University of Arkansas, Fayetteville

ScholarWorks@UARK

Graduate Theses and Dissertations

8-2017

Identifying Three-Way Gene Interactions from Microarray Data using Kolmogorov-Smirnov and Cross-Match Tests

Shubhashree Khadka
University of Arkansas, Fayetteville

Follow this and additional works at: https://scholarworks.uark.edu/etd

 Part of the Genetics Commons, and the Statistics and Probability Commons

Citation

Khadka, S. (2017). Identifying Three-Way Gene Interactions from Microarray Data using Kolmogorov-Smirnov and Cross-Match Tests. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/2449

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, uarepos@uark.edu.

Identifying Three-Way Gene Interactions from Microarray Data
Using Kolmogorov-Smirnov and Cross-Match Tests

A thesis submitted in partial fulfillment
Of the requirements for the degree of
Master of Science in Statistics and Analytics

By

Shubhashree Khadka
Southern Arkansas University
Bachelor of Science in Computer Science, 2015

August 2017
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____

Dr. Qingyang Zhang
Thesis Director

_____ _____

Dr. Mark Arnold
Committee Member

Dr. Avishek Chakraborty
Committee Member

ABSTRACT

Human gene network is much more complex than just pairwise interaction among the genes. Zhang et al. [6] extracted microarray data from International Genomics Consortium (IGC), and presented the detection of three-way gene interactions in their paper using Fisher's z-transformation test. Three-way gene interactions are closer than pairwise correlations in representing the complex gene structures. Additionally, it was more tractable than assessing four or more gene interactions. In this paper, we are simulating different models where Fisher's test might not be as effective. Zhang et al.'s approach utilized Pearson's correlation coefficients and involved detection of linear interactions only. Since gene interactions could show any kind of behavior, their evaluation approach might not work most of the time. Therefore, we are utilizing the dataset Zhang et al. provided in order to detect the three-way gene interaction using non-parametric tests like Kolmogorov-Smirnov and Cross-Match.

ACKNOWLEDGEMENTS

I would like to convey my utmost gratitude to my advisor, Dr. Qingyang Zhang, for his instruction, continued support, and patience throughout this research project. Additionally, I would like to thank the rest of the thesis committee, Dr. Mark Arnold and Dr. Avishek Chakraborty, for committing their time in providing me with their valuable guidance in improving this project. I am indebted to all my professors and instructors for their direction and motivation.

I would also like to thank my father, Mr. Gokarna Bahadur Khadka, for always being there to end all my doubts and worries. I am grateful to Caleb Parks, for without your encouragement, none of this would have been possible. My special thanks to the rest of my family and friends for the abundant support they always bestow upon me.

DEDICATION

To my late grandfather and my first teacher, Tirtha Bahadur Khadka, I still miss you every day. To my parents, Gokarna Khadka and Meena Pandey, for always expecting the best from me. To my brother, Ishan Khadka, whose striving to excel in music always inspires me. To my Gangadi, for your love and sacrifices. To my sixth grade math teacher, Mr. Madan Chand, for influencing my interest in math. Lastly, to Caleb Parks, for being the best-friend I look up to and count on. You bring out the best in me.

TABLE OF CONTENTS

1. INTRODUCTION

A Human Genome provides a vast amount of information about a human being. It is a complete set of nucleic acid sequence, encoded as DNA with 23 chromosome pairs in cell nuclei as well as in a small DNA molecule found within individual mitochondria. Human genome consists about 19000-20000 protein-coding genes. The human genome sequence determines the human development and physiology. It also aides in the advancement of medicine and in understanding evolution. It contains the blueprint of human life [1]. Genetic variations influence different attributes of human body such as eye color, hair color or height, but most importantly, proneness to hereditary diseases like color blindness, cystic fibrosis, or diabetes.

Genetic diseases like Down syndrome, and Hemophilia are cases of chromosomal abnormality whereas, diseases like Type 2 diabetes or cancer could be the result of family history or environmental factors. Cystic Fibrosis is induced by mutation of both of the copies of CFTR (Cystic fibrosis transmembrane conductance regulator**)** gene, and the presence of BRCA1 and BRCA2 genes give rise to breast cancer. Gene-gene interaction has been attributed to understanding the causes of complex disease traits [2].With genome sequencing, such gene interactions can be detected and the genetic abnormalities managed, if not treated.

In the mid to late twentieth century, genome sequencing was done manually using methods like Maxam-Gilbert sequencing and Sanger sequencing. Manual sequencing of even microscopic organisms took years to complete. The manual method was simply writing down all the base pairs in a DNA molecule. It took scientists about 10 years to identify the CFTR gene that mutates and causes cystic fibrosis. Besides being time-consuming, this method also posed a high risk of erroneous data sequencing. It was in the 1990s that the transition of genome sequencing methods from manual to the much faster, automated sequencing was made. Shortly

after that, the Human Genome Project was proposed in order to record the entire human genome. Thus, the largest undertaking in the history of biological science started in 1990 and was conducted in a number of universities all around the world until its completion in 2003 [3]. In the meantime, a parallel project was carried out by Celera Genomics from 1998 and continued through 2003. The improved sequence was then published and has been made freely available for researchers ever since.

Gene Interaction or Epistasis is the influence that a gene has in the presence of one or more controller genes. Steen [4] mentions that gene-gene interaction on traits of interest presents an exponential growth in terms of methodological development as well as translation of statistical gene interactions to biological. Furthermore, gene interactions and genomic complexity are correlated i.e. with more complex gene interaction, the mutational effects tend to strengthen each other rather than cancel out like in the cases of less complicated epistasis. The existence of complex epistasis results in genetic variation in complex diseases like asthma, cancer, diabetes, hypertension, and obesity [2]. It is essential for researchers to model complex gene-gene interactions so as to understand the joint genetic effects that lead to complex diseases.

In her paper, Cordell [5] brings into attention that in most genome-wide association studies, only single-locus analysis strategy is utilized where each variant is tested individually for a specific phenotype. Those studies are often unsuccessful because they fail to account for any interaction between different loci. If complex diseases occur as a result of complex mechanisms that involve multiple genes and environmental factors, then studying each gene in isolation may lead us to miss those genetic interaction effects. In addition, Zhang et al. [6] stresses pairwise correlation to extract gene network information is too simplistic to express the complex relationships among real genetic structures. In their paper, they introduced a three-way

interaction model where they employ a controller gene to demonstrate the dynamic nature of co-expression in gene pairs. They evaluated three-way interactions in each gene triplet by computing Pearson correlation coefficient of the log-scale values between the controlled genes, followed by Fisher's z-transformation to transform the correlation coefficients to a test statistic z.

In this paper, we use the same microarray dataset preprocessed by Zhang et al. that comprised of 1000 genes containing 678 cancer samples. For the gene expressions, instead of targeting linear interactions as in Zhang et al. [6], we have decided to assess the three-way gene interactions in the gene triplets using non-parametric approaches like Kolmogorov-Smirnov and Crossmatch tests. Due to the high-dimensionality of the data as well as high time complexity of the tests, we will evaluate the gene triplets for the first 50 genes of the 1000 genes included in the microarray dataset. We will take log-scale values of all gene expressions and cluster them into low and high expressions before conducting the two tests.

In Chapter 2, we will further discuss the paper by Zhang et al. [6] .After that, in Chapter 3, we will simulate the Fisher's z-transformation, Kolmogorov-Smirnov, and Crossmatch tests for different models and discuss in detail the implementation of the K-S and C-M tests on the microarray dataset. Results for the two non-parametric tests are interpreted in Chapter 4, followed by the Conclusion in Chapter 5.

2. BACKGROUND/LITERATURE REVIEW

2.1 Data pre-processing and clustering

According to Zhang et al. [6], pairwise correlation does not record the dynamic characteristic of genetic co-expression relationships because a gene pair may be co-expressed only in a specific organ or in a particular disease state. Zhang et al. proposed an alternative approach to identify co-expression gene network by introducing a third 'controller gene' that can affect such co-expression associations. Therefore, this approach focuses on three-way gene interactions in lieu of pairwise correlations or two-way interactions.

The human gene network is much more complex than just three-way interactions among the genes. More than one gene and environmental factor may affect the co-expression relationship among a gene pair. Since models that involve more than three genes have higher number of combinations or gene triplets, it is less tractable than three-way interaction models. Hence, assessing interaction among gene triplets is a reasonable compromise between authenticity and tractability.

Zhang et al. [6] obtained the raw microarray data from Gene Expression Omnibus database whereas, the data was generated by International Genomics Consortium (IGC) in its Expression Project for Oncology using Affymetrix human genome array HG-U133 plus 2.0. They used data in IGC batches excluding the batches that showed significant difference from the samples in other batches. The samples were derived from cancer tissues or cell lines. According to Hansen and Irizarry [22], the RNA-sequence data contains some distortions and require data normalization. They proposed a normalization method that improved the precision without loss of accuracy. The normalization method also removed systematic bias brought about by deterministic features such as Guanine-Cytosine content. Quantile Normalization can be defined

as a technique to make two distributions identical in statistical properties so they can be compared efficiently. Quantile Normalization Method was used to normalize the probe level data followed by Positive Dependent Nearest Neighbor (PDNN) Model to obtain the gene expression values. After the application of quantile normalization again to reduce biases, the samples were divided into training and testing sets.

The processed data was then subjected to MCLUST, a software package for cluster analysis. MCLUST implements parameterized Gaussian hierarchical algorithms and EM algorithm (Farley and Raftery[7]). MCLUST also provides a function ‘bic' to compute the Bayesian Information Criterion or BIC given the data and a model along with conditional probability estimates [7]. This model-based clustering algorithm is used to check if the distribution of log-scale values of a gene is a single normal or a mixture of two normal distributions (Zhang et al. [6])

Using the function for BIC, Zhang et al. [6] found out that the logarithmic gene expressions were a mixture of two normal distributions. MCLUST was then used to compute a threshold value to T in order to divide the samples into two subgroups. Only subgroups with at least 60 samples were selected for analyses to ensure sufficient sample size for the subsequent assessments.

2.2 Evaluating three-way gene interactions

Zhang et al. [6] decided on assessing three-way gene interactions because they are closer, in characteristic, to the actual gene networks than two-way interactions or pairwise correlation. In addition, evaluating gene triplets is much more feasible than four or more genes because the number of combinations for three genes is much smaller. While considering three genes for analysis, they supposed the genes in a triplet as A, B, and C where C is the controller

gene. After the pre-processing and clustering of the data, the threshold T was noted for each of the controller genes in the training dataset. Zhang et al. [6] then divided the 339 samples in the training set to low expression group of n_1 sample size and high expression group of n_2 sample size according to the threshold obtained for a particular controller gene C.

Once the samples are divided into low and high expression subgroups, Pearson correlation coefficient of the log-scale values between gene A and gene B are computed for each of the subgroups. Robert [8] defines correlation as a statistical measure of how closely two variables are related. Correlation can either be positive or negative and the degree of correlation strong or weak. Pearson Product Moment Correlation or Pearson correlation is a statistic 'r' that measures the strength of a linear association between gene A and gene B expressions and is calculated using the following formula:

$$r = \frac{S_{ab}}{\sqrt{S_{aa} S_{bb}}}$$

In the above equation, S_{ab} is the covariance between gene A and gene B expressions whereas, S_{aa} is the variance within gene A and S_{bb} is the variance within gene B. Using the formula for r, two Pearson correlations coefficients r_1 and r_2 for log-scale values between gene A and gene B are computed for n_1 and n_2 samples respectively. Since the variances of gene A and gene B expressions are small, the Pearson correlation coefficients tend to be unstable. Hence, all the triplets with either variance of gene A or gene B less than 0.1 are discarded in either of the low expression or high expression subgroups.

The computations of Pearson correlation coefficients is followed by Fisher's z-transformation to convert r_1 and r_2 correlation coefficients into a z_1 and z_2 values. According to Fisher and Belle [9], the Fisher z-transformation is used to transform responses whose range is between -1 and 1. The z-transformation was developed especially for the Pearson product-moment correlation coefficient. The standard normal transformation is used mostly in nonparametric analyses. This method is used mostly for testing purposes rather than estimation. One of the most important benefits of this transformation is that the tables, procedures, and software algorithms for normal transformation procedures are already available. For a particular controller gene C, the low and high expression subgroups of the training set each have the correlation coefficients r_1 and r_2. These coefficients r_1 and r_2 are first transformed into z_1 and z_2 using the following formulas:

$$z_1 = 0.5 * \ln\left[\frac{1 + r_1}{1 - r_1}\right]$$

$$z_2 = 0.5 * \ln\left[\frac{1 + r_2}{1 - r_2}\right]$$

After transforming the two correlation coefficients, for each controller gene C, Zhang et al. [6] determined the z-statistic using two sample z-test and the formula for the statistic is as follows:

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Where, z_1 and z_2 are previously computed values and n_1 and n_2 are the low expression and high expression subgroup sizes respectively. The z-statistic is then used to assess whether there are significant pairwise interaction among Gene A and Gene B when there is a high expression level of controller gene C.

Using MCLUST, Zhang et al. [6] determined that out of the 1000 genes they had previously selected, 796 genes possessed a bimodal expression distributions in the training set. These 796 genes were used as the controller genes and separated into low and high expression subgroups. They tested 0.4 billion possible three-way interactions among 1000 genes that comprised of 796 controller genes. In theory the 0.4 billion z-statistics obtained from the triplets are supposed to have a standard normal distribution. However, Zhang et al. [6] discovered that the z values showed a variance of 1.88. This inflation in the variance indicates that the z-values are not independent of each other and cannot be regarded as resulted from null data. The heavy tails of the distributions of z-statistics also suggested the prevalence of significant gene interactions.

3. METHODOLOGY

3.1 Data Preparation

The data we utilized is obtained from the Zhang et al. [6] resources. The data provided was split into training and testing sets. We combined the two sets and used log-scaled gene expression data. The initial data includes 1000 genes. However, we are only including 50 genes in our study because of the large number of combination of gene triplets and the high time complexity of the tests. Parallel computing technique such as GPU computing will be needed in order to analyze the whole set of 1000 genes. Minimizing the number of genes from 1000 to 50 decreased the number of three gene combinations from 498501000 to 58000 and the processing time from several years to a few days. After that, we applied MCLUST function (discussed in Chapter 2.1) using built in 'mclust' package in R so as to cluster the sample data into low expression and high expression sub-groups for all the genes. For each of the controller genes, we generated and saved all the different cutoffs that separated the two sub-groups. After clustering, for any particular controller gene C, we applied the Kolmogorov-Smirnov and Cross-Match tests and recorded their statistics.

3.2 Data Simulation

Before carrying out the Kolmogorov-Smirnov(K-S) and Cross-Match(C-M) tests to check for three-way gene interactions in the dataset, we simulated the Fisher's z-transformation, K-S and C-M tests on various models to inspect the instances when Fisher's z transformation test is not as effective as K-S and C-M tests in comparing two distributions. First, we generated a logarithmic and an exponential model in order to run the three different detecting methods for

the two distributions. We followed it by generating two bivariate-normal models in order to carry out the same goodness-of-fit tests. And lastly, different polynomials were produced to repeat the same procedure. For the three sets of simulations, three different statistics i.e. p-values of z-transformations, p-values of cross-match test and K-S statistics were documented.

3.3 Kolmogorov-Smirnov Test

Goodness-of-fit tests have been described by Lopes et al. [10] as statistics that measure the likeliness of a dataset or sample to some theoretical probability distribution. One of the most commonly used goodness-of-fit tests is Pearson's Chi-square correlation test where we take the distribution of an observed data and compare it to the expected probability distribution. We are trying to assess the effect of low and high expression of controller gene in the interaction of gene pairs in the dataset obtained by Zhang et al. [6]. Therefore, as our first approach to extract three-way gene interactions, a goodness-of-fit test known as Kolmogorov-Smirnov (K-S) test is performed. K-S test is much like the chi-square test except it is exclusively used for continuous data and powerful than the usual Pearson chi-square test [11]. There is also no loss of data in K-S test unlike the chi-squared test.

Kolmogorov-Smirnov (K-S) test is a non-parametric hypothesis test [12]. The test is named after Andrey Kolmogorov and Nikolai Smirnov. The classical one-dimensional K-S test measures the probability that a particular sample distribution is drawn from the same parent population as a continuous reference model whereas, the two-sample K-S test compares the probability distribution functions of one sample dataset to a second sample dataset. In any case, one of the first things to obtain is the empirical distribution of the sample datasets by taking the integral of their probability density functions. The empirical distribution function (EDF) simply

gives the cumulative probability of a random variable. The K-S test depends on the K-S statistic which measures the greatest distance between the EDF of the dataset and reference model, in case of one-dimensional K-S, or the supremum distance between EDF of dataset one and dataset two in case of the two-dimensional K-S test. Evans et al. [13] mentioned in their paper that, the EDF $F_n(x)$ is the proportion of the observations $X_1,X_2,...,X_n$ that are less or equal to x and is defined as:

$$F_n(x) = \frac{I(x)}{n},$$

where, n is the size of the random sample and I(x) is the number of X_i's less than or equal to x.

For the purpose of our study, we will exclusively discuss a two-dimensional K-S test as the genetic interactions involve two genes in low or high expression of controller gene C. Lopes et al. [10], defined two independent stochastic variables X and Y whose cumulative distribution functions F and G are unknown. $X_1,X_2,...,X_n$ are the observed samples for low controller gene C and $Y_1,Y_2,...,Y_n$ are the observed samples for high controller gene C. After that, they proposed the hypotheses for the K-S test as follows:

$$H_0 : F(x) = G(x), \text{for every } x \, \epsilon \, R^d$$
$$H_1 : F(x) \neq G(x), \text{for some } x \, \epsilon \, R^d$$

Where, H_0 is the null hypothesis that suggests that the two distributions are almost identical and H_1, the general alternative hypothesis, indicates that the two distributions are different for some

random variable x. In our research study, however, null hypothesis states that the low or high expression of controller gene had no effect in the interaction of gene A and gene B, while the alternative hypothesis indicates that there are some samples with gene A and gene B interactions for low or high expression of controller genes.

The K-S test is applicable to continuous, unbinned data samples and uses the supremum absolute difference between sample distributions with the low expression gene C and high expression gene C functions [10]. When comparing the two distribution functions F(x) and G(x) where x is a random sample observation, the K-S statistic is defined as:

$$D_{KS} = \max |F(x) - G(x)|$$

The K-S statistic for a two-dimensional space is difficult to obtain because unlike the one-dimensional K-S test, the direction of the ordering of the data would have to be considered [10]. Therefore, Peacock has put forth the idea of making K-S statistic independent of the any sort of ordering [14]. The same Peacock test was utilized in our research. Even though this method is very efficient, it is also very demanding. Performing the test on 2^{18} points on a 4GHz processor would require several days to execute [10]. It is due to high time complexity that we only consider using the first fifty out of 1000 genes in the dataset.

The Peacock test only outputs K-S statistics but not the p-values. Hence, the hypotheses was tested using only K-S statistics. Based upon a significance level α=0.05, a critical value D_α is found from the standard K-S table. When the sample size is greater than 35, the K-S statistic can be computed by dividing 1.36 by square root of sample size(n) for α=0.05. In our case, using 50 genes, the critical value can be set as $(1.36/\sqrt{50}) = 0.19$. Once the K-S statistic (D_{KS}) is

obtained from the Peacock test, it is compared with the $D_\alpha = 0.19$ in order to draw conclusions about the K-S test. If $D_{KS} < D_\alpha$, we fail to reject the null hypothesis that the expression level of controller gene has no effect on the interaction between gene A and gene B. On the other hand, if $D_{KS} > D_\alpha$ we reject the null hypothesis and conclude that the low/high expression of a particular controller gene has a significant effect on the two-way gene interactions. For the 50 genes, we computed 58800 K-S statistics and sorted them in descending order so as to extract the top-five significant three-way gene interactions.

In [12], Feigelson and Babu mention that the K-S test is used in over 500 articles every year. It is very convenient to use because it is distribution-free, there is no restriction on the size of the sample, the critical values are widely available and it is easy to understand graphically. The article then puts forward instances where KS-test might not be the most sensitive. KS-test is sensitive when the EDFs differ in the center of the distribution. However, in case of repeated deviations in the distributions when the curves cross each other multiple times, the measured deviations is reduced. The Anderson-Darling (AD) test is better because other than center and deviation, AD test is also sensitive to the differences between the curves at the beginning as well as end of the two EDFs. Nevertheless, since we are not much concerned about the tails (outliers), we assess the hypotheses using the K-S test.

3.4 Cross-Match Test

Heller et al. [15] define the Cross-Match (C-M) test as “an exact distribution-free test of no treatment effect on a high-dimensional outcome in a randomized experiment”. C-M test compares two multivariate distributions by using distances between observations [16]. This comparison is done by using optimal non-bipartite matching to pair 2I subjects into I pairs based

on similar outcomes. Rosenbaum [16] introduced optimal non-bipartite matching in his paper as matching the sample data into disjoint pairs to minimize the total distance within pairs. In other words [15], we are concerned about the number of times that a subject from a low expression gene C was paired with the one from a high expression gene C sub-group. The cross-match statistic A is the total number of times such a cross-match is made. Test statistics A indicates the total number of pairs that include one observation from a low expression controller gene sub-group and another observation from a high expression sub-group.

One of the most powerful tools in statistical design and analysis is 'Matching' [17]. While bipartite matching is most commonly used, it is limited to simpler designs. Hence, a non-bipartite pairing can be introduced to take care of multiparty matching situations and to find sets of pair such that they minimize the sum of distances based on a given distance matrix. Non-bipartite matching provides options like multi-group comparisons which brings about greater flexibility than the bipartite matching. The goal of our study is to evaluate the causal effect of low expression of controller gene C versus high expression. One of the best methods of carrying out causal inference is randomized experiment [18]. However, we have no control over whether a sample consists of low expression of controller gene or high and we would have to execute observational studies instead. Therefore, in order to manage any selection bias in the study, utilizing pairing methods is a good choice [19]. Lu et al. [17] mentions a number of benefits of utilizing pairing methods. Well-paired datasets provide easy to understand analyses. Also, some paired analyses do not require parametric assumptions. Additionally, non-overlapped pairs enables the proper use of existing inference models and the method of pairing does not involve the outcome variable information, thus, preventing any data manipulation.

Even though bipartite pairing is the most popular pairing procedure, non-bipartite is a better alternative to perform causal inference in observational studies and carrying out an exact distribution-free test between two multivariate distributions [17]. Lu et al. further mention that optimal pairing is the pairing that minimizes the total distance among all pairs. Bipartite and Non-bipartite pairing algorithms can be determined by the number of disjoint groups in the graph. In bipartite graph, the disjoint pairs are produced from only two disjoint groups, whereas, in non-bipartite graph, there are multiple groups that provide disjoint pairs. The optimal non-bipartite matching has not gained much attention partly because of the complex algorithm. In observational studies like the one we have, there is no control over whether a sample contains low or high expression of controller gene C and carrying out a bipartite pairing is not satisfactory at all. Hence, we cannot replace non-bipartite pairing by the simpler bipartite method. According to Papadimitriou and Steiglitz [20], one of the well-known algorithms of optimal non-bipartite matching is based on Derig's shortest augmentation path algorithm.

Non-bipartite matching is beneficial mostly because it generates unbiased treatment effect estimation in complicated observational studies. Optimal non-bipartite pairing can also be used to construct a distribution-free assessment for comparing two multivariate distributions [17]. Rosenbaum [16] used the optimal non-bipartite pairing to come up with an exact test to check whether two different distributions follow the same parent distribution. The two distributions can either be the treated and controlled groups or like in our case, it can be the low and high expression of controller gene C groups. The hypotheses for the test can be set as follows:

$$H_0 : \ F_L(x) = F_H(x), for\ every\ x\ \epsilon\ R^d$$

$$H_1 : \ F_L(x) \neq \ F_H(x), \text{for some } x\ \epsilon\ R^d$$

where, F_L is the distribution of observations with low expression of controller gene and F_H is the distribution of observations with high expression of controller gene C. First, the observations from the two groups are pooled while ignoring the grouping information. Then, optimal non-bipartite matching was carried out to create matched pairs among all observations. These matched pairs can include observations with low expressions only, ones with high expressions only, or pairs with one low and one high expressions of gene C. If the third category, with pairs containing both low and high expression observations, have really low number of pairs then, it provides significant evidence against the null hypothesis. Hence, suggesting that the two distributions of low/high expressions are not the same and that the controller gene could have some influence on the observations. This test proposed by Rosenbaum is known as the Cross-Match (C-M) test. Rosenbaum also derived the normal approximation version of the C-M test and compared it to the Kolmogorov Test [17].

In [15], Heller et al. have provided a definition of the Cross-Match Statistics as follows: If there are 2I subjects, m= 1, 2,..., 2I, where subject m has low expression of controller gene C if indicator $U_m = 0$ and has high expression of gene C if $U_m = 1$. The number of observations with high expression gene C is given by $n = \sum U_m$ from m = 1 to 2I whereas, the number of observations with low expression gene C is simply 2I-n. According to Rosenbaum [16], a $2I \times 2I$ symmetric distance matrix is defined with row k and column m giving a distance between observations. Then, the 2I subjects are paired into I non-overlapping pairs to minimize the distances within pairs. Non-overlapped pairs are pairs that are matched without replacement [21].

In [15], the subjects are renumbered, j= 1,.., 2I so that subjects 2i are paired for i=1,..,I. As mentioned before test statistic A is the total number of pairs that contain one observation with low expression of gene C and one observation with high expression. The formula for test statistics A is given as follows:

$$A = \sum_{i=1}^{I} \left[U_{2i-1}(1 - U_{2i}) + (1 - U_{2i-1})U_{2i} \right],$$

where, A is the C-M statistic and a small value of A would suggest that the two distributions are different [16]. If instead of 2I subjects, there happened to be an odd 2I+1 number of subjects then, a pseudo-subject is added to the distance matrix at zero distance from everyone else. After that, I+1 pairs are formed and the pair containing the pseudo-subject is discarded. This is done to ensure that the least matchable subject is being discarded.

In order to carry out the C-M test on the 50 genes, for the assessment of three-way gene interactions, we used the data obtained upon clustering by the MCLUST function in R. The observations were clustered into low expression of controller gene and high expressions. For each of the controller gene, using the cutoff obtained from MCLUST, we applied the C-M tests using the 'crossmatch' package in R and obtained 58800 different p-values for all possible gene triplets with 50 genes. Then, these p-values were sorted in ascending order so as to find the top 5 significant three-way gene interactions in our dataset. We assessed it further by looking at the plots of each of the significant gene triplets we obtained.

4. RESULTS AND DISCUSSIONS

4.1 Results

A data simulation process preceded the research analysis. Before conducting the K-S and C-M tests on the data obtained from Zhang et al. [6], we used three different sets of models to simulate low expression and high expression data. Using these models, we checked the instances where K-S and C-M tests were more effective than Fisher's z-transformation. The models used were logarithmic, bivariate normal, and polynomial and had two variations that produced two different sets of data to simulate low and high expressions. Figure 4.1.1 demonstrates the three simulation models:

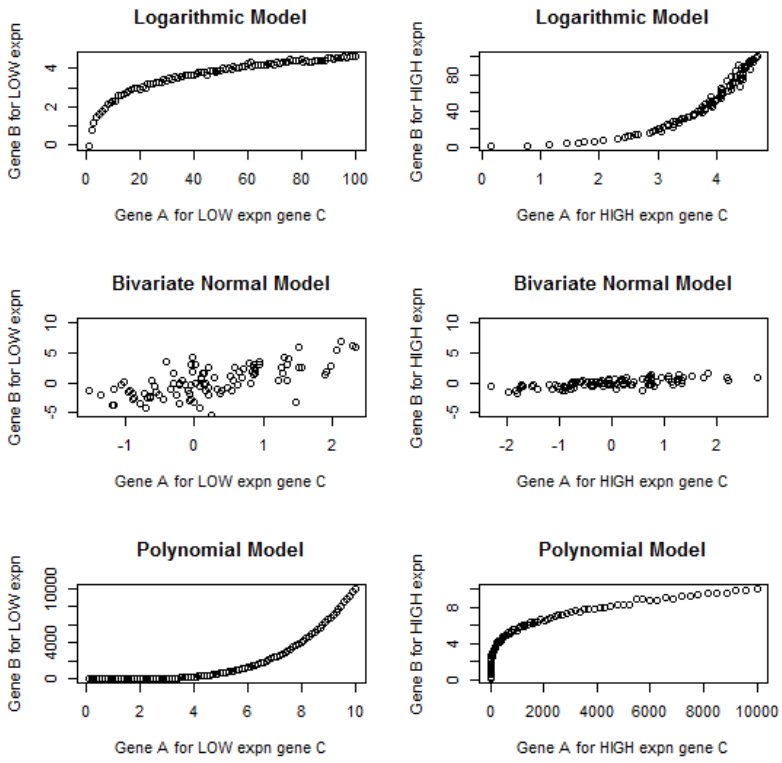

Figure 4.1.1. Sets of simulation models

The three different models and their low and high expressions were simulated using the following equations:

Logarithmic Model	lowA1 = seq(1,100,1) lowB1 = log(lowA1)+rnorm(100,0,1/16) highA1 = log(lowA1)+rnorm(100,0,1/16) highB1 = lowA1
Bivariate Normal Model	lowa2=rnorm(100,0,1) lowb2=2*lowa2+rnorm(100,0,2) lowgc2 = cbind(lowa2,lowb2) higha2=rnorm(100,0,1) highb2=1/2*higha2+rnorm(100,0,1/2)
Polynomial Model	lowa3= seq(0.1,10,0.1) lowb3=(lowa3)^4+rnorm(100,0,1/8) higha3 = (lowa3)^4 highb3= lowa3+rnorm(100,0,1/8)

Table 4.1.1. Equations for Data Simulation Models

In R, using the Peacock.test and crossmatch packages, we obtained the Fisher's p-value, K-S statistic and C-M p-value within low-high expressions of each of the simulation models. We have discussed the hypotheses for the three tests in Chapter 3. Low p-values for the Fisher's, and C-M tests as well as high K-S statistics for the gene triplets suggest that we reject the null hypothesis i.e. the distribution of low expression gene C and high expression gene are not from a same parent distribution. On the other hand, if the p-values were higher and the K-S statistics lower, it would suggest acceptance of the null hypothesis, thus, suggesting that the controller

gene had no effect on the gene-gene interaction. For the three models, we computed the following results:

Simulated Models	Fisher's p-value	K-S statistic	C-M p-value
Logarithmic	0.9478169	0.99	5.920000×10^{-24}
Bivariate Normal	0.6874768	0.38	4.517735×10^{-06}
Polynomial	0.9752655	0.89	3.105823×10^{-22}

Table 4.1.2. Test Results for the simulated models

In Data Simulation Results, using α=0.05 and D_α=0.19, high p-values for Fisher's z-transformation tests, high statistics for K-S tests and low p-values of C-M tests suggested that there were instances were K-S and C-M tests were better than the Fisher's test. It is easier to verify this through simulation because we simulated variations of different models and there is no way the test statistics should have led us to high p-values. The high p-values for Fisher's z-transformation suggests that the test only targets detection of linear correlations and fails to detect non-linear interactions among variables or genes. Therefore, for the 50 genes in the actual data, we applied the MCLUST clustering functions and obtained the various cutoffs for the controller genes. After that, we performed the K-S tests and saved 58,800 K-S statistics using the 'peacock2' command in R. We sorted the K-S in descending order and noted the top-5 significant gene triplets and statistics as follows:

Gene C	Gene A	Gene B	K-S statistics
31	29	30	0.95958082
31	2	30	0.95750182
31	5	30	0.95750182
31	13	30	0.95750182
31	14	30	0.95750182

Table 4.1.3. Top-5 significant Gene Triplets based on K-S statistics

The top-5 significant gene triplets have the highest KS-statistics, thus, leading us to reject the null hypothesis i.e. H_0: The Expression level of Controller Gene has no effect on the interaction been Gene A and B, when critical value D_α=0.19. Following are the graphs for the significant Gene A and Gene B interactions:

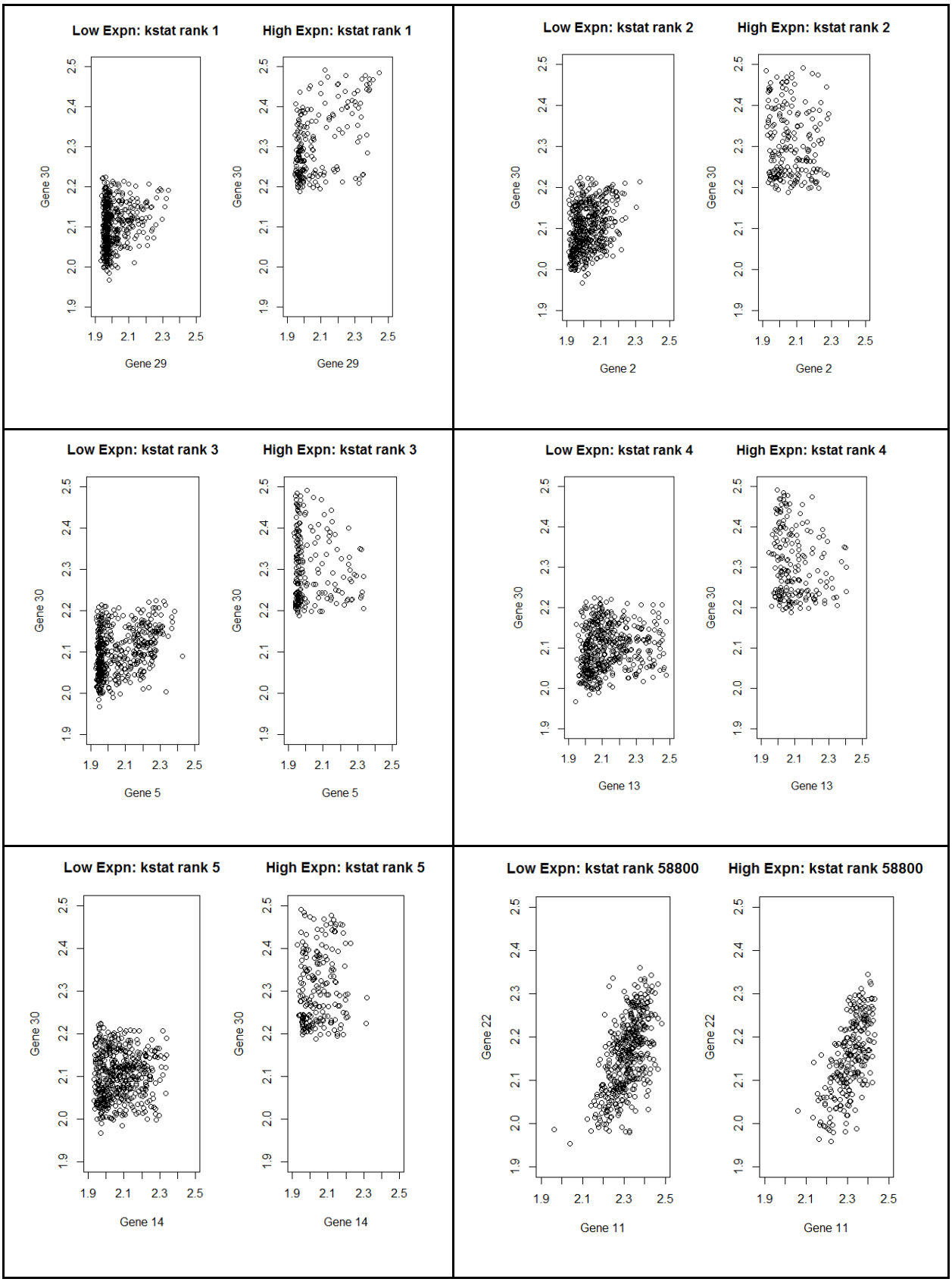

Fig 4.1.2. Gene A-Gene B interaction for K-S stats ranks 1 to 5, & 58800 for comparison

As one can notice in the graphs, top-5 ranked K-S statistics have gene A- gene B interaction graphs that are completely different for low and high expression controller genes C, whereas, in case of the interaction graph for the smallest K-S statistics the expression of controller genes has no effect on the gene interaction.

Similarly, we used the same data as in K-S test in case of C-M test. Cutoffs were generated from MCLUST to determine the high and low expressions of controller gene. After clustering, 'crossmatch' was used in R to compute and save all possible 58800 p-values for the C-M test. There p-values were then sorted in ascending order to gain the top-5 significant gene interactions which are presented in the table below:

Gene C	Gene A	Gene B	C-M p-values
30	31	38	2.864951×10^{-65}
30	10	31	1.537875×10^{-63}
31	30	48	4.329318×10^{-63}
30	1	31	7.813794×10^{-62}
30	2	32	7.813794×10^{-62}

Table 4.1.4. Top-5 significant Gene Triplets based on C-M p-values

Based on the lowest p-values, using $\alpha=0.05$, we reject the null hypothesis, H_0 : The Expression level of Controller Gene has no effect on the interaction been Gene A and B, just like in the case of K-S test. The graph of gene-gene interactions in case of low and high expression are included below:

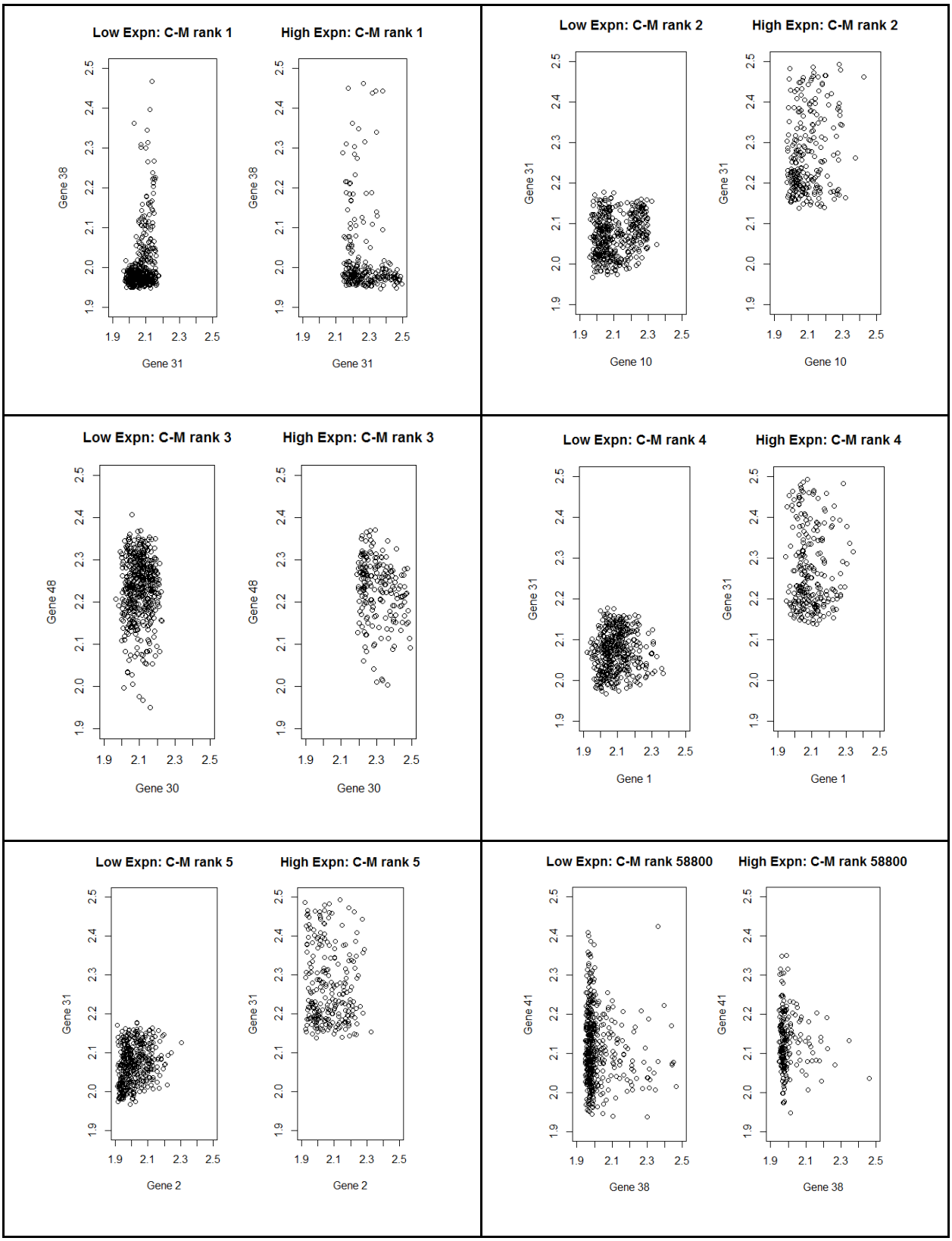

Fig 4.1.3. Gene A-Gene B interaction for C-M p-vals ranks 1:5, & 58800 for comparison

In the graphs, Gene pairs ranked on the top-5 have different interaction graphs for low and high expression of gene C, whereas, the gene pair ranked 58800 shows no difference.

4.2 Discussions

The Data simulation process involved generating separate distributions to simulate low/high distributions of different models and carrying out the three different tests. From the results listed in Table 4.1.2., using critical value D_{α}=0.19 and significance level α=0.05, the high K-S statistic and low C-M p-values lead us to reject the null hypothesis and conclude that the low/high expressions of each models have different distributions. However, it is only in the case of Fisher's z-transformation test that a high p-value is observed and there is strong evidence to support the null hypothesis. Therefore, this suggest that Fisher's z-transformation test only targets linear associations and was not effective in detecting the non-linear associations in the simulated models and it might be appropriate to perform a different hypothesis test.

This brings us to the actual data analysis part. For each of the triplets, we adjusted the two-way interactions between gene C-gene A and gene C-gene B using linear regression model. The top-5 most significant gene triplets for the Kolmogorov-Smirnov tests were mentioned in Table 4.1.3. We considered the 50 genes as controller genes one at a time and carried out the K-S test a total of 58800 times which is the number of all possible three-way gene interactions for 50 genes. Once the 58,800 K-S statistics were computed, we saved and sorted them as mentioned in 4.1 Results. K-S statistics is simply defined as the supremum difference between the empirical distributions of the sample distributions with low and high expressions of gene C. A high K-S statistic supports the null hypothesis whereas, a low K-S statistics is against it. Therefore, we sorted the K-S statistics in descending order and this arranges the top-5 gene triplets to be the

most significant ones. As per the K-S test results, gene 31 as a controller genes presents the most significant gene A- gene B interactions. The graph for the various ranks of K-S statistics were provided in Figure 4.1.2. The graphs serve as a visual representation that if a controller gene C were to affect the gene A- gene B interaction, then the graphs of sample distributions are different in case of low and high expression gene C. However, if the controller gene C had no effect in the pairwise gene correlation and the K-S statistic were very small, then the graphs are very similar if not exact.

Table 4.1.4 presents the top-5 most significant three-way Gene interactions based on the Cross-Match test. Just like in the case of the K-S test, 50 genes were utilized instead of 1000 in order to minimize time complexity. Also, for each of the triplets, the two-way interactions between gene C-gene A and gene C-gene B were adjusted using linear regression model. We used 'crossmatch' in R to compute 58,800 C-M p-value and sorted them in ascending order this time. The test statistic of a C-M test is denoted by A and mentioned in [15], and is simply the total count of disjoint sample observations during non-bipartite pairing. A low p-value and high A statistic suggests a specific controller Gene C encourages pairwise interaction among other two genes whereas, a high p-value/ low A statistic indicates that presence of high expression of Gene C makes no change in gene-gene interaction. Based on the top-5 significant Gene triplets, Gene 30 seems to be the most effective controller Gene and Gene 31 makes a close second. Figure 4.1.3 shows graphs for the Gene A- Gene B interaction for the significant triplets. We also included a rank 58800 Gene triplet so as to show how the low p-value is an indicator that low or high Expression of Gene C causes no effect to the pairwise interaction of the other two genes.

5. CONCLUSIONS

5.1 Summary

In data simulation process, three different models were simulated in order to carry out the Fisher's z-transformation, Kolmogorov-Smirnov and Cross-Match tests. The simulated sets of models were logarithmic, bivariate normal and polynomial models with low and high expression variations. We obtained high Fisher's p-values, high K-S statistics, and low C-M p-values for all of the simulated models. This led us to fail to reject the null hypothesis for the Fisher's test and reject the null hypotheses in case of K-S and C-M tests. Thus, demonstrating that K-S and C-M are more relevant tests for detection of three-way gene interactions.

From the data obtained from Zhang et al. [6], 50 out of 1000 genes were utilized for data analysis in order to decrease the processing time from several years to a few days. The number of gene triplets assessed also decreased from approximately 4.985 billion to 58,800. For each of the controller gene C, we generated cutoffs by clustering them into low and high expression groups. After that, for each controller gene C, the K-S and C-M tests were executed within the two expression groups. In case of K-S test, we generated and saved the statistics, whereas, for C-M test p-values were recorded. The 58800 K-S statistics were sorted in descending and the C-M p-values were arranged in ascending order. These sorted lists gave us the top five significant gene-interaction triplet for both the tests (Table 4.1.3, Table 4.1.4). We followed it by generating the low/high expression graphs (Fig 4.1.2, Fig 4.1.3) for the top 5 most significant gene triplets and comparing it to the least significant gene triplet at rank 58800. From the graphs, it can be perceived that high expression of gene C in top five cases affect the pairwise gene interactions between gene A and gene B.

Therefore, Zhang et al.'s approach only targets the linear association of gene A and gene B. However, with K-S and C-M tests, we will be able to assess gene- interactions more efficiently.

5.2 Future Work

In our research, we only included 50 genes instead of 1000 provided in Zhang et al. [6]. We only took part of the sample so as to decrease the total possible gene triplet combinations we needed to test. Decreasing the number of genes decreased the total number of combinations from 498501000 to 58800, which is almost 99.9% reduction in the total number of gene triplets. The reason to do so was mostly to bring down the time complexity. The time to carry out the K-S as well as C-M tests shows an exponential growth. If we were to execute the tests in a personal computer, the processing time for the test could take several years. Therefore, we chose only 50 genes in our data analysis which conveniently took only about a few days for the K-S test and few hours for the C-M tests.

The study would provide the best results if we were able to include all 1000 genes. Since, we are processing a high volume data, it would be more efficient to utilize 'parallel computing' in our study. Parallel Computing is a computational method in which a number of processes can be executed simultaneously. According to Eugster et al. [23], the statistical programming language R provided parallel computing within computer and also in multicore systems using different packages. A few of the available packages that facilitate parallel computing in R language are multicore, snow, snowfall, and nws.

Parallel computing is a high-performance computing which is very useful in processing high-volume datasets like genomic data, and complex methodologies like bootstrapping. In [23],

the four packages Eugster et al. mentioned, present three different kinds of parallel computing scenarios provided below:

1. Multi-core environment: combines two or more CPUs in one machine
2. Cluster environment: connects a set of computers
3. Cluster environment with huge amounts of data for calculation

Eugster et al [23] also mentioned what scenario the different packages are used for. The 'multicore' package is used for multi-core environment, the 'snow' or enhanced 'snowfall' packages are used in case of computer cluster and computer intensive calculations, and 'nws' is utilized if there is a huge amount of data to be processed at each computer.

The large volume of our data might ask for the usage of the 'nws' package. The nws package uses the "NetWorkSpaces" server (NWS) [24]. The package acts as a client for the NWS technology. Both the package and NWS server are open source, commercial product from REvolution Computing. In order to run the NWS server application, several other software components would have to be installed. However, efficiency of a computation method often comes with the price of code-clarity. Further work needs to be done on how to apply the parallel computing method in our dataset so that we can include all 1000 genes and compare the results to the ones obtained in Zhang et al. [6].

References

1. An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature, 489*(7414), 57-74. Retrieved from http://0-search.proquest.com.library.uark.edu/docview/1069238769?accountid=8361

2. Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson L.S. Tang, and Weichuan Yu. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. The American Journal of Human Genetics, Vol 87, 325-340.

3. Simon Tripp and Martin Gruwber. (2011). Economic Impact of the Human Genome Project. Executive Summary, 1.

4. Steen, K. V. (2011). Travelling the world of gene-gene interactions. Briefings of Bioinformatics, Vol 13 NO 1, 1-19.

5. Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews.Genetics, 10*(6), 392-404.

6. Jiexin Zhang, Yuan Ji and Li Zhang. (2007). Extracting three-way gene interactions from microarray data. Bioinformatics, Vol 23, 2903-2909.

7. Farley,C. and Raftery,A. (1999). MCLUST: Software for Model-Based Cluster Analysis. Journal of Classification, Vol 16 Issue 2, 297-306.

8. Robert, W. E. (2015). Causation and pearson's correlation coefficient. *Journal of Visual Impairment & Blindness (Online), 109*(3), 242. Retrieved from http://0-search.proquest.com.library.uark.edu/docview/1687501470?accountid=8361

9. Fisher,L.D. and Belle,G.V. (1993). Biostatistics. John Wiley & sons Inc, New York, NY.

10. Lopes*, R. H., Reid, I., & Hobson, P. R. (2007, April 23-27). *The two-dimensional Kolmogorov-Smirnov test*. Speech presented at XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research in The Netherlands, Amsterdam. *Speaker

11. Stephens, M. A. (1970, January 15). Kolmogorov-Type Tests for Exponentiality when the scale parameter is unknown. Department of Statistics, Stanford University, CA.

12. Feigelson, E., & Babu, G. J. Beware the Kolmogorov-Smirnov test!. Center for Astrostatistics. Penn State University.

13. Evans, D.L., Drew, J.H., & Leemis, L.M (2008). The Distribution of the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson Darling Test Statistics for Exponential Populations with Estimated Parameters. Communications in Statistics- Simulation and Computation. 1532-4141.

14. Peacock, J.A.(1983). Two-dimensional goodness-of-fit testing in astronomy. Monthly Notices Royal Astronomy Society. Vol 202. 615-627.

15. Heller, R., Jensen, S.T., Rosenbaum, P.R., & Small, D.S. (2010). Sensitivity Analysis for the Cross-Match Test, With Applications in Genomics. Journal of the American Statistical Association. Vol 105. No. 491.

16. Rosenbaum, P.R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of Royal Statistical Society, Vol 67. 515-530.

17. Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal Non-bipartite Matching and Its Statistical Applications. Am Stat. 65(1). 21-30.

18. Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. Journal of the American Statistical Association. Vol 103. 1334-1344.

19. Cochran, W.G. & Chambers, S.P. (1965). The Planning of Observational Studies of Human Populations. Journal of Royal Statistical Society, Ser A. 128:234–266.

20. Papadimitriou, C.H., & Steiglitz, K. (1998). Combinatorial Optimization: Algorithms and Complexity. New York: Dover.

21. Hansen, B.B., & Klopfer, S.O. (2006). Optimal Full Matching and Related Designs via Network Flows. Journal of Computational and Graphical Statistics. Vol 15. 609–627.

22. Hansen, K.D., & Irizarry, R.A. (2012). Removing Technical Variability in RNA-seq data using conditional quantile normalization. Biostatistics. Vol 13(2). 204-216.

23. Eugster, M.J.A.,Knaus, J., Porzelius, C.,Schmidberger, M., & Vicedo, E. (2011). Hands-on tutorial for parallel computing with R. Computational Statistics. 26.2, 219-239.

24. Bjornson R., Carriero N., & Weston S. (2007). Python NetWorkSpaces and parallel programs. Dr Dobb's Journal. 1–7. http://www.ddj.com/web-development/200001971