STATISTICAL METHODS FOR ANALYZING GENOMICS DATA

By

SINJINI SIKDAR

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2017

Dedicated to my husband, Sandipan, who has been a constant source of support and encouragement during the challenges of my PhD life.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor Professor Susmita Datta for her unyielding support and invaluable guidance throughout my graduate work. It would not have been possible for me to reach this point without her support. I also want to express my sincere thanks to Professor Somnath Datta as I learnt a lot through my interactions with him and benefited from his expert suggestions. I am also thankful to my other dissertation committee members Professor Fei Zou, and Professor Lauren McIntyre for all their kind support and constructive comments regarding my dissertation. I want to thank Professor Ryan Gill from University of Louisville for his helpful contributions to my research works.

I also want to thank all the faculty, staff, and students of the Department of Biostatistics at University of Florida as well as all the faculty and students of the Department of Bioinformatics and Biostatistics at University of Louisville who have helped me reach this point.

I would like to thank my sister Shreejata Sikdar for her invaluable friendship and constant support throughout my PhD life. Finally, I want to thank my parents for their immense encouragement that have helped me in moving forward in my academic life.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

STATISTICAL METHODS FOR ANALYZING GENOMICS DATA

By

Sinjini Sikdar

August 2017

Chair: Susmita Datta
Major: Biostatistics

My doctoral dissertation comprises of three different projects that address three different aspects of statistical analysis of genomics data.

The first project describes a novel statistical approach for identification of 'master regulator' transcription factor in a genome. A 'master regulator' transcription factor, being at the top of the hierarchy of the transcriptomic regulation, may control the regulatory activities of the other transcription factors and the associated genes. For example, in cases of systemic disease, cellular function is disrupted. We hypothesize that these may be the result of a regulatory disruption in the most upstream elements of the regulatory cascade, the 'master regulator'. Therefore, it is important to identify and target the master regulator transcription factor for proper understanding of the associated disease process. Through simulated scenario and real dataset analyses, we show that our method performs well in validating the existence of a master regulator, and identifies biologically meaningful master regulators.

The second project involves an integrated analysis using multiple cancer datasets for investigating the significance of the biological pathways which are interrupted by cancer-associated genetic mutations. This dataset consists of expression profiles for genes/proteins of patients receiving treatment, for three types of cancer - Head and Neck Squamous Cell

Carcinoma, Lung Adenocarcinoma and Kidney Renal Clear Cell Carcinoma. We consider pathway analysis to identify all the biological pathways which are active among the patients and investigate the roles of the significant pathways using a differential network analysis of the protein expression datasets for the three cancers separately. We then integrate the pathway based results of all the three cancers which provide a more comprehensive picture of the three cancers.

In the third project, we develop a novel meta-analysis method of combining p-values from different independent experiments involving large-scale multiple testing frameworks. Adhering to the regular statistical assumptions regarding the null distributions of test statistics can lead to incorrect significance testing results and biased inference in large-scale multiple testing frameworks when results from different independent genomic experiments are combined. In order to overcome this, our proposed method takes into account empirical adjustments of the individual test statistics and p-values. Consequently, our method outperforms the standard meta-analysis approach of significance testing as shown in simulation studies and real genomic data analyses.

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEWS

This dissertation is based on three different projects that are linked through the analysis of genomic datasets. The first project involves a novel two step statistical approach for identification of a 'master regulator' transcription factor that controls most of the regulatory activities of the other transcription factors and the associated genes. In the second project, we explore the importance of cancer pathways through the integrative analyses of differential protein expression networks in three different cancers. A third project is also proposed that aims to develop a meta-analysis approach for large-scale multiple hypothesis testing framework in genomic datasets.

1.1 Identification of the Master Regulator Transcription Factor

Through several scientific findings, it has been suggested that cancer is mainly caused by the mutations in certain genes [2-4]. The genes, which can transform a cell into tumor cell, are termed as 'oncogenes'. So, for effective understanding of cancer, identification of these mutated genes (oncogenes) is essential. Genomic studies of different cancer datasets leads to identification of several potential oncogenes which are directly or indirectly responsible for development and progression of cancer [5-7]. With the advent of modern technology, high-throughput data analysis typically yields a list of differentially expressed genes or proteins that may have potential to play important roles in a given phenomenon or phenotype. However, It is a very challenging task to target and individually study all of these genes as they are large in number [8-10]. One way to identify patterns is to group the proteins and genes belonging to the same pathway [8, 10]. These genes and their corresponding pathways form networks that are

Reprinted with permission from Sikdar S, Datta S. A novel statistical approach for identification of the master regulator transcription factor. BMC Bioinformatics. 2017; 18: 79.

hypothesized to control various cellular functions, and there has been sufficient interest in analyzing such pathway based networks [11-14]. However recent findings suggest that oncogenes and tumor suppressor genes may encode “transcription factors”, deregulations of which play key roles in carcinogenesis [15-18]. Cancer signaling pathways seem to identify transcription factors [15, 19], which lead to tumor development, progression and cancer metastasis through the controlling of the gene expression patterns [15, 16]. As suggested by [15, 17, 20-22], some of the major groups of transcription factors, which have been identified for cancer, are the steroid receptors (e.g. estrogen receptors in breast cancer, androgen receptors in prostate cancer), resident nuclear proteins activated by kinase cascades, and the STAT protein family members. Apart from these, the ETS protein family members have also been identified as potential cancer transcription factors for their emerging roles in human cancer [23-25]. It has been shown that direct suppression of transcription factors expressions, which serve as the convergence points of oncogenic signaling and are functionally altered in many cancers, can lead to significant antitumor responses with minimal side-effects, and targeting these transcription factors in tumor-related immune cells can help in recovering from tumor immunoresistance [26]. As a result of these features of the transcription factors, in addition to the facts that they are much smaller in number than the oncogenes and have well-regulated expression and activities, transcription factors are gaining popularity as potential therapeutic targets in anti-cancer drug development [16, 17, 27, 28].

In the recent past, studies have identified a transcription factor or a group of transcription factors as the driving force behind the development of a biological or disease process [29-35]. In order to facilitate such detection there have been attempts to develop statistical methods for accurate identification of transcription factors that regulate large number of genes [36-42]. To

this end methods have been attempted for identification of transcription factors and transcription factor binding sites in cell cycle of yeast and similar organisms using multiple data sources [36-41], and to human cell culture [42]. In addition, there have been efforts in developing statistical tools for identifying a cluster of transcription factors that cooperatively regulates a large number of genes and the associated disease process [43, 44]. Methods have also been developed for identifying differentially regulated gene sets by integrating regulatory networks of transcription factors and gene expression data [45]. Also, transcription factor activities have been estimated through their effect on target genes [46]. The importance of transcription factor regulation is also evident from methods that have been developed for identifying coordinately activated functional modules from gene expression data [47, 48]. These methods assume that the transcription factor regulated target genes are differentially expressed from non-target genes in the same functional module [47, 48]. In fact there have been several studies for identifying transcription factors under the assumption that co-expression indicates co-regulation [49-51]. The main idea behind such transcription factor regulation is that genes regulated by such transcription factors should have, on an average, significantly different expression levels during one or more cell cycle phases [52]. Besides, there have been studies for identifying groups of important transcription factors through integration of different genomic and epigenomic features [53] and integration of transcriptional and protein interaction networks [54]. Most of these recent methods, including that of [55] and [56], have been directed towards identification of a group of candidate driver transcription factors. Despite the fact that in most cases there are a group of transcription factors that regulate the oncogenes and hence the disease process, it has been seen that there is a hierarchical structure in the regulatory activities of these transcription factors where a ‘master regulator’ transcription factor often appears to control most of the

regulatory activities of the other transcription factors and the associated genes [57-59]. According to the definition provided in [58] the "master regulator" transcription factor is at the top of a regulatory hierarchy and must not be under the regulatory influence of any other gene or transcription factor. We use this definition and attempt to finding the "master regulator" transcription factor. This master regulator transcription factor can be targeted for proper understanding of the associated disease process and can be used as a biomarker.

In current literature there is a lack of appropriate statistical methods which use a single data source for accurate identification of such a master regulator among a set of identified transcription factors. Motivated by this, we develop a novel two-step statistical approach to test for the existence of a master regulator transcription factor and for subsequent identification of the master regulator, if it exists, from gene expression data alone.

1.2 Meta-Analysis of Differential Protein Expression Networks of Cancer Signaling Pathways in Three Different Cancers

Several studies have found that there are approximately 140 "driver" genes which can promote the formation of tumors if affected by intragenic mutations. These "driver genes" are known to be directly or indirectly responsible for causing selective growth advantage. These "driver genes" are classified into twelve signaling pathways. According to [60], this selective growth advantage can occur only through these twelve signaling pathways. We refer these twelve signaling pathways as "target pathways". Also the "target pathways" regulate three core cellular processes - "cell fate", "cell survival" and "genome maintenance" [60]. We believe understanding the roles of these twelve pathways can result into novel therapeutic intervention strategy. In this project, we undertake a novel investigation of the roles of these "target pathways" using a differential network analysis of the protein expression datasets on three cancers namely, Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma

(LUAD) and Kidney Renal Clear Cell Carcinoma (KIRC). We find the study of protein expressions is justified as the mutations of the previously mentioned "driver genes" alter the protein products. These datasets are available to us from International Cancer Genomic Consortium (ICGC) as part of the CAMDA 2015 challenge data. We pursue an integrative analysis of protein expressions of all these three cancer datasets to investigate whether each of these "target pathways" plays a significant role in these three cancers. For example, we determine whether the proteins associated with these pathways interact differently between the two clinical groups ("progression" or "complete remission") of patients. The differentially expressed pathways between the two disease groups allow us to gain more insights about the functional working mechanism of the cells than just the individual differentially expressed genes/proteins [61]. The process begins with grouping of the proteins according to their biological pathways. Then, we examine whether the network structures of all the proteins in these "target pathways" have changed from the "complete remission" group of patients to the "progression" group. We also examine whether the connectivity of each single protein in the networks of proteins in these "target pathways" has changed between the two groups. Then we rank the "target pathways" as well as the constituent proteins to get overall ranked lists which would then rank the pathways by their global order of importance with respect to all the three cancers. This ranking may shed light to the regulatory roles of individual proteins in the context of all others in the pathway.

1.3 Meta-analysis Approach for Large-scale Simultaneous Hypothesis Testing in Genomic Experiments

In genomic experiments and association studies, meta-analysis is a popular tool for pooling results from multiple experiments and studies in order to reach an overall decision. In recent times, rapid progress in technology has led to major development of high throughput

genomic assays. This means that hundreds and thousands of genes are now being analyzed at the same time. As a result, the level of simultaneous inference has undergone a huge surge over the last decade. Development of novel meta-analysis approaches is crucial in order for such settings since the sample size of individual experiments are generally small compared to the number of genes leading to low power of statistical detection from them after adjusting for multiplicities. However, there has not been much change in the meta-analysis methods to accommodate this large-scale aspect of the underlying inference and the possibility of underlying hidden factors that act as confounders. For example, in testing for the significance of genes in disease studies, more or less the same meta-analysis methods are being applied to experiments involving hundreds and thousands of genes as were initially developed for experiments involving a few number of candidate genes. One common practice is to use Fisher's method [62] for combining the p-values from different testing problems involving the same overall hypothesis. The individual p-values are calculated for one gene at a time and the negative log-transformed p-values are combined to form a chi-squared test statistic under the assumption that they are individually uniformly distributed under the null hypothesis. However, as pointed out by Efron [63], in large-scale multiple testing problems, the "marginal" or "empirical" distribution of the p-values may not be uniformly distributed. Consequently, the distributional assumption of Fisher's combined test statistic becomes questionable. In order to draw "better" inference, adjustments are needed to ensure that all the p-values from individual experiments are uniformly distributed so that the Fisher's method of combining the individual p-values can be properly implemented.

In single hypothesis test framework the main aim is to reject the null hypothesis in favor of some "interesting" alternative hypothesis with high power, say 80%. But in large-scale multiple testing framework involving, say, 10,000 hypotheses related to 10,000 genes, rejection

of 80% of 10,000, i.e. 8,000, null hypotheses is no longer a desired outcome. Rather, the aim of such large-scale testing framework is to identify a small set of "interesting" cases or genes, usually less than 10%, which can be pursued for further investigation. The advantage of having large number, e.g., 100 or more, of hypotheses over a single hypothesis is that it enables the estimation of empirical null distribution avoiding the dependency on theoretical asymptotic null distribution, as pointed out by Efron [63]. The use of this empirical null is more appropriate for addressing the goal of large-scale hypotheses testing problems. This is particularly relevant in large observational studies which are often characterized by the presence of unobserved variable effects (e.g., batch effects) or unmeasured/missed confounding factors. Unlike the theoretical null, the empirical null distribution, automatically, takes into account the effects of the additional variation (and also small to moderate biases). This has an even more serious consequence in meta-analyses of large scale genomic experiments where a number of potentially low-powered study results are combined in order to achieve significance. However, a p-value combination method such as Fisher's method [62] may lead to incorrect findings if at least one of such studies contains hidden sources of variation leading to a violation of the theoretical null distributional assumption for that component study.

In particular, a possible consequence of combining unadjusted or incorrectly adjusted p-values through Fisher's method [62] in a large-scale multiple testing situations is the prevalence of large number of false discoveries when some underlying hidden variable plays the role of a confounder. Occurrences of false discoveries are common in large scale DNA microarray experiments where the aim is to detect genes that are differentially expressed between two or more biological conditions. A good example in this context can be the well-known study of [64] which aimed to identify the differentially expressed genes between two types of genetic mutation

of breast cancer, namely, “BRCA1” and “BRCA2” mutations. In this study 3266 genes were analyzed, out of which 51 genes turned out to be significant initially at the p-value cut-off of 0.001. Later on it was shown that the chosen cut-off is expected to produce substantial false positives, and the authors had to lower the cut-off resulting in the lower number of significant genes. Details of this study results can be found in [65]. This highlights the possibility of inaccurate scientific conclusions due to the occurrences of false discoveries in large-scale experiments. A possible remedy to this problem is to make adjustments using the false discovery rate (FDR), developed by [66], which have been widely used in methods for analyzing data from genomic experiments [67-70]. However, using the aforementioned breast cancer microarray data, Efron [63] showed that even adjusting for FDR may not be enough to restrict false discoveries if the underlying assumption of the standard normal distribution of the test statistic is under suspicion. Instead, he advocated using the “empirical null” in order to make the calls while using a local false discovery rate calculation. Motivated by this, we develop a meta-analysis method called Empirically Adjusted Meta-analysis (EAMA) that does not combine the raw p-values – they are first transformed, where the amount of transformation depends on the discrepancy between the empirical and the theoretical null (e.g., uniform distribution in case of p-values), before they are Fisher-combined. Of course, a multiple hypothesis method such as Benjamini–Hochberg (1995) [66] is applied at the end to make the significance calls. We show that this procedure is very effective in reducing the FDR and increasing specificity at the expense of a slight reduction in sensitivity in a variety of situations which are affected by the presence of some hidden confounders.

CHAPTER 2
A NOVEL STATISTICAL APROACH FOR IDENTIFICATION OF THE MASTER REGULATOR TRANSCRIPTION FACTOR

2.1 Methods

In this chapter, we present our proposed statistical approach for identification of master regulator transcription factor. We first discuss the biological considerations that motivated the development of our method. We then provide the steps to formulate our test statistic and use it for the identification of the regulatory circuit of the transcription factors and genes. Finally, we identify the master transcription regulator at the top of the regulation hierarchy.

2.1.1 Biological Considerations

Important biological processes can have multiple layers of regulation and control. A transcription factor is known to control not only genes but also other transcription factors. As discussed before in section 1.1, usually there is a hierarchical structure in the regulation of the transcription factors so that the master regulator controls most of the regulatory activities of the other transcription factors and the associated genes. In this article, we aim to identify the master regulator transcription factor which is at the top of the hierarchy for better understanding of the associated disease process. A toy example is shown in Figure 2-1 which shows the possible regulatory network across a set of genes and transcription factors in a genome.

In Figure 2-1, TF_1, TF_2, TF_3 and TF_4 denote the transcription factors and $g_1, g_2, \dots, g_{12}$ denote the set of genes. Suppose TF_1 directly regulates five of the genes, which are g_1, g_2, g_3, g_4 and g_5, and also all the other three transcription factors, TF_2, TF_3 and TF_4. The transcription factor TF_2 regulates the genes g_6, g_7 and g_8. Similarly, the transcription factor TF_3 regulates the genes g_9 and g_{10} and finally, the transcription factor TF_4 regulates the

Reprinted with permission from Sikdar S, Datta S. A novel statistical approach for identification of the master regulator transcription factor. BMC Bioinformatics. 2017; 18: 79.

genes g_{11} and g_{12}. In this example, there exists a hierarchical structure with three layers. We have TF_1 at the top of the hierarchical structure as it directly or indirectly regulates the other transcription factors and the genes. So, TF_1 is considered to be the first layer of the hierarchy. Now, TF_1 directly regulates the other transcription factors, TF_2, TF_3 and TF_4 and the genes $g_1, g_2, \ldots, g_5$. So, TF_2, TF_3 and TF_4 and $g_1, g_2, \ldots, g_5$ are considered to be at the second layer of the hierarchy. TF_1 regulates the genes $g_6, g_7, \ldots, g_{12}$ indirectly through the transcription factors TF_2, TF_3 and TF_4. Thus, the genes $g_6, g_7, \ldots, g_{12}$ form the third layer of the hierarchy. In this example, TF_1 directly or indirectly regulates all the layers of the hierarchy and is not under the regulatory influence of any other gene or transcription factor. Therefore, according to the definition, TF_1 can be considered as the master regulator transcription factor.

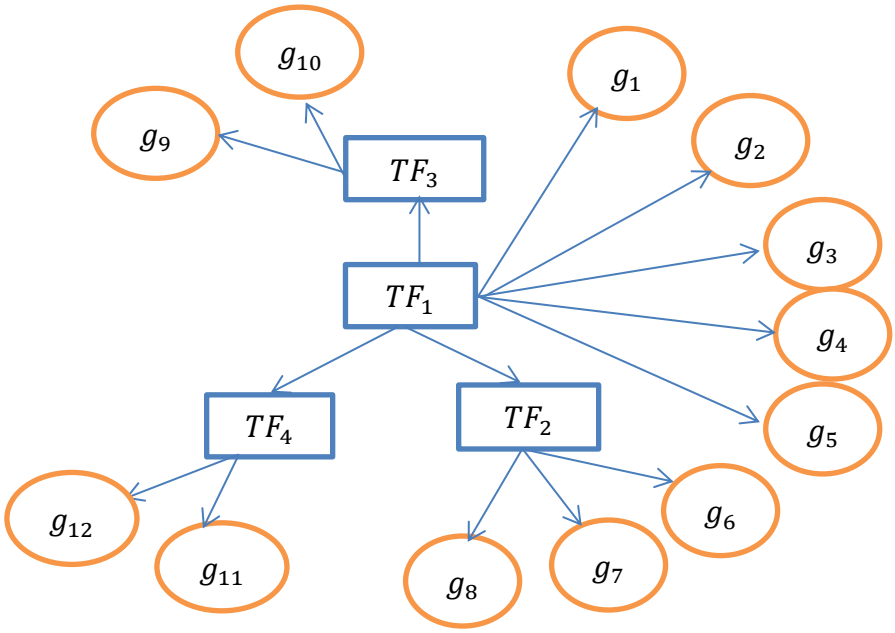

Figure 2-1. A toy example showing possible regulatory network across a set of genes and transcription factors.

Here, in this project, we attempt to develop a test that can check if there exists any transcription factor that acts as a master regulator in a genome, and identify such a master regulator if present. The details of our proposed method are given in the next section.

2.1.2 Identification of the Master Regulator through a Hypothesis Testing Framework

Let M denote the total number of transcription factors present in a genome. Let the transcription factors be denoted by $TF_1, TF_2, \ldots, TF_M$. Let the genes, which are not transcription factors, be denoted by $g_1, g_2, \ldots, g_N$, where N denotes the total number of such genes. So, in total, we have expression data on $M + N$ genes. Let us assume that there are two groups of subjects, for example, the case group (the disease group) and the control group (non-disease group). Let there be r_1 subjects in the case group and r_2 subjects in the control group. So, in other words, we have two groups of subjects with expression levels for $M + N$ features in each group. It is well known that the genes including the transcription factors are expressed differently in the two groups. Additionally, they are connected with one another differently in the networks of the two groups. There are methods to reverse engineer the networks of genes with association measures such as correlations, partial correlations and partial least squares regression scores [71, 72]. These inter genomic connectivity are different in two groups and can be detected by statistical methods such as Differential Network Analysis (DNA) [73].

Since, it is believed that the master regulator maximally controls the other transcription factors as well as the associated genes; it is important to find the regulatory network among the transcription factors and also the degree of regulation of all the transcription factors on the genes. We first measure the degree of regulation of the transcription factors on the genes. The degree of regulation of a transcription factor on the genes is measured by the change in connectivity of the genes it regulates in the two networks. In other words, we find how the connectivity of a

transcription factor with the genes differs between the subjects in case and control groups. For this, we estimate the change in connectivity of a transcription factor with the genes in the two groups of samples using connectivity scores of the given transcription factor with all the genes in the case group with that in the control group. The difference in connectivity is measured using the following statistic [73]:

$$x_i = \frac{1}{N}\sum_{g\in\mathcal{G}} \left| s^{case}_{TF_i,g} - s^{ctl}_{TF_i,g} \right| \quad ; \quad i = 1, 2, \dots, M \tag{2-1}$$

where $\mathcal{G}$ denotes the set of all the genes, i.e., the cardinality of $\mathcal{G}$ is N. Here, $s^{case}_{TF_i,g}$ and $s^{ctl}_{TF_i,g}$ are the connectivity scores between the transcription factor $TF_i,\ i = 1,2, \dots, M,$ and the gene g in the case and control groups, respectively. There are several choices for connectivity scores such as Pearson's correlation scores, partial correlation scores, partial least square based association scores. In this project, we use the Pearson's correlation scores as connectivity scores. So, here, the $x_i's$ give us an idea about the magnitude of the differential regulation of the transcription factors on the genes between the case and control groups.

Next, we find the regulatory structure among the transcription factors. For this, we measure the association between each pair of transcription factors using the Pearson's correlation coefficient scores between them. For each pair (j, k), let y_{jk} denote the absolute value of the Pearson's correlation coefficient score between the transcription factors TF_j and TF_k ; $j, k = 1, 2, \dots, M$, where $y_{jj} = 1$. Note that this calculation is done by pooling the data from both the groups.

At this stage, for a transcription factor TF_j , $j = 1, 2, \dots, M$, we have two measures: a measure of the differential regulation of TF_j on the genes (given by x_j); and a measure of the association of TF_j with all the transcription factors ($y_{jk},\ k = 1, 2, \dots, M$).

We argue that the degree of change in connectivity of the genes in the two networks is controlled by the transcription factors which are correlated amongst themselves in a hierarchical manner. That is, the hierarchical regulation pattern (as measured by the rank order) among the M transcription factors is the same with the differential connectivity of genes in the two groups that they control. In other words, the rank order of the amount of differential connectivity of a transcription factor with other genes it controls is in line (e.g., *concordant*) with its ordered connectivity with the master regulator. Therefore, we consider two ranked lists. One that ranks the transcription factors by the amount of differential connectivity of the genes it controls and another that puts the master regulator in the first position and ranks the remaining transcription factors by their correlation with the master regulator. We evaluate the concordance of these two sets of ranks using a statistical measure which is described in next paragraph. Since we do not know a priori the identity of the master regulator/s, we maximize this measure of concordance over the set of all transcription factors in candidacy for playing the role of the master regulator. In case the maximal *concordance* is statistically significant, we conclude that there is a master regulator. In addition, we declare the transcription factor for which this concordance measure is maximal amongst all transcription factors to be the master regulator.

We construct a concordance statistic K_j for each transcription factor TF_j that is in candidacy for the master regulator; $1 \leq j \leq M$, in the following way:

Step 1. We calculate the Kendall's rank correlation coefficient test statistic given by (2-2) below based on the pairs of data $(x_1, y_{j1}), (x_2, y_{j2}), \ldots, (x_j, y_{jj}), \ldots, (x_M, y_{jM})$. Note that x_i denotes the average difference in connectivity of transcription factor TF_i between the two groups, and y_{ji} is the absolute correlation between the transcription factors TF_i and TF_j. This test statistic K_j below conveys whether the differential connectivity of the genes with the

transcription factor TF_j in the two experimental groups is concordant with the correlations of the transcription factor TF_j with all other transcription factors. In other words, K_j measures whether the differential connectivity is concordant with the hierarchical regulation of the transcription factors amongst themselves. The Kendall's rank correlation coefficient test statistic for the transcription factor TF_j is given as:

$$K_j = \frac{n_{c,j} - n_{d,j}}{n_0} \tag{2-2}$$

where, $n_{c,j} =$ number of concordant pairs in the above paired list,

$n_{d,j} =$ number of discordant pairs in the above paired list,

$n_0 = \frac{M(M-1)}{2}$ = Total number of such paired observations for TF_j.

This statistic can be used to test the null hypothesis that the two sets of ranks produced by differential connectivity x and the correlations with TF_j are non-concordant versus the alternative hypothesis that they are concordant.

Step 2. We repeat step 1 for all such transcription factors, so that we have a concordance test statistic for each of the transcription factors which is a potential master regulator.

We believe that the master regulator has the maximum measure of concordance, among all the transcription factors. Since we do not know the identity of the master regulator, we maximize the measure of concordance, given by K_j, over the set of all transcription factors. So, we define K as the maximum of the statistics given in (2-2) over all the transcription factors $TF_j s$ that are in candidacy for the role of the master regulator, i.e.

$$K = \max_j K_j \tag{2-3}$$

Thus, statistically significant large values of K would indicate the existence of a master regulator.

Significance of K can be assessed by a bootstrap (resampling) based procedure as the sampling distribution of K is not tractable. This will calculate the p-value or the observed level of significance of the value of test statistic K calculated in (2-3). We draw B bootstrap samples from the original sample each of size $r_1 + r_2$ and consider the first r_1 samples as the case group and the remaining r_2 samples as the control group. We compute the test statistic value for each bootstrap sample. Let K_b denotes the value of our test statistic for the b^{th} bootstrap sample, where $1 \leq b \leq B$. In order to estimate the p-value, we calculate the proportion of times the test statistic values based on the bootstrap samples exceed the test statistic value obtained from the original sample, i.e.,

$$\text{p} - \text{value} = \frac{\sum_{b=1}^{B} I(K_b > K)}{B} \tag{2-4}$$

If the p-value obtained from (2-4) is low then the test is significant and we conclude that there exists a master regulator in the system.

In case we conclude that there exists a master regulator the transcription factor T is claimed to be the master regulator if it has the maximum value of the statistic given in (2-2), i.e.

$$T = arg\max_j K_j \tag{2-5}$$

We evaluate the performance of our master regulator identification procedure using a simulation experiment in the next section.

2.2 Simulations

In order to evaluate the performance of our proposed method, we generate synthetic datasets of gene expressions of the case and control groups with the different regulation schemes of the transcription factors. The simulation scheme consists of the following steps:

2.2.1 Data Generation

We consider M transcription factors $TF_1, TF_2, \dots, TF_M$ and N genes $g_1, g_2, \dots, g_N$, as described before in the Methods section. Also, let there be r_1 subjects in the case group and r_2 subjects in the control group. The gene expression data for the two groups of subjects are generated as given below. Note that, the choices of all the design parameters considered below are given in later sections depending on whether we are simulating under the null or under the alternative.

Step 1. We assume that (log-transformed) expression values for TF_1 follows a normal distribution with mean μ and variance 1 i.e. $N(\mu, 1)$ in the case group, and $N(\vartheta, 1)$ in the control group.

We also generate M independent random variables V_i from $N(0,1); i = 1, 2, \dots, M$, that are also independent of TF_1.

Step 2. We want to generate all the transcription factors in such a way that there exists a hierarchical regulatory pattern among them. In other words, we want to generate the remaining $M-1$ transcription factors in such a way that $Corr(TF_j, TF_k) > Corr(TF_j, TF_l)$ $(j = 1, 2, \dots, M;\ k, l = j+1, \dots, M;\ k < l)$, where $Corr(TF_j, TF_k)$ denotes the correlation between the transcription factors TF_j and TF_k. One way of achieving this is to simulate the remaining $M-1$ transcription factors TF_i ; $i \neq 1$ as follows:

$$TF_i = \frac{\rho_i TF_1 + V_i}{\sqrt{1+\rho_i^2}} \quad ; \qquad i \neq 1 \tag{2-6}$$

where, $\rho_i's$ are decreasing in i, $i \neq 1$.

In this case, the correlation structures among all the transcription factors are given by:

$$Corr(\,TF_1\,,\,TF_i\,) = \frac{\rho_i}{\sqrt{1+\rho_i^2}}\;;\;\; i \neq 1$$

and $\;Corr(\,TF_j\,,\,TF_k\,) = \frac{\rho_j \rho_k}{\sqrt{1+\rho_j^2}\sqrt{1+\rho_k^2}}\;\;;\;\; j,k \neq 1\,; j \neq k.$ (2-7)

Step 3. The next step is to generate the genes. We assume that each of the transcription factors $TF_i\,;\; i = 1,2,\dots,M$, regulates m_i genes. Here, $N = m_1 + m_2 + \cdots + m_M$. The genes, $g_1, g_2, \dots, g_{m_1}$, which are directly regulated by TF_1 alone, are generated as given below:

$$g_j = \begin{cases} TF_1\gamma_1 + \epsilon_j & \text{for case group} \\ TF_1\gamma_2 + \epsilon_j' & \text{for control group} \end{cases} \qquad j = 1,2,\dots,m_1 \tag{2-8}$$

where, ϵ_j and ϵ_j' are independent and identically distributed (i.i.d) as $N(0,1)$, and $\gamma_1$and γ_2 are real numbers.

Here, the correlation between the transcription factor TF_1 and the genes $g_k\,, k = 1,2,\dots,m_1$ is given by

$$Corr(\,TF_1\,,\,g_k\,) = \begin{cases} \dfrac{\gamma_1}{\sqrt{1+\gamma_1^2}} & \text{for case group} \\ \dfrac{\gamma_2}{\sqrt{1+\gamma_2^2}} & \text{for control group} \end{cases} \qquad k = 1,2,\dots,m_1 \tag{2-9}$$

The genes, regulated by the remaining $M-1$ transcription factors $TF_i;\; i \neq 1$, are generated as follows:

$$g_j = \begin{cases} V_i r_{1i} + \epsilon_j & \text{for case group} \\ V_i r_{2i} + \epsilon_j' & \text{for control group} \end{cases} \qquad j = m_{i-1}+1,\dots,m_i \tag{2-10}$$

where, ϵ_j and ϵ_j' are i.i.d $N(0,1)$ and r_{1i} and r_{2i} are real numbers, $i \neq 1$.

In this case, the correlation between a transcription factor $TF_i;\; i \neq 1$ and the genes regulated by that transcription factor is given by

$$Corr\left(TF_i, g_k\right) = \begin{cases} \dfrac{r_{1i}}{\sqrt{1+\rho_i^2}\sqrt{1+r_{1i}^2}} & \text{for case group} \\ \dfrac{r_{2i}}{\sqrt{1+\rho_i^2}\sqrt{1+r_{2i}^2}} & \text{for control group} \end{cases} \quad (2\text{-}11)$$

$$k = m_{i-1} + 1, \ldots, m_i$$

Also, the correlations between a transcription factor $TF_i;\ i \neq 1$ and the genes which are not regulated by that transcription factors are zero i.e. $Corr(\ TF_i\ ,\ g_k\ ) = 0$ for $k \neq m_{i-1} + 1, \ldots, m_i$. Furthermore, $Corr(\ TF_1\ ,\ g_k\ ) = 0\ ;\ k \neq 1, 2, \ldots, m_1$.

We calculate the size and power of our test in the following sections.

2.2.2 Size of the Test

Recall that, the null hypothesis of interest is that the rank order of the transcription factors based on their differential connectivity with the genes is not statistically concordant with their rank order based on their correlations with the master regulator. So, the null situation can be created by assuming that there exists a hierarchical regulatory pattern among the transcription factors but there is no differential regulation of the genes in the two experimental groups due to the transcription factors. Hence, there is no such master regulator.

In order to follow the null hypothesis in the simulation setup, we assume $\rho_i s$ to be decreasing in $i\ , i = 2, 3, \ldots, M$ and choose $\gamma_1 = \gamma_2$ and $r_{1i} = r_{2i}, i = 2, 3, \ldots, M$. The decreasing nature of $\rho_i s$ ensures that there exists a hierarchical regulatory pattern among the transcription factors. $\gamma_1 = \gamma_2$ and $r_{1i} = r_{2i}, i = 2, 3, \ldots, M$ ensure that the associations of the transcription factors with the genes remain the same in the two groups i.e. there is no differential connectivities of the transcription factors with the genes between the two groups. We generate r_1 samples for the case group and r_2 samples for the control group using the above described

scheme. We calculate the value of our test statistic, denoted by K, using equation (2-3) and find its p-value as described in the Methods section.

In order to find the size of the test, we use Monte-Carlo method. We repeat the whole process 1000 times and therefore, get 1000 p-values using equation (2-4). Let the p-value for the i^{th} Monte-Carlo iteration be denoted as $p_i, i = 1, 2, \dots, 1000$. The size for the test is given by:

$$\text{Size} = \frac{\sum_{i=1}^{1000} I(p_i < 0.05)}{1000} \tag{2-12}$$

In particular, we consider the following choices of the parameters for calculating the size of the test:

- $M = 10$, $N = 105$, $r_1 = r_2 = 500$, $B = 500$
- $\mu = 50$, $\vartheta = 5$
- $m_1 = 30$, $m_2 = m_3 = \cdots = m_7 = 10$, $m_8 = m_9 = m_{10} = 5$
- $\rho = (\rho_2, \dots, \rho_{10}) = (0.95, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$
- $\gamma_1 = \gamma_2 = 0.5$
- $r_{12} = r_{22} = 0.45$ and $r_{1i} = r_{2i} = r_{1(i-1)} - 0.05$ for $i = 3, \dots, 10$.

For the above choices of the parameters, the empirical size of the test came out to be 0.032 which is close to the nominal size of 0.05.

2.2.3 Power of the Test

To calculate the power of our test, we generate a data under the alternative hypothesis H_1. Here the alternative hypothesis is that the rank order of the transcription factors based on their differential connectivity with the genes is concordant with their rank order based on their correlations with the master regulator. So, we generate the data in such a way that TF_1 acts as the master regulator, that is, the connectivity of TF_1 with other transcription factors are most concordant with the differential connectivity of the genes with the transcription factors. We set $\gamma_1 > \gamma_2$ and $r_{1i} > r_{2i}$; $i = 2, 3, \dots, 10$. Here $\gamma_1 > \gamma_2$ ensures that the connectivity (associations) of TF_1 with the genes, regulated by it, are greater in case group than that in the control group. Similarly, $r_{1i} > r_{2i}$ ensures that the connectivity of TF_i with the genes, regulated

by it, are greater in case group than that in the control group, $i = 2, 3, \dots, 10$. Also, we assume $\rho_i s$ to be decreasing in i , $i = 2, 3, \dots, M$, so that there is a hierarchical regulatory structure among the transcription factors, TF_1 being at the top of the hierarchy. We follow the same steps in calculating the p-value as we did for size calculation in the previous section. We consider the same choices for $M,\ N,\ r_1, r_2,\ B, \mu, \vartheta, \rho$ and m_i ; $i = 1, 2, \dots, 10$ as we consider for size calculation. In particular, we choose $\gamma_2 = 0.5;\ r_{12} = 0.45$ and $r_{1i} = r_{1(i-1)} - 0.05$ for $i = 3, \dots, 10$. We choose $r_{2i} = (1 - \delta) r_{1i}$, $i = 2, 3, \dots, 10$ where $0 \leq \delta \leq 1$. These choices of r_{2i} , $i = 2, 3, \dots, 10$; ensure that increase in the value of δ also increase the difference between r_{1i} and r_{2i} , $i = 2, 3, \dots, 10$. In other words, the differential regulations of the transcription factors on the genes between the two groups increase as δ increases.

For the choice of γ_1, we consider the following relation: $\gamma_1 = \gamma_2 + \delta(r_{12} - r_{22})$, which implies $\gamma_1 = \gamma_2 + \delta^2 r_{12}, 0 \leq \delta \leq 1$. This choice of γ_1 ensures that increase in the value of δ also increase the difference between γ_1 and γ_2. We draw the power curve for different choices of δ, as shown in Figure 2-2.

From Figure 2-2, we see that the power steadily increases as the differential connectivity (regulated by δ) of the genes with the transcription factors between the two groups increase. The power curve starts from 3.2% at $\delta = 0$ (no difference in the connectivity of the genes with the transcription factors in the two groups) and reaches its maximum of 100% at $\delta = 1$ (maximum difference in the connectivity of the genes with the transcription factors in the two groups). The power reaches over 80% with a moderate choice of $\delta = 0.6$. Therefore, we can say that our proposed method is a valid test (e.g., size ≤ 0.05) that performs reasonably well (power reaching 100%) in identifying a significant concordance in the differential connectivity of the genes with

the transcription factors and the connectivity of a transcription factor with master regulator, if one exists.

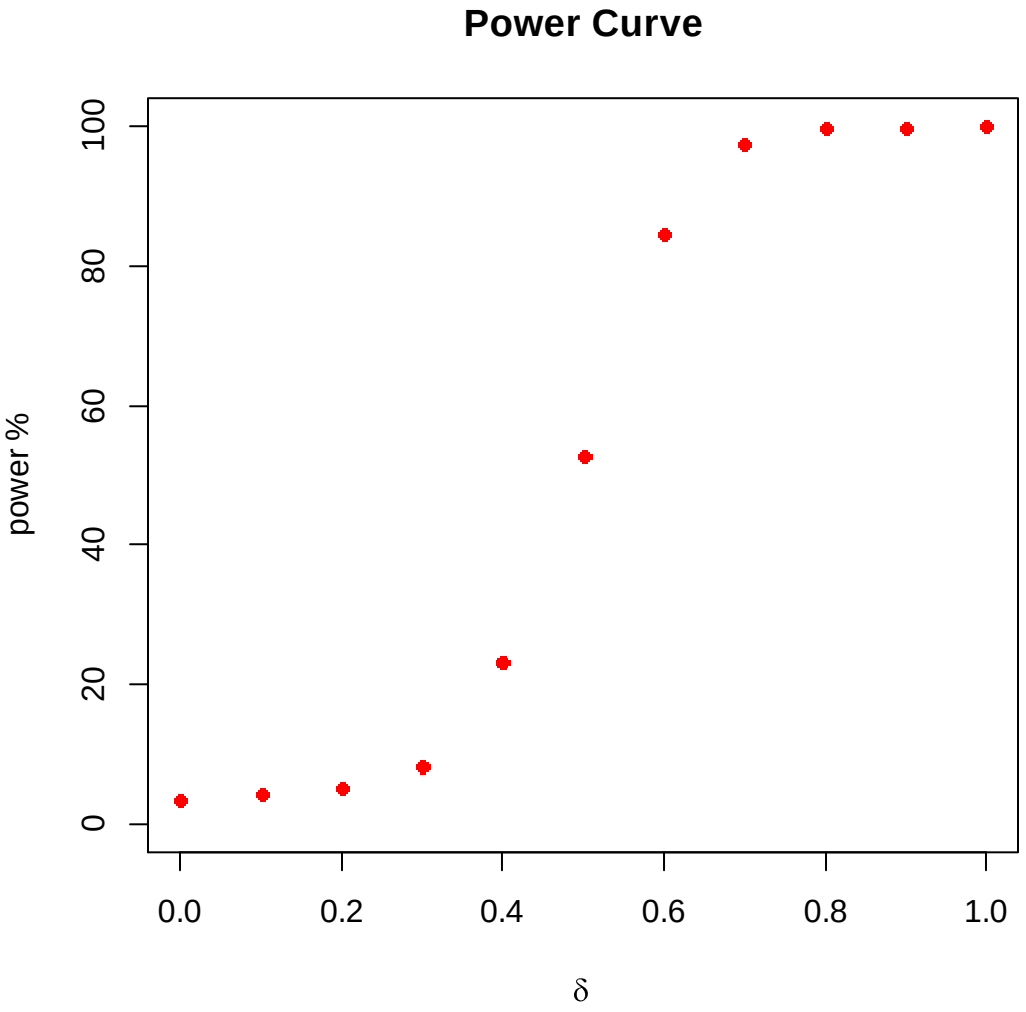

Figure 2-2. The power curve with 500 subjects in each group for several choices of δ

We also consider several other choices of the sample sizes in each of the two groups (case and control), and calculate the size and draw the power curves for each of the following choices of the sample sizes: $r_1 = 100,\ r_2 = 70;\ r_1 = 50,\ r_2 = 40;$ and $r_1 = r_2 = 50,$ representing reduced sample sizes and unequal sample sizes in each treatment group.

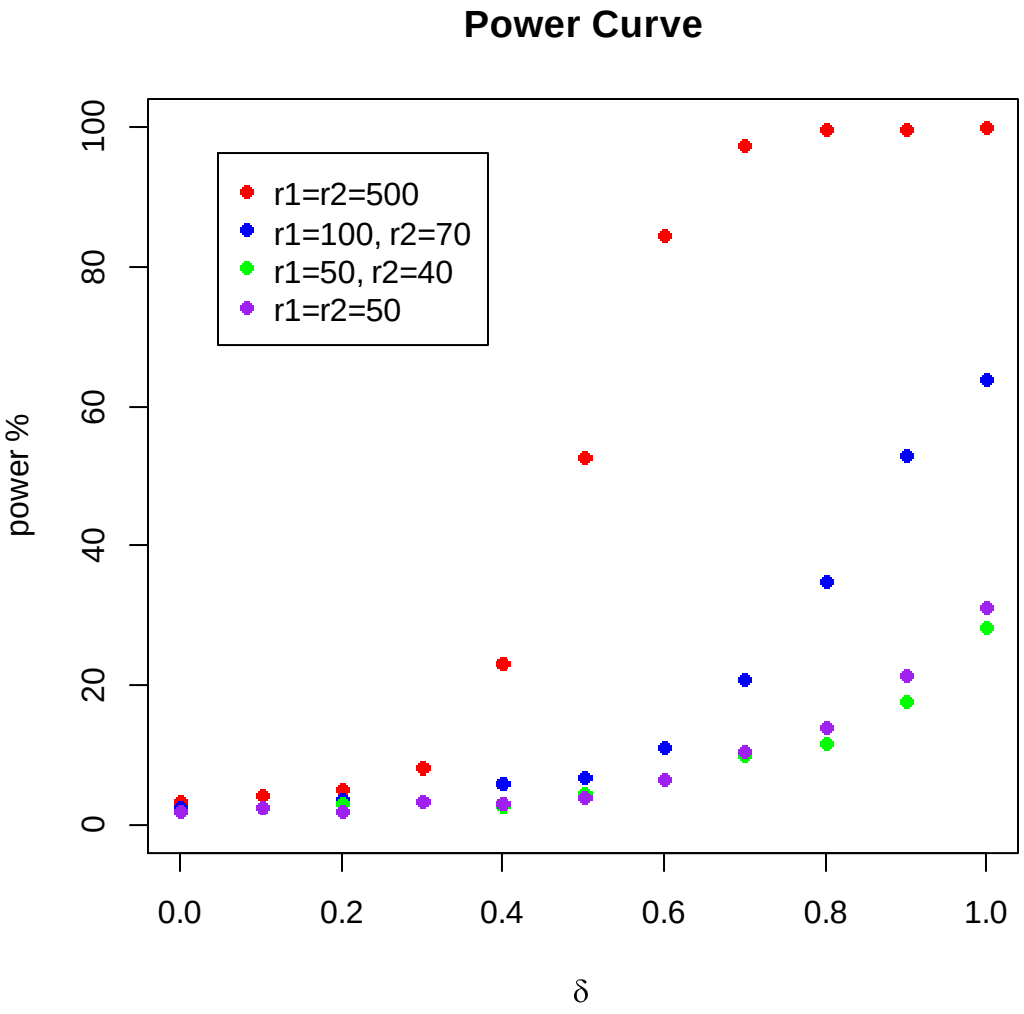

Figure 2-3. Plot of the power curves for different choices of the sample sizes with several choices of δ, using simulated datasets.

Overall, from our analyses with different choices of sample size, we find that the power of our test is increasing with increase in the sample size as well as an increase in the differential connectivity of the genes with the transcription factors in the two groups. Details of the variation of the power with sample size can be found in Figure 2-3 which shows the power curves for each of the above choices of the sample sizes with different choices of δ, $0 \leq \delta \leq 1$.

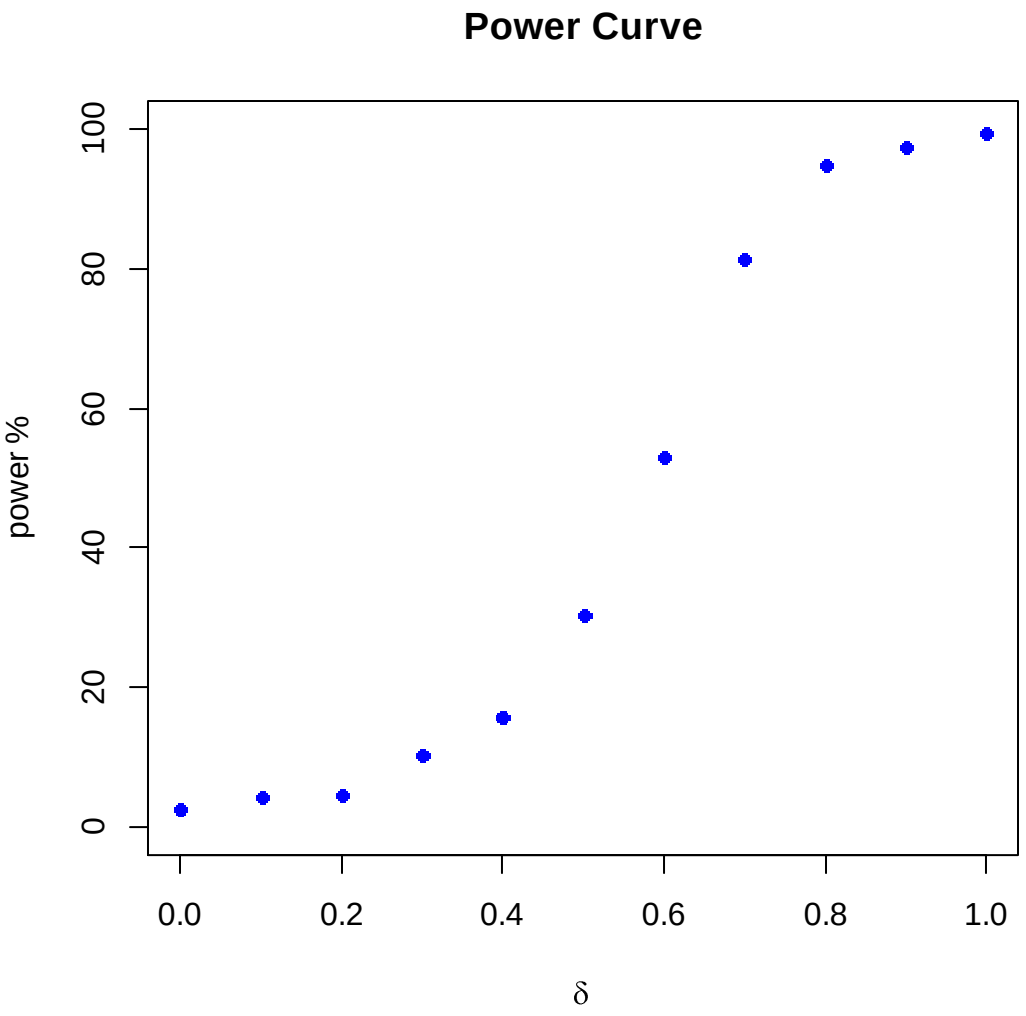

Figure 2-4. Plot showing the power performance of our test in presence of two independent master regulators with varying δ, using simulated datasets.

In order to check the performance of our test in case there are more than one master regulator transcription factors, we have also studied a simulated scenario where there are two independent master regulator transcription factors regulating two independent sets of genes through transcriptional regulatory networks. Figure 2-4 shows the power performance of our test in the presence of two independent master regulators in the system. In this case, too, our test has substantial power performance, similar to the simulated settings of a single master regulator

transcription factor. Note that, here we have considered one of the many possible simulation settings. However, our method can also be generalized for several other complicated scenarios.

2.3 Data Analysis

2.3.1 Prostate Cancer Data

We apply our test statistic, proposed in Methods section, to a human Illumina expression array dataset GSE18684 of androgen regulated gene expression in the LNCaP prostate cancer cell line [74]. It is believed that androgens and the androgen receptor (AR) play significant roles in prostate cancer cell proliferation and invasion. So, this study was conducted by [74] with an aim to identify the androgen receptor (AR) regulated genes. The LNCaP cells were treated with androgen (R1881) or with vehicle (ethanol) control. There are 10 control and 35 androgen treated samples with expression levels for 17182 probes in the dataset. We identify the set of probes which are differentially expressed in the two groups (androgen treated and vehicle control) using the "limma" package in Bioconductor [75]. After adjusting for false discovery rate (FDR) at 5% significance level, 6054 probes are differentially expressed in the two groups, out of which 542 are transcription factors.

Now, we test whether there exists any master regulator in the above mentioned dataset. For this, we compute the value of our test statistic for this dataset using equation (2-3), which turns out to be 0.49 with a bootstrap based p-value of 0.006. Since the p-value is highly significant we conclude that there exists a master regulator transcription factor in the system which is controlling all the other transcription factors and the genes. In order to find the master regulator, we use equation (2-5) as given in the Methods section. For this study, the two transcription factors "PEG3" and "ARNT2" have the same value of the test statistic given in equation (2-3). So, we conclude that these two transcription factors maximally control all the other transcription factors and consequently control the connectivity of the genes differently in

the two groups. Additionally, the Pearson's correlation coefficient value between the two transcription factors "PEG3" and "ARNT2" is 0.8. This high value of the correlation suggests that these two transcription factors are approximately at the same level of transcriptional regulatory hierarchy. Therefore, it can be concluded that both of them are the master regulators. Among these two master regulators, "PEG3" has often been linked to the development of prostate cancer. It is believed that deregulation of WNT/β catenin pathway contributes to prostate cancer progression [76-81] and according to [82], inhibition of the transcription factor "PEG3" can lead to enhanced β catenin expression and proliferation in human glioma stem cells. This function of the transcription factor "PEG3" is relevant to prostate cancer [83]. Further, the expression of the transcription factor "PEG3" is known to be associated with the processes of cancer aggressiveness and angiogenesis [84]. The results from our analysis show consistency with these known roles of "PEG3" in prostate cancer and demonstrate the utility of our proposed method to identify the master regulator transcription factor. Besides, the transcription factor "ARNT2" is known to have a critical role in human renal tract development, thereby showing congenital abnormalities of the kidneys and urinary tract [85]. "ARNT2" is also known to have significant roles in many cancers like NSCLC [86], breast cancer [87], etc.

Table 2-1. The list of top ten transcription factors which are highly correlated with the two master regulators "PEG3" and "ARNT2".

Master Regulators	Top 10 transcription factors correlated with the master regulator
PEG3	WWC1, FOXD4L1, NCOA7, TSHZ3, CTBP1, TCFL5, LHX2, ARID5B, CDCA7L, MAK,
ARNT2	MSRB2, TULP4, TSHZ3, TCFL5, SNAPC5, TFDP1, WWC1, CITED4, NCOA7, GRAMD4

From Table 2-1, it can be seen that the transcription factors "WWC1", "NCOA7", "TSHZ3" and "TCFL5" are highly correlated with both the master regulators. Among these,

"WWC1" is known to be associated with prostate cancer. The expression of "WWC1" is influenced by AR signaling and is increased in prostate cancer [88]. The transcription factor "NCOA7" is known to affect AR-mediated transcription [89]. The expression of "TSHZ3" is known to be downregulated in prostate cancer [90]. FOXD4L1 is also implicated in many cancers [91].

2.3.2 Colorectal Cancer Data

We apply our method to another human microarray dataset GSE4107. This study was conducted by [92] with an aim to identify differentially expressed genes in early onset colorectal cancer (CRC). RNA samples are extracted from colonic mucosa of patients as well as healthy controls and analyzed using GeneChip U133-Plus 2.0 Array. There are 22 subjects involved in the study which included 12 patients and 10 controls. All the patients and the controls in the data are young Chinese who are aged 50 years or less. There are expression levels for 54,675 genes for all the patients in the dataset. We first filter the data in order to find the set of differentially expressed genes between the case and the control groups. For this purpose we use the "limma" package in Bioconductor [75]. After adjusting for FDR at 10% significance level, the number of differentially expressed genes turns out to be 5192, among which 266 are transcription factors.

Next, we apply our method to the filtered dataset. We first test whether there exists a master regulator in the data. The value of our test statistic, given in equation (2-3), is 0.38 for this dataset with a p-value of 0.04 for the bootstrap based test. Since, the p-value is small enough to make the test significant, we conclude that there exists a master regulator in the data. We identify the master regulator using equation (2-5), given in the Methods section. The master regulator in this data is the transcription factor "NFKB2". Hence, we conclude that the transcription factor "NFKB2" maximally controls all the transcription factors and the genes in the data.

The transcription factor “NFKB2” is a subunit of the transcription factor nuclear factor-kappa-B (NFKB). “NFKB” transcription factors are known to be the key regulators of innate immune responses, inflammation, and cell survival [93, 94]. Also, “NFKB” activation has been frequently associated with tumor growth in leukemias and lymphomas, as well as prostate, pancreatic and colorectal cancers [95-97]. It has been widely suggested that “NFKB” activation plays a leading role in regulation of target genes that promote cell proliferation, anti-apoptosis, regulate immune and inflammatory response, and results in pathogenesis of various cancers [94, 98-102]. Further, it has been shown that constitutive activation of “NFKB” instigates strong resistance to chemotherapy and radiotherapy [102], while molecular targeted therapy against “NFKB” activation is believed to be effective in colorectal carcinomas with constitutive “NFKB” activation [94]. According to [101], “NFKB” may contribute to the promotion of the ongoing inflammatory process in the gut mucosa resulting in the progression of colitis associated colorectal cancer. Besides, it is believed that “NFKB” activation is involved in development of not only colitis-associated cancer, but also sporadic colorectal cancer [103].

From our data, we find that the master regulator “NFKB2” is maximally (negatively) correlated with the transcription factor “PPARGC1A (PGC-1alpha)” with an overall correlation value of -0.76. The correlation of “NFKB2” and “PPARGC1A” is -0.72 in the patients group whereas it is -0.39 in the control group. It is known that “NFKB” directly repress the activity of “PPARGC1A” in cardiac cells. This leads to the increase in glucose oxidation which is observed during pro-inflammatory state [104].

2.4 Discussion and Conclusion

In this project, we present a novel approach to identify a master regulator transcription factor in a system using only the gene expression profiles of the patients. We consider a simulation setting which validates our approach with a reasonable power in detecting the

existence of a master regulator. We have also checked the power of our test in the presence of two independent master regulator transcription factors in the simulation setup. We apply our approach to two human microarray datasets and detect the existence of master regulators in those. In order to check the robustness of our method in experiments not typically falling under the 'case-control' category, we have applied our method to an additional dataset, namely, Glioblastoma (GBM) TCGA RNA-seq data [105]. Here we compare the two types of GBM tumors: Mesenchymal and Classical. Our method concludes the existence of a master regulator transcription factor (PPRC1) between the two types of GBM tumors (Mesenchymal and Classical) with a p-value of 0.08 (marginally significant).

One advantage of our proposed method is that it needs only a single data source for accurate identification of a master regulator among a set of transcription factors. In the existing literature, most of the approaches integrate multiple platforms to identify a group of transcription factors or transcription factor binding sites or differentially regulated gene sets. Our method is particularly useful in case there is limitation in data sources and data from multiple platforms are not available. Also, our method is aimed to identify a single master regulator, as opposed to identifying a group of transcription factors associated with the disease process as in the case of other existing methods. The method can identify multiple master regulator transcription factors if they are individually at the top of hierarchy of the transcription regulation. This is advantageous in anti-cancer drug development processes which initially target the most potential transcription factor associated with the disease and can be used as a potential biomarker. However, there is a scope of further improvement of our proposed method by incorporating important platforms like ChIP-Seq data. From simulation settings, we see that the performance of our method gets better with the increase in the number of patients in each group. So, our method is expected to be more

efficient when there is sufficiently large number (around 100) of patients in each group while it may not be very efficient in case the sample size is very small. Although both the data analyzed in this project have lower number of subjects in each group, our test was still successful in identifying master-regulator transcription factors from the data. One important assumption of our method is that the ranking of the transcription factors on the basis of their differential connectivity of the genes between two experimental conditions is concordant with the hierarchical order of their own regulation. The fulfilment of the above mentioned condition is a key indicator to the existence of a master regulator transcription factor and its subsequent detection through our method. However, it may be possible that in certain situations, although there exists a master regulator transcription factor, there is no such clear cut concordance between it's regulation on other transcription factors and differential connectivity with the other genes. In such a case, our method may not perform well. Overall, we believe that our method will give new insight for efficient identification of potential disease biomarker and therapeutic target in drug development processes based on master regulator transcription factor.

CHAPTER 3
EXPLORING THE IMPORTANCE OF CANCER PATHWAYS BY META-ANALYSIS OF DIFFERENTIAL PROTEIN EXPRESSION NETWORKS IN THREE DIFFERENT CANCERS

3.1 Methods

In this chapter, we analyze the preprocessed challenge datasets for CAMDA 2015 provided by the International Cancer Genomic Consortium (ICGC) to explore the importance of cancer pathways. For our study, we consider the protein expression and the clinical profiles of the patients for the three cancers: Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), and Kidney Renal Clear Cell Carcinoma (KIRC). A set of 132 proteins is present in the protein expression profiles of each of the three cancers; the patient sample sizes of HNSC, LUAD and KIRC are 212, 237 and 454 patients, respectively. The number of patients in the clinical profile of HNSC, LUAD and KIRC are respectively 422, 473 and 515. The clinical profile of each of the cancer type represents the disease status ("progression" or "complete remission") of each patient. However, 44, 111 and 28 patients have missing disease status in the clinical data of HNSC, LUAD and KIRC respectively. After removing the patients who have missing disease status, we are left with 185 patients (71 in "progression" group and 114 in "complete remission" group) in HNSC, 172 patients (51 in "progression" group and 121 in "complete remission" group) in LUAD and 430 patients (132 in "progression" group and 298 in "complete remission" group) in KIRC, each with expression values of 132 proteins. In summary, we have two groups of patients for each cancer type and a set of recorded protein expression values of 132 proteins for each of them.

Reprinted with permission from Sikdar S, Datta S, Datta S. Exploring the importance of cancer pathways by meta-analysis of differential protein expression networks in three different cancers. Biol Direct. 2016; 11: 65.

3.1.1 Pathway Analysis

From several studies, it has been found that approximately there are 140 genes that are responsible for selective growth advantage. These genes are termed as "driver" genes. As mentioned before in section 1.1.2, mutations occur in a typical tumor due to two to eight such "driver" genes, while there are only twelve signaling pathways which allow selective growth advantage [60]. These twelve signaling pathways ("target pathways") are: TGF − β, MAPK, STAT, PI3K, RAS, Cell Cycle/Apoptosis, NOTCH, HH, APC, Chromatin modification, Transcriptional regulation and DNA damage control. Among these, TGF − β, MAPK, STAT, PI3K, RAS and Cell Cycle/Apoptosis regulate "cell survival"; NOTCH, HH, APC, Chromatin modification and Transcriptional regulation regulate "cell fate"; while the DNA damage control signaling pathway regulates "genome maintenance". We separately analyze the protein profiles of the three cancer types using "GO" clustering tool [107, 108] and group the proteins according to their biological pathways. As selective growth advantage can occur only through the "target pathways", we consider only the proteins included in the "target pathways" for our analysis.

3.1.2 Differential Network Analysis

For each cancer, in order to identify whether the network structures of the "target pathways" have changed from the "complete remission" group to the "progression" group, we perform differential network analysis [73] using the R package *dna* [109]. This differential network analysis for each pathway is conducted based on connectivity scores between the proteins in these "target pathways". The difference in connectivity between the two groups ("progression" versus "complete remission") is computed mathematically, using the following statistic:

$$\Delta(\mathcal{F}) = \frac{1}{k(k-1)} \sum_{p \neq p' \in \mathcal{F}} \left| s_{pp'}^{pr} - s_{pp'}^{cr} \right| \tag{3-1}$$

where $\mathcal{F}$ denotes the set of proteins present in a "target pathway" and k denotes the number of proteins in $\mathcal{F}$. Here $s_{pp'}^{pr}$ and $s_{pp'}^{cr}$ are the connectivity scores between the proteins p and p' in the "progression" and "complete remission" groups, respectively. For our analysis, the connectivity score of a protein pair in a network is taken to be the Pearson's correlation coefficient of the expression values of the two proteins in the corresponding sample data. A permutation test is then carried out using the test statistic $\Delta(\mathcal{F})$ as follows: let p denote the number of proteins in the sample (132 in this case). Let N_{pr} and N_{cr} denote the number of patients in the "progression" group and in the "complete remission" group, respectively. The expression profiles of the patients in the two groups are given in the matrix form of orders $N_{pr} \times p$ and $N_{cr} \times p$, respectively. Now, a matrix E is constructed by merging the rows of the two profiles, i.e. E is of order $(N_{pr} + N_{cr}) \times p$. Then, the rows of E are randomly permuted and the first N_{pr} patients are considered as one group and the remaining N_{cr} patients as another group. For the i^{th} permutation, the connectivity scores $s_{pp'}^{pr,i}$ and $s_{pp'}^{cr,i}$ between the proteins p and p' are then computed using the expression profiles of the permuted data. Thus, the i^{th} permuted test statistic $\Delta(\mathcal{F}, i)$ is given by

$$\Delta(\mathcal{F}, i) = \frac{1}{k(k-1)} \sum_{p \neq p' \in \mathcal{F}} \left| s_{pp'}^{pr,i} - s_{pp'}^{cr,i} \right| \tag{3-2}$$

We repeat this process 1000 times and obtain the observed level of significance (p-value) using

$$pval(\mathcal{F}) = \frac{1}{1000} \sum_{i=1}^{1000} I\big(\Delta(\mathcal{F}, i) \geq \Delta(\mathcal{F})\big) \tag{3-3}$$

Next, we construct graphical networks for those "target pathways" for which the p-values are significant, so that we get an idea about the network structures in each of the two groups. The graphical networks are constructed by connecting each pair of proteins for which the connectivity score exceeds a threshold.

In addition to testing the overall pathway significance for each cancer, we also test whether the connectivity of each single protein has changed between the two groups ("progression" vs "complete remission") using the following statistic:

$$d(p) = \frac{1}{f-1} \sum_{p' \in \mathcal{G},\ p' \neq p} \left| s_{pp'}^{pr} - s_{pp'}^{cr} \right| \tag{3-4}$$

where $\mathcal{G}$ denotes the set of all proteins and f is the number of proteins in $\mathcal{G}$. Once again, a permutation test is carried out for each protein, in the way described before. The p-value corresponding to each protein is obtained using (3-3), with $\Delta(\mathcal{F})$ and $\Delta(\mathcal{F}, i)$ replaced by $d(p)$ and $d(p,i)$ respectively, where $d(p,i)$ is given by

$$d(p,i) = \frac{1}{f-1} \sum_{p' \in \mathcal{G},\ p' \neq p} \left| s_{pp'}^{pr,i} - s_{pp'}^{cr,i} \right| \tag{3-5}$$

3.1.3 Rank Aggregation

The p-values, obtained using the test statistic given in (3-1), are used to obtain ranked lists of the "target pathways" for each cancer type. Here, ranking is done in such a way that the "target pathway" with the lowest p-value gets rank 1, the next one gets rank 2 and so on. Since, these ranked lists vary according to the cancer type; we need to aggregate them in a meaningful way to get an overall ranked list which would then rank the pathways by their global order of importance. In other words, this overall ranked list may provide us with the most important "target pathways" in all the three cancers. In order to get this overall ranked list, we consider the R package *RankAggreg* [110, 111] based on the Cross-entropy Monte Carlo algorithm [112]. In

the framework of optimization problem, *RankAggreg* [110, 111] minimizes an objective function, so that a final ordered list is obtained which is close to all the individual ordered lists, simultaneously. The objective function is defined as follows

$$O(x) = \sum_{i=1}^{m} w_i d(x, Y_i) \tag{3-6}$$

Here, Y_i is the i^{th} ordered list, x is the proposed ordered list, d is a distance measure and w_i denotes the weight associated with the ordered list Y_i. The aim here is to find x^* for which the objective function $O(x)$, given in (3-6), is minimum. In other words,

$$x^* = argmin \sum_{i=1}^{m} w_i d(x, Y_i) \tag{3-7}$$

For our analysis, we consider equal weights (w_i) for all the three ranked lists. Here, Y_i is the i^{th} ordered list of "target pathways". As a distance measure, we consider weighted Spearman's footrule distance [113]. A brief description of the algorithm of the weighted Spearman's footrule distance, used for our purpose, is as follows: Let $p_i(1);\ p_i(2);\ \dots;\ p_i(5)$ be the p-values (in ascending order) associated with the i^{th} ordered list, Y_i. Let $r_A(Y_i)$ and $r_A(x)$ denote the ranks of the "target pathway" A in the i^{th} ordered list Y_i and the proposed ordered list x , respectively. Then, the weighted Spearman's footrule distance is given by

$$S(x, Y_i) = \sum_{t \epsilon Y_i \cup x} \left| p\left(r_t(Y_i)\right) - p\left(r_t(x)\right) \right| \times |r_t(Y_i) - r_t(x)| \tag{3-8}$$

For our second analysis at the individual protein level, the p-values, obtained using the test statistic given in (3-4), are used to rank the set of 132 individual proteins. An overall ranked list of these proteins is obtained using the R package *RankAggreg* [110, 111] in a similar way.

3.2 Results

3.2.1 Pathway Analysis Results

For each cancer type, we find representation of five out of twelve "target pathways" in our sample of 132 proteins using "GO" clustering tool [107, 108]; they are the PI3K signaling pathway, Cell Cycle, Apoptosis, RAS signaling pathway and MAPK signaling pathway. It is interesting to note that all these five "target pathways" are all related to "cell survival" function.

3.2.2 Differential Network Analysis Results

Next, we determine whether there is any significant difference in the network structures of the "target pathways" between the two groups of patients ("progression" vs "complete remission"). We perform differential network analysis [73, 109] of the network of proteins in these representative "target pathways" between the two groups of patients ("progression" vs "complete remission") using the test statistic given in (3-1), with Pearson's correlation coefficients as connectivity scores and absolute distance measure for each of the three cancer types. The p-values of the analysis are reported in Table 3-1. Based on these results, we have the following findings: the RAS signaling pathway is highly significant (p-value = 0.026) and MAPK signaling pathway is marginally significant (p-value = 0.082) in HNSC; for LUAD, PI3K signaling pathway is highly significant (p-value = 0.013).

Table 3-1. "Target pathways" along with the p-values obtained from differential network analysis for each cancer type.

Target Pathway	Cancer Type		
	HNSC	LUAD	KIRC
RAS signaling pathway	0.026	0.507	0.156
MAPK signaling pathway	0.082	0.759	0.517
PI3K signaling pathway	0.241	0.013	0.774
Apoptosis	0.407	0.417	0.487
Cell Cycle	0.410	0.238	0.997

A graphical representation of the network structure of the proteins in the two groups of patients for RAS signaling pathway in HNSC is shown in Figure 3-1. In this figure, two proteins

are connected if the connectivity score between them is significantly large. Different colors and shades in the figure represent positive or negative correlations and the thickness of the lines represents the strength of the associations. A visual inspection reveals some obvious differences in the network connectivity between the two groups of patients.

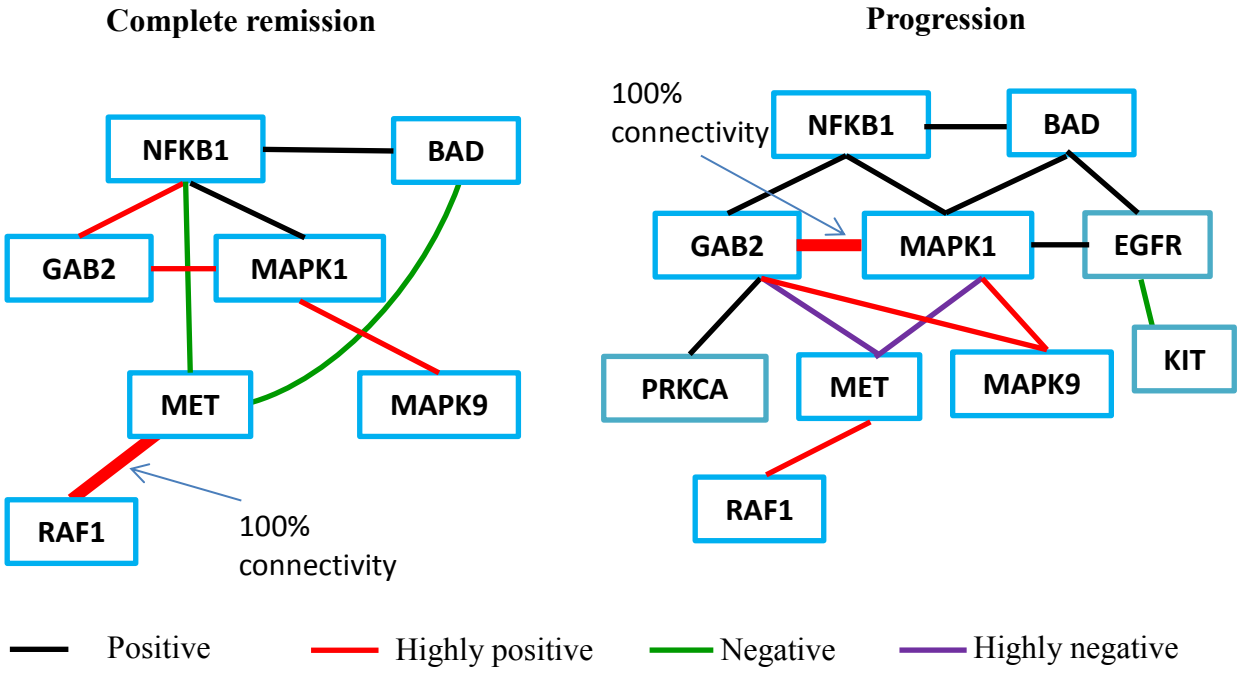

Figure 3-1. Network structure for RAS signaling pathway in Head and Neck Squamous Cell Carcinoma (HNSC)

It can be seen from Figure 3-1 that the protein MET has very high connectivity with the proteins GAB2 and MAPK1 in the "progression" group, whereas no such connectivity can be seen between them in the "complete remission" group. On the other hand, MET can be seen to be connected with NFKB1 and BAD in the "complete remission" group but no such connectivity can be seen in the "progression" group. Also, it is interesting to note that MET has 100 percent connectivity with RAF1 in the "complete remission" group. The protein GAB2 appears to be highly connected with MAPK9 in the "progression" group but no connectivity can be seen

between them in the “complete remission” group. GAB2 is also connected with PRKCA only in the “progression” group. Further, GAB2 has 100 percent connectivity with MAPK1 in the “progression” group. In the “progression” group, MAPK1 and BAD are connected among themselves and also with EGFR, which is further connected with KIT. But no such interesting connections can be seen in the “complete remission” group. Summarizing, GAB2, MAPK1, MET, and BAD show noticeably different connectivities in the two networks. The corresponding genes are known oncogenes. GAB2 is known to be overexpressed in multiple human tumors especially in melanoma. It is altered by MAPK and PI3K signaling pathways [114]. MAPK1 (Mitogen-activated protein kinase 1) is broadly implicated in many cancers [115]. MET is associated with MET signaling process. In many cancers involving solid tumors, inhibiting this signaling has major therapeutic effect [116]. Similarly, it is found that BAD is a pro-apoptotic protein which has been identified as an integrator of several anti-apoptotic signaling pathways in prostate cancer cells [117].

A graphical representation of the network structure of the proteins in PI3K signaling pathway for the two groups of patients in LUAD is shown in Figure 3-2. From the figure, it can be seen that the protein MET has interesting connections with the proteins CASP9, FOXO3, YWHAE and MAPK1 in the “progression” group whereas no such interesting connections can be found in the “complete remission” group. Also, PRKCA is connected with FRAP1, MAPK1, BCL2L11 and NFKB1 in the “progression” group but no such connectivity can be seen in the “complete remission” group. The protein MAPK1 seems to be connected with MET, YWHAE, PRKCA, FRAP1 and TSC2 in the “progression” group whereas it is only connected with TSC2 in the “complete remission” group. Here, MET, PRKCA and MAPK1 show significant differences in the two networks. PRKCA is a serine/threonine-protein kinase that is highly

expressed in a number of cancer cells where it can act as a tumor promoter and is implicated in malignant phenotypes of several tumors such as gliomas and breast cancers [118].

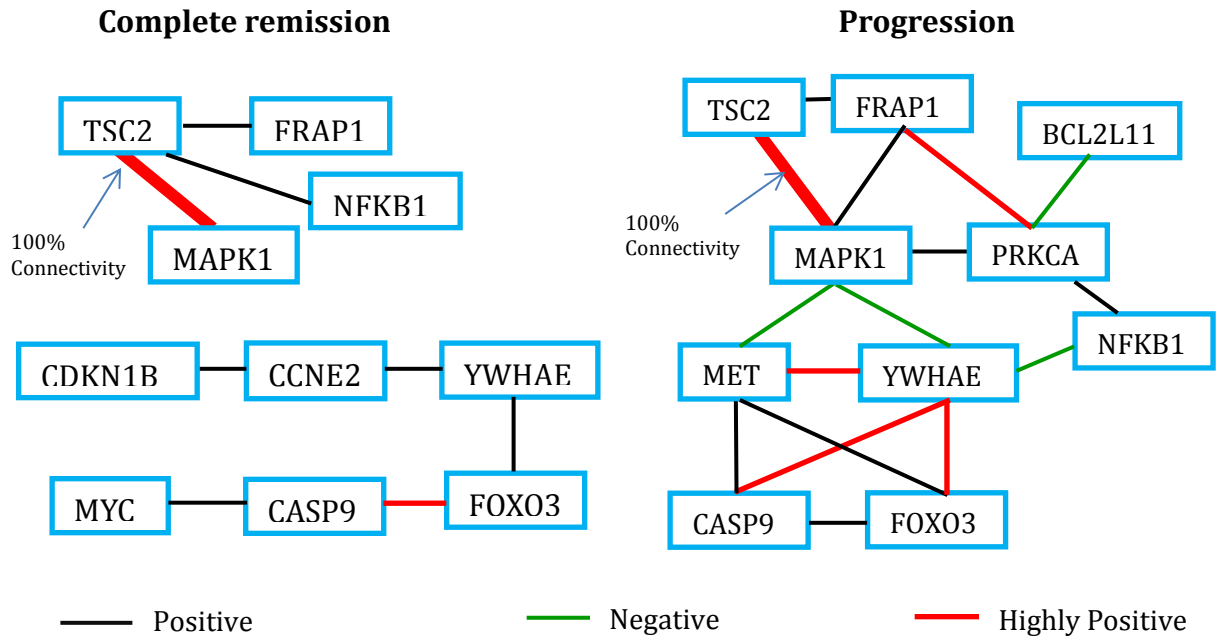

Figure 3-2. Network structure for PI3K signaling pathway in Lung Adenocarcinoma (LUAD)

Our analysis of individual proteins using the test statistic given in (3-4) produces Figure 3-3 (A)-(C). The pie charts represent the proportions of top fifty differentially connected proteins for each of the "target pathways" in the three cancers. This provides a global visual representation of the relative importance of the "target pathways" in all three cancers. For all the three cancers, PI3K and RAS signaling pathways show significant contributions in terms of differential network connectivity.

3.2.3 Rank Aggregation Results

Next, we rank the relative importance of the "target pathways" based on the p-values, obtained using (3-3) from the differential network analysis, so that we can get an idea about the

ordering of importance of the "target pathways". Since, the ordering varies from one cancer to another; we obtain a rank aggregated list of the "target pathways" using *RankAggreg* [110, 111]. Table 3-2 shows the ordering of the five "target pathways" for the three cancers along with the rank aggregated list. Thus overall, the RAS signaling pathway appears to be most important followed by the PI3K signaling pathway, based on our integrative analysis of the available data on three cancers.

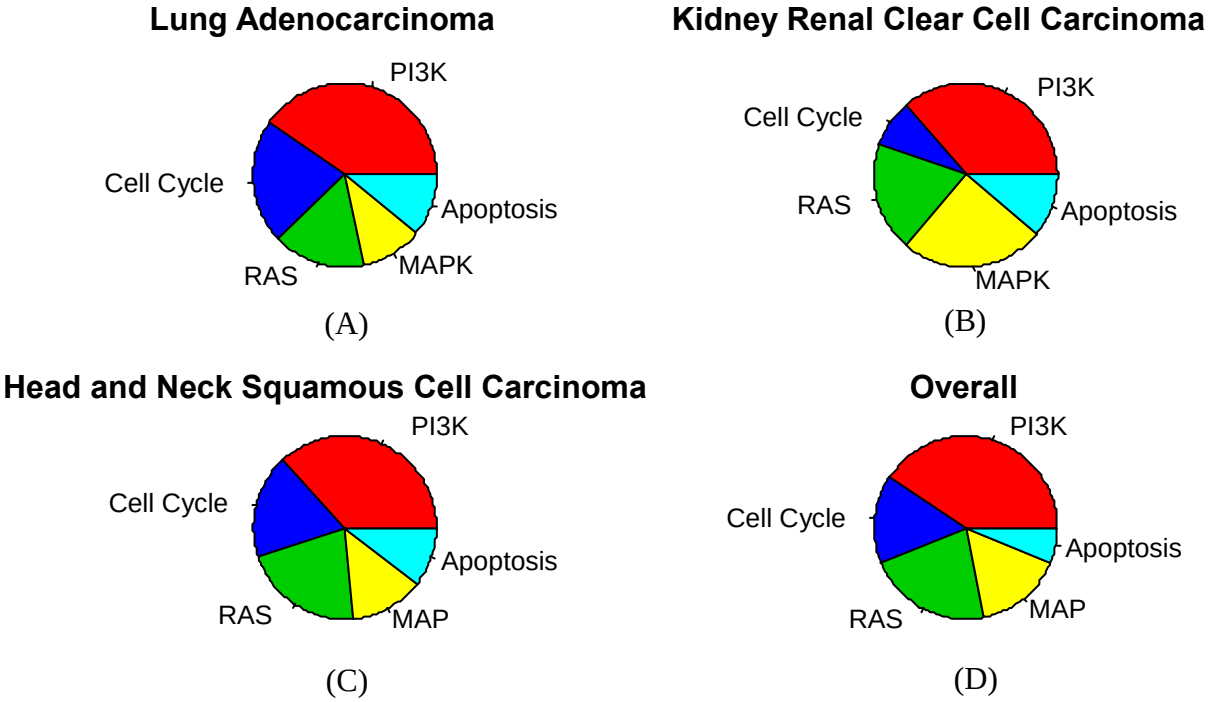

Figure 3-3. Relative contributions of the "target pathways" in the three cancers separately as well as combined.

We also obtain a rank aggregated list of the proteins, based on the p-values of the tests based on differential connectivity of each protein in the two different networks of the two patient groups. The proportions of top fifty differentially connected proteins for each of the "target pathways" are shown in Figure 3-3(D). Once again, PI3K and RAS take the top two most important spots in terms of differential network connectivity.

Table 3-2. "Target pathways" ordered by p-values for each cancer type along with the overall ordering.

Cancer Type	Pathway Ordering by p-values
HNSC	R, M, P, A, C
LUAD	P, C, A, R, M
KIRC	R, A, M, P, C
Overall	R, P, M, A, C

R: RAS Signaling pathway M: MAPK Signaling pathway P: PI3K Signaling pathway
A: Apoptosis C: Cell Cycle

3.3 Discussion and Conclusion

It is known that for most cancers with solid tumors the genes in the above mentioned "target pathways" display somatic mutations and change their protein products [60]. In all human tumors, PI3K is known to be as one of the most frequently targeted pathways. Mutation in PI3K pathway components contributes up to 30 percent of all human cancer and is known to be activated by RAS [119]. It is interesting to know that PI3K is a regulatory subunit, which binds to cell-surface receptors and to the RAS protein. Genes and proteins in PI3K and RAS have been investigated as therapeutic targets for many cancers [120, 121]. Here this purely quantitative analysis of the existing protein expression data of three different cancers also reveals the significant alteration of the proteins in PI3K and RAS pathways. Our findings are consistent with this and suggest that continued future efforts be made in this direction.

Since genes act in consort during a biological process, a network analysis is essential for a system-wide understanding. Thus, a study of differential network connectivity could yield interesting findings that are not possible from a differential expression analysis of individual proteins. In addition, pathway level information should be incorporated in a differential network analysis whenever possible. It can be seen from Figure 3-3 (D) that on the basis of analysis of differential connectivity of individual proteins, PI3K emerged as the most important pathway

overall. However, this could be partly due to the fact that PI3K pathway has the largest number of proteins compared to the other "target pathways". On the other hand, the test statistic based on differential connectivity of pathways is automatically normalized by the size of the pathway and on the basis of this comparison, RAS turned out to be the most significant overall as shown in Table 3-2.

In a recent study [122], multiple molecular profile data of LUAD for the CAMDA 2014 lung adenocarcinoma challenge data provided by ICGC is analyzed and it is noted that EGFR signaling pathway plays a significant role among the patients. Besides, it is known that EGFR activation initiates RAS signaling [123-125], and EGFR induces rapid increase in number of epithelial cells by activating a network of signaling elements, including members of the RAS and PI3K [125, 126]. Thus, it is common for lung cancer patients to have active EGFR mutation. Moreover, the RTK/RAS/RAF pathway, identified as the main route in causing adenocarcinoma, is shown to be activated in patients with "common" pathway mutations, e.g. in KRAS, BRAF, and EGFR [127]. Although the EGFR kinase inhibitor Gefitinib is an effective treatment for lung cancers with EGFR activating mutations, amplification of MET causes Gefitinib resistance by driving ERBB3 (HER3)–dependent activation of PI3K [128], a pathway thought to be specific to EGFR/ERBB family receptors. This fact, along with the fact that MET has been found in the module of the "progression" group but not in that of the "complete remission" group in PI3K signaling pathway of LUAD (see Figure 3-2) in our current study, suggest that the patients under study might have been treated with Gefitinib, but the presence of MET in some of these patients (those in the "progression" group) may have led to the resistance to this drug. However, this is subject to verification as we don't have the information regarding the treatment regime for any of these patients. This highlights the potential role of MET in lung cancer progression.

An interesting observation from Figure 3-3 is that much lesser proportion of proteins in cell cycle pathway is differentially connected between the two groups of patients in KIRC compared to LUAD and HNSC. So underlying molecular mechanisms related to cell cycle pathway may be a little different in KIRC compared to LUAD and HNSC.

In a recent paper [129], patient level information such as mutation profiles is incorporated to identify protein-protein interaction (PPI) interfaces enriched in somatic mutations. It will be interesting to explore how to incorporate patient heterogeneity information into our approach.

CHAPTER 4
EAMA: EMPIRICALLY ADJUSTED META-ANALYSIS FOR LARGE-SCALE SIMULTANEOUS HYPOTHESIS TESTING IN GENOMIC EXPERIMENTS

4.1 Methods

4.1.1 Empirically Adjusted Meta-Analysis (EAMA)

Here, we provide a description of empirically adjusted meta-analysis (EAMA) in details. Let us suppose we have G number of genes in our study. We are interested in finding out which of these G genes contribute significantly to our outcome of interest. Suppose $H = \{H_g : 1 \leq g \leq G\}$ be the collection of null hypotheses where H_g denotes the hypothesis that the gene g has no significant contribution to the outcome of interest, $1 \leq g \leq G$. Also, suppose we have data from multiple independent experiments. Let M be the number of independent genomic experiments. By combining the results of these M independent experiments we aim to identify the genes which contribute significantly to the outcome of interest.

Let us define $P^{(i)} = \left\{\hat{p}_g^{(i)} : 1 \leq g \leq G\right\}$ as the collection of p-values from the i^{th} genomic experiment, where $\hat{p}_g^{(i)}$ is the p-value corresponding to gene g (i.e. corresponding to the null hypothesis H_g) in the i^{th} experiment, $1 \leq i \leq M, 1 \leq g \leq G$.

Using the inverse z-transformation, we get the collection of z-scores as

$$z^{(i)} = \left\{z_g^{(i)} = \Phi^{-1}\left(\hat{p}_g^{(i)}\right) : 1 \leq g \leq G\right\}, 1 \leq i \leq M \tag{4-1}$$

The z-scores given in (4-1) may not follow a $N(0,1)$ distribution under the null hypotheses. Here we modify the z-scores, given in (4-1), using the Efron's technique of estimating an empirical null distribution [63] so that the resulting z-scores follow $N(0,1)$ distribution under the null hypotheses. A brief detail on estimating the empirical null can be found in the next section. Following the steps of [63], suppose the empirical null distribution is

obtained as $\hat{f}_0 = N(\hat{\mu}_0, \hat{\sigma}_0{}^2)$ using the R package "locfdr" [130]. Then the modified z-values are calculated as

$$\tilde{z}_g^{(i)} = \frac{z_g^{(i)} - \hat{\mu}_0}{\hat{\sigma}_0}, \quad 1 \le g \le G, 1 \le i \le M \tag{4-2}$$

These empirically adjusted z-values, given in (4-2), can be assumed to follow a $N(0,1)$ distribution under the appropriate null hypotheses. Finally, we convert the modified z-values into the corresponding p-values as $\tilde{p}_g^{(i)} = \Phi\left(\tilde{z}_g^{(i)}\right), 1 \le g \le G, 1 \le i \le M$.

At this stage, we have a set of M p-values from the M independent experiments for each gene g. But for proper inference on the overall effect of a gene, we need to have a single p-value for that gene. So, for a typical gene g we combine the p-values from all of the M experiments to obtain a single p-value using Fisher's method [62] as given below:

If $\tilde{p}_g^{(1)}, \tilde{p}_g^{(2)}, \dots, \tilde{p}_g^{(M)}$ are the M p-values for the gene g obtained from the M independent experiments, we combine these M p-values to get a single test statistic $T_g = 2\sum_{i=1}^{M}\left\{-log\left(\tilde{p}_g^{(i)}\right)\right\}, 1 \le g \le G$. Under the null hypothesis H_g : gene g does not contribute significantly to the outcome, T_g follows a χ^2 distribution with $2M$ degrees of freedom assuming that the p-values $\tilde{p}_g^{(i)}$ follow uniform distribution. Using these we obtain a resulting set of G p-values $\{\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_G\}$, where $\tilde{p}_g$ is the p-value corresponding to the test statistic T_g. In this way we are able to get a single p-value corresponding to each of the G genes. To account for the large number of hypotheses being tested in the study, we apply the Benjamini-Hochberg [67] method of multiplicity correction to get a set of corrected p-values $\{p_1^*, p_2^*, \dots, p_G^*\}$. It may be noted that other methods for multiplicity correction can also be used. Finally, using the p-value p_g^* we decide whether the gene g has any overall significant contribution to the outcome of interest.

4.1.2 Estimation of the empirical null distribution

A brief discussion on estimating the null distribution, following [63], is as follows. Suppose that there are N z-values which can be classified into two classes, "Uninteresting" or "Interesting", depending on whether or not z_i is generated according to the null hypothesis. Also, assume that the prior probabilities of the z-values belonging to the "Uninteresting" or "Interesting" classes are p_0 and $p_1 = 1 - p_0$ respectively, and that z_i has density either $f_0(z)$ (null density) or $f_1(z)$ (non-null) depending on its class. Then the mixture density of z is given by $f(z) = p_0 f_0(z) + p_1 f_1(z)$. Following Bayes theorem the *a posteriori* probability of belonging to the Uninteresting class given z is obtained as $\text{Prob}\{\text{Uninteresting}|z\} = p_0 f_0(z)/f(z)$. The local false discovery rate is then defined as $f_0(z)/f(z)$. The main idea is to estimate the density f_0 from the central peak of the observed histogram of the z-values. Under the assumption that f_0 is density of a normal distribution with mean δ_0 (not necessarily 0) and standard deviation σ_0 (not necessarily 1), for z close to zero, we can write $logf(z) = -\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2 + \text{constant}$. Then δ_0 and σ_0 can be estimated as $\delta_0 = argmax\{f(z)\}$ and $\sigma_0 = \left[-\frac{d^2}{dz^2} logf(z)\right]_{\delta_0}^{-\frac{1}{2}}$. Around $z = 0$, $logf(z)$ curve is estimated through a quadratic approximation which leads to the final estimation of δ_0 and σ_0. This is done under the assumption that the central peak of the z-value histogram, presumably close to zero, is mainly contributed by the null cases, and this method of estimating the null parameters is termed as "central matching" method (see [131] for more details on this approach).

4.2 Simulation Studies

In order to evaluate the performance of our proposed method (EAMA), described in section 4.1.1, for accurate identification of significant genes, we simulate datasets mimicking

multiple genomic experiments. We simulate continuous expression datasets as obtained from microarray experiments, as well as, count datasets which are found in next generation sequencing experiments. Besides, we also consider the possible presence of some unknown hidden variables or confounders that often impact the results of the genomic experiments. It is our interest to check the performance of EAMA in circumstances affected by the presence of hidden variables or confounders.

We call a gene differentially expressed if the corresponding "Benjamini-Hochberg" [66] adjusted p-value is less than 0.05. The performance of EAMA along with its naïve counterpart is assessed using four performance assessment measures: sensitivity, specificity, false discovery rate (FDR) and false non-discovery rate (FNR) as defined below:

1. Sensitivity (or true positive rate or recall): Proportion of genes that are correctly identified as differentially expressed among all the differentially expressed genes.
2. Specificity (or true negative rate): Proportion of genes that are correctly identified as non-differentially expressed among all the non-differentially expressed genes.
3. False discovery rate (FDR, or 1 minus precision): Proportion of genes that are incorrectly identified as differentially expressed among the set of identified differentially expressed genes.
4. False non-discovery rate (FNR): Proportion of genes that are incorrectly identified as non-differentially expressed among the set of identified non-differentially expressed genes.

The values of all the above four measures are calculated for our method (EAMA) based on 500 independent Monte-Carlo simulations. For comparison, we also obtain the results from a naïve meta-analysis which does not apply the empirical adjustment (null transformation) to the raw p-values. The naïve method combines the raw p-values using Fisher's method [62] and adjust the resulting p-values using "Benjamini-Hochberg" method of multiplicity correction [66].

Details of the data generation methods are described below.

4.2.1 Generation of Continuous Data (Microarray Based Gene Expression)

We generate a number of simulation studies involving multiple genomic experiments. We have datasets obtained from $M = 10$ independently generated experiments, where each experiment has data on $G = 1000$ genes and $N = 20$ subjects distributed equally over two groups (for example, case and control). The first 10 subjects are considered to be in one group and the remaining 10 subjects in the other. The (log) expression levels of G genes for a typical experiment m are generated as follows.

Let Y_{ijk} be the (log) expression value corresponding to the i^{th} gene belonging to the k^{th} subject and j^{th} group. The (log) expression values (Y_{ijk}) of genes are generated using a linear model as given below:

$$Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + W_{ijk} + e_{ijk} \qquad (4\text{-}3)$$

where μ denotes the general mean effect in the model, G_i is the effect of the i^{th} gene, V_j is the effect of the j^{th} group, $(GV)_{ij}$ is the interaction effect of the i^{th} gene and the j^{th} group, W_{ijk} is the effect of a latent confounder on the i^{th} gene of the k^{th} subject in the j^{th} group, e_{ijk} is the error component corresponding to the i^{th} gene of the k^{th} subject in the j^{th} group.

Here, $i = 1,2, \dots, 1000;\ j = 1,2;\ k = \begin{cases} 1, \dots, 10 & \text{for } j = 1 \\ 11, \dots, 20 & \text{for } j = 2 \end{cases}$

For each of the M independent experiments, the gene (log) expression profiles are generated as mentioned above. For our simulation studies on microarray data, the following two simulation schemes are considered.

Scheme 1. In this scheme, we simulate the microarray datasets for each experiment using (4-3) assuming that there is no effect of any hidden confounder in the model. To achieve this, we set $W_{ijk} = 0$ for $1 \le i \le 1000;\ j = 1,2;\ k = 1,2, \dots, 20$ in (4-3).

For simplicity, we further assume that all the main effect terms were zero. That is, we set $\mu = 0,\ G_i = 0$ for $1 \leq i \leq G$ and $V_j = 0$ for $j = 1,2$ in (4-3). The $e_{ijk}s$ denote mean zero random errors. We generate these random errors e_{ijk}, in (4-3), independently from $\mathrm{N}(0, 0.05^2)$ distribution under the assumption that all the genes in the datasets are independent.

In our simulations, we set 70 genes as differentially expressed between the two groups of subjects. In particular, we consider the differences in magnitudes of differential (log) expressions of these 70 genes between the two groups to be 15. For this, the interaction effects between the genes and the groups are generated as given below:

For $1 \leq i \leq 20$, $(GV)_{i1} = -7.5,\ (GV)_{i2} = 7.5$

For $21 \leq i \leq 70$, $(GV)_{i1} = 7.5,\ (GV)_{i2} = -7.5$

For $71 \leq i \leq G$, $(GV)_{i1} = \ (GV)_{i2} = 0$

This data generation is repeated for each of the $M = 10$ independent experiments.

Scheme 2. In this scheme, we want to evaluate the performance of EAMA in the presence of hidden confounder in the model. The effects of the latent variable (W) in (4-3) are generated in this scheme in such a way that it varied not only over the two groups of subjects and different groups of genes but also over different experiments. So, here we generate W_{ijk} as $W_{ijk} = u_{ijk}I(s_{ijk} = 1)$, where $s_{ijk} \sim$ Bernoulli(0.4). When $s_{ijk} = 0$, $W_{ijk} = 0$, which implies that there is no effect of the latent confounder on the i^{th} gene of the k^{th} subject in the j^{th} group. On the other hand, when $s_{ijk} = 1$, $W_{ijk} = u_{ijk}$, i.e., the effect of the latent confounder is given by u_{ijk} for the i^{th} gene of the k^{th} subject in the j^{th} group. Here, u_{ijk} is generated depending on the gene, subject group, and experiment ID (m) as follows:

$$u_{i1k} \sim \begin{cases} N(-3+m, 0.01^2) & \text{for } 1 \le i \le 20; \quad k = 1,2,\dots,10 \\ N(2+m, 0.01^2) & \text{for } 21 \le i \le 70; \quad k = 1,2,\dots,10 \\ N(10+m, 0.01^2) & \text{for } 71 \le i \le 1000; \quad k = 1,2,\dots,10 \end{cases}$$

and (4-4)

$$u_{i2k} \sim \begin{cases} N(3+m, 0.01^2) & \text{for } 1 \le i \le 20; \quad k = 11,12,\dots,20 \\ N(15+m, 0.01^2) & \text{for } 21 \le i \le 70; \quad k = 11,12,\dots,20 \\ N(20+m, 0.01^2) & \text{for } 71 \le i \le 1000; \quad k = 11,12,\dots,20 \end{cases}$$

Generation of W_{ijk} using the above design represents a situation where the gene expression values depend on the group status of the subjects and also on some unobserved features of the experiment (for example, age and gender of the subjects, geographical location of the experiment, etc.) which are often present in observational studies. Apart from W_{ijk}, all the variables in (4-3) are generated in the same way as in Scheme 1 for each of the M independent experiments.

After generating the microarray datasets, the set of differentially expressed genes are identified for each of the M experiments using "limma" in Bioconductor [75] and the corresponding raw p-values of all the genes under study are stored. Note that while identifying the set of differentially expressed genes using "limma" we do not consider the effect of any unmeasured/hidden confounding factors that may be present in the simulation model. This is because, although affecting the outcome, these factors remain unaccounted in practice, the very reason that they are labelled as "hidden" or "latent". We then apply our method EAMA to obtain the set of significant genes.

We consider the results from the setting involving 10 independent experiments and 1000 uncorrelated genes where 70 genes are differentially expressed. The difference in magnitudes of the (log) expressions of these 70 genes between the two groups is considered to be 15.

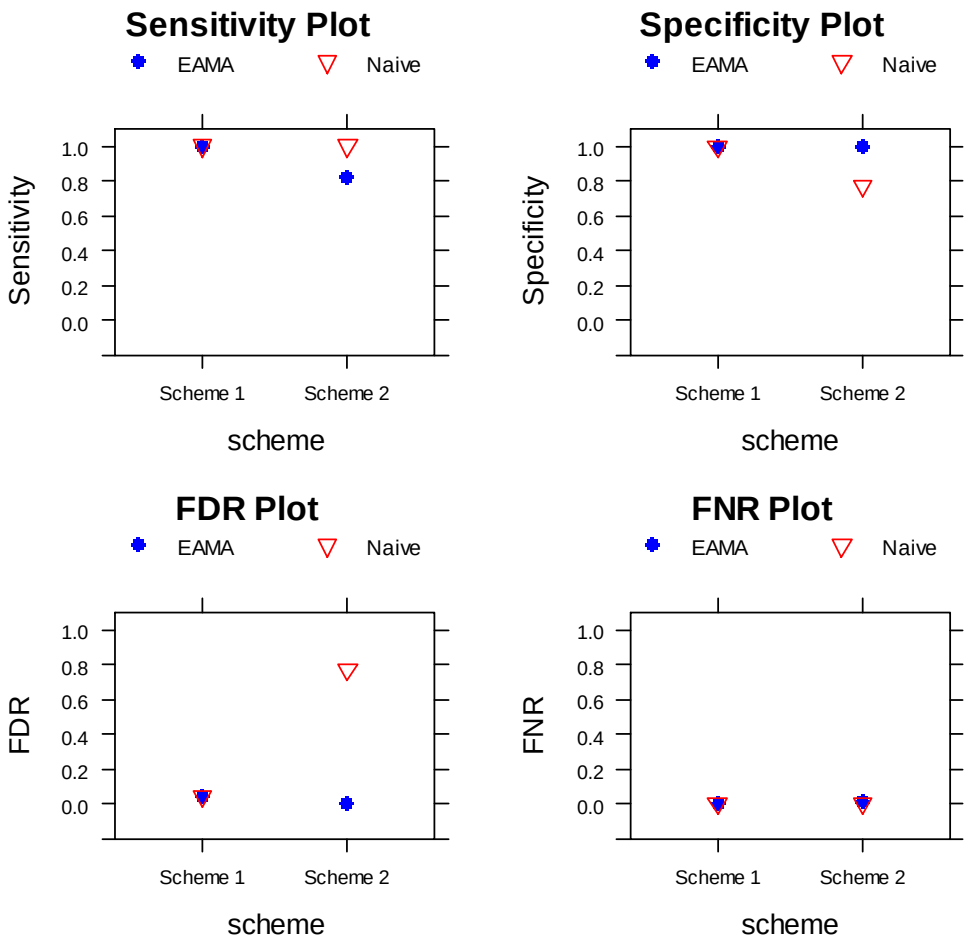

Figure 4-1. Performance assessment with 10 experiments, 1000 uncorrelated genes and absolute differences in differential expressions as 15.

Figure 4-1 shows the plots of sensitivity, specificity, FDR, and FNR of EAMA and that of the naïve method for each of the two simulation schemes. From Figure 4-1, we find that the performances of EAMA and the naïve method are very similar in Scheme 1 (no hidden confounder in the model) in terms of all the four performance measures. However, there is a wide difference between the performances of the two competing methods in Scheme 2 (presence of hidden confounder in the model) as evident from Figure 4-1. The sensitivity of the naïve method is better than that of EAMA, while the specificity of EAMA is much better than that of the naïve method. The most drastic difference is observed in the FDR measure in Scheme 2.

Here EAMA outperforms the naïve method by a large margin in terms of FDR, as the FDR of the naïve method is unacceptably high compared to that of EAMA.

From this simulation study, it appears that the performances of EAMA and the naïve method are similar when there is no latent factor present in the study. But in the presence of some hidden variables which act as confounders and have significant effects on the expression levels of the genes, our method (EAMA) performs reasonably well in terms of having low FDR whereas the naïve method has unacceptably high FDR values. This result justifies our expectation that the EAMA, based on the empirical null distribution adjustment, lead to an accurate inference by accounting for the excess variations caused by the latent factors which are missed by the theoretical null based naïve method.

We also consider some variations of the above simulation schemes by introducing correlation among some of the genes under study, changing the number of independent experiments, varying the number of genes involved, changing the magnitude of differential expression among the set of differentially expressed genes. We also consider a simulation setting by introducing an effect of a hidden variable which does not act as a confounder. Each of the above mentioned schemes are described below.

4.2.1.1 Correlated gene expression levels

In reality, genes having same biological functions are expected to have correlated expression values in the datasets. To study the robustness of EAMA, we generate a simulation setting where there are 4 clusters of correlated genes. We use the same model (4-3) in generating the (log) expression values, but assure correlation among the genes within a cluster through the generation of the random error term (e_{ijk}) in the following way:

Let $C_1 = \{1, 2, \dots, 10\}, C_2 = \{41, 42, \dots, 60\}, C_3 = \{501, 502, \dots, 520\}$ and $C_4 = \{891, 892, \dots, 910\}$ denote the 4 clusters and $C = \{C_1, C_2, C_3, C_4\}$ denotes the union of the 4 clusters.

We generate e_{ijk} as

$$e_{ijk} = \begin{cases} \frac{1}{\sqrt{2}} e^1_{ijk} + \frac{1}{\sqrt{2}} e^2_{ijk} & if\ i \epsilon C \\ e^2_{ijk} & o.w \end{cases} \quad (4\text{-}5)$$

where e^1 are drawn independently from $N(0, 1)$ in such a way that the values of e^1 are same for all the genes belonging to the same cluster. e^2 are generated independently from $N(0, 2^2)$. All the other variables in (4-3) are generated in two different ways, similar to the studies with independent genes, namely Scheme 1 and Scheme 2.

After generating the microarray datasets with correlated gene expression values, we apply our method EAMA to obtain the set of significant genes as before. The results with correlated genes are found to be similar to what we obtained from the study of uncorrelated genes where EAMA performed much better than the naïve method in terms of FDR in the presence of latent factors in the study (see Figure 4-2).

4.2.1.2 Reduction of the number of experiments (M)

We consider simulation settings with reduced number of experiments. In particular, we consider two other choices of the number of experiments M, which are 8 and 4. The (log) expression values of the genes are generated using (4-3) in the same way as before under the two schemes, namely, Scheme 1 and Scheme 2.

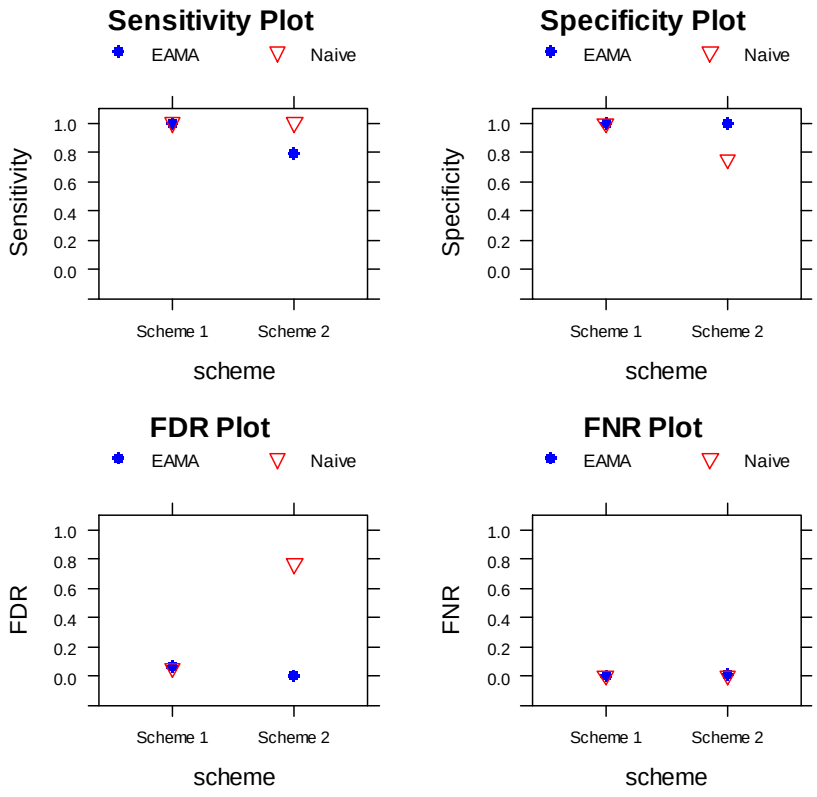

Figure 4-2. Performances of EAMA and the naïve method with 10 experiments, 1000 correlated genes and absolute differences in differential expressions as 15.

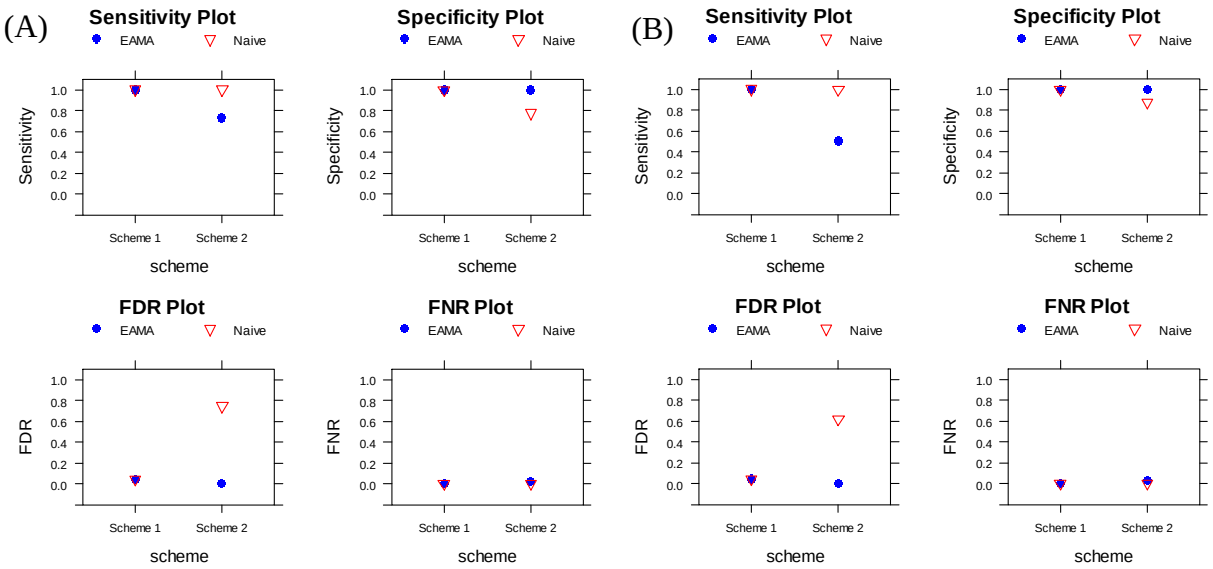

Figure 4-3. Performances of EAMA and the naïve method with 1000 uncorrelated genes, absolute differences in differential expressions as 15, and (A) number of experiments was 8 (B) number of experiments was 4.

Overall, we find that the relative performances of EAMA and the naïve method are similar for different choices of M, the number of experiments, although the performance of the naïve method gets progressively worse and that of EAMA gets better as the number of experiments increases (see Figures 4-3 (A)-(B)).

4.2.1.3 Reduction in the difference in magnitude of the expression levels of the genes

Further, we consider the situation where the magnitude of the difference in the (log) expression levels of the 70 differentially expressed genes between the two groups is reduced. This is reflected through the reduction of magnitudes of the interaction effects between genes and groups in (4-3). Here the interaction effects (GV) are generated as given below:

For $1 \leq i \leq 20$, we set $(GV)_{i1} = -4,\ (GV)_{i2} = 4$

For $21 \leq i \leq 70$, we set $(GV)_{i1} = 4,\ (GV)_{i2} = -4$

For $71 \leq i \leq G$, we set $(GV)_{i1} = \ (GV)_{i2} = 0$

So, the difference in magnitude of the (log) expression levels of the differentially expressed genes between the two groups is 8 instead of 15 as considered in the previous setting. In this case, too, we observe similar performances of EAMA and the naïve method in both the schemes (see Figure 4-4). However, the difference between the FDR values of EAMA and the naïve method is huge where the EAMA outperforms the naïve method by a very large margin. Interestingly, it appears that the performance of the naïve method gets worse as the difference in the magnitudes of differential expression tend to decrease.

Additionally, we also reduce the number of experiments (M) to 8 and 4, where the difference in magnitude of the (log) expression levels of the differentially expressed genes between the two groups is 8. The results are given in Table 4-1. The results of the study with reduced number of experiments are similar to what we obtained with $M = 10$.

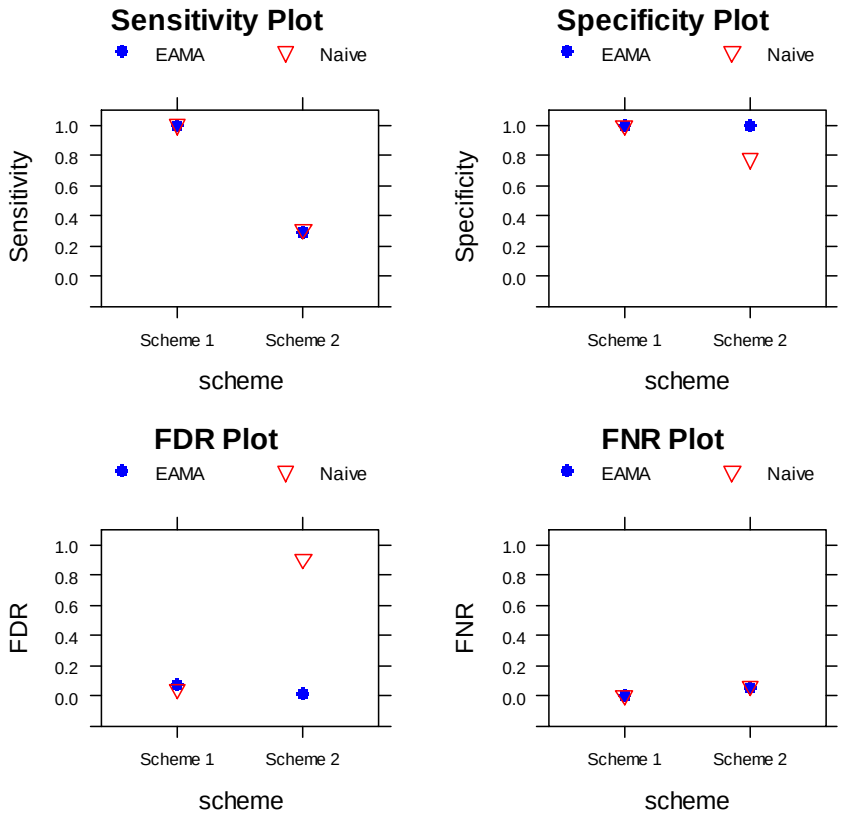

Figure 4-4. Performance assessment of EAMA and the naïve method with 10 experiments, 1000 uncorrelated genes and absolute differences in differential expressions as 8.

Table 4-1. Performances of EAMA and the naïve method with 1000 uncorrelated genes, absolute differences in differential expressions as 8 and reduced number of experiments.

Scheme	No. of experiments	Method	Sensitivity	Specificity	FDR	FNR
		EAMA	1	0.993	0.078	0
	8	Naïve	1	0.996	0.047	0
1		EAMA	1	0.994	0.067	8.63e-06
	4	Naïve	1	0.996	0.045	1.29e-05
		EAMA	0.286	1	0.005	0.051
	8	Naïve	0.315	0.809	0.889	0.06
2		EAMA	0.286	0.999	0.046	0.051
	4	Naïve	0.311	0.907	0.793	0.054

4.2.1.4 Increase in the number of genes (G)

We consider a data generation through (4-3) where the number of genes (G) involved is increased to 5000 such that the proportion of differentially expressed genes is reduced. The

simulation is designed in such a way that, out of these 5000 genes, 200 genes are differentially expressed. So the proportion of overall differentially expressed genes is 4%, which is lower than our previously discussed setting. Here the interaction effects (GV) are generated as given below:

For $1 \leq i \leq 50$, we set $(GV)_{i1} = -7.5,\ (GV)_{i2} = 7.5$

For $51 \leq i \leq 200$, we set $(GV)_{i1} = 7.5,\ (GV)_{i2} = -7.5$

For $201 \leq i \leq G$, we set $(GV)_{i1} = \ (GV)_{i2} = 0$

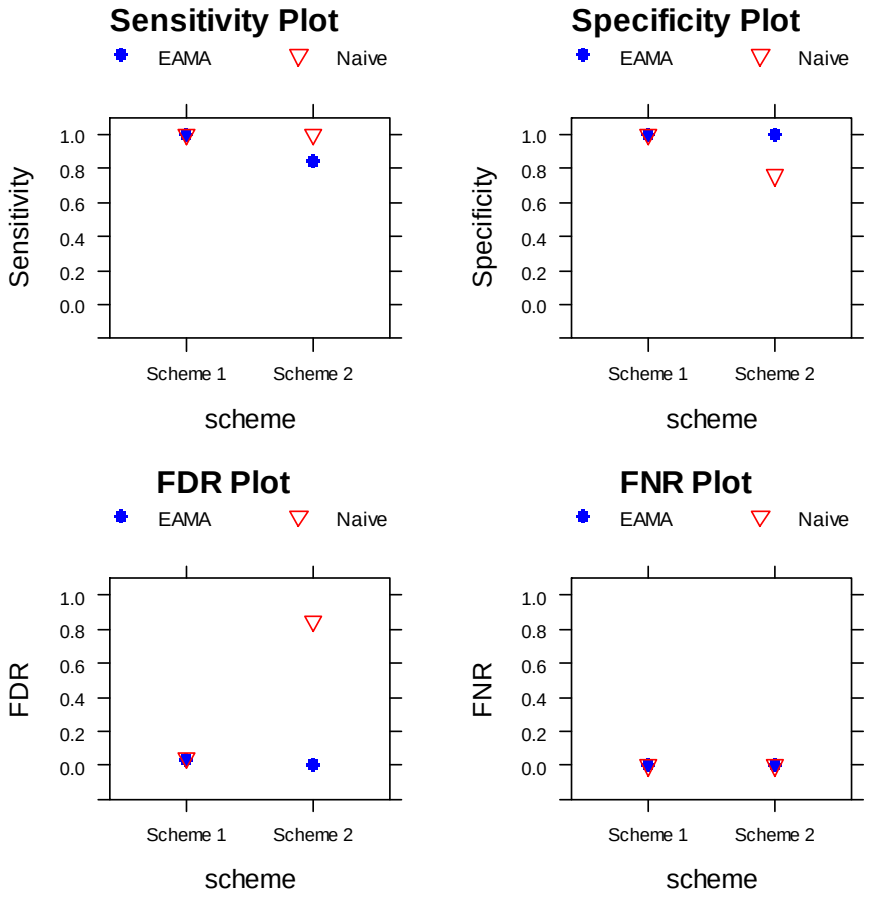

Figure 4-5. Performance assessment of EAMA and the naïve method with 10 experiments, 5000 uncorrelated genes and absolute differences in differential expressions as 15.

As before, we consider two different ways of generating the latent factor (W) leading to the two simulation schemes, namely, Scheme 1 and Scheme 2. The performances of EAMA and the naïve methods are similar to what we obtained from the study of 1000 genes (see Figure 4-5) where EAMA largely outperforms the naïve method in terms of FDR.

Additionally, the performances of the two methods are compared after reducing the number of experiments and also the difference in magnitude of the interaction effects between genes and groups. Overall, we find that there is not much difference in the performances of the two methods when the number of genes is increased from 1000 to 5000 in both the schemes.

4.2.1.5 Presence of a hidden variable which does not act as a confounder

In addition to Scheme 1 and Scheme 2, we also check the situation where a hidden variable, although present and affects the outcome, does not act as a confounder. We refer to this scheme as Scheme 3. Here we consider a simulation setting where the distribution of the latent variable (W) in (4-3) is the same in the case and the control groups of subjects i.e. the effect of the hidden variable on the outcome does not vary significantly between the two groups of subjects. Here, $W_{ijk} = u_{ijk} I(s_{ijk} = 1)$, where $s_{ijk} \sim \text{Bernoulli}(0.4)$ and u_{ijk} is generated for the m^{th} experiment as :

$$u_{ijk} \sim \begin{cases} N(2+m, 0.1^2) & \text{for } 1 \le i \le 20; \quad k = 1,2,\dots,20 \\ N(2+m, 0.1^2) & \text{for } 21 \le i \le 70; \quad k = 1,2,\dots,20 \\ N(2+m, 0.1^2) & \text{for } 71 \le i \le 1000; \quad k = 1,2,\dots,20 \end{cases} ; \quad j = 1,2$$

The differences in magnitudes of differential (log) expressions of the 70 differentially expressed genes between the two groups are 2. We found that in Scheme 3, where the latent variable no longer acts as a confounder, EAMA has higher sensitivity than the naïve method while the naïve method appears a bit conservative with low sensitivity and FDR (see Table 4-2).

Table 4-2. Performance assessment of EAMA and the naïve method where a hidden variable does not act as a confounder.

Method	Sensitivity	Specificity	FDR	FNR
EAMA	0.517	0.995	0.0963	0.0351
Naive	0.337	0.999	0.023	0.0476

4.2.2 Generation of count data (NGS based gene expression)

We also generate realistic NGS-like datasets for our simulation experiments using a popular NGS-simulator called SimSeq [132]. SimSeq generates read counts in two treatment groups for a known set of differentially expressed genes based on a real RNA-sequencing dataset. Here, we generate a count data using SimSeq and kidney renal clear cell carcinoma data as the source dataset [133]. The kidney renal clear cell carcinoma dataset consists of 20,531 genes and 144 paired samples with tumor and non-tumor replicate. We filter the kidney renal clear cell carcinoma dataset by including only those genes which have more than one hundred non-zero read counts so that the simulated dataset did not include all zero read counts. From the reduced source dataset we generate read counts with $G = 5000$ genes and $N = 60$ subjects distributed equally over two groups. Out of these 5000 genes, 1000 genes are considered to be differentially expressed.

Similar to the previous simulated settings involving expression datasets, we also assume that there exists an effect of a hidden confounder in the count dataset. In order to achieve this, we independently generate another set of read counts using SimSeq with 5000 genes and 60 subjects as before where 1000 genes are differentially expressed. For the i^{th} gene, j^{th} group and k^{th} subject we generate a random observation s_{ijk} from $\text{Bernoulli}(0.4)$, $i = 1, \dots, 5000,\ j = 1,2,$ and $k = 1, \dots, 60$. When $s_{ijk} = 1$, we add the two read counts and divide the resulting sum by two in order to maintain the original magnitude of the read counts, for the i^{th} gene in the j^{th} group for the k^{th} subject. We then round the result to nearest integer. If $s_{ijk} = 0$, we retain

the original read count for the i^{th} gene, j^{th} group and k^{th} subject. In this way, s_{ijk} determine whether there exists an effect of the hidden variable on the i^{th} gene in the j^{th} group for the k^{th} subject. We repeat this whole process for $M = 10$ experiments.

After generating the count datasets, the set of differentially expressed genes are identified for each of the $M = 10$ experiments using edgeR in Bioconductor [134, 135] and the corresponding raw p-values of all the genes under study are stored. Then we apply our method EAMA to obtain the set of significant genes. The performances of our method (EAMA) and that of the naïve meta-analysis method are assessed using the simulated count datasets. The results are shown in Table 4-3. From this table we find that the EAMA and naïve method have similar performances in terms of sensitivity, specificity, and FNR, while the major difference lies in the FDR. The FDR of EAMA is much lower than that of the naïve method, and hence EAMA has less false discoveries. The results are similar to what we obtained in the studies involving continuous datasets.

Table 4-3. The performances of EAMA and that of the naïve method using the simulated count datasets.

Method	Sensitivity	Specificity	FDR	FNR
EAMA	0.870	0.968	0.129	0.032
Naïve	0.904	0.942	0.204	0.025

4.3 Data Analysis

We consider five publicly available lung cancer gene expression datasets: Bhattacharjee [136], GSE11969 [137], GSE29016 [138], GSE30219 [139] and GSE43580 [140]. These datasets are previously analyzed by [141] on a classification framework. Each of the datasets is normalized and filtered by [141]. All the five datasets are merged so that each of them has the same set of genes. We use the processed and merged datasets, which are available at https://zenodo.org/record/16006.

Each dataset has normalized expression levels for 7200 genes. Although, information regarding lung cancer type, smoking status, age and gender for the patients is available for our analysis, we only use the information about the cancer type of the patients. Here, we attempt to identify the set of genes, which are differentially expressed between the two lung cancer types: Adenocarcinoma (AD) and Squamous cell carcinoma (SQ). The number of patients in each of the two types of lung cancer within each dataset is shown in Table 4-4.

Table 4-4. The number of patients in each of the two lung cancer types within each dataset.

Dataset	Lung cancer type	
	Adenocarcinoma (AD)	Squamous cell carcinoma(SQ)
Bhattacharjee	60	21
GSE11969	90	35
GSE29016	38	12
GSE30219	85	61
GSE43580	77	73

We fit linear model with the gene expression values of the patients as the response and cancer type of the patients as the predictor, for each experiment separately. Using "limma" in Bioconductor [75] we obtain the p-values corresponding to the main effect term for the lung cancer type. So, we have five sets of p-values for each of the 7200 genes.

We identify the set of differentially expressed genes based on the five experiments using both EAMA and the naïve method. Figure 4-6 shows the histograms of the original z-scores and the modified z-scores (after empirically adjusting the z-scores), obtained using (4.1) and (4.2) respectively, given in the *Methods* section. The curves superimposed on the histograms in Figure 4-6 are the density curves of $N(0,1)$ distribution.

From Figure 4-6, it can be seen that the density of the original z-scores is much deviated from the $N(0,1)$ distribution, whereas the density of the modified z-scores is almost identical to the $N(0,1)$ distribution. So, Figure 4-6 highlights the fact that the test based on the assumption

that the original z-scores are standard normally distributed would lead to incorrect inference and possibly large amount of false discoveries. Hence, it is not surprising that the naïve method, that used the original z-scores, identified 5127 differentially expressed genes (more than 70% of the total number of genes), even after adjusting for the "Benjamini-Hochberg" multiplicity correction method [66]. On the other hand, our proposed method EAMA identified 1541 significantly differentially expressed genes (approximately 21% of the total number of genes) after adjusting for "Benjamini-Hochberg" method of multiplicity correction [66], hence reducing the possibility of gross false discoveries.

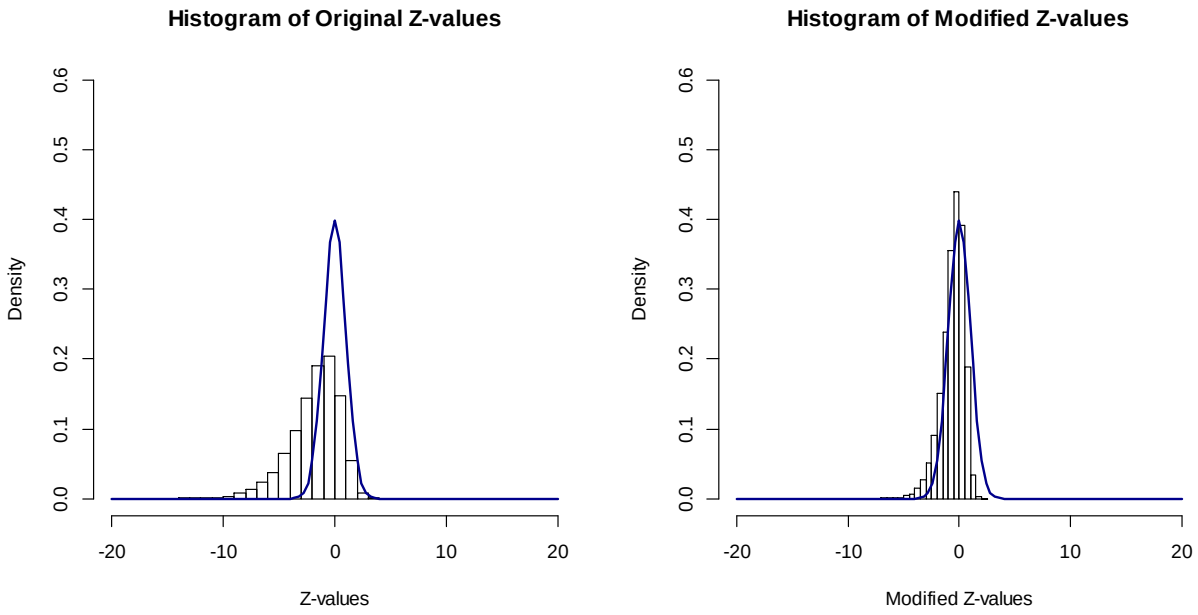

Figure 4-6. The histograms of the original and modified z-scores for the lung cancer datasets.

The genes identified by the methods described above could be further analyzed using pathway analyses of cancer signaling pathways. Several studies had identified 140 "driver" genes that can promote tumor formation if affected by intragenic mutations. These "driver genes", known to be directly or indirectly responsible for tumor formation, are classified into

twelve signaling pathways [60]. These twelve signaling pathways are: TGF − β, MAPK, STAT, PI3K, RAS, Cell Cycle/Apoptosis, NOTCH, HH, APC, Chromatin modification, Transcriptional regulation and DNA damage control. In order to check the association of the significant genes identified through our analysis with these signaling pathways, we functionally group all the significant genes, identified by EAMA as well as the naïve method, using DAVID [67, 142].

Eight out of the twelve possible pathways have been identified through the significant genes obtained from EAMA as well as the naïve method. Table 4-5 shows the grouping of the genes according to the above mentioned twelve signaling pathways. Out of the eight signaling pathways, given in Table 4-5, "Cell Cycle" and "NOTCH" signaling pathways are significant among the genes identified by EAMA at a p-value cutoff of 10%, while "Cell Cycle" and "TGF − β" signaling pathways are significant among the genes identified by the naïve method at 10% significance level. Although the number of genes identified by the naïve method is very large, number of significant signaling pathways, obtained through the identified genes, is same for the naïve method and EAMA. From Table 4-5, we can see that 44 of the genes, identified by EAMA, are in the "Cell Cycle" signaling pathway, whereas 80 of the genes, identified by the naïve method, are in the "Cell Cycle" signaling pathway. All these 80 genes, identified by the naïve method, also include the 44 genes identified by EAMA in the "Cell Cycle" signaling pathway.

We further study some of the genes that have been identified by the naïve method but not by EAMA. For example, the gene with ID 472 is identified by the naïve method but not by EAMA. We study in details the expression pattern of this gene in each of the five datasets. Figure 4-7 shows the violin plots of the gene with ID 472 for the two cancer types in each of the five datasets.

Table 4-5. The grouping of the genes, identified by EAMA and the naïve method, according to the signaling pathways.

Signaling pathway	EAMA	Naive
TGF − β	12	40
MAPK	21	88
STAT	5	34
PI3K	26	99
RAS	22	79
Cell Cycle	44	80
NOTCH	10	17
HH	5	14

From Figure 4-7, we find that the gene with ID 472 is not differentially expressed between the two lung cancer types in four of the five datasets. This suggests that the gene with ID 472, in spite of being identified by the naïve method, is unlikely to be an important factor for the discrimination between the two cancer types. Also, based on the individual analyses, the p-value corresponding to this gene is significant only for the dataset "GSE30219".

Also, from Table 4-5, we can see that there are 40 genes in "TGF − β" signaling pathway under the naïve method. Among these 40 genes, 12 genes are also identified by EAMA as significant whereas the remaining 28 genes are not identified as significant by our proposed method EAMA. For example, the gene with ID 8200 is identified as significant by the naïve method but not by EAMA. Fig 4-8 shows the violin plots for the gene with ID 8200 for the two cancer types in each of the five datasets.

From Figure 4-8, we can see that the gene with ID 8200 is not differentially expressed between the two lung cancer types for four of the five datasets. As before, we can suggest that this gene do not have a considerable impact in distinguishing the two cancer types, and it is reasonable that our proposed method EAMA has not identified such "unimportant" genes.

Figure 4-7. The violin plots of the gene with ID 472 for the two cancer types in each of the five datasets.

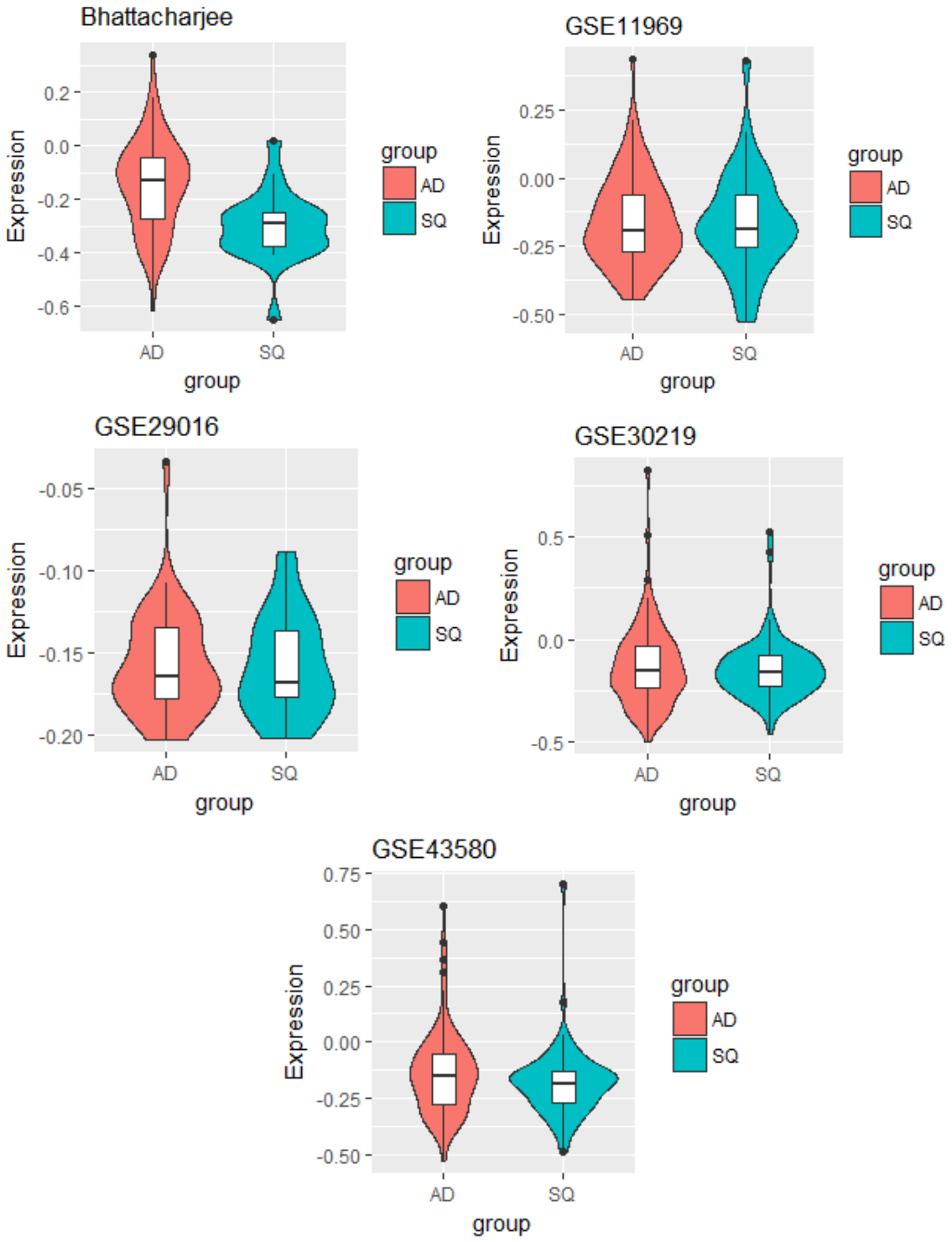

Figure 4-8. The violin plots for the gene with ID 8200 for the two cancer types in each of the five datasets

4.4 Discussion

High throughput technologies have enabled simultaneous analysis of thousands of genes in a single experiment. Combining hypotheses testing results from the multiple genomic experiments is a popular meta-analysis approach of identifying significant genes related to some biological process. However, there is a distinct difference between the aims of meta-analyses involving single hypothesis in each component experiment with that of large-scale multiple hypotheses in each experiment. While the former targets in favoring the alternative 'interesting' hypothesis with high power, the latter is designed for the identification of a small proportion of 'interesting' or 'significant' cases out of a large set of possible candidates. In this project we discuss the possibility of making erroneous conclusions from combining the p-values calculated under standard theoretical assumptions from multiple genomic experiments when each experiment involves simultaneous testing of enormous number of hypotheses in presence of some hidden confounder variable. The presence of some confounder variables can induce over-dispersion and/or bias that remain unaccounted by the theoretical null assumptions. In particular, we show that even adjusting the p-values by taking into account the false discovery rates may not be enough to substantially diminish the false discoveries from the meta-analysis of large-scale multiple testing when hidden confounder variables are present. We propose an alternative approach of modifying the p-values by constructing empirical null distributions and combining these empirically adjusted p-values through proper meta-analysis approach. Through simulation studies and real datasets involving genomic experiments we show that our proposed method has much better performance than the standard meta-analysis approach of combining raw p-values from multiple experiments especially in presence of hidden confounder variables.

This article mainly focuses on developing meta-analysis approach for combining p-values from multiple genomic experiments when the outcome of interest is affected by some hidden

variable that acts as a confounder with different effects between different groups under study. In biological studies, mostly in observational studies, such hidden variables are very common which play the roles of confounders. We show that in such scenarios our proposed method performs much better than the standard naïve method of meta-analysis. However, it may happen in certain situations that a hidden variable affects the outcome but does not have a confounding effect, i.e. the effect of the hidden variable on the outcome does not vary significantly between the different groups under study. We consider multiple simulation settings that addresses the abovementioned scenarios and the results show that the standard meta-analysis approach of combining raw p-values is a bit conservative having lower power and FDR than our proposed method EAMA.

We use empirical null distribution as outlined by [63] where the empirical null distribution are estimated using a "central matching" approach (see section 4.1.2 as well as [131] for details) of the R package "locfdr" [130]. There is an alternative option of estimating the null distribution using the maximum likelihood method in the "locfdr" package (Details can be found in [131]). In various simulation settings, the "central matching" approach of estimating the empirical null distribution appears to perform better than the maximum likelihood approach. However, the overall results obtained through the maximum likelihood approach are not significantly different from that obtained from the "central matching" approach and are not shown here in details. One may note that although we have used FDR-adjusted p-values using Benjamini-Hochberg procedure [66], there are other options for adjustment of false discovery rates including the use of q-values [143].

APPENDIX A
R-SCRIPTS

A.1 R-Script for Identification of Master Regulator Transcription Factor

```
library(dna)

library(MASS)

library(Kendall)


pvalue_boot_tau <- NULL

for(z in 1:500){

### case group

mu_u <- rep(0,10)

Sigma=diag(10)

u <- mvrnorm(100,mu_u,Sigma)

colnames(u) <- paste("u",1:10,sep="")


## TFs 1 to 10


x1 <- rnorm(100,mean=50,sd=1)

rho <- c(0,0.95,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1)

s <- x1

for(i in 2:10){

xi <- (rho[i]*x1+u[,i])/sqrt(1+(rho[i])^2)

s <- cbind(s,xi)

}
```

```
colnames(s) <- paste("TF",1:10,sep="")          ## TF1 to TF10
round(cor(s),4)

## Genes 1 to 30
delta <- 1                                      ## Delta
r1 <- c(0.45,0.4,0.35,0.3,0.25,0.2,0.15,0.1,0.05)
r2 <- (1-delta)*r1
gamma1 <- 0.5+(delta^2)*r1[1]

genes1to30 <- NULL
for(i in 1:30){
e1 <- rnorm(100)
g1 <- x1*gamma1+e1
genes1to30 <- cbind(genes1to30,g1)
}
colnames(genes1to30) <- paste("G",1:30,sep="")        ## Genes1 to 30

m <- c(10,10,10,10,10,10,5,5,5)
r1 <- c(0.45,0.4,0.35,0.3,0.25,0.2,0.15,0.1,0.05)
oth_genes <- NULL
for(i in 2:10){
for(j in 1:m[i-1]){
e <- rnorm(100)
```

```
gj <- u[,i]*r1[i-1]+e
oth_genes <- cbind(oth_genes,gj)
}
}
colnames(oth_genes) <- paste("G",31:105,sep="")        ## Remaining genes

pop_exp <- cbind(s,genes1to30,oth_genes)              ## case group : TFs, Genes

### control group
mu_u <- rep(0,10)
Sigma=diag(10)
u <- mvrnorm(70,mu_u,Sigma)
colnames(u) <- paste("u",1:10,sep="")

## TFs 1 to 10

x1 <- rnorm(70,mean=5,sd=1)
rho <- c(0,0.95,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1)
s <- x1
for(i in 2:10){
xi <- (rho[i]*x1+u[,i])/sqrt(1+(rho[i])^2)
s <- cbind(s,xi)
}
```

```
colnames(s) <- paste("TF",1:10,sep="")              ## TF1 to TF10
round(cor(s),4)

## Genes 1 to 30

gamma2 <- 0.5
genes1to30 <- NULL
for(i in 1:30){
e1 <- rnorm(70)
g1 <- x1*gamma2+e1
genes1to30 <- cbind(genes1to30,g1)
}
colnames(genes1to30) <- paste("G",1:30,sep="")          ## Genes 1 to 30

m <- c(10,10,10,10,10,10,5,5,5)
r2 <- (1-delta)*r1
oth_genes <- NULL
for(i in 2:10){
for(j in 1:m[i-1]){
e <- rnorm(70)
gj <- u[,i]*r2[i-1]+e
oth_genes <- cbind(oth_genes,gj)
}
```

```
}
colnames(oth_genes) <- paste("G",31:105,sep="")          ## Remaining genes

pop_ctl <- cbind(s,genes1to30,oth_genes)                 ## control group: TFs and genes

#############################################################################
TFs <- paste("TF",1:10,sep="")
genes <- paste("G",1:105,sep="")

## Measuring differential connectivity scores for all TFs
s <- NULL
for(l in 1:length(TFs)){
sub <- c(TFs[l],genes)
pop_exp_new <- pop_exp[,colnames(pop_exp)%in%sub]
pop_ctl_new <- pop_ctl[,colnames(pop_ctl)%in%sub]
s1=cornet(pop_exp_new,rescale.scores=FALSE)
s2=cornet(pop_ctl_new,rescale.scores=FALSE)
s[l] <- sum(abs(s1[,TFs[l]]-s2[,TFs[l]]))/(length(genes))
}
DEgenes_pop <- data.frame(TFs,s)
DEgenes_pop
TfTestStat_pop=DEgenes_pop[with(DEgenes_pop, order(-DEgenes_pop$s)),]
TfTestStat_pop
```

```
## Pearson's correlation between each pair of TFs
data_req_pop=rbind(pop_exp,pop_ctl)
TFexpress_pop=data_req_pop[,which(colnames(data_req_pop)%in%TFs)]
cormatTF_pop=cor(TFexpress_pop,method="pearson")

## Kendall's statistic for each TF
stat_tau=NULL
for(t in 1:length(TFs)){
TranInterested=as.character(TfTestStat_pop[t,1])
orderCorrTF=sort(abs(cormatTF_pop[TranInterested,]),decreasing=TRUE)
d1=data.frame(gene=TfTestStat_pop[,1],TfTestStat_pop)
d2=data.frame(orderCorrTF)
d22=data.frame(d2,gene=rownames(d2))
data=merge(d22, d1, by="gene")
test_tau=Kendall(data[,2], data[,4])
stat_tau[t]=test_tau$tau
}
allTFsStat_pop=data.frame(TfTestStat_pop[,1],stat_tau)
Stat_pop_tau=max(allTFsStat_pop[,2])                    ## Final test statistic K

############################  Sampling  ############################
```

```
full_pop <- rbind(pop_exp,pop_ctl)
p_tau <- NULL
for(k in 1:500){
## Bootstrap samples
x <- seq(1:170)
sam <- sample(x,170,replace=TRUE)
sam_exp=full_pop[sam[1:100],]
sam_ctl=full_pop[sam[101:170],]

## Differential connectivity scores for all TFs for a bootstrap sample
s <- NULL
for(l in 1:length(TFs)){
sub <- c(TFs[l],genes)
sam_exp_new <- sam_exp[,colnames(sam_exp)%in%sub]
sam_ctl_new <- sam_ctl[,colnames(sam_ctl)%in%sub]
s1=cornet(sam_exp_new,rescale.scores=FALSE)
s2=cornet(sam_ctl_new,rescale.scores=FALSE)
s[l] <- sum(abs(s1[,TFs[l]]-s2[,TFs[l]]))/(length(genes))
}
DEgenes_sam <- data.frame(TFs,s)
DEgenes_sam

TfTestStat=DEgenes_sam[with(DEgenes_sam, order(-DEgenes_sam$s)),]
```

```
TfTestStat

## Pearson's correlation among the TFs for a bootstrap sample
data_req=rbind(sam_exp,sam_ctl)
TFexpress=data_req[,which(colnames(data_req)%in%TFs)]
cormatTF=cor(TFexpress,method="pearson")

## Kendall's test statistic value for each TF for a bootstrap sample
stat_tau=NULL
for(t in 1:length(TFs)){
TranInterested=as.character(TfTestStat[t,1])
orderCorrTF=sort(abs(cormatTF[TranInterested,]),decreasing=TRUE)
d1=data.frame(gene=TfTestStat[,1],TfTestStat)
d2=data.frame(orderCorrTF)
d22=data.frame(d2,gene=rownames(d2))
data=merge(d22, d1, by="gene")
test_tau=Kendall(data[,2], data[,4])
stat_tau[t]=test_tau$tau
}
allTFsStat=data.frame(TfTestStat[,1],stat_tau)
Stat_obs_tau=max(allTFsStat[,2])          ## Bootstrap based test statistic value
p_tau[k] <- Stat_obs_tau
}
```

```
pvalue_boot_tau[z]=mean(p_tau > Stat_pop_tau)  ## p-value calculation based on 500 bootstrap
samples
}
datafin <- data.frame(seq(1:500),pvalue_boot_tau)
sum(datafin[,2]<0.05)
```

A.2 R-Script for Meta-Analysis of Differential Protein Expression Networks

A.2.1 Pathway-based Analysis

```
LUAD_prot <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Protein_LUAD\\GO_KEGG_LUAD.csv"
,header=T)  ## Pathway results
LUAD_prot1 <- LUAD_prot[,c(3,4)]

Prot_mod_LUAD <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Protein_LUAD\\Prot_LUAD.csv",header
=T)  ## Expression Data

Clinical <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Clinical_LUAD\\Clinical_mod_LUAD.cs
v",header=T) ## Clinical data
Clinical1 <- cbind(as.character(Clinical[,1]),ifelse(Clinical[,6]=="progression","p","cr"))

common_ids <- intersect(Clinical1[,1],Prot_mod_LUAD[,1])
Prot_Clin_LUAD <- Prot_mod_LUAD[which(Prot_mod_LUAD[,1]%in%common_ids),]
```

```
        ## Prot_LUAD dataset after merging with Clinical with 172 ids
Clin_Prot_LUAD <- Clinical1[which(Clinical1[,1]%in%common_ids),]
        ## Clinical_LUAD dataset after merging with Prot_LUAD with 172 ids

Prot_Clin_LUAD_stat <-
cbind(Prot_Clin_LUAD[,1],Clin_Prot_LUAD[,2],Prot_Clin_LUAD[,2:ncol(Prot_Clin_LUAD)])
dim(Prot_Clin_LUAD_stat)

#### pathway 1 -     PI3K-Akt signaling pathway    ##################

GR_prot <- levels(droplevels(LUAD_prot1[1,2]))
GR_prot_split <- unlist(strsplit(GR_prot,","))  ## gene list for P13K

prot_path1 <-
cbind(Prot_Clin_LUAD_stat[,1:2],Prot_Clin_LUAD_stat[,GR_prot_split[1:length(GR_prot_spli
t)]])
                    ## expression data with PI3K proteins only
dim(prot_path1)
prot_path1.cr <- subset(prot_path1,prot_path1[,2]=="cr")
dim(prot_path1.cr)
prot_path1.p <- subset(prot_path1,prot_path1[,2]=="p")
dim(prot_path1.p)
```

```
library(dna)
ourgenelist <- colnames(prot_path1)[-c(1,2)]
t=test.class.genes(prot_path1.cr[,-c(1,2)],prot_path1.p[,-c(1,2)],genelist=ourgenelist,
scores="cor",distance="abs",rescale.scores=TRUE,num.permutations=1000)
get.results(t)

par(mfrow=c(1,2))
s=cornet(prot_path1.cr[,-c(1,2)],rescale.scores=TRUE)
network.modules(s,m=3,epsilon=.7,plot=TRUE,interactive=FALSE,main="complete
remission")
s=cornet(prot_path1.p[,-c(1,2)],rescale.scores=TRUE)
network.modules(s,m=3,epsilon=.7,plot=TRUE,interactive=FALSE,main="progression")

#### pathway 2 -       Cell cycle      ##################

GR_prot <- levels(droplevels(LUAD_prot1[18,2]))
GR_prot_split <- unlist(strsplit(GR_prot,","))  ## gene list for Cell cycle

prot_path2 <-
cbind(Prot_Clin_LUAD_stat[,1:2],Prot_Clin_LUAD_stat[,GR_prot_split[1:length(GR_prot_spli
t)]])
                        ## expression data with Cell cycle proteins only
dim(prot_path2)
```

```
prot_path2.cr <- subset(prot_path2,prot_path2[,2]=="cr")
dim(prot_path2.cr)
prot_path2.p <- subset(prot_path2,prot_path2[,2]=="p")
dim(prot_path2.p)

ourgenelist <- colnames(prot_path2)[-c(1,2)]
t=test.class.genes(prot_path2.cr[,-c(1,2)],prot_path2.p[,-c(1,2)],genelist=ourgenelist,
scores="cor",distance="abs",rescale.scores=TRUE,num.permutations=1000)
get.results(t)

#### pathway 3 -      Ras signaling pathway      ##################

GR_prot <- levels(droplevels(LUAD_prot1[32,2]))
GR_prot_split <- unlist(strsplit(GR_prot,","))  ## gene list for Ras signaling pathway

prot_path3 <-
cbind(Prot_Clin_LUAD_stat[,1:2],Prot_Clin_LUAD_stat[,GR_prot_split[1:length(GR_prot_spli
t)]])
                        ## expression data with Ras signaling pathway proteins only
dim(prot_path3)
prot_path3.cr <- subset(prot_path3,prot_path3[,2]=="cr")
dim(prot_path3.cr)
prot_path3.p <- subset(prot_path3,prot_path3[,2]=="p")
```

```
dim(prot_path3.p)

ourgenelist <- colnames(prot_path3)[-c(1,2)]
t=test.class.genes(prot_path3.cr[,-c(1,2)],prot_path3.p[,-c(1,2)],genelist=ourgenelist,
scores="cor",distance="abs",rescale.scores=TRUE,num.permutations=1000)
get.results(t)

#### pathway 4 -       MAPK signaling pathway      ##################

GR_prot <- levels(droplevels(LUAD_prot1[56,2]))
GR_prot_split <- unlist(strsplit(GR_prot,","))  ## gene list for  MAPK signaling pathway

prot_path4 <-
cbind(Prot_Clin_LUAD_stat[,1:2],Prot_Clin_LUAD_stat[,GR_prot_split[1:length(GR_prot_spli
t)]])
                        ## expression data with  MAPK signaling pathway proteins only
dim(prot_path4)
prot_path4.cr <- subset(prot_path4,prot_path4[,2]=="cr")
dim(prot_path4.cr)
prot_path4.p <- subset(prot_path4,prot_path4[,2]=="p")
dim(prot_path4.p)

ourgenelist <- colnames(prot_path4)[-c(1,2)]
```

```
t=test.class.genes(prot_path4.cr[,-c(1,2)],prot_path4.p[,-c(1,2)],genelist=ourgenelist,
scores="cor",distance="abs",rescale.scores=TRUE,num.permutations=1000)
get.results(t)

#### pathway 5 -       Apoptosis      ##################

GR_prot <- levels(droplevels(LUAD_prot1[71,2]))
GR_prot_split <- unlist(strsplit(GR_prot,","))  ## gene list for  Apoptosis

prot_path5 <-
cbind(Prot_Clin_LUAD_stat[,1:2],Prot_Clin_LUAD_stat[,GR_prot_split[1:length(GR_prot_spli
t)]])
                              ## expression data with  Apoptosis proteins only
dim(prot_path5)
prot_path5.cr <- subset(prot_path5,prot_path5[,2]=="cr")
dim(prot_path5.cr)
prot_path5.p <- subset(prot_path5,prot_path5[,2]=="p")
dim(prot_path5.p)

ourgenelist <- colnames(prot_path5)[-c(1,2)]
t=test.class.genes(prot_path5.cr[,-c(1,2)],prot_path5.p[,-c(1,2)],genelist=ourgenelist,
scores="cor",distance="abs",rescale.scores=TRUE,num.permutations=1000)
get.results(t)
```

```
##########################################################################
library(RankAggreg)

x <- matrix(c("Ras signaling pathway", "MAPK signaling pathway","PI3K-Akt signaling
pathway","Apoptosis","Cell cycle","PI3K-Akt signaling pathway","Cell cycle",
"Apoptosis","Ras signaling pathway","MAPK signaling pathway", "Ras signaling
pathway","Apoptosis", "MAPK signaling pathway","PI3K-Akt signaling pathway", "Cell
cycle"), byrow=T,ncol=5)
w <- matrix(c(0.026,0.082,0.241,0.407,0.410,0.013,0.238,0.417,0.507,0.759,
0.156,0.487,0.517,0.774,0.997),byrow=T,ncol=5)

RankAggreg(x, 5, weights=w, method="CE", distance="Spearman",seed=NULL, maxIter =
1000)
```

A.2.2 Individual Protein-based Analysis

```
####LUNG Cancer

### Merging protein data with clinical

Prot_mod_LUAD <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Protein_LUAD\\Prot_LUAD.csv",header
=T)
```

```
Clinical <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Clinical_LUAD\\Clinical_mod_LUAD.csv",header=T)
Clinical1 <- cbind(as.character(Clinical[,1]),ifelse(Clinical[,6]=="progression","p","cr"))
common_ids <- intersect(Clinical1[,1],Prot_mod_LUAD[,1])

Prot_Clin_LUAD <- Prot_mod_LUAD[which(Prot_mod_LUAD[,1]%in%common_ids),]
        ## Prot_LUAD dataset after merging with Clinical with 172 ids

Clin_Prot_LUAD <- Clinical1[which(Clinical1[,1]%in%common_ids),]
    ## Clinical_LUAD dataset after merging with Prot_LUAD with 172 ids

Prot_Clin_LUAD_stat <-
cbind(Prot_Clin_LUAD[,1],Clin_Prot_LUAD[,2],Prot_Clin_LUAD[,2:ncol(Prot_Clin_LUAD)])
dim(Prot_Clin_LUAD_stat)

prot_path.cr <- subset(Prot_Clin_LUAD_stat,Prot_Clin_LUAD_stat[,2]=="cr")
dim(prot_path.cr)
prot_path.p <- subset(Prot_Clin_LUAD_stat,Prot_Clin_LUAD_stat[,2]=="p")
dim(prot_path.p)

#################### testing for individual proteins ##############################
```

```
library(dna)
tig.results1=test.individual.genes(prot_path.cr[,-c(1,2)],prot_path.p[,-c(1,2)],scores="cor",
distance="abs",rescale.scores=TRUE,num.permutations=1000)
summary(tig.results1)
proteins1 <- get.results(tig.results1)[1:50,]
dim(proteins1)
prots1 <- rownames(proteins1)

###################   pathway data ########################################

LUAD_prot <-
read.csv("C:\\Users\\sinjini8\\Desktop\\CAMDA2015\\Protein_LUAD\\GO_KEGG_LUAD.csv"
,header=T)
LUAD_prot1 <- LUAD_prot[,c(3,4)]

GR_prot1_P13K <- levels(droplevels(LUAD_prot1[1,2]))
GR_prot1_split_P13K <- unlist(strsplit(GR_prot1_P13K,","))  ## gene list for P13K
Tot_P13K1 <- length(GR_prot1_split_P13K)

P13K1 <- GR_prot1_split_P13K[GR_prot1_split_P13K%in%prots1] ## P13K proteins
significant at 10% level

GR_prot1_CC <- levels(droplevels(LUAD_prot1[18,2]))
```

```
GR_prot1_split_CC <- unlist(strsplit(GR_prot1_CC,","))  ## gene list for Cell cycle

Tot_CC1 <- length(GR_prot1_split_CC)

CC1 <- GR_prot1_split_CC[GR_prot1_split_CC%in%prots1] ## cell cycle proteins significant
at 10% level

GR_prot1_RAS <- levels(droplevels(LUAD_prot1[32,2]))
GR_prot1_split_RAS <- unlist(strsplit(GR_prot1_RAS,","))  ## gene list for Ras signaling
pathway
Tot_RAS1 <- length(GR_prot1_split_RAS)

RAS1 <- GR_prot1_split_RAS[GR_prot1_split_RAS%in%prots1] ## RAS proteins significant
at 10% level

GR_prot1_MAPK <- levels(droplevels(LUAD_prot1[56,2]))
GR_prot1_split_MAPK <- unlist(strsplit(GR_prot1_MAPK,","))  ## gene list for  MAPK
signaling pathway
Tot_MAPK1 <- length(GR_prot1_split_MAPK)

MAPK1 <- GR_prot1_split_MAPK[GR_prot1_split_MAPK%in%prots1] ## MAPK proteins
significant at 10% level
```

```
GR_prot1_Apop <- levels(droplevels(LUAD_prot1[71,2]))
GR_prot1_split_Apop <- unlist(strsplit(GR_prot1_Apop,","))  ## gene list for  Apoptosis
Tot_Apop1 <- length(GR_prot1_split_Apop)

Apoptosis1 <- GR_prot1_split_Apop[GR_prot1_split_Apop%in%prots1] ## Apoptosis proteins
significant at 10% level

par(mfrow = c(2, 2),     # 2x2 layout
oma = c(2, 2, 2, 0), # two rows of text at the outer left and bottom margin
mar = c(1, 4, 1, 3), # space for one row of text at ticks and to separate plots
mgp = c(2, 1, 0),    # axis label at 2 rows distance, tick labels at 1 row
xpd = NA)

require(grDevices)
pie_prot1 <- c(length(P13K1),length(CC1),length(RAS1), length(MAPK1),length(Apoptosis1))
names(pie_prot1) <- c("PI3K","Cell Cycle","RAS","MAPK","Apoptosis")

pie(pie_prot1,col = c("red", "blue", "green3","yellow","cyan"), radius = 0.7, main = "Lung
Adenocarcinoma")

####Rank Aggregation   ########################################

lung <- get.results(tig.results1)
```

```
kirc <- get.results(tig.results2)
hnsc <- get.results(tig.results3)

x <- matrix(c(rownames(lung),rownames(kirc),rownames(hnsc)),nrow=3,byrow=T)
w <- matrix(c(lung$p.value,kirc$p.value,hnsc$p.value),nrow=3,byrow=T)

library(RankAggreg)

R <- RankAggreg(x, 50, weights=w, method="CE",distance="Spearman",seed=NULL, maxIter
= 1000)

proteinsR <- c("EGFR","NFKB1","ATM","SRC","ACACA.ACACB","LCK","NF2",
"BCL2L11", "PRKCA", "FRAP1", "CCNB1", "PCNA", "YBX1", "CDKN2A", "YWHAE",
"FN1", "EIF4EBP1", "SMAD1", "BECN1", "DVL3", "ERBB2", "MRE11A", "ACACA",
"RAF1", "CDH2", "AXL", "KRT5", "TSC2", "EEF2", "RAD50", "PTGS2", "XBP1",
"WWTR1", "KRAS", "MACC1", "GAB2", "CLDN7", "NOTCH1", "MAP2K1", "NFE2L2",
"MET", "FOXO3", "CDH3", "STAT3", "ESR1", "CTNNA1", "AKT1.AKT2.AKT3",
"GSK3A.GSK3B", "PRKCD", "PGR")

p13k_proteins <- c(GR_prot1_split_P13K,GR_prot2_split_P13K,GR_prot3_split_P13K)
p13k_proteins_un <- unique(p13k_proteins)
Tot_P13KR <- length(p13k_proteins_un)
```

```
ras_proteins <- c(GR_prot1_split_RAS,GR_prot2_split_RAS,GR_prot3_split_RAS)
ras_proteins_un <- unique(ras_proteins)
Tot_RASR <- length(ras_proteins_un)

cc_proteins <- c(GR_prot1_split_CC,GR_prot2_split_CC,GR_prot3_split_CC)
cc_proteins_un <- unique(cc_proteins)
Tot_CCR <- length(cc_proteins_un)

apop_proteins <- c(GR_prot1_split_Apop,GR_prot2_split_Apop,GR_prot3_split_Apop)
apop_proteins_un <- unique(apop_proteins)
Tot_ApopR <- length(apop_proteins_un)

mapk_proteins <- c(GR_prot1_split_MAPK,GR_prot2_split_MAPK,GR_prot3_split_MAPK)
mapk_proteins_un <- unique(mapk_proteins)
Tot_MAPKR <- length(mapk_proteins_un)

P13K_R <- proteinsR[proteinsR%in%p13k_proteins_un]
RAS_R <- proteinsR[proteinsR%in%ras_proteins_un]
CC_R <- proteinsR[proteinsR%in%cc_proteins_un]
Apop_R <- proteinsR[proteinsR%in%apop_proteins_un]
MAPK_R <- proteinsR[proteinsR%in%mapk_proteins_un]

require(grDevices)
```

```
pie_protR <- c(length(P13K_R),length(CC_R),length(RAS_R),
length(MAPK_R),length(Apop_R))
names(pie_protR) <- c("PI3K","Cell Cycle","RAS","MAPK","Apoptosis")

pie(pie_protR,col = c("red", "blue", "green3","yellow","cyan"),
radius = 0.7, main = "Overall")
```

A.3 R-Script for EAMA

```
library(limma)
library(locfdr)
M <- 10
NG <- 1000
NSG <- 70
N <- 20

GV_con <- matrix(rep(c(rep(-7.5,20),rep(7.5,50),rep(0,NG-NSG)),10),
nrow=NG,ncol=10,byrow=FALSE)
GV_case <- matrix(rep(c(rep(7.5,20),rep(-7.5,50),rep(0,NG-NSG)),10),
nrow=NG,ncol=10,byrow=FALSE)
GV <- cbind(GV_con,GV_case)

false_prop <- NULL
DE_prop <- NULL
false_other <- NULL
```

```
DE_other <- NULL



for(i in 1:500){
Ztrans <- NULL
rpval <- NULL
for(k in 1:M){
Wijk <- NULL
eijk <- NULL
for(j in 1:20){
if(j <= 10){
s <- rbinom(NG,1,0.4)
z1 <- rnorm(20,-3+(1*k),0.01)
z2 <- rnorm(50,2+(1*k),0.01)
z3 <- rnorm(NG-NSG,10+(1*k),0.01)
z <- c(z1,z2,z3)
W <- z*s
Wijk <- cbind(Wijk,W)
e <- rnorm(NG,0,0.05)
eijk <- cbind(eijk,e)
}
else{
s <- rbinom(NG,1,0.4)
z1 <- rnorm(20,3+(1*k),0.01)
z2 <- rnorm(50,15+(1*k),0.01)
z3 <- rnorm(NG-NSG,20+(1*k),0.01)
z <- c(z1,z2,z3)
W <- z*s
Wijk <- cbind(Wijk,W)
e <- rnorm(NG,0,0.05)
eijk <- cbind(eijk,e)
}
```

```
}
yijk <- GV+Wijk+eijk
rownames(yijk) <- paste("G",1:NG,sep="")
design <- cbind(Grp1=1,Grp2vs1=c(rep(0,10),rep(1,10)))
options(digits=3)
fit <- lmFit(yijk,design)
fit <- eBayes(fit)
t <- topTable(fit,coef=2,number=NG,p.value=1)
pval <- data.frame(genes=rownames(t),t[,4])
colnames(pval) <- c("genes",paste("col",k,sep=""))
if(k==1){rpval <- pval}
else{
rpval <- merge(rpval,pval,by="genes")
}
}
rownames(rpval) <- rpval[,1]
raw_pval <- t(rpval[,-1])

t1=apply(raw_pval,2,FUN=function(x) pchisq(2*sum(-log(x)),2*M,lower.tail=FALSE))
adj_pval1=p.adjust(t1,method="BH")
Ztrans <- qnorm(raw_pval)
Zvec <- as.vector(Ztrans)
Efron <- locfdr(Zvec,bre=120,nulltype=2)
mu.hat <- Efron$fp0[5,1]
sig.hat <- Efron$fp0[5,2]
Mod_Ztrans <- (Ztrans-mu.hat)/sig.hat
pval_til <- pnorm(Mod_Ztrans)
t2=apply(pval_til,2,FUN=function(x) pchisq(2*sum(-log(x)),2*M,lower.tail=FALSE))
adj_pval2=p.adjust(t2,method="BH")

prop_meth <- adj_pval2[adj_pval2<0.05]
without_trans <- adj_pval1[adj_pval1<0.05]
```

```
DE_prop[i] <- length(prop_meth[names(prop_meth)%in%paste("G",1:NSG,sep="")])
false_prop[i] <- length(prop_meth[names(prop_meth)%in%paste("G",(NSG+1):NG,sep="")])

DE_other[i] <- length(without_trans[names(without_trans)%in%paste("G",1:NSG,sep="")])
false_other[i] <-
length(without_trans[names(without_trans)%in%paste("G",(NSG+1):NG,sep="")])
}
TP <- DE_prop
FN <- NSG-DE_prop
Sens <- mean(TP/(TP+FN))
NNSG <- NG-NSG
TN <- NNSG-false_prop
FP <- false_prop
Spec <- mean(TN/(TN+FP))
FDR <- mean(FP/(TP+FP))
FNR <- mean(FN/(TN+FN))
##
TP2 <- DE_other
FN2 <- NSG-DE_other
Sens2 <- mean(TP2/(TP2+FN2))
TN2 <- NNSG-false_other
FP2 <- false_other
Spec2 <- mean(TN2/(TN2+FP2))
FDR2 <- mean(FP2/(TP2+FP2))
FNR2 <- mean(FN2/(TN2+FN2))
```

LIST OF REFERENCES

[1] Sikdar S, Datta S. A novel statistical approach for identification of the master regulator transcription factor. BMC Bioinformatics. 2017; 18: 79.

[2] Hahn WC, Counter CM, Lundberg AS, Beijersbergen RL, Brooks MW, Weinberg RA. Creation of human tumour cells with defined genetic elements. Nature. 1999; 400: 464-8.

[3] Hahn WC, Weinberg RA. Rules for making human tumor cells. N Engl J Med. 2002; 347: 1593-603.

[4] Cancer.Net. http://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer.

[5] Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, et al. High-resolution genomic profiles of human lung cancer. Proc Natl Acad Sci USA. 2005; 102(27): 9625-30.

[6] Park H, Cho SY, Kim H, Na D, Han JY, Chae J, Park C, Park OK, Min S, Kang J, et al. Genomic alterations in BCL2L1 and DLC1 contribute to drug sensitivity in gastric cancer. Proc Natl Acad Sci USA. 2015; 112(40): 12492-7.

[7] Sweet-Cordero A, Mukherjee S, Subramanian A, You H, Roix JJ, Ladd-Acosta C, Mesirov J, Golub TR, Jacks T. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet. 2005; 37(1): 48-55.

[8] Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current approaches and outstanding challenges. PLoS Comput Biol. 2012; 8(2): e1002375.

[9] Myers JS, von Lersner AK, Robbins CJ, Sang QX. Differentially expressed genes and signature pathways of human prostate cancer. PloS One. 2015; 10(12): e0145322.

[10] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102(43): 15545-50.

[11] Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L. GeneMSA Consortium, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet. 2009; 18(11): 2078-90.

[12] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3: 140.

[13] Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PloS One. 2009; 4(2): e4346.

[14] Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21(7): 1109-21.

[15] Libermann TA, Zerbini LF. Targeting transcription factors for cancer gene therapy. Curr Gene Ther. 2006; 6(1): 17-33.

[16] Zerbini LF. Oncogenic transcription factors: target genes. In: eLS. John Wiley & Sons Ltd. 2007. http://www.els.net [doi: 10.1002/9780470015902.a0006049].

[17] Darnell JE Jr. Transcription factors as targets for cancer therapy. Nat Rev Cancer. 2002; 2(10): 740-9.

[18] Wasylyk B, Wasylyk C, Flores P, Begue A, Leprince D, Stehelin D. The c-ets proto-oncogenes encode transcription factors that cooperate with c-Fos and c-Jun for transcriptional activation. Nature. 1990; 346(6280): 191-3.

[19] Downward J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer. 2003; 3(1): 11-22.

[20] Huang B, Warner M, Gustafsson JÅ. Estrogen receptors in breast carcinogenesis and endocrine therapy. Mol Cell Endocrinol. 2015; 418: 240-4.

[21] Sharifi N. Steroid receptors aplenty in prostate cancer. N Engl J Med. 2014; 370(10): 970-1.

[22] Yu H, Lee H, Herrmann A, Buettner R, Jove R. Revisiting STAT3 signalling in cancer: new and unexpected biological functions. Nat Rev Cancer. 2014; 14(11): 736-46.

[23] Seth RB, Sun L, Ea CK, Chen ZJ. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. Cell. 2005; 122(5): 669-82.

[24] Federman N, Meyers PA, Daw NC, Toretsky J, Breitmeyer JB, Singh AS, Miller LL, Oltersdorf T, Jezior D, Jessen KA, et al. A phase I, first-in-human, dose escalation study of intravenous TK216 in patients with relapsed or refractory ewing sarcoma. J Clin Oncol. 2017; 35: TPS11626.

[25] Mahajan N. Signatures of prostate-derived ets factor (PDEF) in cancer. Tumour Biol. 2016; 37(11): 14335-40.

[26] Yeh JE, Toniolo PA, Frank DA. Targeting transcription factors: promising new strategies for cancer therapy. Curr Opin Oncol. 2013; 25(6): 652-8.

[27] Redmond AM, Carroll JS. Defining and targeting transcription factors in cancer. Genome Biol. 2009; 10(7): 311.

[28] Bhagwat AS, Vakoc CR. Targeting Transcription Factors in Cancer. Trends in cancer. 2015; 1(1): 53-65.

[29] Tovar H, García-Herrera R, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional master regulator analysis in breast cancer genetic networks. Comput Biol Chem. 2015; 59: 67-77.

[30] Bae T, Rho K, Choi JW, Horimoto K, Kim W, Kim S. Identification of upstream regulators for prognostic expression signature genes in colorectal cancer. BMC Syst Biol. 2013; 7: 86.

[31] Sawle AD, Kebschull M, Demmer RT, Papapanou PN. Identification of master regulator genes in human periodontitis. J Dent Res. 2016; 95(9): 1010-7.

[32] Gubelmann C, Schwalie PC, Raghav SK, Röder E, Delessa T, Kiehlmann E, Waszak SM, Corsinotti A, Udin G, Holcombe W, et al. Identification of the transcription factor ZEB1 as a central component of the adipogenic gene regulatory network. Elife. 2014; 3: e03346.

[33] Medvedovic J, Ebert A, Tagoh H, Busslinger M. Pax5: a master regulator of B cell development and leukemogenesis. Adv. Immunol. 2011; 111: 179–206.

[34] Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010; 463(7279): 318-25.

[35] Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell. 2011; 147(3): 565–76.

[36] Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. Proc Int Conf Intell Syst Mol Biol. 2000; 8: 344-54.

[37] Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003; 301(5629): 102-5.

[38] Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci USA. 2005; 102(38): 13532-7.

[39] Tsai HK, Huang GT, Chou MY, Lu HH, Li WH. Method for identifying transcription factor binding sites in yeast. Bioinformatics. 2006; 22(14): 1675-81.

[40] Cheng C, Li LM. Systematic identification of cell cycle regulated transcription factors from microarray time series data. BMC Genomics. 2008; 9: 116.

[41] Wu WS, Li WH. Systematic identification of yeast cell cycle transcription factors using multiple data sources. BMC Bioinformatics. 2008; 9: 522.

[42] Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res. 2005; 33(10): 3154-64.

[43] Banerjee N, Zhang MQ. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res. 2003; 31(23): 7024-31.

[44] Hu H. An efficient algorithm to identify coordinately activated transcription factors. Genomics. 2010; 95(3): 143-50.

[45] Ma S, Jiang T, Jiang R. Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data. Bioinformatics. 2015; 31(4): 563-71.

[46] Schacht T, Oswald M, Eils R, Eichmüller SB, König R. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics. 2014; 30(17): i401-7.

[47] Petti AA, Church GM. A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. Genome Res. 2005; 15(9): 1298-306.

[48] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature. 2004; 431(7006): 308-12.

[49] Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 1998; 16(10): 939-45.

[50] Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. 2004; 5(1): 18.

[51] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999; 22(3): 281-5.

[52] Das D, Banerjee N, Zhang MQ. Interacting models of cooperative gene regulation. Proc Natl Acad Sci USA. 2004; 101(46): 16234-9.

[53] Gevaert O, Plevritis S. Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features. Pac Symp Biocomput. 2013; 123-34.

[54] Padi M, Quackenbush J. Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. BMC Syst Biol. 2015; 9: 80.

[55] Piao G, Saito S, Sun Y, Liu ZP, Wang Y, Han X, Wu J, Zhou H, Chen L, Horimoto K. A computational procedure for identifying master regulator candidates: a case study on diabetes progression in Goto-Kakizaki rats. BMC Syst Biol. 2012; 6(1): S2.

[56] Saito S, Zhou X, Bae T, Kim S, Horimoto K. Identification of master regulator candidates in conjunction with network screening and inference. Int J Data Min Bioinform. 2013; 8(3): 366-80.

[57] Yang J, Mani SA, Donaher JL, Ramaswamy S, Itzykson RA, Come C, Savagner P, Gitelman I, Richardson A, Weinberg RA. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. Cell. 2004; 117(7): 927-39.

[58] Chan SS, Kyba M. What is a master regulator? J Stem Cell Res Ther. 2013; 3: 114.

[59] De D, Jeong MH, Leem YE, Svergun DI, Wemmer DE, Kang JS, Kim KK, Kim SH. Inhibition of master transcription factors in pluripotent cells induces early stage differentiation. Proc Natl Acad Sci USA. 2014; 111(5): 1778-83.

[60] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127): 1546-58.

[61] Emmert-Streib F, Glazko GV. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. PLoS Comput Biol. 2011; 7(5): e1002053.

[62] Fisher RA. Statistical methods for research workers. London: Oliver and Boyd, 1932.

[63] Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. JASA. 2004; 99: 96-104.

[64] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Raffeld M, et al. Gene-expression profiles in hereditary breast cancer. N Engl J Med. 2001; 344: 539-48.

[65] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. USA. 2003; 100: 9440-5.

[66] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 57: 289-300.

[67] Huang Da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37: 1-13.

[68] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559-75.

[69] Anders Simon, Huber W. Differential expression analysis for sequence count data. Genome biol. 2010; 11: R106.

[70] Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001; 125: 279-84.

[71] Wold H. Estimation of principal components and related models by iterative least squares. In Krishnaiaah, P.R. Multivariate Analysis. New York: Academic Press. 1966. 391–420.

[72] Datta S. Exploring relationships in gene expressions: a partial least squares approach. Gene Expr. 2001; 9(6):249-55.

[73] Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010; 11: 95.

[74] Massie CE, Lynch A, Ramos-Montoya A, Boren J, Stark R, Fazli L, Warren A, Scott H, Madhu B, Sharma N, et al. The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. EMBO J. 2011; 30(13): 2719-33.

[75] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7), e47.

[76] Cronauer MV, Schulz WA, Ackermann R, Burchardt M. Effects of WNT/beta-catenin pathway activation on signaling through T-cell factor and androgen receptor in prostate cancer cell lines. Int J Oncol. 2005; 26(4): 1033–40.

[77] Li H, Kim JH, Koh SS, Stallcup MR. Synergistic effects of coactivators GRIP1 and beta-catenin on gene activation: cross-talk between androgen receptor and Wnt signaling pathways. J Biol Chem. 2004; 279(6): 4212–20.

[78] Song LN, Gelmann EP. Interaction of beta-catenin and TIF2/GRIP1 in transcriptional activation by the androgen receptor. J Biol Chem. 2005; 280(45): 37853–67.

[79] Song LN, Herrell R, Byers S, Shah S, Wilson EM, Gelmann EP. Beta-catenin binds to the activation function 2 region of the androgen receptor and modulates the effects of the N-terminal domain and TIF2 on ligand-dependent transcription. Mol Cell Biol. 2003; 23(5): 1674–87.

[80] Terry S, Yang X, Chen MW, Vacherot F, Buttyan R. Multifaceted interaction between the androgen and Wnt signaling pathways and the implication for prostate cancer. J Cell Biochem. 2006; 99(2): 402–10.

[81] Yang X, Chen MW, Terry S, Vacherot F, Bemis DL, Capodice J, Kitajewski J, de la Taille A, Benson MC, Guo Y, et al. Complex regulation of human androgen receptor expression by Wnt signaling in prostate cancer cells. Oncogene. 2006; 25(24): 3436–44.

[82] Jiang X, Yu Y, Yang HW, Agar NY, Frado L, Johnson MD. The imprinted gene PEG3 inhibits Wnt signaling and regulates glioma growth. J Biol Chem. 2010; 285(11): 8472–80.

[83] Ribarska T, Bastian KM, Koch A, Schulz WA. Specific changes in the expression of imprinted genes in prostate cancer–implications for cancer progression and epigenetic regulation. Asian J Androl. 2012; 14(3): 436-50.

[84] Su ZZ, Goldstein NI, Jiang H, Wang MN, Duigou GJ, Young CS, Fisher PB. PEG-3, a nontransforming cancer progression gene, is a positive regulator of cancer aggressiveness and angiogenesis. Proc Natl Acad Sci USA. 1999; 96(26): 15115-20.

[85] Webb EA, AlMutair A, Kelberman D, Bacchelli C, Chanudet E, Lescai F, Andoniadou CL, Banyan A, Alsawaid A, Alrifai MT, et al. ARNT2 mutation causes hypopituitarism, post-natal microcephaly, visual and renal anomalies. Brain. 2013; 136(10): 3096-105.

[86] Yang B, Yang E, Liao H, Wang Z, Den Z, Ren H. ARNT2 is downregulated and serves as a potential tumor suppressor gene in non-small cell lung cancer. Tumour Biol. 2015; 36(3): 2111-9.

[87] Martinez V, Kennedy S, Doolan P, Gammell P, Joyce H, Kenny E, Prakash Mehta J, Ryan E, O'Connor R, Crown J, et al. Drug metabolism-related genes as potential biomarkers: analysis of expression in normal and tumour breast tissue. Breast Cancer Res Treat. 2008; 110(3): 521-30.

[88] Stauffer S, Chen X, Zhang L, Chen Y, Dong J. KIBRA promotes prostate cancer cell proliferation and motility. FEBS J. 2016; 283(10): 1800-11.

[89] Heemers HV, Regan KM, Schmidt LJ, Anderson SK, Ballman KV, Tindall DJ. Androgen modulation of coregulator expression in prostate cancer cells. Mol Endocrinol. 2009; 23(4): 572-83.

[90] Yamamoto M, Cid E, Bru S, Yamamoto F. Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. PLoS One. 2011; 6(3): e17149.

[91] Jackson BC, Carpenter C, Nebert DW, Vasiliou V. Update of human and mouse forkhead box (FOX) gene families. Hum Genomics. 2010; 4(5): 345-52.

[92] Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res. 2007; 13(4): 1107-14.

[93] Karin M, Lin A. NF-kappaB at the crossroads of life and death. Nat Immunol. 2002; 3(3): 221-7.

[94] Sakamoto K, Maeda S, Hikiba Y, Nakagawa H, Hayakawa Y, Shibata W, Yanai A, Ogura K, Omata M. Constitutive NF-kappaB activation in colorectal carcinoma plays a key role in angiogenesis, promoting tumor growth. Clin Cancer Res. 2009; 15(7): 2248-58.

[95] Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ Jr, Sledge GW Jr. Constitutive activation of NF-kB during progression of breast cancer to hormone-independent growth. Mol Cell Biol. 1997; 17(7): 3629- 39.

[96] Rayet B, Gelinas C. Aberrant rel/nfkb genes and activity in human cancer. Oncogene. 1999; 18(49): 6938-47.

[97] Tai DI, Tsai SL, Chang YH, Huang SN, Chen TC, Chang KS, Liaw YF. Constitutive activation of nuclear factor kB in hepatocellular carcinoma. Cancer. 2000; 89(11): 2274-81.

[98] Luo JL, Kamata H, Karin M. IKK/NF-kappaB signaling: balancing life and death--a new approach to cancer therapy. J Clin Invest. 2005; 115(10): 2625-32.

[99] Karin M, Ben-Neriah Y. Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity. Annu Rev Immunol. 2000; 18: 621-63.

[100] Hayden MS, Ghosh S. Signaling to NF-kappaB. Genes Dev. 2004; 18(18): 2195-224.

[101] Wang S, Liu Z, Wang L, Zhang X. NF-kappaB signaling pathway, inflammation and colorectal cancer. Cell Mol Immunol. 2009; 6(5): 327-34.

[102] Wang CY, Cusack JC Jr, Liu R, Baldwin AS Jr. Control of inducible chemoresistance: enhanced antitumor therapy through increased apoptosis by inhibition of NF-kappaB. Nat Med. 1999; 5(4): 412-7.

[103] Sakamoto K, Maeda S. Targeting NF-kappaB for colorectal cancer. Expert Opin Ther Targets. 2010; 14(6): 593-601.

[104] Alvarez-Guardia D, Palomer X, Coll T, Davidson MM, Chan TO, Feldman AM, Laguna JC, Vázquez-Carrera M. The p65 subunit of NF-kappaB binds to PGC-1alpha, linking inflammation and metabolic disturbances in cardiac cells. Cardiovasc Res. 2010; 87(3): 449-58.

[105] Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, et al. The somatic genomic landscape of glioblastoma. Cell. 2013; 155(2): 462–77.

[106] Sikdar S, Datta S, Datta S. Exploring the importance of cancer pathways by meta-analysis of differential protein expression networks in three different cancers. Biol Direct. 2016; 11: 65.

[107] Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007; 35: W193-W200.

[108] Reimand J, Arak T, Vilo J. g: Profiler—a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011; 39: W307-W315.

[109] Gill R, Datta S, Datta S. dna: An R package for differential network analysis. Bioinformation. 2014; 10: 233-34.

[110] Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. Bioinformatics. 2007; 23: 1607-15.

[111] Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics. 2009; 10: 62.

[112] Rubinstein R. The cross-entropy method for combinatorial and continuous optimization. Methodol Comput Appl. 1999; 1: 127-90.

[113] Fagin R, Kumar R, Sivakumar D. Comparing top k lists. SIAM J Discrete Math. 2003; 17: 134-60.

[114] Yang Y, Wu J, Demir A, Castillo-Martin M, Melamed RD, Zhang G, Fukunaga-Kanabis M, Perez-Lorenzo R, Zheng B, Silvers DN et al. GAB2 induces tumor angiogenesis in NRAS-driven melanoma. Oncogene. 2013; 32: 3627-37.

[115] The Human Protein Atlas. http://www.proteinatlas.org/ENSG00000100030-MAPK1/cancer.

[116] Gherardi E, Birchmeier W, Birchmeier C, Vande Woude G. Targeting MET in cancer: rationale and progress. Nat Rev Cancer. 2012; 12: 89-103.

[117] Smith AJ, Karpova Y, D'Agostino R Jr, Willingham M, Kulik G. Expression of the Bcl-2 protein BAD promotes prostate cancer growth. PLoS One. 2009; 4: e6224.

[118] UniProt. http://www.uniprot.org/uniprot/P17252.

[119] Luo J, Manning BD, Cantley LC. Targeting the PI3K–Akt pathway in human cancer: rationale and promise. Cancer Cell. 2003; 4: 257–62.

[120] Knight ZA, Shokat KM. Chemically targeting the PI3K family. Biochem Soc Trans. 2007; 35: 245-9.

[121] Gysin S, Salt M, Young A, McCormick F. Therapeutic strategies for targeting Ras proteins. Genes Cancer. 2011; 2: 359–72.

[122] Sikdar S, Choo Wosoba H, Abdia Y, Dutta S, Gill R, Datta S, Datta S. An integrative exploratory analysis of –omics data from the ICGC cancer genomes lung adenocarcinoma study. Syst Biomed. 2014; 2: 54-62.

[123] Chakravarti A, Loeffler JS, Dyson NJ. Insulin-like growth factor receptor I mediates resistance to anti-epidermal growth factor receptor therapy in primary human glioblastoma cells through continued activation of phosphoinositide 3-kinase signaling. Cancer Res. 2002; 62: 200-7.

[124] Rojas M, Yao S, Lin YZ. Controlling epidermal growth factor (EGF)-stimulated Ras activation in intact cells by a cell-permeable peptide mimicking phosphorylated EGF receptor. J Biol Chem. 1996; 271: 27456-61.

[125] She QB, Solit DB, Ye Q, O'Reilly KE, Lobo J, Rosen N. The BAD protein integrates survival signaling by EGFR/MAPK and PI3K/Akt kinase pathways in PTEN-deficient tumor cells. Cancer Cell. 2005; 8: 287-97.

[126] Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev Mol Cell Biol. 2001; 2: 127-37.

[127] CrownBio. http://www.crownbio.com/finding-right-path-lung-cancer-therapy/.

[128] Engelman JA, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, Lindeman N, Gale CM, Zhao X, Christensen J et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science. 2007; 316: 1039-43.

[129] Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Computat Biol. 2015; 11: e1004518.

[130] Efron B, Turnbull BB, Narasimhan B. locfdr Vignette Complete Help Documentation Including Usage Tips and Simulation Example 2015.

[131] Efron B. Size, power and false discovery rates. Ann. Stat. 2007; 35: 1351-77.

[132] Benidt S, Nettleton D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. Bioinformatics. 2015; 31: 2131-140.

[133] Network TCGAR. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499: 43-9.

[134] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139-40.

[135] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288-97.

[136] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci. USA. 2001; 98: 13790-5.

[137] Takeuchi T, Tomida S, Yatabe Y, Kosaka T, Osada H, Yanagisawa K, Mitsudomi T, Takahashi T. Expression profile–defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. J Clin Oncol. 2006; 24: 1679-88.

[138] Staaf J, Jönsson G, Jönsson M, Karlsson A, Isaksson S, Salomonsson A, Pettersson HM, Soller M, Ewers SB, Johansson L, et al. Relation between smoking history and gene expression profiles in lung adenocarcinomas. BMC Med Genom. 2012; 5: 22.

[139] Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D, Brichon PY, Lantuejoul S, Hainaut P, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013; 5: 186ra66.

[140] Tarca AL, Lauria M, Unger M, Bilal E, Boue S, Kumar Dey K, Hoeng J, Koeppl H, Martin F, Meyer P, et al. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. Bioinformatics. 2013; 29: 2892-9.

[141] Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 2015; gkv229.

[142] Huang Da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat Protoc. 2009; 4: 44-57.

[143] Storey JD. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol. 2002; 64: 479-98.

BIOGRAPHICAL SKETCH

Sinjini Sikdar was born in India. She received her undergraduate degree in 2010 from Presidency College, University of Calcutta, with a major in statistics and minor in mathematics and economics. She received her master's degree in statistics from University of Calcutta in 2012. She ranked first in the master's program at University of Calcutta.

She joined the graduate degree program in the Department of Bioinformatics and Biostatistics at University of Louisville in 2013 to pursue a PhD. She, along with her supervisor, moved to the PhD program of the Department of Biostatistics at University of Florida in 2015 and continued her PhD degree there. Apart from her dissertation, she also worked on several projects during her PhD program which led to a number of publications in peer reviewed journals. Some of her works have been published in journals like *Briefings in Bioinformatics, BMC Bioinformatics, Biology Direct,* among others. She also presented her research works in several conferences. She was awarded a 2nd prize for excellent research presented at the 13th Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), held at Boston in 2014. She received the University of Florida PHHP Dean's PhD travel award for presenting her research at the ENAR 2017 conference at Washington D.C. She also received one of the best paper awards at Biostatistics Workshop on Statistical Inference on Biomedical Big Data, held at University of Florida in 2017. During her PhD program, she worked as a graduate research assistant and had opportunities to work on several collaborative projects.