.1093/nar/gkq537

Warnat, P., Eils, R., & Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics, 6*, 265. doi:10.1186/1471-2105-6-265

Wei, N. (2015, April 1). *Next Generation Sequencing Analysis of Wild Type and SRSF10-/- embryonic day 13.5 heart Transcriptomes*. Retrieved from Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66965

Wim, T., Marike, B. H., Tjasso, B., H, G. J., J, T. M., C, B. M., . . . Anke, v. D. (2007). Microarray amplification bias: loss of 30% differentially expressed genes due to long probe – poly(A)-tail distances. *BMC Genomics, 8*(1), 277. doi:doi.org/10.1186/1471-2164-8-277

Xie, L., Hoffmann, A. d., Burnicka-Turek, O., Friedland-Little, J. m., Zhang, K., & Moskowitz, I. p. (2012). Tbx5-Hedgehog Molecular Networks Are Essential in the Second Heart Field for Atrial Septation. *Developmental Cell, 23*(2), 280-291. doi:10.1016/j.devcel.2012.06.006

Xing, F., Ning, F., Song, G., Zheng, Y., Ying, X., Hao, H., . . . Philipp, K. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics, 10*(1), 161. doi:10.1186/1471-2164-10-161

Zehetmayer, S., Graf, A. C., & Posch, M. (2015). Sample size reassessment for a two-stage design controlling the false discovery rate. *Statistical Applications in Genetics and Molecular Biology, 14*(5), 429-442. doi:0.1515/sagmb-2014-0025

Zhang, B., Madden, P., Gu, J., Xing, X., Sankar, S., Flynn, J., . . . Wang, T. (2017). Uncovering the transcriptomic and epigenomic landscape of nicotinic receptor genes in non-neuronal tissues. *BMC genomics, 18*(1), 439. doi:10.1186/s12864-017-3813-4

Zhang, K. K., Xiang, M., Zhou, L., Liu, J., Curry, N., Heine Suñer, D., . . . Xie, L. (2016). Gene-network and familial analyses uncover a gene network involving Tbx5/Osr1/Pcsk6 interaction in the second heart field for atrial septation. *Human molecular genetics, 25*(6), 1140-1151. doi:10.1093/hmg/ddv636

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE, 9*(1), e78644. doi:10.1371/journal.pone.0078644

Zhou, L., Liu, J., Olson, P., Zhang, K., Wynne, J., & Xie, L. (2015). Tbx5 and Osr1 interact to regulate posterior second heart field cell cycle progression for cardiac septation. *Journal of Molecular and Cellular Cardiology, 85*, 1-12. doi:10.1016/j.yjmcc.2015.05.005

Zhu, J., He, F., Hu, S., & Yu, J. (2008). On the nature of human housekeeping genes. *Trends in Genetics, 24*(10), 481-484.

Zhu, X., Wolfgruber, T. K., Tasato, A., Arisdakessian, C., Garmire, D. G., & Garmire, L. X. (2017). Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Medicine, 9*(1), 108. doi:10.1186/s13073-017-0492-3

Zien, A., Fluck, J., Zimmer, R., & Lengauer, T. (2004). Microarrays: how many do you need? *Journal of Computational Biology, 10*(3-4), 653-667. doi:https://doi.org/10.1089/10665270360688246

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Statistical Methodology Series B, 67*, 301-320. doi:10.1111/j.1467-9868.2005.00503.x

APPENDIX A. SUPPORTING TABLES FOR SIMULATION DATA SETS

Table A.1. The MR of BSP with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.642	0.000	0.000	0.000	0.000	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000
0.5	0.005	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002	0.003	0.002	0.002	0.002	0.002	0.346	0.005	0.003	0.002	0.003	0.003
0.6	0.015	0.015	0.016	0.015	0.014	0.015	0.015	0.015	0.015	0.016	0.014	0.015	0.015	0.015	0.016	0.015	0.015	0.015	0.016	0.015
0.7	0.029	0.030	0.028	0.028	0.031	0.027	0.031	0.028	0.028	0.029	0.031	0.031	0.027	0.032	0.031	0.028	0.031	0.028	0.030	0.031
0.8	0.048	0.048	0.048	0.049	0.048	0.048	0.048	0.053	0.048	0.048	0.049	0.047	0.049	0.048	0.056	0.048	0.049	0.048	0.047	0.049
0.9	0.087	0.092	0.090	0.089	0.097	0.090	0.090	0.087	0.098	0.089	0.091	0.087	0.087	0.092	0.087	0.091	0.095	0.091	0.091	0.093
1.0	0.118	0.114	0.108	0.107	0.108	0.109	0.112	0.110	0.108	0.114	0.121	0.132	0.109	0.106	0.109	0.112	0.109	0.124	0.111	0.115
1.1	0.145	0.145	0.155	0.147	0.147	0.158	0.162	0.166	0.156	0.155	0.148	0.155	0.147	0.145	0.154	0.149	0.159	0.146	0.146	0.145
1.2	0.181	0.181	0.179	0.178	0.202	0.180	0.178	0.184	0.178	0.180	0.180	0.187	0.177	0.175	0.178	0.180	0.179	0.181	0.190	0.177
1.3	0.197	0.211	0.190	0.217	0.192	0.208	0.198	0.204	0.189	0.192	0.193	0.192	0.192	0.189	0.192	0.199	0.233	0.190	0.190	0.192
1.4	0.220	0.239	0.240	0.245	0.221	0.249	0.269	0.267	0.240	0.220	0.245	0.252	0.257	0.221	0.239	0.225	0.221	0.249	0.224	0.245
1.5	0.304	0.283	0.293	0.276	0.283	0.247	0.315	0.310	0.243	0.331	0.305	0.288	0.313	0.312	0.310	0.283	0.315	0.317	0.352	0.276

Table A.2. The MR of K-means with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.235	0.250	0.000	0.250	0.235	0.000	0.235	0.235	0.000	0.235	0.000	0.250	0.250	0.235	0.250	0.235	0.235	0.235	0.250	0.000
0.2	0.000	0.235	0.235	0.000	0.235	0.250	0.250	0.000	0.000	0.000	0.250	0.000	0.250	0.235	0.000	0.235	0.250	0.250	0.250	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.235	0.000	0.000	0.250	0.235	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.000	0.250	0.000
0.4	0.236	0.250	0.236	0.000	0.236	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.5	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
0.6	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014
0.7	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028
0.8	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
0.9	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094	0.094
1.0	0.107	0.106	0.106	0.107	0.107	0.107	0.106	0.107	0.107	0.107	0.107	0.106	0.106	0.106	0.106	0.106	0.107	0.106	0.106	0.106
1.1	0.150	0.150	0.147	0.150	0.150	0.147	0.150	0.150	0.147	0.150	0.147	0.150	0.147	0.150	0.147	0.150	0.150	0.147	0.150	0.150
1.2	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176	0.176
1.3	0.200	0.198	0.200	0.198	0.198	0.200	0.200	0.200	0.198	0.200	0.198	0.198	0.198	0.200	0.198	0.198	0.198	0.198	0.198	0.200
1.4	0.231	0.232	0.232	0.232	0.231	0.231	0.232	0.232	0.231	0.232	0.231	0.232	0.231	0.231	0.232	0.232	0.231	0.231	0.231	0.232
1.5	0.274	0.263	0.263	0.263	0.263	0.274	0.263	0.263	0.274	0.274	0.263	0.274	0.274	0.264	0.263	0.263	0.263	0.263	0.263	0.263

Table A.3. The Number of DMGs of BSP with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
0.2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	970	10	10	10	10	10
0.3	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
0.4	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
0.5	10	8	8	6	6	8	9	8	6	8	8	6	7	7	501	12	7	8	7	7
0.6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.8	4	4	5	5	3	5	4	5	4	5	5	5	5	4	5	5	5	3	3	3
0.9	2	2	2	2	3	2	1	2	1	1	1	1	2	1	2	1	1	1	2	4
1.0	3	3	3	3	3	3	3	3	3	3	3	4	3	3	3	3	3	3	3	3
1.1	3	3	3	5	3	3	2	7	5	5	3	5	3	3	3	3	3	3	3	3
1.2	5	2	5	5	3	6	5	2	5	2	5	2	5	5	5	6	5	5	5	5
1.3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1.4	1	1	1	1	1	1	3	3	1	1	1	1	1	1	1	1	1	3	1	1
1.5	2	2	2	2	2	1	2	2	1	2	2	2	2	2	2	2	2	2	2	2

Table A.4. The Number of DMGs of K-means with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	972	1033	14	1033	972	14	973	973	14	972	14	1036	1033	972	1033	974	974	973	1033	14
0.2	10	974	974	10	970	1032	1032	10	10	10	1032	10	1034	970	10	974	1032	1032	1032	10
0.3	7	7	7	7	7	951	7	7	1016	951	7	7	7	951	7	7	7	7	1016	7
0.4	754	822	754	14	754	14	14	14	14	14	14	822	14	14	14	14	14	14	14	14
0.5	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
0.6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0.9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1.1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1.2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.5	2	1	1	1	1	2	1	1	2	2	1	2	2	2	1	1	1	1	1	1

Table A.5. The Accurate Number of Clusters of BSP with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
0.6	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1	1	1	0	1	0
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	0	1	1	1	1
0.9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1
1.1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0
1.2	0	1	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0
1.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.4	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1
1.5	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	1	1	1	0	1

Table A.6. The Accurate Number of Clusters of K-means with Different σ Values for Normal Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
0.2	1	0	0	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	1
0.3	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0.4	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
0.5	1	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1
0.6	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	0	1	1	1	1	0	0	0	1	1	0	0	0	1	1	1	0	1	0	1
1.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.7. The MR of BSP with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.003	0.005	0.004	0.003	0.004	0.003	0.003	0.003	0.004	0.004	0.004	0.004	0.003	0.004	0.003	0.004	0.003	0.004	0.005	0.003
0.5	0.011	0.012	0.011	0.013	0.012	0.013	0.011	0.012	0.014	0.011	0.014	0.010	0.011	0.012	0.010	0.012	0.012	0.011	0.011	0.010
0.6	0.021	0.023	0.022	0.022	0.023	0.023	0.021	0.021	0.021	0.021	0.023	0.021	0.023	0.022	0.022	0.022	0.021	0.021	0.022	0.022
0.7	0.044	0.042	0.043	0.043	0.044	0.043	0.046	0.045	0.043	0.041	0.043	0.043	0.043	0.043	0.043	0.044	0.043	0.043	0.043	0.043
0.8	0.052	0.052	0.052	0.055	0.054	0.056	0.051	0.051	0.052	0.052	0.053	0.054	0.053	0.052	0.053	0.052	0.052	0.052	0.052	0.055
0.9	0.076	0.079	0.083	0.079	0.084	0.079	0.077	0.076	0.078	0.076	0.076	0.077	0.082	0.081	0.076	0.076	0.078	0.078	0.078	0.077
1.0	0.112	0.122	0.114	0.117	0.116	0.113	0.115	0.129	0.114	0.124	0.117	0.118	0.115	0.117	0.119	0.111	0.114	0.129	0.118	0.115
1.1	0.121	0.117	0.117	0.115	0.128	0.117	0.118	0.116	0.118	0.122	0.118	0.117	0.119	0.117	0.117	0.117	0.117	0.118	0.118	0.117
1.2	0.149	0.153	0.146	0.151	0.149	0.148	0.150	0.153	0.150	0.152	0.150	0.152	0.152	0.152	0.155	0.149	0.146	0.152	0.151	0.150
1.3	0.171	0.164	0.193	0.186	0.177	0.166	0.173	0.170	0.176	0.173	0.177	0.164	0.185	0.170	0.171	0.172	0.177	0.167	0.174	0.181
1.4	0.199	0.191	0.190	0.197	0.198	0.197	0.200	0.197	0.211	0.200	0.199	0.201	0.199	0.196	0.198	0.198	0.191	0.201	0.199	0.205
1.5	0.224	0.241	0.220	0.222	0.215	0.219	0.267	0.219	0.262	0.221	0.256	0.214	0.213	0.225	0.238	0.239	0.220	0.219	0.219	0.265

Table A.8. The MR of K-means with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.235	0.250	0.000	0.250	0.235	0.000	0.235	0.235	0.000	0.235	0.000	0.250	0.250	0.235	0.250	0.235	0.235	0.235	0.250	0.000
0.2	0.000	0.235	0.235	0.000	0.235	0.250	0.250	0.000	0.000	0.000	0.250	0.000	0.250	0.235	0.000	0.235	0.250	0.250	0.235	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.235	0.000	0.000	0.250	0.235	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.000	0.250	0.000
0.4	0.238	0.250	0.238	0.003	0.238	0.003	0.003	0.003	0.003	0.003	0.003	0.250	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
0.5	0.010	0.010	0.010	0.242	0.010	0.010	0.010	0.253	0.010	0.010	0.010	0.010	0.242	0.010	0.010	0.010	0.242	0.010	0.010	0.010
0.6	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
0.7	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
0.8	0.050	0.051	0.050	0.051	0.051	0.051	0.051	0.051	0.050	0.050	0.050	0.050	0.051	0.051	0.051	0.051	0.051	0.050	0.051	0.051
0.9	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077
1.0	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112
1.1	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120	0.120
1.2	0.144	0.143	0.143	0.143	0.143	0.144	0.143	0.143	0.144	0.144	0.143	0.144	0.144	0.143	0.144	0.144	0.143	0.144	0.144	0.143
1.3	0.163	0.166	0.163	0.166	0.166	0.163	0.163	0.163	0.166	0.163	0.166	0.166	0.166	0.163	0.166	0.166	0.166	0.166	0.166	0.163
1.4	0.191	0.190	0.190	0.190	0.192	0.192	0.190	0.190	0.192	0.190	0.192	0.190	0.192	0.190	0.190	0.190	0.192	0.192	0.192	0.190
1.5	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221	0.221

Table A.9. The Number of DMGs of BSP with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
0.2	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23
0.3	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
0.4	27	28	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	28	27
0.5	30	32	31	34	32	35	30	33	36	30	34	31	30	32	30	31	32	30	34	30
0.6	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.7	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.9	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1.0	3	4	3	3	3	3	3	4	3	4	3	3	3	3	4	3	3	4	3	3
1.1	9	8	8	8	10	8	8	8	8	9	8	8	8	8	8	8	8	8	8	8
1.2	9	9	9	10	9	9	9	10	9	10	10	9	10	10	10	10	8	9	10	10
1.3	4	1	6	4	4	3	3	4	4	4	4	1	4	4	4	4	4	3	4	4
1.4	7	7	7	7	7	7	7	7	7	6	7	7	7	7	7	6	6	6	7	7
1.5	7	7	7	7	7	7	7	7	7	7	7	7	7	8	7	7	7	7	7	7

Table A.10. The Number of DMGs of K-means with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	975	1036	17	1036	973	17	973	973	17	973	17	1037	1036	973	1036	973	973	973	1036	17
0.2	23	979	979	23	969	1040	1040	23	23	23	1040	23	1039	969	23	976	1040	1040	979	23
0.3	18	18	18	18	18	952	18	18	1024	952	18	18	18	952	18	18	18	18	1024	18
0.4	825	848	825	27	825	27	27	27	27	27	27	848	27	27	27	27	27	27	27	27
0.5	29	29	29	518	29	29	29	570	29	29	29	29	518	29	29	29	518	29	29	29
0.6	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.7	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.9	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1.1	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
1.2	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
1.3	2	1	2	1	1	2	2	2	1	2	1	1	1	2	1	1	1	1	1	2
1.4	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
1.5	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7

Table A.11. The Accurate Number of Clusters of BSP with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	1	0	1
1.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.3	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.12. The Accurate Number of Clusters of K-means with Different σ Values for Exponential Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
0.2	1	0	0	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	1
0.3	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0.4	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
0.5	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0.6	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1
0.7	1	1	1	0	1	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0
0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.3	0	1	0	1	1	0	0	0	1	0	1	1	1	0	1	1	1	1	1	0
1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.13. The MR of BSP with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.003	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.003	0.003	0.004	0.003	0.003
0.5	0.009	0.009	0.009	0.008	0.008	0.007	0.007	0.008	0.009	0.008	0.007	0.008	0.008	0.007	0.008	0.008	0.008	0.009	0.008	0.007
0.6	0.027	0.027	0.026	0.028	0.027	0.027	0.027	0.030	0.027	0.030	0.027	0.027	0.027	0.026	0.026	0.029	0.027	0.027	0.027	0.027
0.7	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.042	0.037	0.038	0.041	0.038	0.038	0.039	0.038	0.038	0.038	0.038	0.038	0.041
0.8	0.056	0.053	0.054	0.053	0.053	0.057	0.053	0.053	0.053	0.056	0.054	0.053	0.054	0.048	0.056	0.053	0.053	0.053	0.052	0.048
0.9	0.083	0.079	0.079	0.078	0.080	0.078	0.078	0.079	0.079	0.078	0.078	0.078	0.079	0.080	0.081	0.078	0.080	0.081	0.079	0.079
1.0	0.097	0.099	0.096	0.098	0.103	0.105	0.100	0.097	0.113	0.101	0.099	0.096	0.103	0.105	0.095	0.104	0.100	0.098	0.097	0.103
1.1	0.117	0.119	0.116	0.117	0.115	0.119	0.117	0.117	0.117	0.121	0.116	0.117	0.121	0.119	0.119	0.124	0.118	0.119	0.121	0.116
1.2	0.148	0.148	0.143	0.146	0.143	0.148	0.145	0.146	0.145	0.145	0.146	0.144	0.147	0.146	0.147	0.143	0.145	0.160	0.143	0.162
1.3	0.178	0.174	0.177	0.176	0.178	0.177	0.177	0.175	0.176	0.183	0.176	0.183	0.177	0.176	0.177	0.179	0.177	0.175	0.172	0.178
1.4	0.214	0.202	0.201	0.200	0.209	0.197	0.198	0.203	0.196	0.202	0.210	0.215	0.197	0.212	0.206	0.210	0.198	0.201	0.211	0.234
1.5	0.223	0.222	0.224	0.223	0.222	0.228	0.219	0.222	0.225	0.220	0.225	0.229	0.240	0.225	0.224	0.228	0.220	0.220	0.224	0.232

Table A.14. The MR of K-means with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.235	0.250	0.000	0.250	0.235	0.000	0.235	0.235	0.000	0.235	0.000	0.250	0.250	0.235	0.250	0.235	0.235	0.235	0.250	0.000
0.2	0.000	0.237	0.237	0.000	0.235	0.250	0.250	0.000	0.000	0.000	0.250	0.000	0.250	0.235	0.000	0.235	0.250	0.250	0.250	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.235	0.000	0.000	0.250	0.235	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.000	0.250	0.000
0.4	0.237	0.251	0.237	0.003	0.237	0.003	0.003	0.003	0.003	0.003	0.003	0.251	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
0.5	0.008	0.008	0.008	0.241	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.241	0.008	0.008	0.008	0.241	0.008	0.008	0.008
0.6	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027
0.7	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
0.8	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
0.9	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081
1.0	0.097	0.098	0.098	0.097	0.097	0.097	0.098	0.097	0.097	0.098	0.097	0.098	0.098	0.098	0.098	0.098	0.097	0.098	0.098	0.098
1.1	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
1.2	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148
1.3	0.172	0.167	0.172	0.167	0.167	0.172	0.172	0.172	0.167	0.172	0.167	0.167	0.167	0.172	0.167	0.167	0.167	0.167	0.167	0.172
1.4	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195	0.195
1.5	0.213	0.215	0.215	0.215	0.215	0.213	0.215	0.215	0.213	0.213	0.215	0.213	0.213	0.213	0.215	0.215	0.215	0.215	0.215	0.215

Table A.15. The Number of DMGs of BSP with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
0.2	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23
0.3	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28
0.4	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
0.5	14	13	13	12	12	12	12	12	14	13	12	12	12	12	12	13	12	13	12	12
0.6	6	6	6	7	6	6	6	6	6	7	6	6	6	6	6	6	6	6	6	6
0.7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
0.8	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	3	3
0.9	7	7	7	7	6	7	6	6	7	6	6	7	7	7	5	6	7	7	6	6
1.0	4	4	4	4	6	6	6	4	6	6	4	4	6	6	4	4	6	4	4	4
1.1	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
1.2	8	8	6	8	7	8	7	8	8	8	8	5	8	6	8	8	8	8	5	8
1.3	8	7	8	8	8	8	7	7	8	8	7	8	7	7	8	8	7	7	7	8
1.4	3	2	2	2	2	2	2	2	2	2	2	3	2	3	2	2	2	2	2	3
1.5	8	8	8	7	9	8	8	8	8	8	8	8	8	8	9	9	8	8	8	8

Table A.16. The Number of DMGs of K-means with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	976	1040	25	1040	976	25	980	980	25	976	25	1036	1040	976	1040	975	975	980	1040	25
0.2	23	976	976	23	975	1036	1036	23	23	23	1036	23	1038	975	23	975	1036	1036	1036	23
0.3	28	28	28	28	28	958	28	28	1030	958	28	28	28	958	28	28	28	28	1030	28
0.4	793	820	793	11	793	11	11	11	11	11	11	820	11	11	11	11	11	11	11	11
0.5	12	12	12	520	12	12	12	12	12	12	12	12	519	12	12	12	520	12	12	12
0.6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
0.7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
0.8	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0.9	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1.1	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
1.2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1.3	7	6	7	6	6	7	7	7	6	7	6	6	6	7	6	6	6	6	6	7
1.4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1.5	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7

Table A.17. The Accurate Number of Clusters of BSP with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.9	0	0	0	0	1	0	1	1	0	1	0	0	0	1	1	1	0	0	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0
1.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.4	0	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.18. The Accurate Number of Clusters of K-means with Different σ Values for Gamma Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
0.2	1	0	0	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	1
0.3	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0.4	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
0.5	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1
0.6	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	0	0	0	1	1	0
1.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
1.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.19. The MR of BSP with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000
0.5	0.006	0.006	0.005	0.006	0.005	0.007	0.005	0.008	0.008	0.007	0.006	0.006	0.005	0.008	0.005	0.005	0.006	0.005	0.005	0.007
0.6	0.028	0.031	0.028	0.027	0.029	0.027	0.026	0.028	0.026	0.026	0.029	0.029	0.028	0.030	0.030	0.029	0.026	0.024	0.030	0.028
0.7	0.047	0.050	0.046	0.047	0.047	0.046	0.045	0.045	0.048	0.051	0.046	0.046	0.050	0.052	0.051	0.045	0.052	0.045	0.046	0.048
0.8	0.087	0.082	0.081	0.087	0.085	0.079	0.081	0.086	0.082	0.083	0.082	0.081	0.082	0.083	0.082	0.081	0.082	0.083	0.082	0.080
0.9	0.105	0.104	0.118	0.105	0.118	0.103	0.103	0.108	0.108	0.104	0.105	0.104	0.108	0.104	0.108	0.107	0.108	0.108	0.103	0.104
1.0	0.169	0.174	0.171	0.223	0.193	0.174	0.172	0.181	0.173	0.196	0.225	0.182	0.173	0.181	0.165	0.228	0.171	0.180	0.171	0.169
1.1	0.176	0.178	0.178	0.188	0.185	0.193	0.178	0.177	0.188	0.179	0.190	0.177	0.189	0.184	0.189	0.179	0.190	0.189	0.192	0.178
1.2	0.215	0.220	0.219	0.215	0.216	0.208	0.217	0.214	0.222	0.215	0.217	0.229	0.218	0.207	0.215	0.214	0.207	0.214	0.215	0.215
1.3	0.253	0.248	0.250	0.273	0.248	0.249	0.261	0.270	0.247	0.249	0.247	0.247	0.262	0.274	0.260	0.257	0.248	0.242	0.253	0.259
1.4	0.266	0.278	0.273	0.297	0.270	0.272	0.286	0.292	0.305	0.280	0.278	0.302	0.289	0.292	0.287	0.268	0.281	0.275	0.269	0.291
1.5	0.327	0.329	0.322	0.313	0.332	0.318	0.310	0.316	0.327	0.319	0.334	0.332	0.306	0.319	0.315	0.304	0.318	0.317	0.300	0.355

Table A.20. The Accuracy of K-means with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.235	0.250	0.000	0.250	0.235	0.000	0.235	0.235	0.000	0.235	0.000	0.250	0.250	0.235	0.250	0.235	0.235	0.235	0.250	0.000
0.2	0.000	0.235	0.235	0.000	0.235	0.250	0.250	0.000	0.000	0.000	0.250	0.000	0.000	0.235	0.000	0.235	0.250	0.250	0.235	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.235	0.000	0.000	0.250	0.235	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.000	0.250	0.000
0.4	0.236	0.253	0.236	0.000	0.236	0.000	0.000	0.000	0.000	0.000	0.000	0.253	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.5	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
0.6	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026
0.7	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053
0.8	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
0.9	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104
1.0	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162	0.162
1.1	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178
1.2	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212
1.3	0.247	0.254	0.247	0.254	0.254	0.247	0.247	0.247	0.254	0.247	0.254	0.254	0.254	0.247	0.254	0.254	0.254	0.254	0.254	0.247
1.4	0.272	0.274	0.274	0.274	0.272	0.272	0.274	0.274	0.272	0.274	0.272	0.274	0.272	0.274	0.274	0.274	0.272	0.272	0.272	0.274
1.5	0.306	0.304	0.304	0.304	0.304	0.306	0.304	0.304	0.306	0.306	0.304	0.306	0.306	0.306	0.304	0.304	0.304	0.304	0.304	0.304

Table A.21. The Number of DMGs of BSP with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
0.2	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
0.3	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
0.4	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
0.5	3	2	2	4	2	4	2	6	7	4	4	2	2	5	2	2	2	2	2	3
0.6	4	5	3	3	4	4	3	3	3	3	4	3	3	4	4	4	3	3	4	3
0.7	2	1	1	1	1	1	2	1	2	1	1	1	2	1	1	2	1	2	2	2
0.8	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0.9	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1.0	4	3	3	3	3	3	3	3	3	3	3	3	3	8	4	3	3	3	4	2
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	1	2	0	1	0	0	0	2	2	0	0	0	0	1	2	1	0	0	0	0
1.3	2	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1
1.4	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.22. The Number of DMGs of K-means with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	974	1034	12	1034	974	12	972	972	12	974	12	1032	1034	974	1034	970	970	972	1034	12
0.2	9	971	971	9	971	1033	1033	9	9	9	1033	9	9	971	9	971	1033	1033	971	9
0.3	11	11	11	11	11	902	11	11	965	902	11	11	11	902	11	11	11	11	965	11
0.4	659	654	659	8	659	8	8	8	8	8	8	654	8	8	8	8	8	8	8	8
0.5	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
0.6	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0.9	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.23. The Accurate Number of Clusters of BSP with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	1	0
0.6	0	0	1	1	0	0	1	1	1	1	0	1	1	0	0	0	1	1	0	1
0.7	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	0	1
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	1	0	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1
1.3	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
1.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table A.24. The Accurate Number of Clusters of K-means with Different σ Values for Gaussian Mixture Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
0.2	1	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	0	0	0	1
0.3	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0.4	0	0	0	1	0	1	0	0	1	1	1	0	1	1	0	0	1	0	1	0
0.5	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1
0.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9	1	0	1	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.3	0	0	1	1	0	1	1	0	0	0	1	1	0	1	0	1	1	0	1	0
1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table A.25. The MR of BSP with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table A.26. The Accuracy of K-means with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0.235	0.250	0.000	0.250	0.235	0.000	0.235	0.235	0.000	0.235	0.000	0.250	0.250	0.235	0.250	0.235	0.235	0.235	0.250	0.000
0.2	0.000	0.236	0.236	0.000	0.235	0.250	0.250	0.000	0.000	0.000	0.250	0.000	0.250	0.235	0.000	0.235	0.250	0.250	0.236	0.000
0.3	0.000	0.000	0.000	0.000	0.000	0.235	0.000	0.000	0.250	0.235	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.000	0.250	0.000
0.4	0.235	0.250	0.235	0.000	0.235	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.5	0.000	0.000	0.000	0.235	0.000	0.000	0.000	0.238	0.000	0.000	0.000	0.000	0.238	0.253	0.000	0.000	0.235	0.000	0.000	0.000

Table A.27. The Number of DMGs of BSP with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24
0.2	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
0.3	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
0.4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.28. The Number of DMGs of K-means with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	973	1039	24	1039	974	24	975	975	24	974	24	1042	1039	974	1039	978	978	975	1039	24
0.2	19	973	973	19	973	1036	1036	19	19	19	1036	19	1038	973	19	975	1036	1036	973	19
0.3	6	6	6	6	6	947	6	6	1018	947	6	6	6	947	6	6	6	6	1018	6
0.4	766	795	766	2	766	2	2	2	2	2	2	795	2	2	2	2	2	2	2	2
0.5	0	0	0	471	0	0	0	559	0	0	0	0	471	568	0	0	471	0	0	0

Table A.29. The Accurate Number of Clusters of BSP with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table A.30. The Accurate Number of Clusters of K-means with Different σ Values for Beta Distribution

σ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1
0.2	1	0	0	1	0	0	0	1	1	1	0	1	0	0	1	0	0	0	0	1
0.3	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0.4	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
0.5	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1

APPENDIX B. THE LIST OF THE DIFFERENTIALLY MEASURED GENES

GENE_SYMBOL	GENE_NAME	Alignment
Agpat4	1-acylglycerol-3-phosphate O-acyltransferase 1 (lysophosphatidic acid acyltransferase, delta)	1
Hacl1	2-hydroxyacyl-CoA lyase 1	1
Adam22	a disintegrin and metallopeptidase domain 22	1
Akap12	A kinase (PRKA) anchor protein (gravin) 12	2
Abhd4	abhydrolase domain containing 4	1
Anp32a	acidic (leucine-rich) nuclear phosphoprotein 32 family, member A	1
Arpc1b	actin related protein 2/3 complex, subunit 1B	1
Actb	actin, beta, cytoplasmic	1
Actg2	actin, gamma 2, smooth muscle, enteric	NA
Actg-ps1	actin, gamma, pseudogene 1	1
Ascc1	activating signal cointegrator 1 complex subunit 1	1
Atf5	activating transcription factor 5	1
Acss1	acyl-CoA synthetase short-chain family member 1	1
Acadvl	acyl-Coenzyme A dehydrogenase, very long chain	1
Ap3b1	adaptor-related protein complex 3, beta 1 subunit	2
Adssl1	adenylosuccinate synthetase like 1	1
Aldh7a1	aldehyde dehydrogenase family 7, member A1	1
Ambp	alpha 1 microglobulin/bikunin	NA
Atrx	alpha thalassemia/mental retardation syndrome X-linked homolog (human)	1
Angptl4	angiopoietin-like 4	1
Ankrd13b	ankyrin repeat domain 13b	1
Apobec2	apolipoprotein B editing complex 2	1
Rnpepl1	arginyl aminopeptidase (aminopeptidase B)-like 1	1
Arntl2	aryl hydrocarbon receptor nuclear translocator-like 2	NA
Asns	asparagine synthetase	1
Aste1	asteroid homolog 1 (Drosophila)	1
Astn2	astrotactin 2	1
Atxn7l1	ataxin 7-like 1	1

GENE_SYMBOL	GENE_NAME	Alignment
Atxn7l3	ataxin 7-like 3	NA
Atp5g2	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2	3
Atp5h	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit d	2
Atp13a2	ATPase type 13A2	1
Btg1	B-cell translocation gene 1, anti-proliferative	3
Bbc3	Bcl-2 binding component 3	1
Bok	Bcl-2-related ovarian killer protein	NA
Bmp10	bone morphogenetic protein 10	1
Bche	butyrylcholinesterase	1
Celsr1	cadherin EGF LAG seven-pass G-type receptor 1	1
Cdh24	cadherin-like 24	1
Cnn1	calponin 1	1
Cnr1	cannabinoid receptor 1 (brain)	1
Chsy1	carbohydrate (chondroitin) synthase 1	1
Cbl	Casitas B-lineage lymphoma	1
Casp2	caspase 2	1
Cdx1	caudal type homeo box 1	1
Cebpd	CCAAT/enhancer binding protein (C/EBP), delta	1
Cd63	Cd63 antigen	1
Cables1	Cdk5 and Abl enzyme substrate 1	1
BC003965	cDNA sequence BC003965	1
BC017158	cDNA sequence BC017158	1
BC037034	cDNA sequence BC037034	1
BC048507	cDNA sequence BC048507	1
Cdc5l	cell division cycle 5-like (S. pombe)	5
Crabp2	cellular retinoic acid binding protein II	1
Cep68	centrosomal protein 68	1
Cspp1	centrosome and spindle pole associated protein 1	1
Cct6b	chaperonin subunit 6b (zeta)	1
Cmklr1	chemokine-like receptor 1	NA
Coq10a	coenzyme Q10 homolog A (yeast)	1

GENE_SYMBOL	GENE_NAME	Alignment
Coq3	coenzyme Q3 homolog, methyltransferase (yeast)	1
Cfl1	cofilin 1, non-muscle	2
Cc2d1b	coiled-coil and C2 domain containing 1B	1
Ccdc12	coiled-coil domain containing 12	2
Ccdc15	coiled-coil domain containing 15	1
Ccdc22	coiled-coil domain containing 22	1
Ccdc42	coiled-coil domain containing 42	1
Ccdc73	coiled-coil domain containing 73	1
Ccdc85b	coiled-coil domain containing 85B	1
Csf2ra	colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)	2
Ccnk	cyclin K	NA
Cdk4	cyclin-dependent kinase 4	1
Cyyr1	cysteine and tyrosine-rich protein 1	1
Creld2	cysteine-rich with EGF-like domains 2	1
Cox4i2	cytochrome c oxidase subunit IV isoform 2	NA
Cox5b	cytochrome c oxidase, subunit Vb	3
Cox6b1	cytochrome c oxidase, subunit VIb polypeptide 1	2
Cycs	cytochrome c, somatic	2
Ddx23	DEAD (Asp-Glu-Ala-Asp) box polypeptide 23	1
Ddx54	DEAD (Asp-Glu-Ala-Asp) box polypeptide 54	1
Dhx30	DEAH (Asp-Glu-Ala-His) box polypeptide 30	1
Degs2	degenerative spermatocyte homolog 2 (Drosophila), lipid desaturase	1
Dnase1	deoxyribonuclease I	1
Dhdh	dihydrodiol dehydrogenase (dimeric)	1
Dab2ip	disabled homolog 2 (Drosophila) interacting protein	2
Dvl3	dishevelled 3, dsh homolog (Drosophila)	NA
D17H6S53E	DNA segment, Chr 17, human D6S53E	1
Dnajc19	DnaJ (Hsp40) homolog, subfamily C, member 19	2
Dpm3	dolichyl-phosphate mannosyltransferase polypeptide 3	NA
Dusp6	dual specificity phosphatase 6	1
Dctn3	dynactin 3	1

GENE_SYMBOL	GENE_NAME	Alignment
E2f2	E2F transcription factor 2	1
E2f4	E2F transcription factor 4	1
Elf4	E74-like factor 4 (ets domain transcription factor)	1
Ehd1	EH-domain containing 1	1
Ehd4	EH-domain containing 4	1
Elk1	ELK1, member of ETS oncogene family	1
Elk4	ELK4, member of ETS oncogene family	1
Sil1	endoplasmic reticulum chaperone SIL1 homolog (S. cerevisiae)	1
Erp29	endoplasmic reticulum protein 29	1
Erdr1	erythroid differentiation regulator 1	NA
Eef1a1	eukaryotic translation elongation factor 1 alpha 1	9
Eif2b3	eukaryotic translation initiation factor 2B, subunit 3	1
Eif4ebp3	eukaryotic translation initiation factor 4E binding protein 3	1
Exoc8	exocyst complex component 8	1
Exo1	exonuclease 1	1
Nme2	expressed in non-metastatic cells 2, protein	3
AU020206	expressed sequence AU020206	1
F11r	F11 receptor	1
Fance	Fanconi anemia, complementation group E	1
Daxx	Fas death domain-associated protein	1
Fabp3	fatty acid binding protein 3, muscle and heart	2
Fabp5	fatty acid binding protein 5, epidermal	NA
Ferd3l	Fer3-like (Drosophila)	1
Fbl	fibrillarin	2
Fau	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived)	2
Fkbp10	FK506 binding protein 10	1
Foxa2	forkhead box A2	1
Foxg1	forkhead box G1	1
Fmnl3	formin-like 3	2
Frem1	Fras1 related extracellular matrix protein 1	1
Frat1	frequently rearranged in advanced T-cell lymphomas	1

GENE_SYMBOL	GENE_NAME	Alignment
Fyb	FYN binding protein	1
Gps1	G protein pathway suppressor 1	2
Gpr63	G protein-coupled receptor 63	1
Gabpb2	GA repeat binding protein, beta 2	NA
Galt	galactose-1-phosphate uridyl transferase	1
Gjc1	gap junction membrane channel protein chi 1	1
Gata4	GATA binding protein 4	1
Gsn	gelsolin	1
Gtf2f1	general transcription factor IIF, polypeptide 1	1
Gpi1	glucose phosphate isomerase 1	2
Gstm1	glutathione S-transferase, mu 1	2
Gstm3	glutathione S-transferase, mu 3	1
Grrp1	glycine/arginine rich protein 1	1
Gsk3a	glycogen synthase kinase 3 alpha	1
Grhpr	glyoxylate reductase/hydroxypyruvate reductase	1
Gpc1	glypican 1	1
Gnas	GNAS (guanine nucleotide binding protein, alpha stimulating) complex locus	NA
Golga2	golgi autoantigen, golgin subfamily a, 2	1
Gpsm1	G-protein signalling modulator 1 (AGS3-like, C. elegans)	1
Gab2	growth factor receptor bound protein 2-associated protein 2	NA
Gnb1l	guanine nucleotide binding protein (G protein), beta polypeptide 1-like	1
Gnal	guanine nucleotide binding protein, alpha stimulating, olfactory type	NA
H19	H19 fetal liver mRNA	1
H2afj	H2A histone family, member J	3
Hsp90ab1	heat shock protein 90kDa alpha (cytosolic), class B member 1	1
Herc1	hect (homologous to the E6-AP (UBE3A) carboxyl terminus) domain and RCC1 (CHC1)-like domain (RLD) 1	1
Hba-x	hemoglobin X, alpha-like embryonic chain in Hba complex	1
Hbb-y	hemoglobin Y, beta-like embryonic chain	1
Hbb-bh1	hemoglobin Z, beta-like embryonic chain	1

GENE_SYMBOL	GENE_NAME	Alignment
Hs6st1	heparan sulfate 6-O-sulfotransferase 1	1
Hps4	Hermansky-Pudlak syndrome 4 homolog (human)	1
Hk2	hexokinase 2	1
H2-T10	histocompatibility 2, T region locus 10	2
H2-T24	histocompatibility 2, T region locus 24	NA
Hist1h2af	histone 1, H2af	13
Hist1h2ai	histone 1, H2ai	5
Hist1h2ak	histone 1, H2ak	3
Hist1h4f	histone 1, H4f	11
Hist2h2ac	histone 2, H2ac	4
Hist2h3c1	histone 2, H3c1	3
Hist2h4	histone 2, H4	1
Hist3h2a	histone 3, H2a	1
Hdac1	histone deacetylase 1	NA
Hoxd9	homeo box D9	1
Hunk	hormonally upregulated Neu-associated kinase	1
Hcn1	hyperpolarization-activated, cyclic nucleotide-gated K+ 1	1
Ier5	immediate early response 5	1
Imp4	IMP4, U3 small nucleolar ribonucleoprotein, homolog (yeast)	1
Incenp	inner centromere protein	1
Impdh2	inosine 5'-phosphate dehydrogenase 2	2
Itpr3	inositol 1,4,5-triphosphate receptor 3	1
Insr	insulin receptor	1
Itgb5	integrin beta 5	1
Ifitm7	interferon induced transmembrane protein 7	1
Il11ra1	interleukin 11 receptor, alpha chain 1	1
Il12rb2	interleukin 12 receptor, beta 2	1
Ift122	intraflagellar transport 122 homolog (Chlamydomonas)	1
Irx3	Iroquois related homeobox 3 (Drosophila)	1
Jmjd4	jumonji domain containing 4	1
Kpna2	karyopherin (importin) alpha 2	7
Klhl17	kelch-like 17 (Drosophila)	1
Krt8	keratin 8	2

GENE_SYMBOL	GENE_NAME	Alignment
Ldhb	lactate dehydrogenase B	2
Lats1	large tumor suppressor	1
Lime1	Lck interacting transmembrane adaptor 1	1
Lemd2	LEM domain containing 2	1
Lrfn1	leucine rich repeat and fibronectin type III domain containing 1	1
Lrrc45	leucine rich repeat containing 45	1
Lrrc58	leucine rich repeat containing 58	5
Lzts2	leucine zipper, putative tumor suppressor 2	1
Letm1	leucine zipper-EF-hand containing transmembrane protein 1	1
Lrch3	leucine-rich repeats and calponin homology (CH) domain containing 3	1
Lrig3	leucine-rich repeats and immunoglobulin-like domains 3	1
Lars2	leucyl-tRNA synthetase, mitochondrial	3
Lmx1b	LIM homeobox transcription factor 1 beta	NA
Lsr	lipolysis stimulated lipoprotein receptor	NA
Lrp12	low density lipoprotein-related protein 12	1
Lsm7	LSM7 homolog, U6 small nuclear RNA associated (S. cerevisiae)	3
Lfng	lunatic fringe gene homolog (Drosophila)	1
Lyl1	lymphoblastomic leukemia	1
Smad9	MAD homolog 9 (Drosophila)	NA
Mgrn1	mahogunin, ring finger 1	1
Mfhas1	malignant fibrous histiocytoma amplified sequence 1	2
Mlycd	malonyl-CoA decarboxylase	1
Man2a1	mannosidase 2, alpha 1	1
Maml3	mastermind like 3 (Drosophila)	1
Mnt	max binding protein	3
Mia3	melanoma inhibitory activity 3	NA
Mbd5	methyl-CpG binding domain protein 5	1
Mical1	microtubule associated monoxygenase, calponin and LIM domain containing 1	1
Mdn1	midasin homolog (yeast)	1

GENE_SYMBOL	GENE_NAME	Alignment
Mdk	midkine	1
Mfge8	milk fat globule-EGF factor 8 protein	1
Mrpl11	mitochondrial ribosomal protein L11	1
Mrpl23	mitochondrial ribosomal protein L23	2
Map3k1	mitogen activated protein kinase kinase kinase 1	1
Map3k2	mitogen activated protein kinase kinase kinase 2	1
Mapk8ip3	mitogen-activated protein kinase 8 interacting protein 3	1
Mkl2	MKL/myocardin-like 2	1
Mogat2	monoacylglycerol O-acyltransferase 2	1
Morf4l1	mortality factor 4 like 1	10
Mmrn2	multimerin 2	1
Msi1	Musashi homolog 1(Drosophila)	1
Mcc	mutated in colorectal cancers	1
Mybpc3	myosin binding protein C, cardiac	1
Myo18b	myosin XVIIIb	1
Myh4	myosin, heavy polypeptide 4, skeletal muscle	1
Myh7	myosin, heavy polypeptide 7, cardiac muscle, beta	1
Myh8	myosin, heavy polypeptide 8, skeletal muscle, perinatal	1
Myh9	myosin, heavy polypeptide 9, non-muscle	1
Myl6	myosin, light polypeptide 6, alkali, smooth muscle and non-muscle	6
Ndufa13	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 13	1
Ndufb4	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 4	3
Ndufs5	NADH dehydrogenase (ubiquinone) Fe-S protein 5	2
Nkd1	naked cuticle 1 homolog (Drosophila)	1
Naca	nascent polypeptide-associated complex alpha polypeptide	2
Nell1	NEL-like 1 (chicken)	1
Ntng2	netrin G2	NA
Neurod4	neurogenic differentiation 4	1
Nenf	neuron derived neurotrophic factor	1
Nkx2-9	NK2 transcription factor related, locus 9 (Drosophila)	1
Nomo1	nodal modulator 1	1
Nr2f1	nuclear receptor subfamily 2, group F, member 1	1

GENE_SYMBOL	GENE_NAME	Alignment
Ofd1	oral-facial-digital syndrome 1 gene homolog (human)	1
Osbpl5	oxysterol binding protein-like 5	1
Pappa2	pappalysin 2	NA
Pask	PAS domain containing serine/threonine kinase	1
Peg10	paternally expressed 10	1
Pdzrn3	PDZ domain containing RING finger 3	1
Pelo	pelota homolog (Drosophila)	1
Ppard	peroxisome proliferator activator receptor delta	1
Ebp	phenylalkylamine Ca2+ antagonist (emopamil) binding protein	1
Ppap2b	phosphatidic acid phosphatase type 2B	1
Pik3r4	phosphatidylinositol 3 kinase, regulatory subunit, polypeptide 4, p150	1
Pitpnm1	phosphatidylinositol membrane-associated 1	1
Pgm2	phosphoglucomutase 2	1
Plcb4	phospholipase C, beta 4	1
Phyhipl	phytanoyl-CoA hydroxylase interacting protein-like	NA
Phlda3	pleckstrin homology-like domain, family A, member 3	1
Plag1	pleiomorphic adenoma gene 1	1
Paox	polyamine oxidase (exo-N4-amino)	2
Polr2a	polymerase (RNA) II (DNA directed) polypeptide A	1
Polr2k	polymerase (RNA) II (DNA directed) polypeptide K	4
Polr2l	polymerase (RNA) II (DNA directed) polypeptide L	2
Kctd15	potassium channel tetramerisation domain containing 15	1
Kcnmb4	potassium large conductance calcium-activated channel, subfamily M, beta member 4	1
Kcna6	potassium voltage-gated channel, shaker-related, subfamily, member 6	1
Pou3f2	POU domain, class 3, transcription factor 2	1
Pbxip1	pre-B-cell leukemia transcription factor interacting protein 1	1
Pfdn4	prefoldin 4	1
Pdss2	prenyl (solanesyl) diphosphate synthase, subunit 2	1
Pfn1	profilin 1	1
Pdcd10	programmed cell death 10	2

GENE_SYMBOL	GENE_NAME	Alignment
Pdcd5	programmed cell death 5	2
Ptger1	prostaglandin E receptor 1 (subtype EP1)	NA
Psmd3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3	1
Psma5	proteasome (prosome, macropain) subunit, alpha type 5	3
Psmb10	proteasome (prosome, macropain) subunit, beta type 10	1
Psmb3	proteasome (prosome, macropain) subunit, beta type 3	3
Pin4	protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting, 4 (parvulin)	2
Prmt2	protein arginine N-methyltransferase 2	1
Prmt7	protein arginine N-methyltransferase 7	1
Ptprz1	protein tyrosine phosphatase, receptor type Z, polypeptide 1	15
Ptprs	protein tyrosine phosphatase, receptor type, S	1
Pcmtd1	protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1	2
Ptma	prothymosin alpha	5
Rab15	RAB15, member RAS oncogene family	1
Rabep2	rabaptin, RAB GTPase binding effector protein 2	1
Rdm1	RAD52 motif 1	1
Rad54l2	Rad54 like 2 (S. cerevisiae)	1
Rapgef3	Rap guanine nucleotide exchange factor (GEF) 3	1
Rhoc	ras homolog gene family, member C	1
Rasal2	RAS protein activator like 2	1
Rasl2-9	RAS-like, family 2, locus 9	1
Rgs12	regulator of G-protein signaling 12	1
Rexo4	REX4, RNA exonuclease 4 homolog (S. cerevisiae)	1
Rft1	RFT1 homolog (S. cerevisiae)	1
Arhgap19	Rho GTPase activating protein 19	NA
Arhgef17	Rho guanine nucleotide exchange factor (GEF) 17	1
Rhbdd3	rhomboid domain containing 3	1
Rpl12	ribosomal protein L12	13
Rpl17	ribosomal protein L17	7
Rpl18a	Ribosomal protein L18A	4
Rpl23	ribosomal protein L23	1

GENE_SYMBOL	GENE_NAME	Alignment
Rpl23a	ribosomal protein L23a	16
Rpl27	ribosomal protein L27	10
Rpl28	ribosomal protein L28	6
Rpl3	ribosomal protein L3	4
Rpl35	ribosomal protein L35	5
Rpl36	ribosomal protein L36	10
Rpl36a	ribosomal protein L36a	5
Rpl37	ribosomal protein L37	1
Rpl37a	ribosomal protein L37a	1
Rpl38	ribosomal protein L38	5
Rpl41	ribosomal protein L41	7
Rpl5	ribosomal protein L5	6
Rpl7a	ribosomal protein L7a	14
Rpl9	ribosomal protein L9	10
Rps12	ribosomal protein S12	1
Rps13	ribosomal protein S13	NA
Rps16	ribosomal protein S16	4
Rps21	ribosomal protein S21	1
Rps23	ribosomal protein S23	10
Rps27	ribosomal protein S27	6
Rps28	ribosomal protein S28	6
Rps29	ribosomal protein S29	5
Rps5	ribosomal protein S5	1
Rps7	ribosomal protein S7	21
Rps9	ribosomal protein S9	2
Rplp2	ribosomal protein, large P2	2
0610040J01Rik	RIKEN cDNA 0610040J01 gene	1
1110002L01Rik	RIKEN cDNA 1110002L01 gene	1
1700049G17Rik	RIKEN cDNA 1700049G17 gene	1
1700073E17Rik	RIKEN cDNA 1700073E17 gene	1
1810022K09Rik	RIKEN cDNA 1810022K09 gene	2
2610042L04Rik	RIKEN cDNA 2610042L04 gene	35
2610307P16Rik	RIKEN cDNA 2610307P16 gene	1

GENE_SYMBOL	GENE_NAME	Alignment
2700029M09Rik	RIKEN cDNA 2700029M09 gene	1
2700060E02Rik	RIKEN cDNA 2700060E02 gene	1
2810468N07Rik	RIKEN cDNA 2810468N07 gene	NA
3000002C10Rik	RIKEN cDNA 3000002C10 gene	1
3300002I08Rik	RIKEN cDNA 3300002I08 gene	1
4930480K23Rik	RIKEN cDNA 4930480K23 gene	1
4930481A15Rik	RIKEN cDNA 4930481A15 gene	2
9930104L06Rik	RIKEN cDNA 9930104L06 gene	1
A430005L14Rik	RIKEN cDNA A430005L14 gene	1
B230118H07Rik	RIKEN cDNA B230118H07 gene	1
B230369F24Rik	RIKEN cDNA B230369F24 gene	1
D130017N08Rik	RIKEN cDNA D130017N08 gene	1
D130040H23Rik	RIKEN cDNA D130040H23 gene	1
D330041H03Rik	RIKEN cDNA D330041H03 gene	NA
D430019H16Rik	RIKEN cDNA D430019H16 gene	1
E030024N20Rik	RIKEN cDNA E030024N20 gene	1
E030030I06Rik	RIKEN cDNA E030030I06 gene	2
Rfwd3	ring finger and WD repeat domain 3	2
Rnf150	ring finger protein 150	1
Rnf167	ring finger protein 167	NA
Rbm10	RNA binding motif protein 10	1
Rpusd2	RNA pseudouridylate synthase domain containing 2	1
Robo4	roundabout homolog 4 (Drosophila)	1
Rspo3	R-spondin 3 homolog (Xenopus laevis)	NA
Rspo1	R-spondin homolog (Xenopus laevis)	1
Rufy2	RUN and FYVE domain-containing 2	1
Runx1t1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)	1
Ruvbl1	RuvB-like protein 1	2
Ryr2	ryanodine receptor 2, cardiac	NA
Sap30l	SAP30-like	1
Slfn9	schlafen 9	2
Scrib	scribbled homolog (Drosophila)	1

GENE_SYMBOL	GENE_NAME	Alignment
Scyl1	SCY1-like 1 (S. cerevisiae)	1
Sec61b	Sec61 beta subunit	2
Sec61g	SEC61, gamma subunit	2
Selenbp2	selenium binding protein 2	2
Selk	selenoprotein K	2
Serpina1d	serine (or cysteine) peptidase inhibitor, clade A, member 1d	2
Spint2	serine protease inhibitor, Kunitz type 2	NA
Srrm1	serine/arginine repetitive matrix 1	1
Stk32c	serine/threonine kinase 32C	NA
Sh3pxd2a	SH3 and PX domains 2A	1
Shroom4	shroom family member 4	1
Sirt3	sirtuin 3 (silent mating type information regulation 2, homolog) 3 (S. cerevisiae)	1
Ssb	Sjogren syndrome antigen B	1
Srgap1	SLIT-ROBO Rho GTPase activating protein 1	1
Snrpf	small nuclear ribonucleoprotein polypeptide F	3
Snrpg	small nuclear ribonucleoprotein polypeptide G	3
Snapc4	small nuclear RNA activating complex, polypeptide 4	1
Smcr8	Smith-Magenis syndrome chromosome region, candidate 8 homolog (human)	1
Snai2	snail homolog 2 (Drosophila)	1
Scn5a	sodium channel, voltage-gated, type V, alpha	1
Slc24a2	solute carrier family 24 (sodium/potassium/calcium exchanger), member 2	1
Slc25a23	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 23	1
Slc26a11	solute carrier family 26, member 11	1
Slc35b2	solute carrier family 35, member B2	1
Slc35e3	solute carrier family 35, member E3	2
Slc4a4	solute carrier family 4 (anion exchanger), member 4	1
Slc6a9	solute carrier family 6 (neurotransmitter transporter, glycine), member 9	1

GENE_SYMBOL	GENE_NAME	Alignment
Slc7a1	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1	1
Slco3a1	solute carrier organic anion transporter family, member 3a1	1
Slco5a1	solute carrier organic anion transporter family, member 5A1	1
Sorbs1	sorbin and SH3 domain containing 1	1
Spock1	sparc/osteonectin, cwcv and kazal-like domains proteoglycan 1	1
Sparcl1	SPARC-like 1 (mast9, hevin)	1
Spata5l1	spermatogenesis associated 5-like 1	NA
Smpd1	sphingomyelin phosphodiesterase 1, acid lysosomal	1
Sf3a2	splicing factor 3a, subunit 2	1
Spry4	sprouty homolog 4 (Drosophila)	1
Sart1	squamous cell carcinoma antigen recognized by T-cells 1	NA
Sart3	squamous cell carcinoma antigen recognized by T-cells 3	1
Sox12	SRY-box containing gene 12	1
Sox3	SRY-box containing gene 3	NA
Sox5	SRY-box containing gene 5	1
Sfn	stratifin	NA
Strn3	striatin, calmodulin binding protein 3	1
Sod2	superoxide dismutase 2, mitochondrial	1
Synpo2l	synaptopodin 2-like	1
Syt7	synaptotagmin VII	NA
Ss18	synovial sarcoma translocation, Chromosome 18	2
Taf10	TAF10 RNA polymerase II, TATA box binding protein (TBP)-associated factor	1
Tbc1d10b	TBC1 domain family, member 10b	1
Tbx1	T-box 1	1
Tead1	TEA domain family member 1	1
Tgif2	TGFB-induced factor 2	3
Traip	TRAF-interacting protein	1
Tle4	transducin-like enhancer of split 4, homolog of Drosophila E(spl)	1
Tmed1	transmembrane emp24 domain containing 1	1

GENE_SYMBOL	GENE_NAME	Alignment
Tmed10	transmembrane emp24-like trafficking protein 10 (yeast)	2
Tmem109	transmembrane protein 109	1
Tmem158	transmembrane protein 158	1
Tsen54	tRNA splicing endonuclease 54 homolog (SEN54, S. cerevisiae)	1
Tnni2	troponin I, skeletal, fast 2	1
Trub2	TruB pseudouridine (psi) synthase homolog 2 (E. coli)	1
Tysnd1	trypsin domain containing 1	1
Tufm	Tu translation elongation factor, mitochondrial	2
Tsc2	tuberous sclerosis 2	1
Tbcc	tubulin-specific chaperone c	1
Tnfsf12	tumor necrosis factor (ligand) superfamily, member 12	1
Ttyh2	tweety homolog 2 (Drosophila)	1
Ywhae	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide	1
Ube2d1	ubiquitin-conjugating enzyme E2D 1, UBC4/5 homolog (yeast)	1
Uxt	ubiquitously expressed transcript	1
B4galt5	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5	1
Unc45b	unc-45 homolog B (C. elegans)	1
Ulk1	Unc-51 like kinase 1 (C. elegans)	2
Upp2	uridine phosphorylase 2	1
Utp14b	UTP14, U3 small nucleolar ribonucleoprotein, homolog B (yeast)	NA
Abl2	v-abl Abelson murine leukemia viral oncogene 2 (arg, Abelson-related gene)	NA
Vps13d	vacuolar protein sorting 13D (yeast)	1
Vegfc	vascular endothelial growth factor C	1
Ralb	v-ral simian leukemia viral oncogene homolog B (ras related)	1
Wasf3	WAS protein family, member 3	1
Wdr20	WD repeat domain 20	1
Wdr76	WD repeat domain 76	1

GENE_SYMBOL	GENE_NAME	Alignment
Wtip	WT1-interacting protein	1
Zic2	Zic finger protein of the cerebellum 2	1
Zc3hav1l	zinc finger CCCH-type, antiviral 1-like	1
Zfp109	zinc finger protein 109	1
Zfp146	zinc finger protein 146	1
Zfp286	zinc finger protein 286	1
Zfp36l2	zinc finger protein 36, C3H type-like 2	1
Zfp395	zinc finger protein 395	1
Zfp428	zinc finger protein 428	1
Zfp503	zinc finger protein 503	1
Zfp521	zinc finger protein 521	1
Zfp58	zinc finger protein 58	NA
Zfp597	zinc finger protein 597	1
Zfp691	zinc finger protein 691	1
Zfp697	zinc finger protein 697	1
Zic5	zinc finger protein of the cerebellum 5	2
Zfand1	zinc finger, AN1-type domain 1	2
Zbed3	zinc finger, BED domain containing 3	1