DIFFERENTIAL CORRELATION IN HIGH-THROUGHPUT DATA

by

CHARLOTTE SISKA

B.A., University of Colorado, Boulder, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Computational Bioscience Program

2016

This thesis for the Doctor of Philosophy degree by

Charlotte Siska

has been approved for the

Computational Bioscience Program

by

Michael Strong, Chair

Katerina Kechris, Advisor

Sonia Leach

Tzu Phang

Russell Bowler

Date: 12/16/2016

Siska, Charlotte (Ph.D. Computational Bioscience)

Differential Correlation in High-Throughput Data

Thesis directed by Associate Professor Katerina Kechris

ABSTRACT

Differential correlation or coexpression occurs when feature pairs (e.g. transcripts, proteins, or metabolites) have different types of associations between biological groups. Differentially correlated feature pairs may indicate processes or interactions that are unique to disease. Differential correlation can be categorized into different types. *Disrupted* differential correlation is when there is no association in one group, but an association in the other group. *Cross* differential correlation is when there is a positive association in one group, but a negative association in the other group. Both types of differential correlation are relevant in biology, but most differential correlation methods are better suited to identify cross differential correlation, but not disrupted differential correlation. However, cases of differential correlation discovered in low-throughput experiments are often disrupted. In this thesis I present a novel approach for determining differential correlation called Discordant, which uses Gaussian mixture models and the EM algorithm. In simulations for continuous data, Discordant identifies disrupted differential correlation at a much higher rate than leading methods. Discordant identified experimentally-validated feature pairs in -omics data sets of Glioblastoma multiforme and Chronic Obstructive Pulmonary Disorder to be more significant than competing methods. We also determined if Discordant could be applied to non-normal data, such as counts from sequencing data. Since correlation metrics for sequencing data

are not well established, multiple correlation metrics were compared. Using simulations and breast cancer data it was demonstrated that Spearman's correlation metric performed the best over other metrics. We also examined extensions to Discordant to determine how they affected its performance. First, we manipulated Discordant so that it could identify features with elevated differential correlation, which is when the feature pair has an association in both groups, but which is stronger in one of the groups. Second, we developed an approach that addresses the independence assumption and decreases computational complexity. In summary, we report on the Discordant method and corresponding R package, which is a powerful and flexible tool to discover differential correlation on a variety of -omics data types.

The form and content of this abstract are approved. I recommend its publication.

Approved: Katerina Kechris

To my grandparents, Emil and Marie Siska, for their example of a long, happy life

ACKNOWLEDGMENTS

It would be an understatement to say that this thesis would not be possible without my advisor, Dr. Katerina Kechris. When I began working with her, I had little training in statistics and needed to be molded. Through her guidance and kind toughness I have become a good scientist. She is brilliant and is well-regarded within the scientific community - I tell people in academia who I work with and they immediately remark on her work with respect and admiration. I have fond memories of our meetings where I would present a problem I was having, and she would whip out a seemingly random statistical theory that would solve everything. She has been a wonderful mentor, and I am grateful.

I would also like to thank the members of my committee: Dr. Michael Strong, Dr. Sonia Leach, Dr. Tzu Phang and Dr. Russell Bowler; through their guidance and constructive criticism my science has improved. I would like to recognize the hard work and good attitudes of Kathy Thomas and Elizabeth Wethington, the Computational Bioscience administrators, because without them we would all be lost. During my time here, the Computational Bioscience graduate students and Post-Doctorate Fellows have always supported each other. I am grateful for this, and I hope it continues. Dr. Larry Hunter started the Computational Bioscience program 16 years ago, and many more students after me will benefit because of him.

My family has always been supportive regardless what I do (as long as it's not illegal – and if it is, it's still negotiable). I would especially like to acknowledge my parents, Robert and Jane Siska, for telling me how proud they are and helping me to maintain my motivation. A special call-out to some of my great friends, Louis Cicchini

and Laura Griffin, for making me laugh when I needed it. My boyfriend, Eric Nguyen, for being my rock and taking my stress and angst in stride these last five months. Lastly, I would like to thank my cat, Teddy, who has turned me into a cat lady.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 High-throughput Data Collection and Analysis

1.1.1. –Ome and –Omics

The suffixes -ome and -omics are used to construct terms that define various levels of biological systems. The suffix -ome was first applied by Hans Winkler in 1920 when he used the word “genome” to describe a haploid chromosome set (Lederberg, 2001). “Genome” lead to “genomics,” a term that was coined by Tom Roderick and associates at an international meeting in 1986 on the potential to map the entire human genome (Kuska, 1998). Today, the suffix -ome is used to describe an organism-wide set of features and their characteristics. The most commonly mentioned –ome, the genome, is a set of genes and their structure and function. The suffix -omics is used to describe disciplines that investigate various –omes, e.g. genomics is the collection of technologies and bioinformatic tools used to characterize genes.

The most common types of –omics found in biological research are genomics, transcriptomics, proteomics and metabolomics (Choi and Pavelka, 2012) as shown in Figure 1). All of these –omics relate to each other: genes are transcribed into transcripts, transcripts are translated into protein, proteins bind to genes to either upregulate or downregulate transcription, metabolites are catalyzed by protein enzymes and metabolites can bind to proteins to either inhibit or activate them (Voet and Voet, 2009). Commonly these –omics are interpreted separately from each other, limiting the potential of a systems-level analysis.

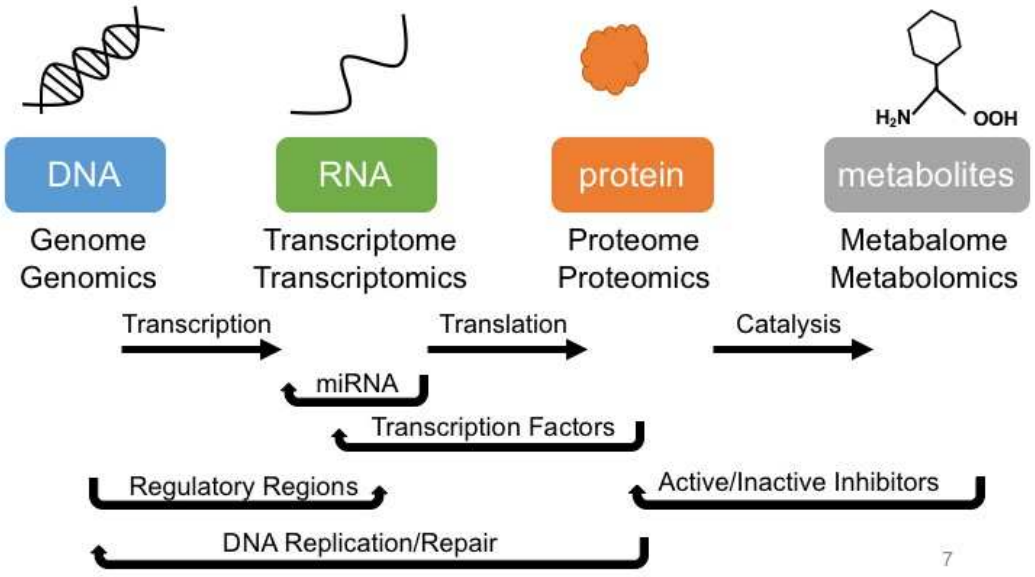


Figure 1. –Omics and the Central Dogma

–Omics datasets are now used more frequently in systems-level analysis because the different types of –omics data represent various levels of a biological system (Kitano, 2002). Recent efforts in systems biology have been made to integrate the different types of –omics data (Choi and Pavelka, 2012). Projects such as the the Cancer Genome Atlas (http://cancergenome.nih.gov/), Human Microbiome Project (http://hmpdacc.org/), NCI-60 cell lines (http://discover.nci.nih.gov/cellminer/), and ENCODE (http://www.genome.gov/encode/) are examples of efforts to collect and curate diverse types of data from either controlled samples or a wide range of subjects with various backgrounds. All of these datasets encompass several levels of biological systems (such as genomic, transcriptomic, proteomoic, metabolomic etc.) and have been used repeatedly for systems-level analysis (Bussey, 2006; Kellis et al., 2014; McLendon et al., 2008; Weiss et al., 2016).

1.1.2. Platforms

There are multiple platforms to collect –omics data. Transcriptomics data is collected traditionally using microarrays and more recently with RNA sequencing. Microarrays contain probes with complementary base pairing to a set of transcripts. Extracted mRNA is labeled with fluorescent dyes and are hybridzed to the arrays. The intensity of the fluorescent dye is used to measure gene expression (Malone and Oliver, 2011). In RNA-Seq, protein-coding RNA is commonly selected based on the absence or presence of a polyA tail, and then cleaved into fragments. The RNA is then amplified, reversed transcribed into complementary cDNA, and sequenced by binding of fluorescent-tagged nucleotides. Once the RNA sequences, called reads, are obtained they are mapped against a reference genome. However, if a reference genome is not available or alternative splicing is being examined, RNA transcripts are reconstructed using *de novo* assembly (Li et al., 2010). The number of reads mapping to a gene (or other genomic feature) can then be used to calculate gene counts (Wang et al., 2009).

Sequencing methods have also been applied to epigenetics, such as chemical changes to the DNA or histones (Zhu, 2008) and post-transcriptional modifications to RNA (Liu and Pan, 2015). DNA methylation is measured by treating DNA with a solution that will facilitate distinction between unmethylated and methylated DNA. One solution is bisulfite treatment, which converts cystosine nucleotides to uracil except for 5-methylcytosine. Sequencing is performed on samples with and without bisulfite treatment, allowing DNA methylation locations to be mapped (Chatterjee et al., 2012).

Chromatography and mass spectrometry (MS) can be used to collect metabolomic data. Chromatography is used to separate analytes based on physical properties and mass spectrometry is used for characterization. There are two different types of chromatography: liquid and gas. In liquid chromatography (LC) analytes are contained in a solution and passed through a column (Patti et al., 2012), while in gas chromatography (GC) analytes are vaporized and passed though a coated, fused capillary. Volatile metabolites, such as fatty acids and sterols, are better suited to GC while less volatile metabolites and ionic compounds, such as amino acids and sugars, are examined with LC (Agilent Technologies, 2007). LC-MS is often used for metabolic profiling since the molecular ion is usually still present and can be characterized using its m/z ratio (mass over charge ratio) (Patti et al., 2012).

MS-MS is a technology used to collect proteomic data. Proteins are broken into peptides and then put through MS to determine the m/z. The m/z peaks are separated, selected and the products are examined again with MS (hence MS-MS) to look at their spectra, which are used to determine peptide sequences (Aebersold and Mann, 2003). Proteomics can also be performed with immunoassays, which are much like transcriptomic microarrays except the probes are antibodies that proteins bind to (Borrebaeck and Wingren, 2007) and targets are often protein biomarkers.

1.1.3. Challenges

Various challenges arise from characteristics of –omics data: multiple hypotheses, dependence between features, large variance, high dimension and small sample size compared to the number of features (small n, large p). For most

models, the false assumption is made that features are independent of each other even though features are largely dependent on each other in a biological systems. Unfortunately, this assumption holds in most models because otherwise analytical methods accounting for the dependencies would be computationally expensive.

Statistical power is determined partly by effect size and sample size (Cohen, 1992). Effect size is a quantitative measurement of the phenomenon of interest (e.g. the mean gene expression difference between two groups). However, if the variance is large it can mask the observed true difference between groups. In general, large sample size improves statistical power of observing small effect sizes. However, what is reported repeatedly in –omics data is large effect size and small sample size (Button et al., 2013). This problem is most common with human data, because unlike classic experiments human phenotypes cannot be selected and controlled, therefore many unknown variables unrelated to the scientific question are introduced. Cell lines and mouse models tend to have smaller variance, but sometimes the results may not directly translate to humans.

In high dimensional data sets, a hypothesis test is often performed for each feature by their associations with outcome or phenotype, which creates a multiple testing problem when there are thousands or more features. False positive rates or Type I error is defined as the probability that the null hypothesis, or $H_{0,}$ will be determined to be false when it is actually true, a value that is normally set to 0.05. With multiple testing, the false positive rate increases dramatically. For example, if there are 1000 tests at least 50 are false positives given a p-value of 0.05 (or, are expected to have $p < 0.05$ when the null hypothesis H_0 is true). Multiple hypothesis

(or comparisons) methods have been developed to reduce the false positive rate by determining a new confidence level or convert p-values into q-values, which have been adjusted to control the false positive rate (the percentage of false positives amongst all predictions). Popular methods are Bonferroni and False Discovery Rate (FDR) (Dudoit et al., 2003). Another way to diminish the effects of multiple hypotheses is to increase power with larger sample size or reduce the number of dimensions by filtering or using lower-dimensional alternatives from methods such as Principal Components Analysis (PCA) (Lay et al., 2006).

Although concerns about variability and high dimensional data can be alleviated with increased sample size, studies are usually hindered by lack of funding and available samples. For example, the cost for a microarray per replicate is $150 to $500, and sequencing per replicate is $600 to $1000 (costs can vary by sequencing center). The cost of a study can increase rapidly with sample size. Furthermore, human studies are often limited by the numbers of subjects for a variety of reasons, including needing informed consent, obtaining a representative selection of socioeconomic status, gender, location, or obtaining control samples which may not be easily available (Greely, 2001; McDermott et al., 2013). Some cell lines, while easily controlled, may only produce small volumes of biological material because they are difficult to grow (e.g. HPV-infected NIKS cell lines (Griffin et al., 2013)). Mice from inbred panels are another option, but can be expensive (Flint and Eskin, 2012).

1.1.4. Types of Analyses

Types of analyses performed on –omics datasets are outlined in Figure 2. The most common strategy to analyze –omics data is differential expression. Differential expression is when a molecular feature has significantly disparate levels of expression or abundance between biological groups (Malone and Oliver, 2011; Oshlack et al., 2010). Statistical methods to determine differential expression depend on the nature of the data. For example, microarrays have a continuous distribution, so a simple Student t-test or Empirical Bayes alternative is used (Ritchie et al., 2015). Several methods, such as `limma`, `DiffSeq` and the Tuxedo suite, have been developed to determine differential expression in sequencing data, which is often modeled with a negative binomial distribution (Anders and Huber, 2010; Robinson et al., 2010; Trapnell et al., 2012).

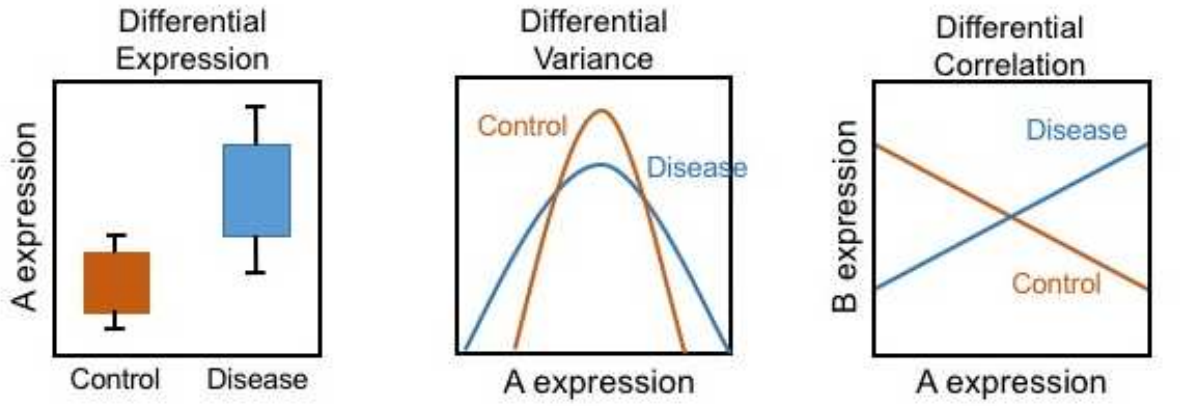


Figure 2. Different Types of –Omics Analyses. Control and Disease are two biological groups. In panels 1 and 2, only feature A is examined for differences between groups control and disease. In panel 3, features A and B are examined for differences between groups control and disease.

Another analysis is differential variance or covariance, where molecular features are identified that have dissimilar variance or covariance between groups (Ho et al., 2008; Hu et al., 2009, 2010). Differential variance has been used to study

methylation data, where it is assumed that significant differences in variation reflect adaptation (Xu et al., 2013).

Another approach is differential correlation, which is the change of association of molecular features between biological groups (i.e. healthy and disease). These differential associations may indicate molecular interactions that characterize or reflect biological or disease state. Molecular feature pairs that experience differential correlation are most likely involved in a similar mechanism or biological pathway that behaves differently between biological groups. For example, differential correlation of features can result from a break in regulation from the loss of a regulator gene (Shedden and Taylor, 2005). Differential correlation identifies a different type of biological complexity than that identified by differential expression and differential variance. Therefore, features that are differentially expressed may not experience differential correlation, and vice versa. Identifying differential correlation in –omics datasets will contribute to the understanding of the distinct complexity that exists in biological data sets.

1.2. Differential Correlation

1.2.1. Examples

Examples of differential correlation can be found in both low and high-throughput studies. One study using chromatin immunoprecipitation determined the effect of mutant p53 on wild-type p53 in the cell (Willis et al., 2004). The gene p53 is a transcription factor responsible for activating and inhibiting many pathways that inhibits the progression of cancer, making it a tumor suppressor (Lodish, 2008). Understanding the mechanism of p53 is one of the main focuses in cancer research.

Mutated p53 reduces the binding of wild-type p53 to the p53 response element of p21, MDM2 and PIG3, causing differential correlation of p53 and these targets between samples with wild-type p53 and mutant p53 (Willis et al., 2004). Many cancerous cells express mutant p53; understanding the mechanism behind the carcinogenic effects of mutant p53 and its targets may be useful for developing therapeutics.

Differential correlation in another low-throughput study was also used to understand the mechanism in paracoccidioidomycosis (PCM). PCM is a fungal infection that causes lesions in the skin and lung disease (Marques, 2012). The study used a enzyme-linked immunosorbent assay, or ELISA and a lymphoproliferation assay to determine that patients with treated PCM had no correlation between interleukins and tumor necrosis factor, but there was correlation in untreated patients (Silva et al., 1995). From these results, they were able to conclude that PCM broke down immunologic regulation.

Large-scale studies on the influence of transcription factors on transcript expression have also identified differential correlation. Using microarrays, a study determined differential correlation of transcription factors and cell cycle genes between hyperdiploid myeloma and non-hyperdiploid myeloma (Wang et al., 2014). Hyperdiploid myeloma is characterized by trisomies on several chromosomes and non-hyperdiploid has translocations at several critical loci. Patients with hyperdiploid myeloma have better survival rates than those with non-hyperdiploid myeloma (Anderson and Carrasco, 2011). Therefore, understanding the differences in molecular behavior is critical to understanding the pathogenesis and response to

treatment. Differentially correlated gene pairs were identified using Spearman's correlation and Hotelling's test (Wang et al., 2014). It was found that the cell cycle transcription factors SP1 and CDK2 have a positive correlation in hyperdiploid myeloma, but no correlation in non-hyperdiploid myeloma, supporting the hypothesis that transcription factor targeting of the cell cycle is disregulated in non-hyperdiploid myeloma (Wang et al., 2014).

Another study used differential correlation to understand the molecular processes that are unique to obesity (Walley et al., 2012). Obesity is a concern within the health community because it can exacerbate numerous physical conditions. Identifying genomic regions or biomarkers that are linked to obesity would be beneficial, but it has been a struggle replicating results. A transcriptomic study that examined expression differences between lean and obese siblings found that NEGR1 is a central hub in obesity-related differential correlation networks (Walley et al., 2012), using permutations to determine gene pairs that had correlations that were significantly different between the two groups. NEGR1 is a cell adhesion molecule, and other cell adhesion molecules have been implicated in obesity.

1.2.2. Disrupted vs. Cross Differential Correlation

There are two different types of differential correlation: cross and disrupted (Figure 3). To illustrate this, let us assume we have molecular features A and B and biological groups 1 and 2 (healthy vs. disease, control vs. experimental, etc.). Molecular features A and B have a positive correlation (+) in group 1. The other types of correlation are negative (-) or no correlation (0). There are three different

scenarios of differential correlation given the type of correlation A and B have in group 2.

(1) Group 1: +, Group 2: -

(2) Group 1: +, Group 2: 0

(3) Group 1: +, Group 2: +

Example 1 is an extreme version of differential correlation, where the correlation is in opposite directions between groups. Example 2 also illustrates differential correlation, except that in Group 2 the correlation is zero. In Example 3 there is no differential correlation because the correlation is in the same direction for both groups. Most methods are well suited to detect molecular feature pairs with a pattern similar to Example 1 (i.e., cross), but they are less likely to identify differential correlation molecular feature pairs with a pattern similar to Example 2 (i.e. disrupted). Molecular feature pairs in Example 2 could be biologically relevant since they indicate an interaction in one group that is disrupted in the other group. While Example 3 is interesting since both groups have existing associations, they do not pertain to what is unique in one group compared to another.

Current differential correlation methods are designed to identify the most extreme differences in correlation coefficients between groups, i.e. cross differential correlation in example 1. However, examples in biology from section 1.2.1 illustrate cases of disrupted differential correlation in example 2. A method that incorporates the identification of both cross and disrupted differential correlation could capture more associations that are relevant to disease.

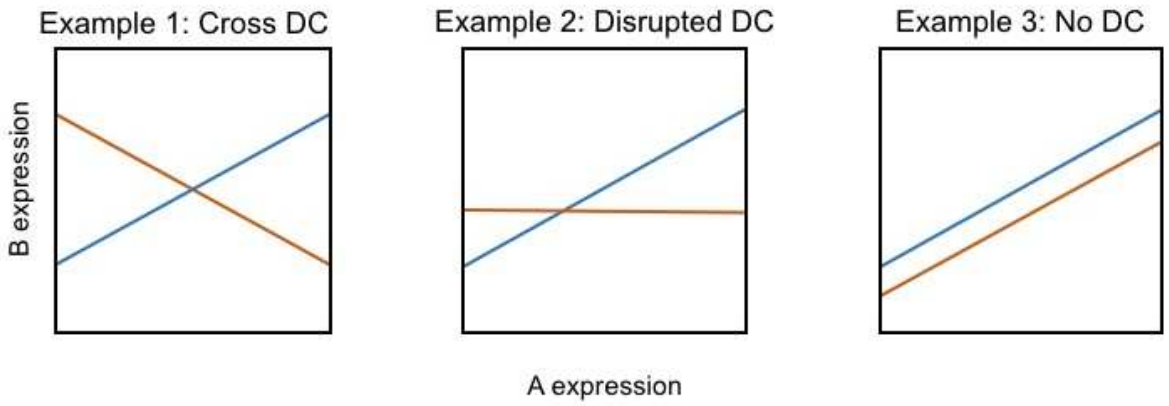


Figure 3. Types of Differential Correlation (DC). Blue and orange lines represent two different groups. Differential correlation between features A and B are being examined in two hypothetical groups (represented by blue and orange lines).

1.3. Current Differential Correlation Models

1.3.1. Classical Frequentist

The most well-known classical frequentist method for differential correlation is the statistic developed by Fisher. First, the Pearson's correlation coefficient *r* between two features in a group is converted into a z score using Fisher's transformation (Fisher, 1915).

$$z = \frac{1}{2} ln \frac{(1+r)}{(1-r)} \quad (1)$$

The Fisher's transformation is also used in other differential correlation models (Dawson and Kendziorski, 2012; Siska et al., 2015) and has an approximately normal distribution (Hotelling, 1953). To test the null hypothesis that the correlations between two groups are equal (H_0: $r_1 = r_2$), the test statistic is (Fisher, 1915):

$$z^* = \frac{z_2 - z_1}{\frac{1}{(n_2-3)^2} - \frac{1}{(n_1-3)^2}} \quad (2)$$

In equation 1, z_1 and z_2 are the z scores, n_1 and n_2 are the sample sizes and z^* is the statistic that measures the dissimilarity between z_1 and z_2. Feature pairs that have a higher absolute difference between Fisher-transformed z scores will be considered to be the most significant, and therefore by design this method is the most suited to identify cross differential correlation (Figure 3). The statistic z^* follows the normal distribution under the null hypothesis. Software implementing this method has been published (Fukushima, 2013). Another R package, DECODE, using the Fisher method is also available which integrates differential correlation and differential expression (Lui et al., 2015).

1.3.2. Bayesian

Bayesian methods are characterized by using prior information in conjunction with likelihood of the data to estimate posterior probabilities of events. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information has been used as a prior to determine pathways differentially correlated between carbon starved and nitrogen starved *Saccharomyces cerevisiae* (Bradley et al., 2009; Ogata et al., 1999). Another differential correlation application that uses Bayesian modeling is EBCoexpress (Dawson and Kendziorski, 2012; Dawson et al., 2012a). EBCoexpress is a hierarchical model that uses Empirical Bayes to estimate the posterior probability of differential correlation. The model begins with z_1 and z_2, the Fisher-transformed correlation coefficients for group 1 and group 2. z_1 has distribution $N(\lambda_1, \Sigma)$ and z_2 has distribution $N(\lambda_2, \Sigma)$. They each have a prior distribution specified by a mixture model with 1 to 3 components. In the hierarchical model, λ_1 and λ_2 are unobserved parameters. The classes of the model relate to the relationship of λ_1 and

λ_2, where the equivalent correlation (EC) class occurs when $\lambda_1 = \lambda_2$ and the DC class occurs when $\lambda_1 \neq \lambda_2$. The posterior probabilities of the classes EC or DC is determined by integrating over λ_1 and λ_2 using Empirical Bayes to estimate the hyperparameters and using the Expectation-Maximization (EM) algorithm to estimate the parameters of the mixture components. Since EBCoexpress only classifies feature pairs by being differentially or equivalently correlated, cross differential correlation will be selected more since that case results in the biggest difference between correlations. Using prostate cancer data, EBcoexpress was able to determine an enriched pathway that had been previously identified in literature.

1.3.3. Linear Interaction Models

In the linear model context, two variables x and y are expression or abundance values from –omics data. Then, linear models are fit by regression of y on main effects of x, disease group g and the interaction between x and the disease group g. The variables x and y in –omics data can be molecular features, and contain corresponding expression or abundance values. The follow linear model is used:

$$y = \alpha + x\beta_1 + g\beta_2 + xg\beta_3 + \varepsilon \quad (3)$$

where α is the y-intercept, β_1 is the linear parameter for feature x, β_2 is the group effect, β_3 is the interaction term and ε is error. The parameter β_1 represents the effect x has on y, and the group effect β_2 explains how y changes based on the group (i.e. y increases in group 1, but decreases in group 2). The interaction term β_3 is the difference in the effect x has on y between the two groups. Significance of x and y interactions between groups is evaluated by determining if the interaction β_3 has a

significant contribution to the model. This term indicates group specific slopes and would reflect differential correlation. A large absolute difference of group specific slopes results in more significant interactions of feature pairs. Therefore, as in previous methods, cross differential correlation is identified more than disrupted differential correlation.

In one study integrating transcriptomic and metabolomics data, the independent variable was a gene and the dependent variable was a metabolite (Jauhiainen et al., 2012). The linear model also incorporated variables with prior information on whether the gene and metabolite were in a particular pathway. Active genes, or genes assumed to be functioning, are selected based on rate distortion, or the minimization of the overall distortion. Once active genes have been determined and modeled, active metabolites are determined by fitting them to the active genes in one linear model where metabolites are inactive, and a series of linear models where the metabolites are active. Finally, a R^2 for each pathway is determined based on the extracted residual sum of squares and total sum of squares. Using NCI-60 cell lines, the method was able to identify many key pathways such as glycerophospholipid metabolism and nitrogen metabolism.

Another study used linear models to investigate ligand-receptor pairs in ovarian cancer using survival analysis (Eng and Ruggeri, 2015). In this case, the independent variable is the correlation between the ligand and receptor and the dependent variable is the survival time. They validated their method using an ovarian cancer data set, where several ligand-receptor pairs were identified that have already been implicated.

While linear models have been shown to be effective, there are deficiencies when there are large differences in variability between groups, which may be relevant when examining –omics data from different types of platforms and/or data from humans or non-modeled experimental systems. It has been shown that large variability results in incorrect slope estimates (Cornbleet and Gochman, 1979; Ludbrook, 2010). Furthermore, slope estimates can be different depending on what feature is considered the dependent or independent variable in the linear model.

1.3.4. Other Differential Correlation Methods

ROS-DET is a model that uses a similar framework to Fisher's method but with some deviations for better performance (Kayano et al., 2011). Associations of feature pairs are measured using biweight midcorrelation (see section 2.1.1.3), and a score is generated that is the difference of the correlation coefficients multiplied by a constant that reflects group variances. A test statistic T is developed which is based on the sum of the weighted correlation coefficients. The test statistic T follows a Chi-square distribution with one degree of freedom, and p-values are generated based on if the null hypothesis is true (H_0: $\rho_1 = \rho_2$) or false (H_1: $\rho_1 \neq \rho_2$) for each pair of correlation coefficients ρ. This method is much like Fisher, where large absolute differences in correlation coefficients are significant, and cross differential correlation is identified. Kayano et al demonstrate that their new score has better statistical power compared to other correlation metrics, such as Pearson's or Spearman's.

Expected conditional F-statistic modifies the F-statistic from analysis of variance for multiple groups (Kayano et al., 2014; Lai et al., 2004). The F-statistic was adapted to determine molecular feature pairs that share the least variance

instead of single features that share dissimilar mean across groups. To determine the expected conditional F-statistic, first a modified F-statistic is estimated based on the assumption of normality and the principle that as sample size increases, the ratio of sample size in a group compared to total sample size will be greater than 0. The new modified statistic is λ. To identify differentially correlated genes, it is assumed that genes x and y are normally distributed, and that λ now is a function of the conditional distribution of x given y, or $\lambda_{X|Y=y}$. The expectation of y is determined by integrating over $\lambda_{X|Y=y}$, as well as the expectation of x is the integral of $\lambda_{X|Y=x}$. The equations and further explanation is outlined in Lai, et al, 2004. They validated their method by identifying genes that were significantly differentially correlated to tumor suppressor genes in prostate cancer microarrays.

1.3.5. Differential Correlated Modules

There are also methods for identifying differentially correlated modules in –omics data, rather than pairs of features. Some methods use pathway databases to determine feature sets, such as the gene ontology, KEGG and molecular signatures database (Ashburner et al., 2000; Liberzon et al., 2011; Ogata et al., 1999). Once feature sets are obtained, statistical analysis is applied to determine if the correlation matrices between two groups are significantly different. In one study, feature sets are given a dispersion index to test how dispersed correlation coefficients in a feature set are between groups (Choi and Kendziorski, 2009). Another study applies cross correlation on each feature to all other features in the dataset, resulting in a final weight vector (Rahmatallah et al., 2014). Each group has its own unique weight

vector, and differential correlation of a feature set is determined by the differences between the weight vectors.

There are also methods that build feature modules rather than determine them *a priori*, e.g. use annotated information like KEGG pathways (Ogata et al., 1999). This is achieved in a variety of ways. Some feature modules are built on one feature at a time. In Fang et al, the score of a feature set is related to the percentage of features in the set that are differentially correlated. The Apriori algorithm (not to be confused with *a priori*) is used to search for the local minimum, adding and removing features one step at at time (Fang et al., 2009). Kostka et al created subsets for each group by measuring correlation based on the mean squared regression of an additive model. Features are added and removed based on a threshold using the greedy stochastic downhill search algorithm. Once the subsets are determined, the mean squared errors for each group are compared to assess differential correlation (Kostka and Spang, 2004). The issue with trying to find a local minima in Fang et al is that it not always reflects the global minima, but using arbitrary thresholds in Kostka et al is not optimal either.

Methods also use hierarchical clustering to determine modules. In the R package DICER, hierarchical clustering is used to create two subgraphs based on pairs that are up regulated and down regulated (Amar et al., 2013). Another graph is created that contains feature pairs that are consistently correlated (regardless of direction) in both groups. The up and down regulated subgraphs are paired with the consistently correlated subgraph to find metamodules. DiffCoEx is an R package which first determines a score that represents the dissimilarity of a feature pair's

correlation between groups (Tesson et al., 2010). This score is used as a metric in hierarchical clustering. Another method, CoXpress, determines correlated feature subsets in one group and determines if the gene set in the other group has no correlation by using permutations (Watson, 2006).

1.3.6. Discordant Method

We have developed the Discordant method to determine differentially correlated feature pairs (Siska et al., 2015). The Discordant model is based on a method developed previously (Lai et al., 2007, 2014) which aims to identify microarray experiments that are “concordant” and can be integrated. Concordance is measured by the approximate equivalence of z scores derived from Student t-tests in two microarrays. A three component mixture model for each microarray is used to categorize z scores based on upregulation, downregulation and no difference. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is then used to estimate the posterior probability that the microarrays are similar based on similar z scores between the two microarrays. The Discordant model modifies this algorithm in that the z scores are now Fisher-transformed z scores of the correlation coefficients of feature pairs and the mixture model is for each biological group instead of each microarray. We are interested in cases where the z scores are “discordant” instead of “concordant,” hence the name Discordant.

1.4. Novelty

The low-throughput examples outlined in section 1.2.1 provide evidence of disrupted differential correlation. The experiments identified cases where genes or proteins either associated or did not, such as p53 and its response elements and

interleukins in PCM (Silva et al., 1995; Willis et al., 2004). Low-throughput experiments are used to validate the findings in high-throughput experiments because they give more accurate results (Wilkins, 2009), i.e. RT-PCR or western blots. Since the feature pairs are demonstrating disrupted differential correlation instead of cross differential correlation in the low-throughput experiments, it is vital that methods be developed that can identify both types of differential correlation at an equal rate.

By design of the mixture model, disrupted and cross differential correlation are selected at an equal rate, since differences between positive and negative z scores are just as significant as differences between 0 and positive/negative z scores. The main novelty of our approach lies here, since in other competing methods cross differential correlation is easier to detect than disrupted differential correlation.

1.5. Outline of Dissertation

In Chapter Two, “Discordant,” the model and implementation of the model will be discussed. In the model section of Chapter Two, several correlation metrics will be described: Pearson, Spearman, Biweight Midcorrelation and SparCC (Friedman and Alm, 2012). Next, the design of the three component normal mixture model and EM algorithm in Discordant will be outlined along with corresponding issues. Finally, multiple hypothesis testing correction will be discussed. The implementation section will discuss how outliers were detected in the data, along with the approach to compare Discordant to other competing methods and assess the application of different correlation metrics to Discordant. We will also be introducing extensions to

Discordant. One of the extensions is to apply a 5-component mixture model instead of 3-component mixture model in order to identify more types of differential correlation. The other extension is to use subsampling in the EM algorithm in order to solve the issue of independence and decrease run-time.

Chapter Three is titled “Simulations and Biological Data” and will outline the design of simulations and the details of the biological data. Two different simulations were constructed in order to replicate continuous and count data. The process of synthesizing data will be discussed, including types of distributions, covariance structure and generating true positives. For each biological dataset, the preprocessing, normalization, sample size and feature size of each biological dataset used for validation and discovery will be listed. The biological datasets are Chronic Obstructive Pulmonary Disorder from COPDGene (http://www.copdgene.org/) and Glioblastoma multiforme and Breast Cancer from the Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/).

In Chapter Four, “Results”, results from the simulations and biological datasets are outlined. Evaluation of model assumptions are examined first, such as choice of initial parameters for the EM algorithm and the design of the mixture model. Next, the results from the continuous data and comparison of Discordant to other competing methods are investigated. Methods are compared based on Receiving Operating Characteristic (ROC) curves generated from simulations, and identification of phenotype-related features in the biological data by significance. Next, the discovery of novel and known targets demonstrate that Discordant can generate viable hypotheses. The same analysis is applied to count data, except the

application of different correlation metrics is evaluated instead. Finally, subsampling and 5-component mixture model extensions to Discordant are explored in both continuous and count data using both simulations and biological data.

Chapter Five, “Discussion,” contains the discussion of all preceding chapters. First, the conclusions derived from continuous and count analysis are outlined. Next, the simulations and biological validation using GBM and Breast Cancer data are investigated. Then limitations are examined, such as assumptions of the model and sample size. Finally, the future direction of module building is discussed.

CHAPTER 2

DISCORDANT

2.1. Model

The Discordant model is adapted from the Lai et al model which was developed to test for concordance between microarrays (Lai et al., 2007, 2014). We modified it to determine discordance of correlation coefficients between groups. The Discordant model uses one or two –omics datasets as input (in Figure 4, we give the example of two –omics datasets). All possible correlation coefficients are determined for each group, and then are Fisher-transformed (Fisher, 1915) which is explained in section 1.3.1, equation 1. Our method is based on a mixture model with three classes: 0, - and + as seen in Figure 4. The marginal density for one feature pair, with Fisher's transformed correlations z_1 and z_2 of feature pair k for group 1 and group 2 respectively, is:

$$f\left[z_{1,k}, z_{2,k}\right] = \sum_{i=0}^{2}\sum_{j=0}^{2} \pi_{ij}\varphi_{\mu_i,\sigma_i^2}\left[z_{1,k}\right]\varphi_{\eta_j,\tau_j^2}\left[z_{2,k}\right] \tag{4}$$

where φ_{μ,σ^2} is the normal probability distribution function (pdf) for group 1 with mean μ and variance σ^2, φ_{η,τ^2} is the normal pdf for group 2 with mean η and variance τ^2 and π_{ij} is the frequency that the feature pair is in class *i* for group 1 and class *j* for group 2 where:

$$\sum_{i=0}^{2}\sum_{j=0}^{2} \pi_{ij} = 1 \tag{5}$$

The three classes (represented by *i* and *j*) are 0 (*i* or *j* = 0), – (*i* or *j* = 1), and + (*i* or *j* = 2). Class 0 correlations are distributed around 0, class – correlations are

distributed around an unknown negative mean and class + correlations are distributed around an unknown positive mean.

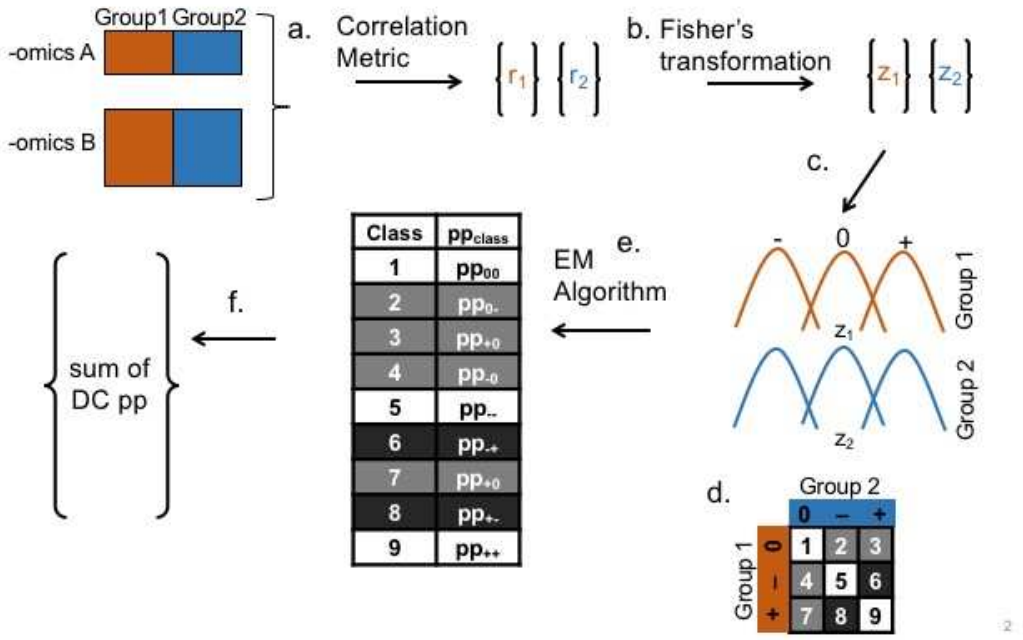


Figure 4. Discordant Method. (a) Pearson's correlation coefficients for all –omics A and B pairs. (b) Fisher's transformation (c) Mixture model based on z scores (d) Class matrix describing between group relationships (e) EM Algorithm used to estimate posterior probability (pp) of each class for each pair (f) Final output is sum of DC pp for each pair.

The parameters of the mixture model are estimated using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). In the mixture model, the true class membership is unobserved, which is represented by $w_{ij}^{(k)}$ = 1 if feature pair *k* was sampled from class *i* for group 1 and class j for group 2, otherwise $w_{ij}^{(k)}$ =0. A 3 by 3 class matrix in Figure 4d is used to explain all possible combinations of i and j (Figure 5). The cases of differential correlation (when $i \neq j$) are those on the off diagonal of the class matrix. Specifically, boxes shaded in white have no differential correlation, boxes shaded in darker gray are cross differential correlation and lighter gray disrupted differential correlation.

Using the observed data, the likelihood function is given as:

$$L(z|\theta) = \prod_{k=1}^{K} f\left(z_{1,k}, z_{2,k}\right) \tag{6}$$

where θ is the set of parameters $[\mu_0, \mu_1, \mu_2, \sigma_0, \sigma_1, \sigma_2, \eta_0, \eta_1, \eta_2, \tau_0, \tau_1, \tau_2]$ for the mixture components. The "complete likelihood" given the observed data and the unobserved class membership $w_{ij}^{(k)}$ is:

$$L(z, w|\theta) = \prod_{k=1}^{K} \prod_{i=0}^{2} \prod_{j=0}^{2} (\pi_{ij} \varphi_{\mu_i, \sigma_i^2}[z_{1,k}] \varphi_{\eta_j, \tau_j^2}[z_{2,k}])^{1(w_{ij}^{(k)}=1)} \tag{7}$$

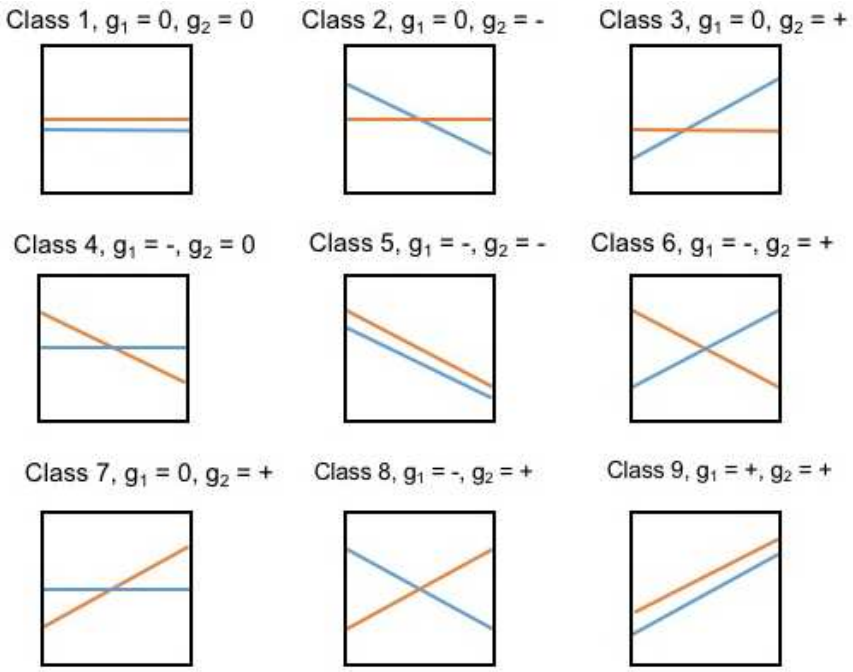


Figure 5. Visualization of Classes from Class Matrix in Figure 4d. Group 1 orange, group 2 blue.

In the E-step, the expectation of feature pair k being in class i for group 1 and class j for group 2 is:

$$E[w_{ij}^{(k)}|\hat{\theta}, z] = \frac{\hat{\pi}_{ij}\varphi_{\hat{\mu}_i,\hat{\sigma}_i^2}[z_{1,k}]\varphi_{\hat{\eta}_i,\hat{\tau}_i^2}[z_{2,k}]}{\sum_{i=0}^{2}\sum_{j=0}^{2}\hat{\pi}_{ij}\varphi_{\hat{\mu}_i,\hat{\sigma}_i^2}[z_{1,k}]\varphi_{\hat{\eta}_i,\hat{\tau}_i^2}[z_{2,k}]} \quad (8)$$

where $\hat{\theta}$ is the estimate from the previous iteration *r-1*. We drop the *r-1* notation for readability.

The updated estimates of class membership from the E-step are used for the M-step to update the mixture component parameters. Each parameter contained in θ and weights π are estimated in a similar way. The symbols *i* and *j* are classes -, 0 and + for groups 1 and 2 respectively and K is the total number of feature pairs.

Mixture Weight Parameters: (9)

$$\hat{\pi}_{ij} = \frac{\sum_{k=1}^{K} E[w_{ij}^{(k)}|\hat{\theta}, z]}{K}$$

Group 1 Parameters:

$$\hat{\mu}_i = \frac{\sum_{k=1}^{K}\sum_{j=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]\cdot z_{1,k}}{\sum_{k=1}^{K}\sum_{j=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]}$$

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{K}\sum_{j=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]\cdot \left(z_{1,k} - \hat{\mu}_i\right)^2}{\sum_{k=1}^{K}\sum_{j=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]}$$

Group 2 Parameters:

$$\hat{\eta}_j = \frac{\sum_{k=1}^{K}\sum_{i=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]\cdot z_{2,k}}{\sum_{k=1}^{K}\sum_{i=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]}$$

$$\hat{\tau}^2 = \frac{\sum_{k=1}^{K}\sum_{i=0}^{2} E\left[w_{ij}^{(k)}|\hat{\theta}, z\right]\cdot (z_{2,k} - \hat{\eta}_j)^2}{\sum_{k=1}^{K}\sum_{i=0}^{2} E[w_{ij}^{(k)}|\hat{\theta}, z]}$$

The mixture weight and distribution parameters in equation 9 are those determined for iteration *r*. Similar to equation 8, we drop the *r* notation in the formulas of equation 9 for simplicity.

Once the parameters are re-estimated, the likelihood is determined using the equation 6. After convergence of the EM algorithm (difference in log likelihood < 0.0001 or squared difference in parameters < 0.01), we report the summed differential coexpressed posterior probabilities (i.e., off-diagonal in Figure 4d):

$$p(DC_k) = \sum_{i \neq j} E[w_{ij}^{(k)} | \hat{\theta}, z] \quad (10)$$

Presented here are the steps of the method:

1. Determine initial parameters of θ^0, π^0, and $E[w_{ij}^{(k)} | \theta^0, z]$ (explained in section 2.1.3.1).
2. For iteration r:

   E-step: Determine expectation $E[w_{ij}^{(k)} | \hat{\theta}^{(r-1)}, z]$ (equation 8) of each molecular feature pair k in class w_{ij} using parameters determined in the last iteration r -1, $\hat{\theta}^{(r-1)}, \hat{\pi}_{ij}^{(r-1)}$.

   M-Step: Update $\hat{\theta}^{(r)}$ and $\hat{\pi}_{ij}^{(r)}$ based on formulas in equation 9 with the expectation in the E-step.
3. Check if likelihood has converged using equation 6 or if squared difference in parameters is less than 0.01. If not, continue to iteration $r+1$ and repeat Step 2. Otherwise, determine the posterior probability of differential correlation by summing for each molecular feature pair the posterior probability when i does not equal j (equation 10).

2.1.1. Correlation Metrics

Correlation metrics are used to measure the relationships between two variables, such as x and y. The metrics produce correlation coefficients, *r,* which are between -1 and 1. Relationships are indicated if:

1. $r > 0 \rightarrow$ Positive relationship
2. $r < 0 \rightarrow$ Negative relationship
3. $r \sim 0 \rightarrow$ No relationship

The significance of a relationship is assessed by testing the null hypothesis (H_0: there is no relationship) against the alternative hypothesis (H_1: there is a relationship). Correlation metrics differ by their underlying assumptions, such as the distribution of the data to be applied or the type of relationships identified.

2.1.1.1. Pearson's Correlation. Pearson's correlation assumes that both variables, X and Y, are normally distributed, and is optimal for identifying linear relationships (Pearson, 1895). The population coefficient for Pearson's correlation is represented by ρ, and is defined by:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_Y} \tag{11}$$

2.1.1.2. Spearman's Rank Correlation. Spearman's correlation is considered a non-parametric alternative to Pearson's (Myers and Well, 2003). Its population coefficient is represented by ρ, similar to Pearson. It is defined by:

$$r_{X,Y} = \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{12}$$

where d_i is the difference in ranks for each observation.

Since Spearman is rank-based, it can be applied to non-normal distributions and measure monotonic relationships (Spearman, 1904). However, Spearman's correlation requires that all ranks for each variable are distinct, which can result in tied ranks if there are repeated values.

2.1.1.3. Biweight midcorrelation. Biweight midcorrelation is much like Pearson's correlation, except it is median-based rather than mean-based. Biweight midcorrelation is considered robust because it is median-based, meaning it does not assume normality and the application of weights make it less sensitive to outliers (Kayano et al., 2014). Weights based on the medians and median absolute deviation of X and Y are determined. The value of the weights depends on the variance of the data, where the weights are heavy if there is large variance and the weights are light if there is small variance. The weights are used to minimize the effect of outliers on the final value of correlation (Langfelder and Horvath, 2012).

2.1.1.4. Sparse Compositional Correlation SparCC (Sparse Compositional Correlation) was originally designed to identify correlated species in microbial data (Friedman and Alm, 2012). Since microbial data is compositional and sparse, it is difficult to measure the absolute abundances because the variables are fractions of the total abundance. The matrix of the absolute abundances follow a multivariate logarithm normal distribution, which makes it possible to apply an additive normal distribution to it (Atchison and Shen, 1980) i.e., the variance of absolute abundances of a pair of species is equal to the variance log difference of the measured abundances of a pair of species. Using this relation and the approximation that the data is sparse and has large feature size, it is possible to estimate the correlation

between a species pair. Although developed for microbial compositional data, SparCC can also be applied to RNA-Seq data which is also count-based and sometimes sparse.

2.1.2. Three Component Mixture Model

2.1.2.1. Comparison of Normal and Pearson VII Distributions. One of the assumptions of Discordant is that the Fisher-transformed z scores follow a normal 3 component mixture model. It is possible that the model could have fewer or more components, or that the distribution is non-normal. A non-normal continuous distribution is Pearson VII, which is characterized by having long tails (Pearson, 1916). The distributions of the Fisher-transformed z scores could have long tails if there are extremely negative or positive correlations. The Pearson VII distribution was compared to the normal distribution to determine which fit the data best.

Another concern is that the Fisher-transformed z scores have a better fit with a number of mixture components other than 3. Mixture models with 1 to 5 components were compared using both normal and Pearson VII distribution. They were evaluated based on the Bayesian Information Criterion (BIC), which is defined as:

$$BIC = -2ln\hat{L} + kln(n) \tag{13}$$

where L is the likelihood of the model, k is the number of parameters being estimated, and n is the feature size (Findley, 1991). BIC is favored over using the likelihood because it accounts for parameter size. The larger the parameter size, the higher the risk of over fitting. BIC introduces a penalty term, k, which accounts for parameter size. R packages `mclust` and `lcmix` were used to measure BIC

(Dvorkin et al., 2013; Fraley and Raftery, 1999). Unfortunately `lcmix` does not measure BIC of Pearson VII mixture models with 1 component, so only 2 to 5 mixture components were examined.

2.1.2.2. Extend to 5 Components. An extension to the Discordant model is to increase the observable differential correlation classes. Currently, the only types of differential correlation observed are cross (associations are opposite in between groups) or disrupted (association is present in one group but not the other). In previous studies, cases where there was an increase in association in one group versus another has been observed. For example, antigen coexpression increased in women 3 days after vaginal delivery (Juretic et al., 2004) and eotaxin and interleukin-5 coexpression was increased in blister fluid of patients with bullous pemphigoid compared to healthy patients (Shrikhande, T. Hunziker, L. R. Braa, 2000).

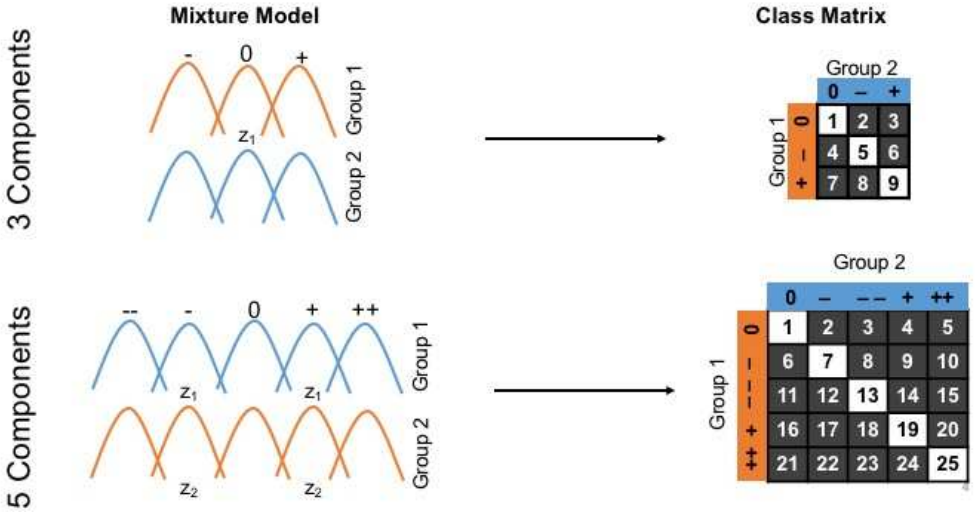


Figure 6. Increasing from Three to Five Components Changes Class Matrix.

In the simplest model, a three component mixture model is used to define whether correlations are not present (0), are positive (+) or are negative (-) (Figure 4cd). We offer an extension that increases the number of components to 5, which isolates the more extreme correlations, i.e. associations that are either very positive (++) or very negative (--). This increases the parameter size from 21 to 35 and the number of classes from 9 to 25 (Figure 6).

2.1.3. EM Algorithm

2.1.3.1. Initial Parameters. The parameters used to separate the class components are denoted by *b* in Figure 7 in the mixture model represented in Figure 4c. Group v=1,2 has a unique parameter b_v to allow for different mixture model variances for each group. The parameter b_v is the standard deviation of the Fisher transformed z scores. Observations between $-b_v$ and b_v are set to component 0, observations to the left of $-b_v$ are set to component – and observations to the right of b_v are set to component +. Based on these assignments, the mean and variance of the observations in each component were used to determine the initial parameters $\mu_0, \mu_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$ for group 1 and $\eta_1, \eta_2, \tau_0, \tau_1, \tau_2$ for group 2. The EM algorithm (Dempster et al., 1977) is then used to iteratively update parameters in the M-step and the posterior probability for each class and group in the E-step (Figure 4).

A similar approach is taken when developing the 5-component mixture model. The ++ component initial distribution is determined by values to the right of $2b_v$ and the – – component initial distribution is determined by values to the left of $-2b_v$. The + component initial distribution is determined by values between b_v and $2b_v$ and the – component initial distribution is determined by values between $-b_v$ and $-2b_v$. The 0

component initial distribution is determined by values between $-b_v$ and $-b_v$, similar to the 3-component mixture model.

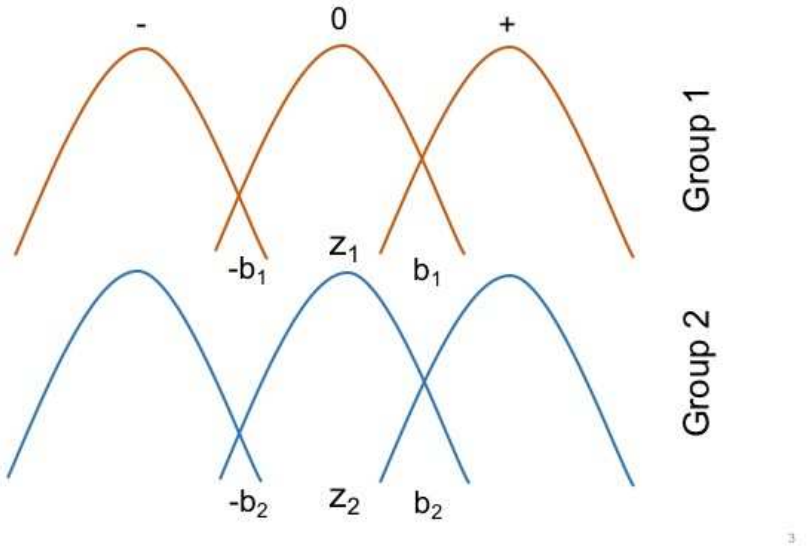


Figure 7. Setting Initial Parameters of Mixture Components. Boundary b_v where v = 1,2 dictates initial distributions for each component. Values to the right of b_v belong to the + component, values to the left of $-b_v$ belong to the – component and values between $-b_v$ and b_v belong to the 0 component.

Adjustments were made from the original Lai, et. al algorithm. In Lai et al., z scores derived from differential expression in microarray experiments have an approximate $N(0,1)$ distribution, whereas the Fisher-transformed z scores of correlation coefficients tend to have a smaller standard deviation. The initial parameters are determined by evaluating the Receiving Operating Curve (ROC) with different values for b_v.

2.1.3.2. Subsampling. Like other methods (Dawson and Kendziorski, 2012), the Discordant model makes a false assumption that molecular feature pairs are independent of each other, but features are in multiple different pairs which violates the independence assumption. A subsampling option is included to address the assumption and also cut down run-time. By default, the EM algorithm determines

parameters across all molecular feature pairs until the EM algorithm converges (Dempster et al., 1977). With the subsampling option, a subsample of correlation coefficients independent of each other (or features that are only present in one pair) are input into the EM algorithm. This is repeated for a number of iterations (default is 100), and the parameters of each mixture component from each iteration are summarized by their mean (Figure 8a-e). Once the summarized parameters of the mixture components are determined, the posterior probabilities of all molecular features are determined (Figure 8f).

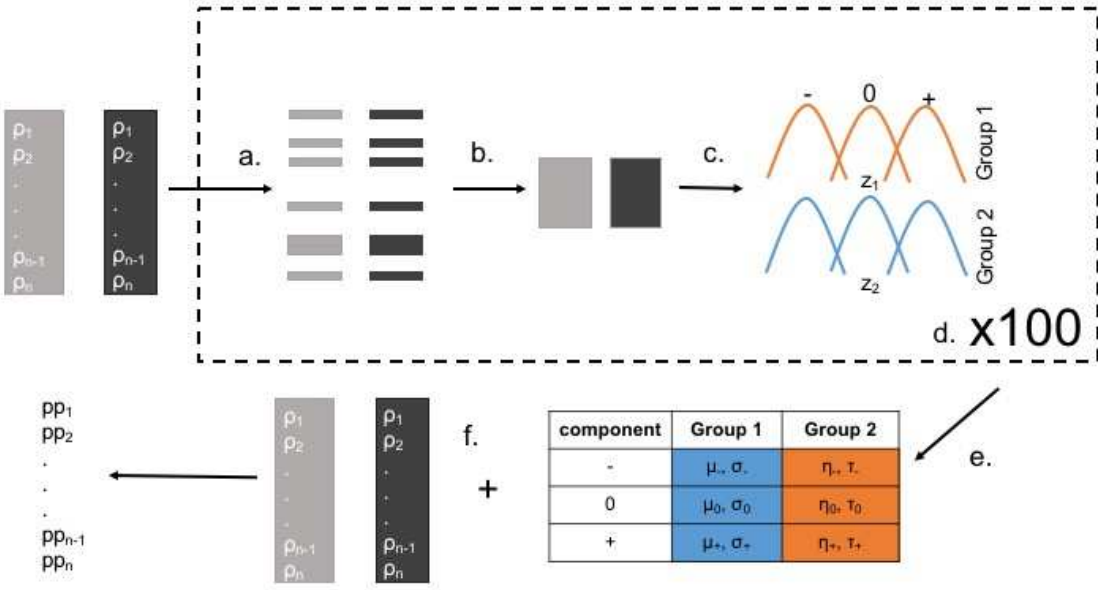


Figure 8. Subsampling. (a) Extract independent correlation coefficients. (b) Take independent correlation coefficients and create subset of correlation vectors. (c) Determine parameters of EM algorithm using subset of correlation coefficients. (d) Repeat steps a-c for 100 iterations. (e) Take average of parameters across runs. (f) Apply parameters to all features to obtain posterior probabilities (E-step of EM algorithm).

2.1.4. Multiple Testing

The number of hypotheses tested in a differential correlation analysis is much greater than a differential expression analysis. There is a hypothesis for each

possible feature pair in differential correlation, resulting in millions of hypotheses to test with even as few as 1500 features. Therefore, applying multiple hypothesis testing methods to the p-value or posterior probability is critical. The `p.adjust` function from the R `stats` package was used to determine FDR for the p-values and the `crit.fun` function from R package `EBarrays` was used to determine q-values for the posterior probabilities (Kendziorski et al., 2005).

2.2 Implementation

In this section, the application of continuous and count data is discussed. Continuous data is characterized by real numbers. In our context, continuous data is assumed to follow a normal distribution. Count data is characterized by integers greater than or equal to 0. There are different approaches in some steps of the analyses depending on whether continuous or count data is modeled. Using the normal distribution to model the continuous data makes many of the statistical analyses more straightforward. In contrast, the count data can be more challenging to model since sequencing data often results in greater density towards 0 and large dispersion. For this reason, performance of the Discordant method was first assessed using continuous data. Once it was established that Discordant had better performance than other methods, the application of count data was then examined.

2.2.1. Outliers

Outliers can skew correlation and create false positives. Filtering out features may result in lost information, but may also reduce false positives and improve power. Also, filtering features based on outliers reduces the dimension of the data which eases computational and multiple testing burdens.

In the case of normal data, Grubbs' outlier test can be used (Grubbs, 1969a). The null hypothesis is that there are no outliers in the data, and the alternative hypothesis is that there are outliers in the data. Therefore, if a feature has a Grubbs' p-value less than 0.05 in either group, it is deemed to contain an outlier and the feature is filtered out. From the `outliers` R package `grubbs.test` was used to determine features with outliers (Komsta, 2006).

Non-normal datasets may be filtered using cutoffs based on the median absolute deviation (MAD) (Leys et al., 2013). Even after pre-processing and normalizing, the distribution of sequencing data still is asymmetrical, where there is large density around zero and long tails to the right. To determine outliers, the values for each feature are split by being greater or lesser than the median (Figure 9). The two sets of features are tested for outliers by the difference they have with their respective MAD (Magwene et al., 2011). The maximum distance of all features from their MAD is used to determine if the feature has an outlier. The standard threshold is two or three times the MAD outside the median (Leys et al., 2013), but since the distribution in the sequencing data we explored is more extreme we used larger thresholds. For the voom-transformed sequencing data (see section 3.4), a threshold of 7 is used to retain most features and filter out those that were most problematic (Law et al., 2014). Non-transformed data has even larger dispersions, so a threshold of 20 was used.

The leading methods, Fisher, linear interaction models and EBcoexpress, were chosen to compare to the Discordant method (Dawson and Kendziorski, 2012; Dawson et al., 2012a; Fisher, 1915). These methods have a similar output to

Discordant, which is a p-value or posterior probability of a molecular feature pair being DC. They were compared based on q-values and ranks from simulations and biological validations. Ranks were used to compare methods because p-values and posterior probabilities are difficult to compare since they represent two different kinds of probability (Käll et al., 2008). The p-value is the probability that an event this extreme or more extreme would occur if the null hypothesis is true (or a false positive), and the posterior probability is the probability that an event occurs given the data.

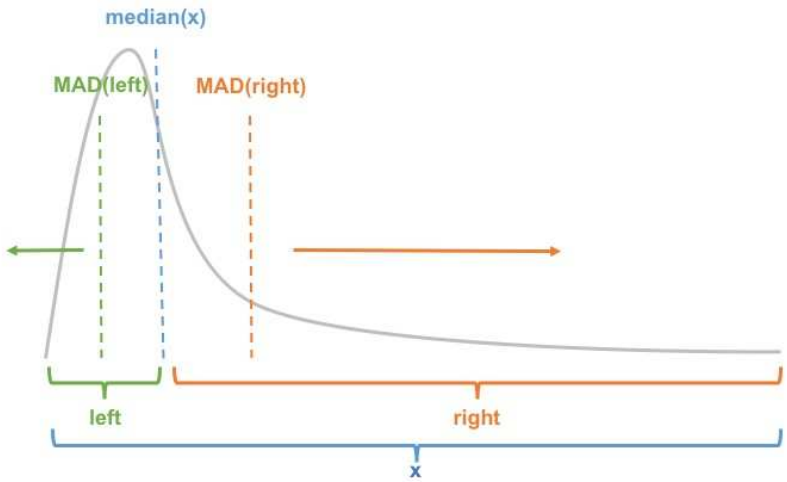


Figure 9. Split MAD Outlier Detection. Values of x are split based on the median of x, or median(x). Values to the left of median(x) are part of the left distribution, and values to the right of the median(x) are part of the right distribution. The median absolute deviation (MAD) is determined for the left and right distribution. Values that are outside a factor of the MAD(left) and MAD(right) are considered outliers.

2.2.2. Compare Discordant to Other DC Methods

Initial hyperparameters in EBcoexpress can either be determined by using the normal mixture modeling function `mclust` by default in the EBcoexpress R package (Dawson et al., 2012b; Fraley and Raftery, 1999), or by a grid approach. We did the latter for the simulations since the hyperparameters determined by `mclust` produced posterior probabilities that had a small range and were non-informative.

The hyperparameters we determined by the grid approach were based on three components and produced a more even distribution of posterior probabilities. Since EBcoexpress takes long periods to run with large datasets, only ten percent of the data was randomly selected for ten iterations when determining appropriate hyperparameters. The final output of EBcoexpress is either the posterior probability of differential correlation or equivalent correlation.

For linear interaction models, we used the `lm` function in the `stats` R package. Linear interaction models are directional unlike other methods. It is possible that the linear models will have different results if the independent and dependent variable are switched. We explored this effect with the GBM miRNA-mRNA data, since it is known that miRNA affects the expression of transcripts (Cannell et al., 2008).

2.2.3. Comparison of Correlation Metrics Applied to Discordant

In count simulations and discrete biological data, it is unclear which correlation metric best fits. It is more straight forward to apply correlation metrics to continuous normal data because many correlation metrics assume normality (Pearson, 1895). Also, the correlation metrics ability to identify relationships in continuous data has been studied in normal continuous data, but not count data (de Siqueira Santos et al., 2014; Song et al., 2012). Correlation metrics Spearman, Pearson, Biweight Midcorrelation (BWMC) and SparCC applied to the Discordant model along with generalized linear models with interaction terms were compared in terms of statistical power and ability to identify experimentally validated features in significant pairs.

CHAPTER III

SIMULATIONS AND BIOLOGICAL DATA

3.1. Simulation Design

3.1.1. Continuous Data

Bivariate normal n by m matrices with n features and m samples were first simulated using the function `mvrnorm` from R package `MASS` (Venables and Ripley). The means were set to 0 and the covariance matrix was a diagonal matrix of 1. We assumed independence for all samples in groups and across all features. The features were separated into two different sections, where these sections were treated as different types of –omics data (Figure 10a). The Pearson's correlation coefficients were calculated (Figure 10b) and then they were reorganized to create pairs that simulate the nine different situations of Figure 4d within the data (Figure 10c). This resulted in known DC pairs, so we could observe how categorizing association types in Discordant affected power compared to the other models.

The simulations were altered to take into account how the methods were affected by feature size, sample size, proportion of forced DC and correlation method. The standard was 1000 feature pairs, sample size of 20 in each group, 0.2 pairs forced to be DC and Pearson's correlation. The effect of these parameters on performance was assessed. All combinations examined are listed in Table 1.

All methods were run on the simulated data and compared using a Receiver Operating Characteristic (ROC) curve and sensitivity/specificity by rank of p-values or 1 - posterior probabilities. Simulations were run 100 times and results were averaged over the runs.

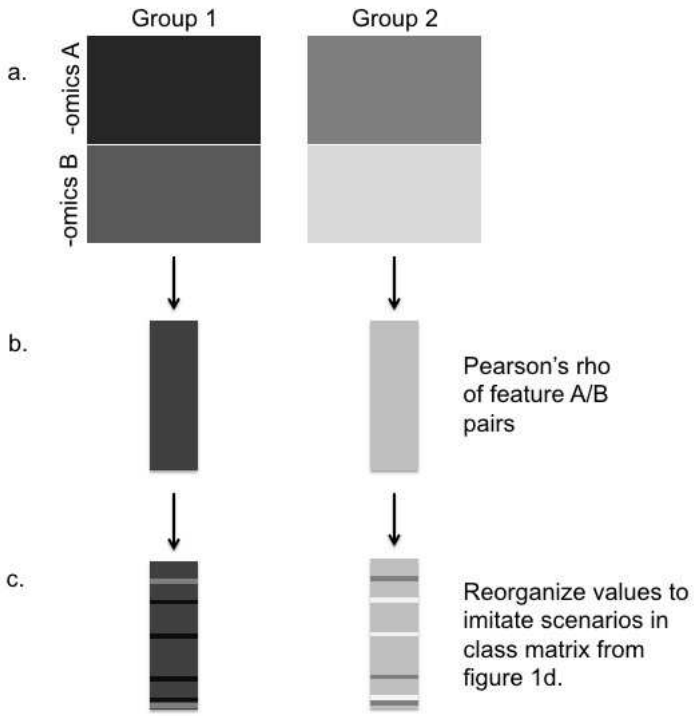


Figure 10. Generation of Data for Simulations. (a). Create a data matrix and separate into two different type of omics. (b). Determine correlation coefficients between features in –omics A and –omics B. (c). Swap correlation coefficients in vectors to simulate scenarios from class matrix.

Table 1. Summary of Simulation Adjustments. Shaded in red are standard parameters for simulations.[1]

Sample Size	Forced DC Pairs	Feature Size	Correlation Method
10 10	0.1	500	Spearman
20 20	0.2	1000	Pearson
50 50	0.3	2000	Biweight MidCorrelation
10 20		5000	
20 50			

3.1.2. Count Data

There are no available R functions that simulate bivariate negative binomial count data for a fixed covariance structure. Therefore, we used information from the TCGA breast cancer data to first generate a data set without any structure between

[1] Portions of this chapter have been reprinted with permission from Bioinformatics.

miRNA and mRNA. Figure 11 shows the pipeline to create simulations. First, parameters are generated based on the TCGA breast cancer data (Figure 11a, Step1) and then they are used to simulate data (Figure 11b, Step 2).

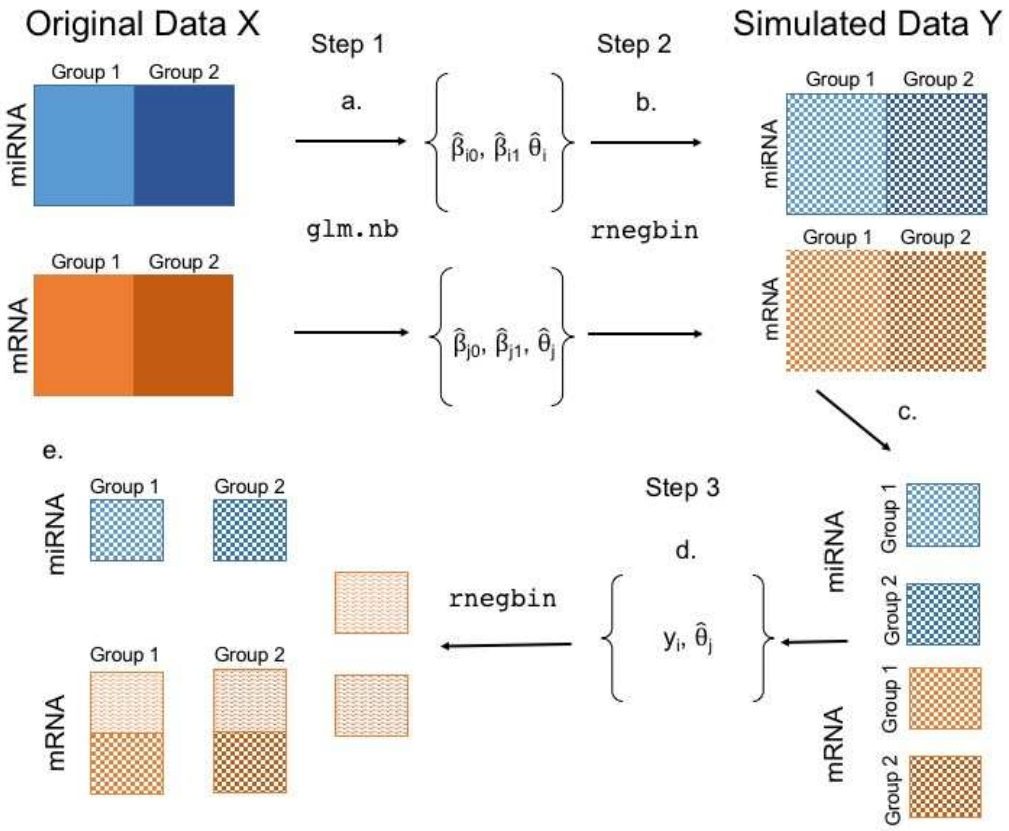


Figure 11. Generating Simulations from TCGA Breast Cancer Data. (a) Determine negative binomial paramters θ and β of each feature using `glm.nb` (b) Use the same θ and β from (a) to create simulated data row by row with `rnegbin` (c) Choose randomly 200 pairs from miRNA and 200 pairs from mRNA (d) Simulate mRNA features that are positively or negatively correlated to miRNA (e) Stack generated mRNA features on top of mRNA simulated subset.

For each feature in the TCGA miRNA and mRNA sequencing data, we estimate the two group (control vs. tumor) means and common dispersion using R function `glm.nb` from library `MASS` (Figure 11a).

Step 1

$$\texttt{glm.nb(x\_i\sim groups)} \rightarrow \hat{\beta}_{0i}, \hat{\beta}_{1i}, \hat{\theta}_i$$

$$\texttt{glm.nb(x\_j\sim groups)} \rightarrow \hat{\beta}_{0j}, \hat{\beta}_{1j}, \hat{\theta}_j$$

Variables `x_i` and `x_j` contains the counts for the i^{th} and j^{th} features in the miRNA and mRNA data sets respectively. The parameters for feature i are means $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ and dispersion $\hat{\theta}_i$ and the parameters for feature j are means $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ and dispersion $\hat{\theta}_j$. The variable `groups` signify which samples are in group 1 and group 2.

Next in Step 2, the R function `rnegbin` and parameters generated from Step 1 are used to simulate data, for each feature y (Figure 11b).

Step 2

$$\hat{\beta}_{0i}, \hat{\theta}_i \rightarrow \texttt{rnegbin} \rightarrow y_{i1}$$

$$\hat{\beta}_{0j}, \hat{\theta}_j \rightarrow \texttt{rnegbin} \rightarrow y_{j1}$$

$$\hat{\beta}_{0i}+\hat{\beta}_{1i}, \hat{\theta}_i \rightarrow \texttt{rnegbin} \rightarrow y_{i2}$$

$$\hat{\beta}_{0j}+\hat{\beta}_{1j}, \hat{\theta}_j \rightarrow \texttt{rnegbin} \rightarrow y_{j2}$$

The variables y_{i1} and y_{i2} are the simulated features for group 1 and group 2 for the i^{th} miRNA feature and variables y_{j1} and y_{j2} are the simulated features for group 1 and group 2 for the j^{th} mRNA feature. For faster simulations, a subset of 200 features of the miRNA and mRNA were extracted (Figure 11c).

In Step 3 dependencies between the miRNA and mRNA are created. The mRNA features (correlated_mRNA) whose values were correlated with miRNA features were generated using the negative binomial distribution and paramters from Step 2. Positive and negative associations were included to capture both indirect and direct effects of miRNA on mRNA. Although not reflecting the common relationships of miRNAs and mRNAs, positive correlations have been observed in

previous studies (Pasquinelli, 2012). This was performed using the `glm.nb` function from library `MASS` with parameters from Step 2 where the mean is defined by the simulated miRNA values $\{y_i\}$ and the mRNA dispersion $\hat{\theta}_j$ to create 200 correlated pairs (Figure 11d). Positive associations were created since correlated_mRNA has a similar pattern to the $miRNA_i$ with added variance. Negative assocations were based on positive associations, which were generated by first obtaining the index of ordered $miRNA_i$ values and then matching the mRNA values in reversed order.

Step 3 mean = $\{y_i\}$, dispersion = $\hat{\theta}_j$ → `rnegbin` → correlated_mRNA

Since the data has been normalized, no constant was used to scale the miRNA value in the generalized linear model. We use miRNA as the independent variable in the linear model since miRNA can target 3'UTR of genes and affect mRNA expression, and not vice versa (Cannell et al., 2008). This creates a subset of data with miRNA→mRNA relationships (orange squiggly lines in Figure 11e), and we also include a set of independent mRNA (orange checkered pattern in Figure 11e) which would be in miRNA-mRNA pairs that are non-correlated pairs (negative cases).

The data are then converted into correlation coefficients for each group (Figure 4b). The highest correlations are reorganized in the Fisher-transformed z vectors to match correlations in each the 9 classes in the class matrix of the Discordant model (Figure 4d). These will be the true positives and there are 16 in each class.

3.1.3. Extensions Simulations

Simulations to test the statistical power of extensions were performed. Both continuous and count simulations were used. In the count simulations, the correlation metric that demonstrated the most power in the simulation analysis was used. For the 3 versus 5 component mixture model analysis, simulations with the standard set of parameters (Table 1) was used. In the subsampling simulation, 100 by 200 features were generated, where 100 feature pairs were used in the subsampling.

3.2. The Cancer Genome Atlas Glioblastoma Multiforme miRNA and mRNA Microarrays

From The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) we accessed normalized GBM miRNA and mRNA expression data that had matched subjects (McLendon et al., 2008b). This dataset was selected because it had the largest sample size of organ-specific control samples between the two arrays on TCGA (Appendix A.1). The miRNA data was generated on an Agilent miRNA array and was normalized using quantile normalization and is available at TCGA. The mRNA data was generated on custom Agilent 244K array and normalized using loess normalization. In the datasets, there are 470 miRNA and 90797 mRNA. Grubbs' outlier test (Grubbs, 1969b) was used to eliminate any molecular features with outliers that could skew correlation, which reduced the feature size to 331 miRNA and 72656 mRNA (Grubbs' p-value > 0.05). The number of matched samples between the –omics datasets are 10 control samples and 21 tumor samples.

Cancer-related miRNAs were accessed from multiMiR and miRcancer (Ru et al., 2014; Xie et al., 2013). We collected miRNAs on four cancers, including GBM as well as breast cancer, prostate cancer and melanoma as negative controls. There were 47 total cancer-related miRNA for GBM, but only 4 were unique to GBM and not occurring in any of the other cancers (Appendix A.2). Since the results are in the form of molecular feature pairs, miRNA may occur in more than one pair. Therefore, the first occurrence of the GBM miRNAs in the top ranked list is reported. After running each method, the top rank, and respective p-value/posterior probability and q-value of unique GBM-related miRNA-transcript pair was interpreted.

3.3. COPDGene Transcriptomic and Metabolomic Data

Through COPDGene (http://www.copdgene.org/), a nation-wide genetic epidemiologic study, we were able to acquire metabolomic and transcriptomic data from COPD patients. The peripheral blood mononuclear cell (PBMC) transcriptomic data was generated on the Affymetrix HGU133 Plus 2.0 array (Gene Expression Omnibus GSE42057) and normalized using RMA (Bahr et al., 2013). Metabolomic data from plasma was processed and generated using LC/MS Agilent software and tools and pre-processed and filtered using MSPrep (Bowler et al., 2015; Hughes et al., 2014). Both datasets were filtered based on Grubbs' outlier test, leaving 38852 transcripts and 1640 metabolites (Grubbs' p-value > 0.05). COPDGene subjects were separated by spirometry, which indicates the severity of COPD in a patient. The control group contained subjects with normal spirometry (FEV_1/FVC > 0.7 and FEV_1 percent predicted > 80% after bronchodilator) and the disease group contained subjects with abnormal spirometry (FEV1/FVC < 0.7 and FEV1 percent predicted <

50% after bronchodilator). The final sample size for each group was control: 39 and COPD: 39.

Previous studies by COPDGene have implicated sphingolipids and their related pathways in COPD (Bowler et al., 2015). Sphingolipid-related metabolites were determined using ID Browser in Mass Profiler Professional (MPP) software (Agilent). The Gene Ontology (GO) database was used to collect transcripts with a GO term related to sphingolipids, and the probes were acquired from Ensembl BioMart. The final number of sphingolipid-related metabolites and transcripts is 37 and 188 respectively (Appendix A.3). We examined the top ranks and respective p-values/posterior probability and q-values of the sphingolipid-related feature pairs. Since the result are in the form of molecular feature pairs, sphingolipid-related features occur in more than one pair. Similar to our GBM analysis, we report the first occurrence of the sphingolipid-related feature in the pairs ranked by p-value or posterior probability.

3.4. TCGA Breast Cancer miRNA-Seq and RNA-Seq

From the Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), we accessed miRNA-Seq and RNA-Seq breast cancer data with matched subjects. Of these, there are 15 samples with normal (or control) tissue and 42 samples with tumor tissue (Appendix A.4). This dataset was selected because it was one of the few datasets on TCGA that had matched samples with miRNA-Seq and RNA-Seq data, and had sufficient number of samples for control and tumor groups. Both datasets was normalized using HTSeq filtering and TMM normalization (Rau et al., 2013; Robinson and Oshlack, 2010) and were transformed using voom for

Pearson's correlation and BWMC (Law et al., 2014). The number of features remaining were 212 miRNA and 19414 mRNA.

Datasets were filtered using median absolute deviation (MAD) (Leys et al., 2013). Features were further filtered by the presence of outliers using the MAD outlier method outlined in section 2.2.1. The number of features after filtering for outliers is 16656 for RNA-Seq data and 200 for miRNA-Seq data for the voom-transformed data and 17972 for RNA-Seq and 200 for miRNA-Seq for non-transformed data.

Results were validated using experimentally validated miRNAs, much like with the COPD analysis GBM analysis in sections 4.2 and 4.3. A total of 8 unique breast cancer miRNAs were found to be in the TCGA breast cancer data (Appendix A.5).

CHAPTER IV

RESULTS

4.1. Evaluating Assumptions

4.1.1. Initial Parameters

The initial parameters to define the mixture component distributions were determined by assigning different values to parameters and comparing their ROC curves from simulations. Parameters are defined by θ, which is the set $[\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \eta_1, \eta_2, \eta_3, \tau_1, \tau_2, \tau_3]$. Parameters μ_i and σ_i are the mean and standard deviation for mixture components i in group 1, and η_i and τ_i are the mean and standard deviation for mixture components j in group 2 (where i, j = {0,+,-}). These parameters are determined by the boundary b. The z scores within $-b$ and b are used to define the prior distribution for mixture component 0, z scores greater than b define the prior distribution for mixture component + and z scores less than $-b$ define the prior distribution for mixture component -. The mean and standard deviation of each of these distributions are defined by the *b* cutoffs to assign initial values of θ.

The model in Lai et al was used to determine concordance of two microarray experiments (Lai et al., 2007, 2014). In their application, the z scores generated from multiple t-tests are assumed to have a *N*(0,1) distribution, so they set b equal to 1. It is evident from simulations and the biological data that Fisher-transformed z scores do not necessarily have a *N*(0,1) distribution, and the variance of the distribution can vary depending on the sample size.

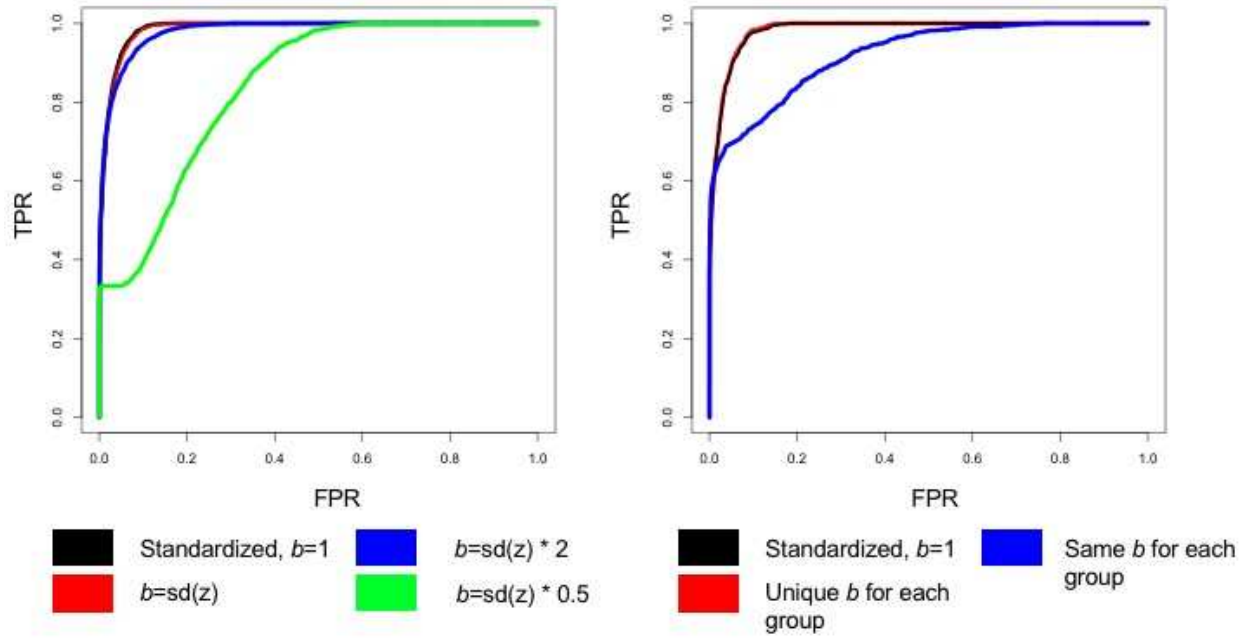


Figure 12. ROC Curves of Initial Parameters. (a) Effect of *b* with same sample size. AUC of standardized, *b* = 1 is 0.985. AUC of *b*=sd(z) is 0.985. AUC of *b*=sd(z)*2 is 0.982. AUC of *b*=sd(z)*0.5 is 0.838. (b) Effect of *b* with different sample size. AUC of standardized, *b* = 1 is 0.990. AUC of b_1=sd(z_1) and b_2=sd(z_2) is 0.990. AUC of *b*=sd(z) for both groups is 0.950.

In Figure 12a, the effects of *b* are examined with the same sample size between groups. ROC curves were plotted based on simulations described in section 3.1.1 b and different manipulations of z scores were evaluated. In one case, z scores were standardized to have a *N*(0,1) distribution with b equal to 1 as in the original model in Lai et al. Another case tested unaltered z scores and b equal to the standard deviation of the z scores. Standardizing z scores and using b equal to 1 produced a similar result when the z scores were unaltered and b equaled the standard deviation of the z scores (Figure 12a). This demonstrates that the best way to generate prior distributions of mixture components is to use the standard deviation of the z scores, and that b equal to 1 was chosen in the Lai et al model because z scores generated from t-tests were close to a *N*(0,1) distribution. The sensitivity of b

equal to the standard deviation of z scores was tested further by increasing or decreasing b by a factor of 2. These ROC curves are plotted in Figure 12a as well, and it is shown that increasing or decreasing b by a factor of two hinders performance.

In many –omics experiments the sample size may be different between groups. The effect of this on the Discordant model in regard to the initial parameters was examined. For example, if the sample size was different between groups the z score distribution for group 1 and group 2 could be dissimilar. It was investigated if each group should have its own unique b or if z scores from both distributions should be pooled together to determine b. The results are in Figure 12b. Several cases were analyzed: standardized z scores with initial parameter b equal to 1, z scores with b equal to the standard deviation of the pooled z scores from both groups, and z scores with unique b for each group. Unfortunately, cases of b equal to twice or half the standard deviation of z scores would result in a segmentation fault and could not be evaluated. It was found that the standardized z scores and the unique b for each group v had similar performance in Figure 12b, but the pooled b had decreased performance. Therefore, we use b_v as the standard deviation of the z scores for group v.

4.1.2. Mixture Model

We explored the number of components and distribution assumptions using the Bayesian Information Criterion (BIC) to compare alternative models. Plots of BIC with the number of components from 1 through 5 for normal and Pearson VII distributions are shown in Figure 13 and 14 for continuous GBM and COPD datasets

and different correlation metrics in breast cancer sequencing data (Spearman, Pearson, SparCC and biweight midcorrelation). The more positive the BIC, the better the model fits the data. In all cases, mixture models with a normal distribution have a more positive BIC than models with the Pearson VII distribution. In GBM and COPD, the difference of BIC between mixture components is negligible for the normal distribution (Figure 13). Figure 14 shows that regardless of the correlation metric, more than one mixture component is favored in the breast cancer data, and the difference in BIC when the component size is greater than two is small. Therefore, we assume that the data can be represented by a 3-component mixture model, because models with a normal distribution had a more positive BIC and the 3 component normal mixture models did no worse than models with component size 2, 4 or 5. Tables of BIC values plotted in Figures 13 and 14 are in Appendices B.1 and B.2.

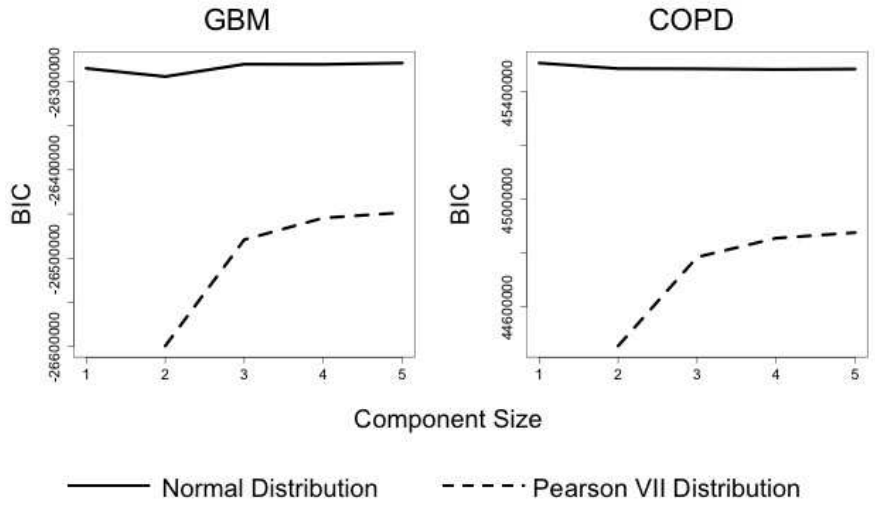


Figure 13. Comparison of Distributions (Normal vs. Pearson VII) and the Number of Mixture Components for the GBM and COPD Data.

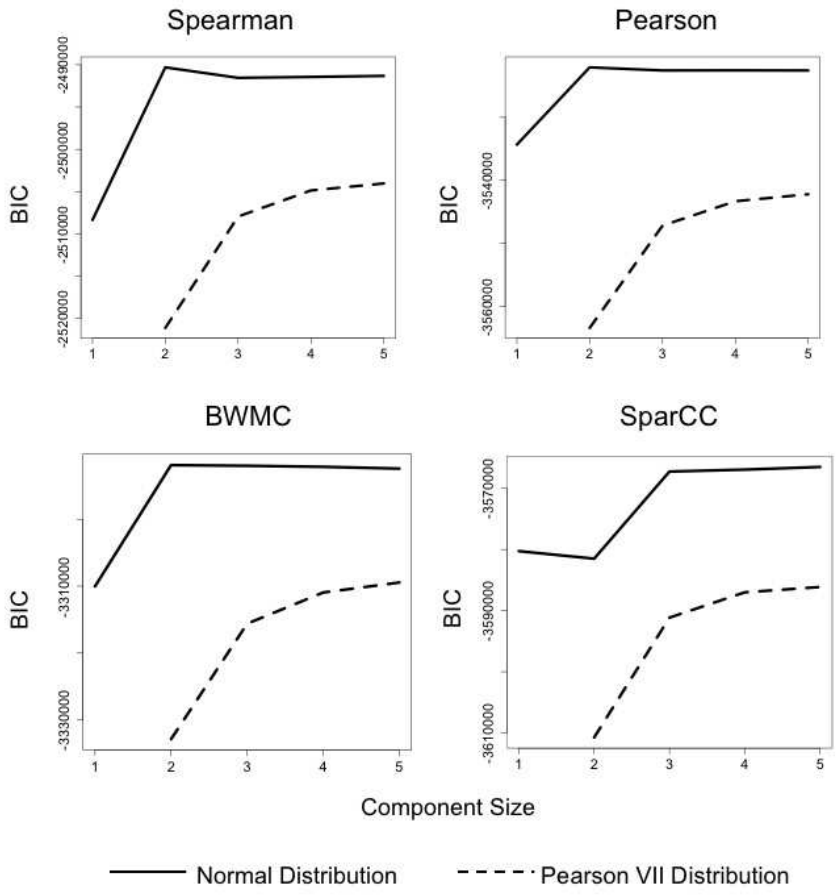


Figure 14. Comparison of Distributions (Normal vs. Pearson VII) and the Number of Mixture Components for the Breast Cancer Data and Various Correlation Metrics.

4.2. Continuous Data and Comparison to Other DC Methods

4.2.1. Simulations

Performances of the methods were evaluated using simulations to observe the concordance of predicted positives and negatives and true positives and true negatives. In the ROC curve, Discordant has more area under the curve (AUC) than any of the methods, and Fisher, linear interaction models and EBcoexpress have similar AUC (Figure 15a). The sensitivity and specificity were plotted to determine why the Discordant method has a better ROC curve (Figure 15b). While specificity is

the same for all three methods, Discordant method performs better with respect to sensitivity demonstrating that the Discordant method identifies more true positives than the other methods.

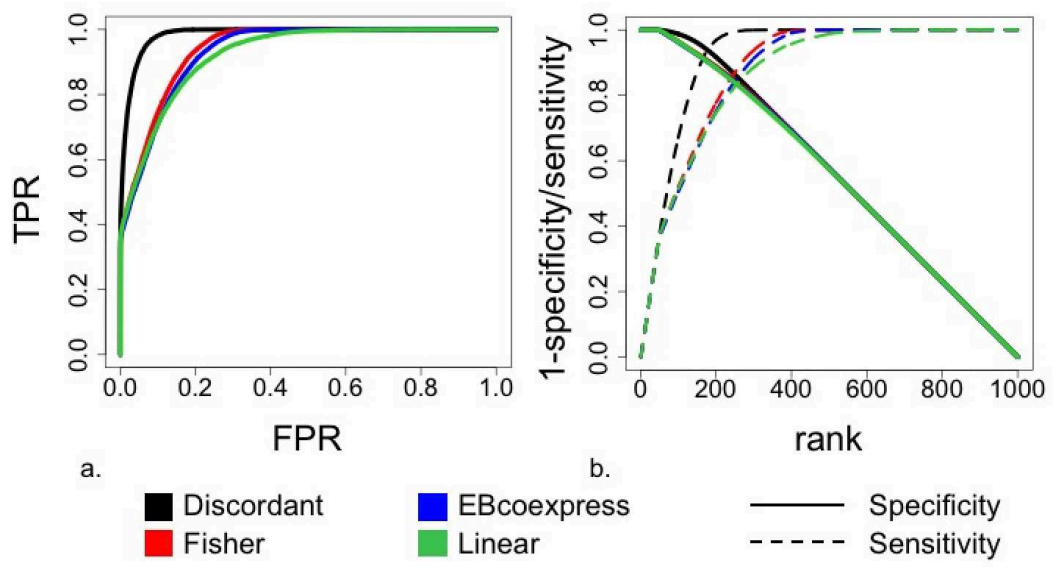


Figure 15. Simulations of Continuous Data. (a) ROC curve. Discordant AUC = 0.985, EBcoexpress AUC = 0.931, Fisher AUC = 0.940, Linear AUC = 0.930. (b) Senstivity/1-Specifcity plot.

The ROC curves and plots of sensitivity and specificity for adjusted simulation parameters are in Appendices C.1 and C.2. From the plots, change in sample size, the type of correlation used and the number of simulated differentially correlated pairs and feature pairs in the simulation did not affect performance except for disparate sample size for the linear interaction models.

To explore the predictions of paired correlation scenarios in the class matrix (Figure 4d), the distribution of the ranks for each class was displayed for each method. The smaller the rank, the more significant the method determined the feature pair to be. Class 3 is an example of disrupted differential correlation, where group 1 has a positive correlation and group 2 has a correlation close to 0, while

class 6 is an example of cross differential correlation, where group 1 has a positive correlation and group 2 has a negative correlation (Figure 5). In Figure 16a, the distribution of ranks for feature pairs that belong in class 3 are shown, where the ranks for Discordant are much smaller than Fisher or EBcoexpress, but in Figure 16b which shows molecular feature pairs that belong in class 6, the distribution of ranks is similar across all three methods. This confirms that binning, or the categorization of types of differential correlation, in Discordant achieves greater power for identifying disrupted DC, whereas all methods identify cross DC pairs at similar significance.

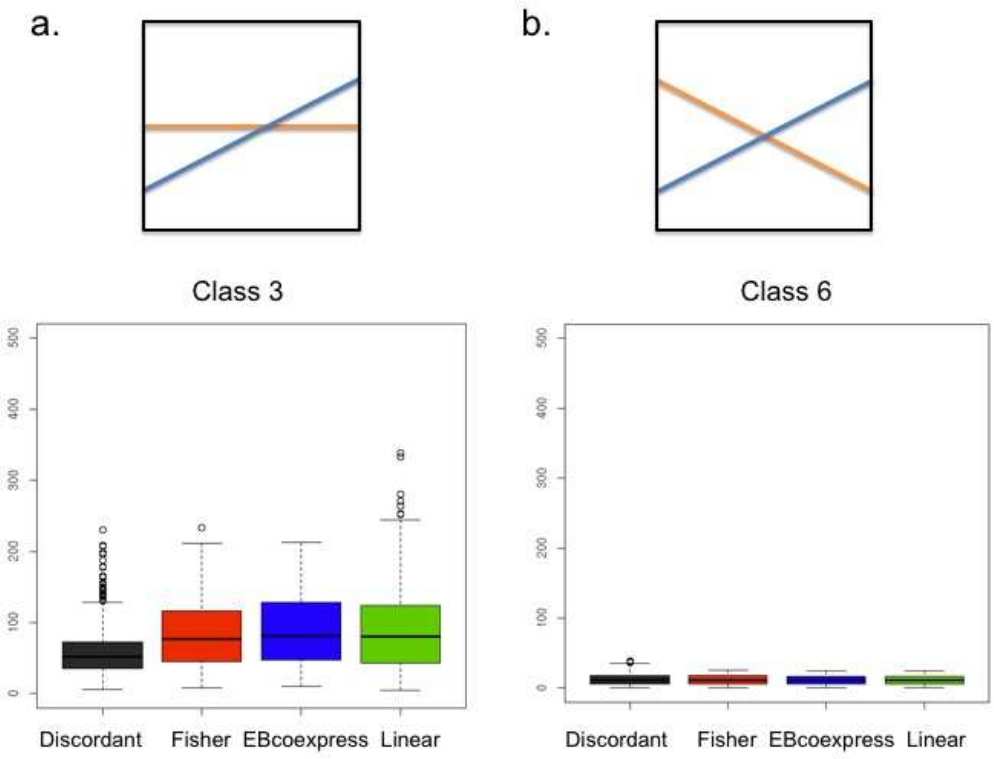


Figure 16. Distribution of Ranks for Classes 3 and 6 for all Methods. Groups 1 and 2 are orange and teal respectively. Black is Discordant, red Fisher, blue EBcoexpress and green Linear. (a) Class 3. (b) Class 6.

Boxplots for all classes are in Appendix C.3. Similar to Figure 16, Discordant has smaller distribution of ranks for all disrupted DC classes but similar distribution of ranks for all cross DC classes in comparison to other methods. Linear interaction models showed the worse performance in Figure 15a and 15b, which is found in the larger distribution of ranks in classes with disrupted DC and lower distribution of ranks in classes with no DC. EBcoexpress has much tighter, higher distribution of ranks for classes 5 and 9, which occurs when both groups have an association between molecular feature pairs, but they are in the same direction.

4.2.2. Biological Validation with Experimentally Validated Features

4.2.2.1. GBM miRNAs. The top ranks, p-values and q-values of the four unique GBM-related miRNAs pairs in Discordant, EBcoexpress, Fisher, miRNA-independent and transcript-independent linear interaction models were examined (Appendix D.1). The mean and median of these ranks are found in Table 2 and complete information is in Appendix D.1. It was found that Discordant had a smaller mean and median rank than the other methods, indicating that overall Discordant identifies unique GBM-related miRNAs more significant than any other method. Furthermore, at q-value < 0.05 Discordant identified all 4 GBM-related miRNAs, while EBcoexpress, Fisher and linear interaction models identify 3, 1 and 1 respectively. The top unique GBM-related miRNA pair, hsa-miR-92b and Agilent probe A_32_P56375, is plotted in Figure 17.

The linear interaction models identify miRNAs at a more significant rank than Discordant but the results are inconsistent between the linear interaction models when the independent and dependent variables are swapped (Appendix D.1). This

was further confirmed using a Wilcoxon Signed-rank test on the $-\log_{10}$(p-values) between the two models (p-value < 0.05).

Table 2. Summary Ranks of Experimentally Validated Features in GBM and COPD. Cells Highlighted in Grey Indicate Best Result.

GBM	Discordant	464.75	347.5
	EBcoexpress	815	607
	Fisher	781	801
	Linear (miRNA-Independent)	1095	532.5
	Linear (transcript-Independent)	2596.5	787.5
COPD	Discordant	5.08e5	2.14e5
	EBcoexpress	4.91e5	3.21e5
	Fisher	5.42e5	4.41e5

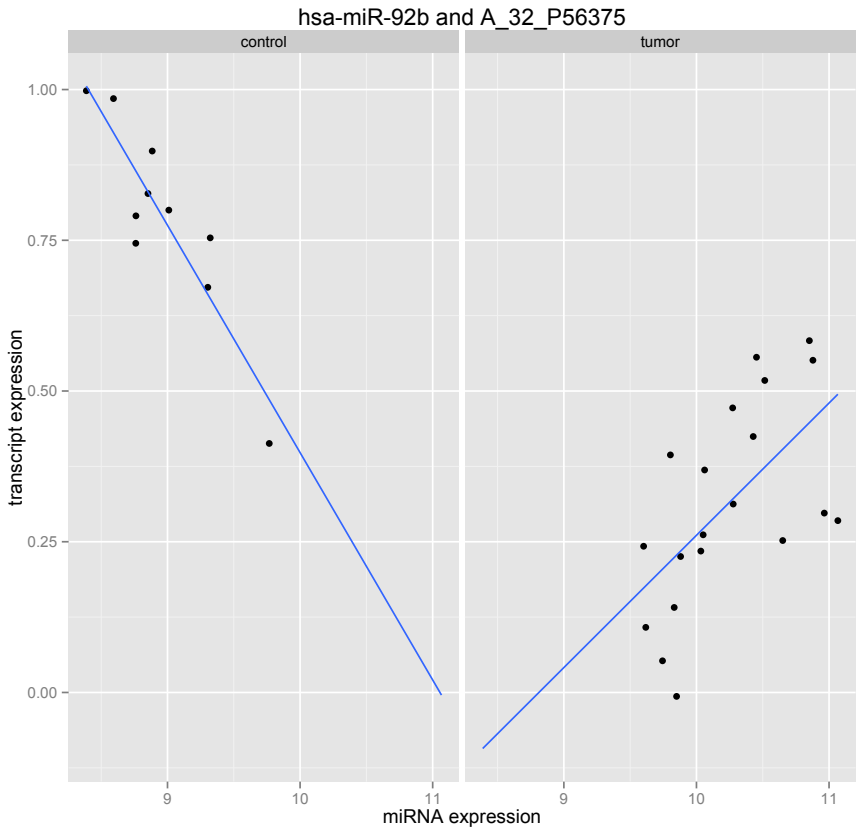


Figure 17. Top Example of Unique GBM-related miRNA Differential Correlation in GBM Data in Discordant. The gene name associated with the probe is unavailable since the probe is unannotated. Control samples are in the left panel, tumor samples on the right panel.

The frequency of GBM-related miRNAs and their associated classes were compared in Discordant and Fishers to determine the effect of binning on the analysis. It was found that the differentially correlated pairs with a GBM-related miRNA were more likely to be class 2 or 3, or disrupted DC, in Discordant (Figure 18a.1), in contrast to Fishers and EBcoexpress where there were relatively more pairs that were class 6 or 8, or cross DC (Figure 18a.2 and 18a.3).

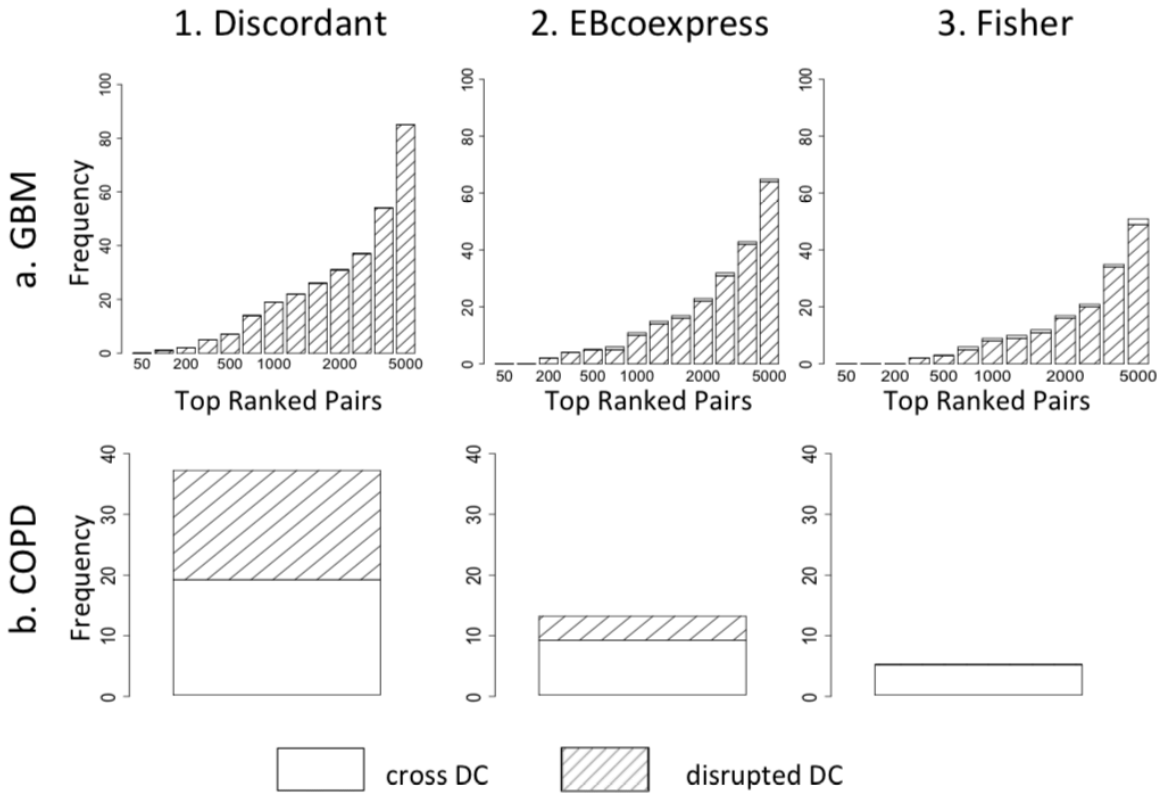


Figure 18. Disrupted vs. Cross DC found in GBM and COPD by DC Methods. a. Increasing frequency of classes in GBM. b. Classes of sphingolipid-related metabolite and gene pairs in top ranked 100,000 pairs (Discordant q-value = 0.08, EBcoexpress q-value = 0.35, Fisher FDR = 1).

4.2.2.2. COPD Sphingolipid-Related Features. The sphingolipid pathway has been previously implicated in COPD (Bowler et al., 2015). A list of sphingolipid-related metabolites and genes was acquired (Appendix A.3.), and the top rank and respective p-value or posterior probability and q-value when sphingolipid-related

pairs identified by the three methods was evaluated (Appendix D.2). The top ranked sphingolipid-related metabolite-transcript pair determined by Discordant, a sphingenine and PSAPL1, is plotted in Figure 19. PSAPL1 codes for a prosaposin, which is a precursor for saposins that cleave glycosphingolipids (Schnabel et al., 1992).

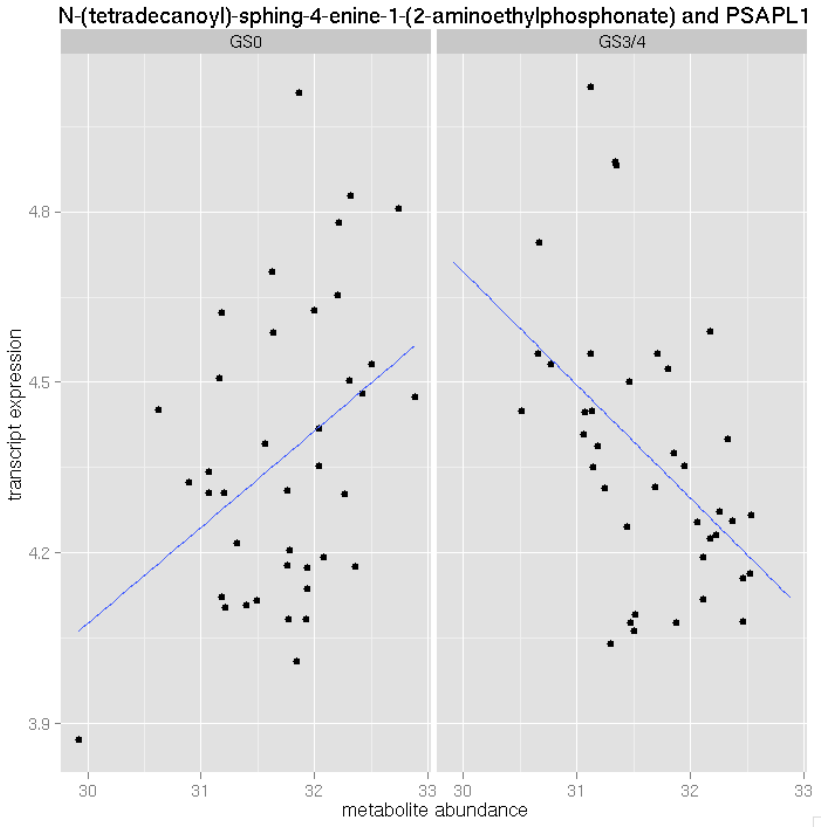


Figure 19. Top example of sphingolipid-related differentially correlated pair in Discordant. Control subjects (GOLD Stage 0) are on the left panel and more severe COPD (GOLD Stage 3/4) are on the right panel.

In Table 2, it was found that the median sphingolipid-related pair rank is smaller for Discordant compared to EBcoexpress and Fisher. The EBcoexpress' mean rank is smaller than Discordant, but only by 2e4 where the median rank between Discordant and EBcoexpress differs by 1e5. At q-value < 0.10 Discordant identified 146 sphingolipid pairs, whereas EBcoexpress identified 1 and Fisher 0.

The findings here indicate that overall Discordant identifies sphingolipid-related feature pairs earlier than the other two methods. However, the ranks are much later in the hundred thousands. This may indicate that although the sphingolipid pathway could be relevant to COPD, there may be other pathways that contribute to the complexity of the disease that may appear more significant in relation to the phenotype.

Since the sphingolipid pathway is not as significantly differentially correlated in COPD, we examined the classes of sphingolipid-related metabolite and gene pairs that were in the top ranked 100,000 pairs. We found that Discordant identified relatively more disrupted classes than EBcoexpress or Fisher (Figure 18b).

4.2.3. Novel and Known Targets

4.2.3.1. GBM miRNAs. Pairs with Discordant posterior probability greater than 0.99 were used to investigate which features had the most connections, or hubs. The top 4 genes that were the biggest hubs with over 30 connections are: AGAP2, CRY2, GRIN1, and UPF3A (Table 3). Most of these genes have functions that are central to the brain, where GBM occurs. AGAP2 is an Arf GAP that has anti-apoptotic effects of nerve growth factor (Inoue and Randazzo, 2007), CRY2 is a circadian rhythm gene that principally is localized in the brain, GRIN1 is a ligand-gated ion channel that facilitates signals through neurons (Wahlsten, 1999). UPF3A is found in the UPF complex that is implicated in pathways altered in cancer such as post-splicing, mRNA decay and nuclear export (Dreyfuss et al., 2002). None of these genes have previously been implicated in GBM.

The miRNA hsa-miR-545 was the biggest hub connected to 39 genes, which is visualized in Figure 20. hsa-miR-545 has not been found to be involved in GBM. Ten of the connected genes are annotated as being transmembrane proteins, and 3 of these are serine/threonine kinases (CDC2L2, PDPK1 and BMPR2). Tyrosine kinases have been found to be involved in GBM and are similar to serine/threonine kinases (Hamza and Gilbert, 2014).

Table 3. Summary of Gene Hubs with Most Connections in Pairs in GBM Analysis (q-value < 1.0e-4).

Gene	Connections	Annotations
AGAP2	60	Anti-apoptotic effects of nerve growth factor
CRY2	39	Circadian protein
GRIN1	34	Glutamate receptor, form ligand-gated ion channel
UPF3A	34	Part of post-splicing multiprotein complex involved in mRNA decay and nuclear export.

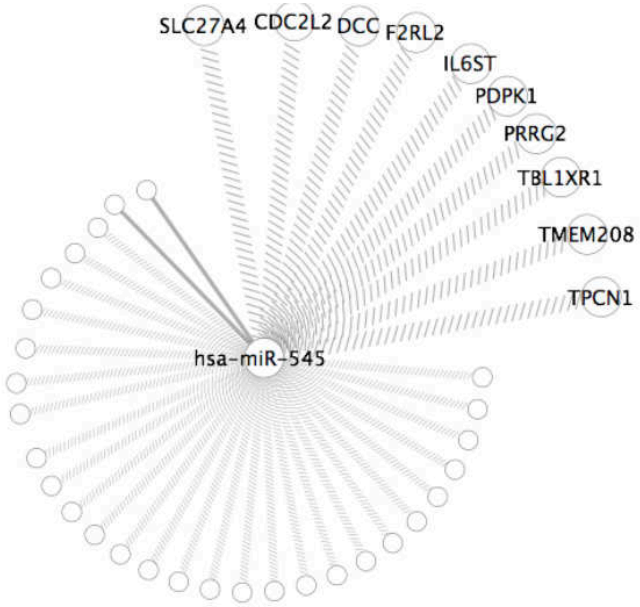


Figure 20. GBM Network of miRNA with Most Significant Connections to Genes. Solid edges are cross DC, dashed edges are disrupted DC. Transmembrane genes are labeled.

4.2.3.2. COPD Sphingolipid-related Features. Molecular features that had the largest hubs were identified and listed in Table 4. IGHG1, or immunoglobulin heavy constant gamma 1 is considered a relevant result since immunity plays a central role in COPD (Rovina et al., 2013). Another gene identified as a hub is SARDH, or sarcosine dehydrogenase which has been implicated previously in COPD (Ubhi et al., 2012). The metabolite that has the largest hub has yet to be formally annotated; its chemical formula is $C_{20}H_{33}N_9P_2S$. The other metabolite that was a large hub is L-Valine, a metabolite involved in multiple biochemical pathways.

Table 4. Summary of Metabolite and Gene Hubs with Most Connections in COPD Analysis (q-value 2.0e-3).

Type	Name	Connections	Annotation
Gene	LOC284561	247	unknown
Gene	SARDH	265	Sarcosine dehydrogenase
Gene	IGHG1	294	Immunoglobulin
Metabolite	C20 H33 N9 P2 S	2222	Unknown
Metabolite	L-valine	1667	Amino acid

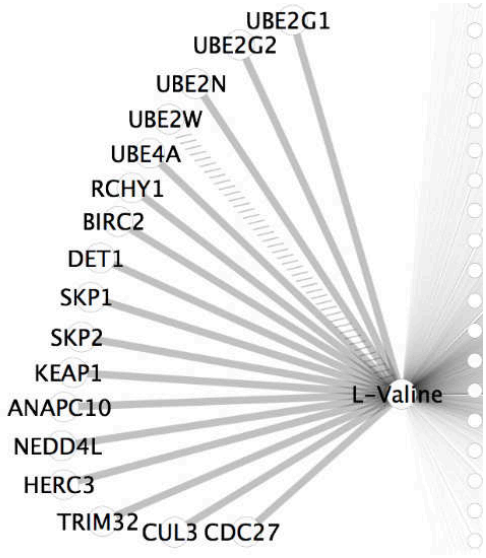


Figure 21. COPD Network of Metabolite with Most Significant Connections to Genes. Solid edges are cross DC, dashed edges are disrupted DC. There are 17 genes involved in ubiquitin mediated proteolysis connected to L-Valine in COPD, L-Valine connected in total to 1667 genes.

Genes connected to L-Valine were investigated using DAVID to determine if they were enriched in a biological pathway that is implicated in COPD (Huang et al., 2008a, 2009). The ubiquitin mediated proteolysis KEGG pathway was most enriched in the L-Valine differential correlated gene set with q-value = 0.001. In Figure 21 the genes involved in this pathway are highlighted from the rest of the other genes, which total to 17 out of 1667 in the gene set. In previous studies, the ubiquitin protease degradation pathway has been associated with COPD (Ottenheijm et al., 2006).

4.3. Count Data and Comparison of Correlation Metrics

4.3.1. Simulations

In our context and others, count data consist of the number of reads mapped to each protein coding region and is commonly modeled with a negative binomial distribution. Correlation metrics have not been thoroughly investigated in count data, so to determine if sequencing data could be applied to Discordant several multiple correlation metrics were assessed. We also investigated Generalized Linear models for negative binomial distributions (NBGLM) with an interaction score, similar to the linear interaction models examined in the continuous data. Fisher and EBCoexpress differential correlation were not run on the sequencing data, since the continuous analysis demonstrated that Discordant had better performance than those methods.

Four correlation metrics (Spearman, Pearson, SparCC, BWMC) and NBGLM were applied to simulated data to assess performance in identifying DC molecular feature pairs. Three methods were applied to count data (Spearman, SparCC, NBGLM) and two methods were applied to transformed data (Pearson, BWMC), as

explained in section 3.4. Figure 22 shows the sensitivity and specificity of the methods.

NBGLM has less area under the curve and lower sensitivity than all other correlation metrics applied to Discordant. Pearson and BWMC perform similarly, most likely because the mean and median are close to each other and Pearson is mean-based and BWMC is median-based. SparCC has the second highest performance. Out of all correlation metrics, Spearman had greater area under the curve in both the ROC curve and the sensitivity curve. Observing rank distributions of each class show similar results (Appendix C.4.). Since Spearman's correlation was the metric that had the best performance, it is used in the simulations and biological validation for the extensions that were explored.

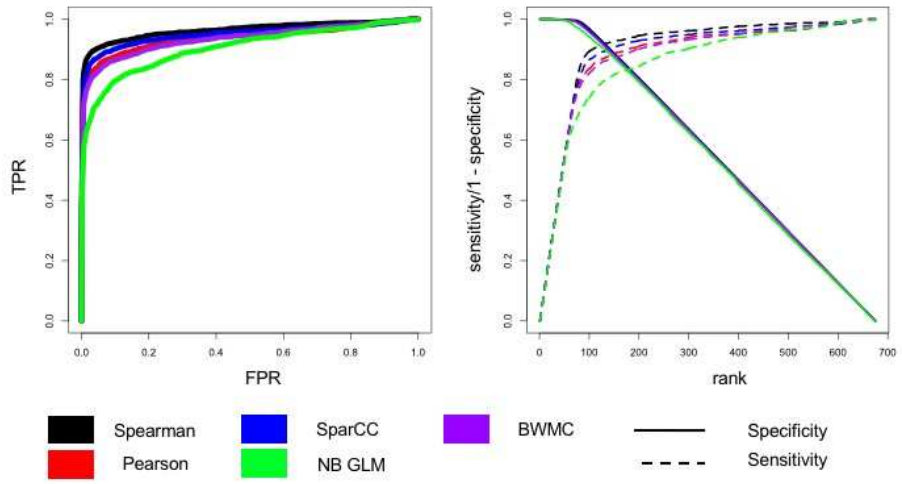


Figure 22. Simulations of Count Data. (a) ROC curve. Spearman AUC = 0.96. Pearson AUC = 0.94. SparCC AUC = 0.95. BWMC AUC = 0.94. GLMNB AUC = 0.90. (b) Rank vs. Sensitivity/Specificity.

4.3.2. Biological Validation with Breast Cancer miRNAs

To identify interacting miRNA-mRNA pairs that may change due to tumor status, we evaluated miRNA and mRNA sequencing data from the TCGA database for breast cancer. Discordant was run with four different correlation methods (Spearman, Pearson, BWMC and SparCC) and the model based method NBGLM. In Table 5, the average ranks and 1 – posterior probability (1-PP) of the most significant pairing breast cancer miRNA with a gene is shown (complete information in Appendix D.3). We report 1-posterior probability instead of posterior probability so differences between values are more distinct. For almost all our benchmark breast cancer miRNA, Spearman correlation finds them to be ranked lower than any other method. The results are similar to those in simulations except for SparCC, and NBGLM. In the simulations, SparCC performed second to Spearman but in the biological validation it is relatively worse. NBGLM performs better in the biological validation than in the simulations, performing only behind Spearman and BWMC. Also, the average p-value for NBGLM is small, but it was found in the simulations that the distribution of p-values derived from NBGLM were generally skewed towards 0 compared to other methods.

Table 5. Average of Ranks and 1-PP or p-value of Top Results of Feature Pairs with Breast Cancer miRNA. Results shaded in grey indicate top performing metric.

Correlation Metric/Method	Rank	1-PP or p-value
Spearman	89	0.0099
SparCC	543	0.0215
BWMC	294	0.0376
Pearson	626	0.0190
NBGLM	472	0.0037

It was also determined that the breast cancer dataset followed the same trend of identifying more disrupted classes than cross classes. This was performed using

Spearman's correlation metric. Figure 23 demonstrates that most significant molecular feature pairs have disrupted DC instead of cross DC. The top result with a breast cancer miRNA is hsa-miR-152 vs. TMC22 (Figure 24).

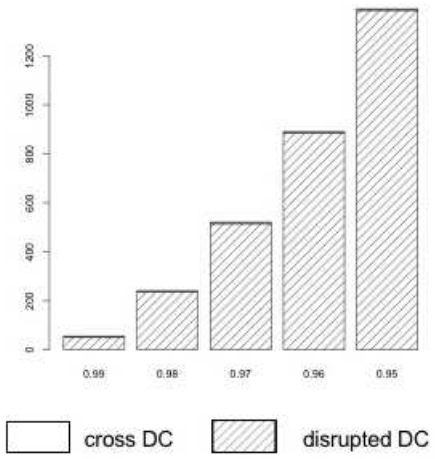


Figure 23. Disrupted vs. Cross DC found in Breast Cancer Using Spearman's Correlation. Posterior probability < 0.95.

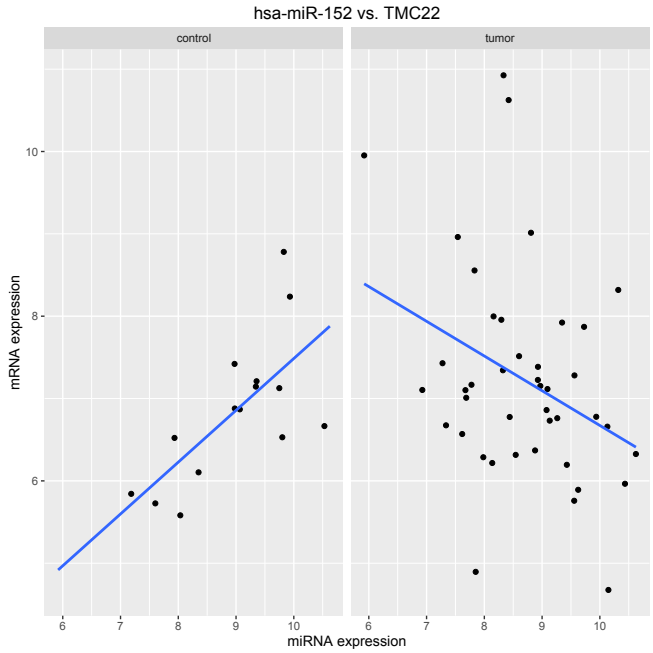


Figure 24. Top Example of Feature Pair with Breast Cancer miRNA Differentially Coexpressed Pair in Discordant.

4.3.3. Novel and Known Targets

Molecular feature pairs with q-value less than 0.05 were investigated for hubs. The top hubs are summarized in Table 6. The top hub is VAMP1, which previously has not been implicated in breast cancer. RNU11 has been found to have significant upregulation in breast tissue before and after hormone therapy in female-to-male transsexuals (Bentz et al., 2010). SESN3 is a subunit of a protein complex that activates a signal cascade that turns on an influential tumor suppressor (Sanli et al., 2012). LYPD1 has been shown to be overexpressed in bone cells that have metastasized from breast cancer (Burnett et al., 2015).

Table 6. Summary of Gene Hubs with Most Connections in Breast Cancer Data. q-value < 0.05.

Gene	Connections	Annotation
VAMP1	32	Vesicle-Associated Membrane Protein 1 (Synaptobrevin 1), part of SNARE complex
RNU11	24	splicing
SESN3	24	Sestrin, stress-induced protein
LYPD1	23	NA

The miRNA with the most significant pairs with q-value < 0.1 is hsa-mir-664, with 327 gene connections. DAVID enrichment analysis was performed and identified that 19 of these genes were annotated with the keyword cell adhesion in the SwissProt/UniProt database (Bairoch, 2004), with Bonferonni adjusted enrichment p-value < 0.1 (Figure 25). Cell adhesion genes facilitate angiogenesis, or a process where new blood vessels are formed from pre-existing vessels. The new blood vessels propagate metastasis (Horak et al., 1992) and metastasis in breast cancer patients is commonly fatal (Li and Feng, 2011).

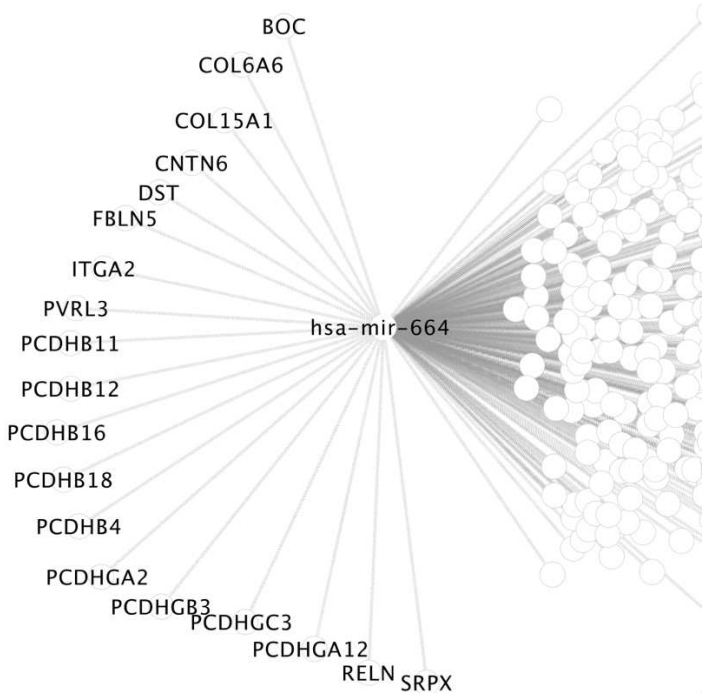


Figure 25. Breast Cancer Network of miRNA with Most Significant Connections to Genes. hsa-mir-664 and significant connected genes with q-value < 0.1 are shown. All pairs are disrupted DC. Genes involved in cell adhesion, Bonferroni p-value < 0.1 have labels.

4.4. Extensions

4.4.1. Subsampling

The standard EM algorithm and subsampling version of the EM Algorithm were compared to assess performance effects. The ROC curves and the sensitivity/specificity are similar for the two versions (Figure 26). This is also evident when looking at each class posterior probability separately for both continuous and count simulations (Appendix C.5 and C.6).

We evaluated the results from the subsampling version on the TCGA Breast Cancer and Glioblastome Multiforme datasets. Appendix D.4 shows that the subsampling with EM had similar ranks for breast cancer miRNAs, but higher posterior probabilities than the standard EM implementation. GBM had ranks that

were close between standard EM and subsampling EM, but not as close as the Breast Cancer analysis.

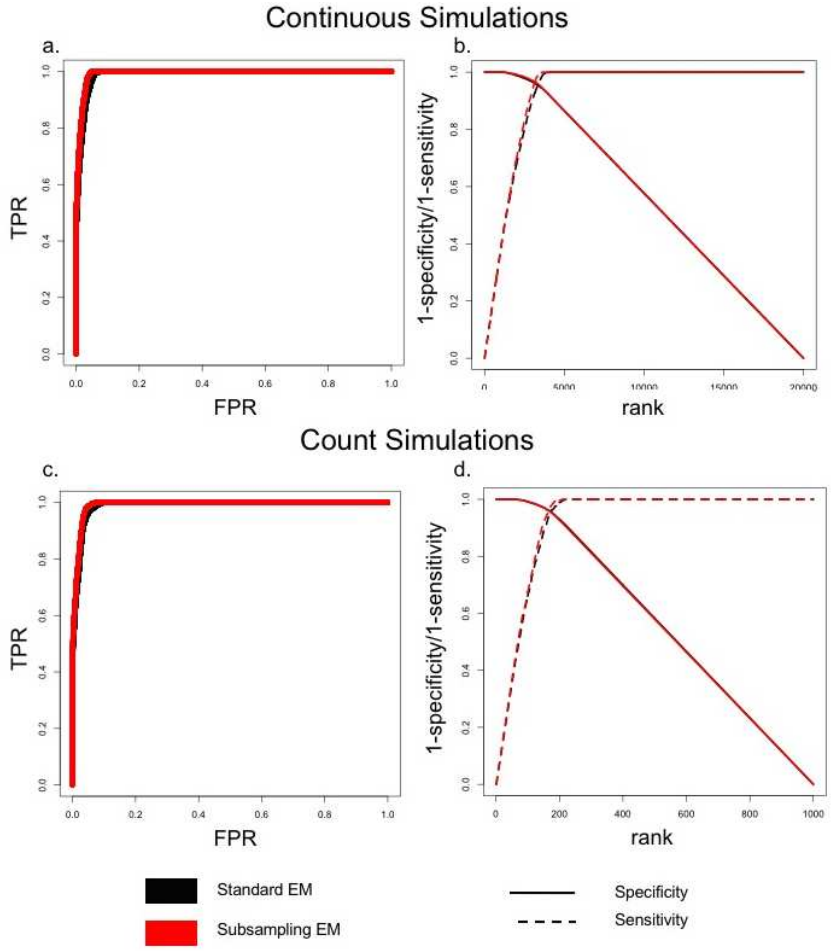


Figure 26. Analysis of Continuous and Discrete Simulations with Subsampling Optional Argument. Continuous (a) ROC. Standard EM AUC = 0.998, Subsampling EM AUC = 0.993. (b) Rank vs. Sensitivity/1-Specificity. Count (a) ROC. Standard EM AUC = 0.990, Subsampling EM AUC = 0.993 (b) Rank vs. Sensitivity/1-Specificity.

4.4.2. Three vs. Five Component Mixture Model

The 3-component normal mixture model has greater power than 5-component mixture model for both continuous and discrete simulations (Figure 27). In Appendix C.7. and C.8. the posterior probability distributions are plotted for each class for both continuous and count simulations. For cross and disrupted DC, the 5-component

mixture model produces higher posterior probabilities. In the 3-component mixture model, 6 out of 9 of the classes are DC, whereas for 5-component mixture models there are 20 out of 25. The posterior probabilities for 5-component mixture models may be larger because a greater proportion of the total class posterior probabilities are used to summarize the final DC posterior probability.

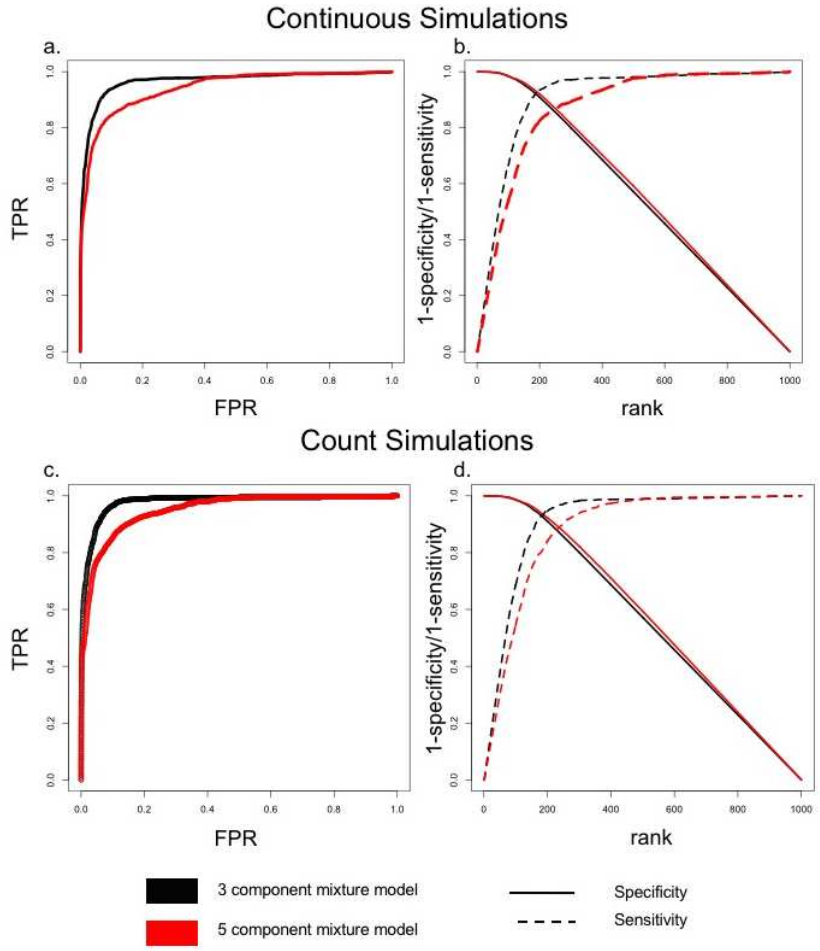


Figure 27. Analysis of Continuous and Discrete Simulations of 3-Component vs. 5-Component Mixture Models. Continuous (a) ROC. 3 Component Mixture Model AUC = 0.97, 5 Component Mixture Model AUC = 0.96. (b) Rank vs. Sensitivity/1-Specificity. Discrete (a) ROC (b) Rank vs. Sensitivity/1-Specificity.

Elevated DC is the new type of DC introduced in the 5-component mixture model, which occurs when there is an association between feature pairs in both groups but it is stronger in one of the groups. Elevated DC is no DC in the three

component mixture model. In the elevated DC boxplots, we expect that the posterior probability distributions to be closer to 0 for the 3-component mixture model, but closer to 1 for the 5-component mixture model. From the boxplots in Appendices C.7. and C.8. it is apparent that the 5 component mixture models are unable to make clear distinctions between the – and – – components and the + and ++ components. The 3-component mixture model has tight rank distributions close to 0 for all classes that are elevated or no DC, but the 5-component mixture model has rank distributions that have larger variation. Although our simulations show reduced performance, this option is included for users interested in these types of associations.

Appendix D.5 shows the ranks and posterior probability of the most significant breast cancer miRNA gene pair when Discordant is run with the 3- or 5-component mixture models. The 3-component mixture model has lower ranks but the 5-component mixture models have higher posterior probabilities. This is the same for both the GBM and Breast cancer data. As mentioned for the simulation results, we expect that the final summarized posterior probabilities to be higher for 5-component mixture model than the 3-component mixture model since a great proportion of classes are used to summarize the posterior probability of DC.

CHAPTER V

DISCUSSION

5.1. Conclusions

In this Chapter, we will discuss the improved performance of Discordant compared to other methods in the continuous analysis, the comparison of correlation metrics in the count analysis, and finally the effects of subsampling and increasing mixture components to 5 in the mixture model. Evidence presented in Chapter 4, the Results section, will be used to achieve this.

5.1.1. Continuous Simulations and Biological Data

A fundamental feature of Discordant is the ability to to categorize the paired correlation scenarios, or “binning”, enabling Discordant to determine more differentially correlated pairs than the other methods and improves power of detecting disrupted interactions. Binning not only improves performance for the Discordant method but also facilitates biological interpretation of results since it categorizes the different types of dysregulation between biological groups. As seen in Figure 18, Discordant identifies more disrupted differentially correlated pairs than EBcoexpress and Fisher, a trend also found in the simulations (Appendix C.2). Discordant also identifies more significant phenotype-related feature pairs in general for both GBM and COPD.

The GBM dataset produced more significant DC results for phenotype-related features than the COPD dataset. The GBM validation set is well curated because there are experimentally validated miRNAs involved in GBM, whereas for COPD there is less known about the molecular pathways. The sphingolipid-related genes

and metabolites were determined by annotation for being in sphingolipid pathways, because there is limited experimental data for specific genes and metabolites. Despite the challenges of the COPD dataset, we did observe that sphingolipid metabolite-gene pairs were identified as more significant in Discordant than EBcoexpress and Fisher (Table 2) and that there were more sphingolipid metabolite-gene pairs in the top 100,000 pairs in Discordant than EBcoexpress and Fisher (Figure 18). Also, at q-value < 0.05 all four GBM miRNAs were identified in a significant pair by Discordant, while three were identified by EBcoexpress, and 1 by both Fisher and linear interaction models. At q-value < 0.10, Discordant identified 146 sphingolipid pairs to be significant, while EBcoexpress identified 1 sphingolipid pair and Fisher identified no sphingolipid pars.

Applications to both GBM and COPD have promising results of known and novel targets from Discordant. This confirms Discordant's ability to identify phenotype-related biological processes and indicates the potential that Discordant can produce further testable hypotheses.

A related method is to apply linear models with interaction terms. One of the benefits of linear models is that it assumes conditional normality instead of joint normality, meaning that the independent variable can be non-normal. Linear models identified GBM-related miRNA pairs in earlier ranks than Discordant in the GBM data, but linear models can be difficult to use since it is unclear what should be the dependent and independent variable. We explored this by switching miRNA and transcript as the dependent and independent variable and we found it changed the results. We also found that the ranks of unique GBM-related miRNA pairs were

different between the two analyses. It is highly suggested to only use linear models if the independent and dependent variable are known in advance, such as miRNA and transcript respectively.

In terms of run-time, Fisher is notably faster than the rest of the methods, EBcoexpress is the slowest and Linear and Discordant only differ slightly (Table 7). The computational complexity notation for Fisher is linear, O(n), where n is the number of feature pairs. For the linear interaction model and Discordant it is polynomial, $O(n^2)$ and $O(2n + 3n^2)$ respectively. The complexity for EBcoexpress is not as simple to identify since there are nested EM algorithms. EBcoexpress requires about about three fold the run-time as Discordant in the GBM and COPD datasets, and it also requires a grid approach to determine hyperparameters. While Discordant does not run faster than Fisher and its run-time is comparable to linear interaction models, it still performs either equally or better with consistent results.

Table 7. Run-time of Methods for GBM and COPD data. Analyses run on 11 Intel(R) Xeon(R) CPU E5-2640 0 2.50GHz processors, 90 GB of memory.

Method	GBM	COPD
Discord	19.26 hours	2.16 days
EBcoexpress	10.42 days	39.53 days
Fisher	6.35 seconds	18.85 seconds
Linear	16.15 hours	N/A

5.1.2. Count Simulations and Biological Data

The Discordant method was tested for its applicability to sequencing data and other platforms that produce discrete or count data. Correlation metrics were compared in simulations and TCGA breast cancer data. BWMC and Pearson have similar ROC curves but BWMC demonstrated more power in the biological

validations. This may be because BWMC is more robust to the presence of outliers and non-symmetric distributions.

NBGLM has increased performance in the biological validations compared to the simulations. In generalized linear models, the interpretation of interaction terms can be challenging because the interaction term is modeled on the scale determined by the link function (e.g. log scale for the negative binomial). (Tsai and Gill, 2013). This makes the results from the NBGLM modeling different than the other correlation metrics examined, which are not scale transformed. Other limitations of the NBGLM modeling include pre-specifying the dependent and independent variables. In the miRNA-mRNA example, this was straightforward because targeting by miRNA is known to promote mRNA degradation and therefore miRNA can be considered the independent variable and mRNA the dependent variable. However, for other types of comparisons (e.g,. transcriptomics and metabolomics) the direction of the effect may not be as straightforward to make those specifications. Furthermore, as discussed above, we found that mis-specifying that relationship decreases performance (Siska et al., 2015).

SparCC demonstrated better performance in the simulations compared to the biological validation. In the simulations, the correlated mRNAs to miRNAs were generated based on the mean of the miRNA and the dispersion of the mRNA (Step 3 in section 3.1.2). The distributions were more similar in the correlated pairs in the simulations than the feature pairs in the biological validation since their distributions shared similar parameters. SparCC predicts the actual values based on the

dispersion of the observed values; therefore using two different types of –omics with their own unique variation may not be suitable.

Spearman's correlation metric demonstrated the most power compared to all other metrics in both the simulations and the biological validation. Spearman's correlation is a non-parametric rank-based metric that makes it well suited for non-normal distributions. Using non-parametric methods when integrating datasets with different variation is favorable, and may explain why Spearman has greater power in both simulations and biological validation.

Although we found improved performance with Spearman's correlation in our simulations and sequencing data, the preferred correlation metric depends on the type of data and study. For example, if the user wants a more conservative method SparCC should be used, but for normal continuous data Pearson's correlation is the natural choice. For these reasons, the correlation metric is a user-defined option in the Discordant R package.

Similar to the qualitative analysis in COPD and GBM continuous datasets, breast cancer count data was able to identify genes that either have already been implicated in breast cancer or were involved in pathways that are connected to breast cancer biology. This further demonstrates Discordant's ability to identify phenotype-related biological processes and generate testable hypotheses in both normally and non-normally distributed data.

5.1.3. Extensions

The Discordant R package provides additional modeling and implementation options compared to existing software DiffCorr and EBcoexpress (Dawson et al.,

2012a; Fukushima, 2013). The subsampling extension to the EM algorithm makes the model more computationally tractable (with run-time decreased by 40 fold), but also solves the problem of dependencies between pairs. There were some inconsistencies, such as the higher posterior probabilities for subsampling compared to the posterior probabilities with no subsampling even though the ranks between subsampling and no subsampling were similar. The posterior probabilities may be larger for two reasons: the parameters for each mixture component are better estimated since the independence assumption is no longer violated and the parameters are averaged over 100 iterations, making the standard error very small. GBM had similar results, but the ranks were more variable. One reason for this is that there are less mRNA in the Breast Cancer dataset compared to the GBM dataset (72656 in GBM dataset compared to 17414 mRNA in breast cancer dataset), but similar numbers of miRNA (313 in GBM dataset to 200 miRNA in Breast Cancer dataset) making the proportion of subsampled pairs to total pairs smaller. This demonstrates that the number of independent pairs are limited by the dimensions of the data, and having fewer independent pairs to subsample from hinders performance. There is also the issue of selection bias, since only correlation coefficients that are independent of each other are used.

Users applying subsampling should also be aware of selection bias, since only correlation coefficients that are independent of each other are used. Another limitation of subsampling is the feature size of independent pairs is limited by the dimensions of the data.

Expanding the normal mixture model from 3 components to 5 components gives the user the opportunity to explore additional and subtle types of DC. The addition of extra classes does reduce power for Discordant and increase run time by a factor of three, so we advise users to consider the 5-component mixture model if elevated DC is relevant to their study and a 5-component mixture model is justifiable based on model selection criteria such as Bayesian Information Criteria (BIC).

5.2. Limitations

There are some limitations to Discordant. We assume independence between pairs, which is inaccurate since features show up in multiple pairs. This assumption is critical to reducing computational complexity, and has been made by others (Dawson et al., 2012). The subsampling extension aims to solve the problem of dependencies between pairs, however it has issues which are outlined in section 5.1.3.

Appropriate sample size is necessary for Discordant or any other DC method to work effectively to accurately estimate *r*, or the correlation coefficient, between two features. An adequate number of samples should be available in order to measure associations. A power analysis using function `pwr.r.test` from library `pwr` for correlation coefficient equal to 0.5, significance level 0.05 and power 0.8 requires at least 29 samples. Unfortunately, sometimes it is difficult to acquire this many samples. Control samples in cancer studies are hard to obtain. For example, the control groups from breast cancer and GBM have 15 and 10 samples respectively. Discordant is capable of handling datasets with different sample sizes

since the initial parameters are set by b_v (where v = 1,2 for Group 1 or 2), making the mixture components comparable regardless of scale.

Lastly, the model assumes there are three Gaussian components in the mixture model. To explore the Gaussian assumption, we measured the BIC of 1 to 5 components with normal or Pearson VII distributions using R packages `mixtools` or `lcmix` (Benaglia et al., 2009, Dvorkin et al., 2013). Across these three datasets, normal mixture models had better fit than Pearson VII mixture models. Also, there was better fit with 3 components or it was negligibly different. All three of the biological datasets do not violate the 3 component normal mixture model, and this may be true for many other biological datasets, but should be evaluated before applying the method.

5.3. Interpretation and Module Building

There are not many tools to interpret feature pair lists derived from differential correlation as compared to feature lists derived from differential expression. We interpreted our analysis by identifying hubs, or features that have many significant connections. This process determined features that were already implicated in the phenotype, or could easily be related. We aim to enhance interpretation by developing a module-building extension and incorporating it into the Discordant R package. Since Discordant produces posterior probabilities for each feature pair, the approach would use algorithms that add or remove one feature at a time, such as the Fang et al. and Kostka et al. approaches, rather than using clustering methods.

Fang et al develops a score that measures differential correlation of a feature set, and then uses the Apriori algorithm to increase the score by either removing or

adding features (Fang et al., 2009). The Apriori algorithm uses breadth first search, which begins at the root node of a tree. It iterates through all nodes in the tree by searching the neighbors of the node it is currently at. Features are added or removed based on if they meet a threshold *d*. Kostka et al. has a similar framework to Fang et al. Differential correlation is measured using the mean squared regression of an additive model. A greedy stochastic downhill search algorithm is used, which is much like the Apriori algorithm except features are either added or removed to reach a local minima (Kostka and Spang, 2004).

Since the Discordant algorithm produces posterior probabilities, the Kostka et al. approach may be better suited for module building. Module building is based on an arbitrary threshold in Fang et al. whereas Kostka et al determines feature subsets that are locally optimal. Our future aim is to create a module building extension based on the Kostka et al. algorithm.

5.4. Multiple Groups

Currently, Discordant only determines significant differences between two groups (e.g. control vs. disease). In some studies, there may be multiple biological groups examined. For example, the clinical variables for each subject in the COPD data provides many opportunities to have more than two groups. The GOLD stages could be separated into five groups (0, 1, 2, 3, 4) instead of two groups (0, 3/4). Unfortunately, when trying to create a study with more than two groups in the COPD data the sample size in at least one group was too small. Other datasets will be used for examining Discordant's ability to identify differences in more than one group.

Some options are studies with time-series data, developmental profiles or with multiple treatments.

Another approach is to treat the phenotype as a continuous variable instead of an ordinal variable. In the COPD data, the FEV_1/FVC ratios are continuous and used to categorize subjects into GOLD stages based on thresholds. A future direction is to modify Discordant so that it can be applied to continuous phenotypes instead of categorical ones.

5.5. R Package

The Discordant R package is available on github at github.com/siskac/discordant. It has been designed to be flexible to the user's preference, and also to be computationally efficient. For example, the EM algorithm is contained in a C wrapper since C is faster than R. Several options have been implemented, such as alternative correlation metrics and the subsampling and 5-component mixture model discussed here.

5.6. Overall Summary

In this thesis, we have demonstrated that Discordant performs better than other competing methods, can generate testable relevant hypotheses and is usable with low computationally complexity and a released R package. For the first time in the current literature, an R package that measures differential correlation has been proven to be applicable to sequencing data. Also, the limitations of Discordant verify known concerns and reveal new ones; information that will strengthen future studies.

Differential correlation is not as popular as differential expression, but it has the same potential of generating testable hypotheses that can lead to new

discoveries (de la Fuente, 2010). The two analyses explore different types of biological complexity. Differential expression identifies features that have large shifts in expression or abundance, whereas differential correlation identifies feature pairs that have different associations between groups. Both of these analyses can provide clues to the biological processes that affect the phenotype. Differential expression identifies features that are the “low-hanging fruit,” or the biomarkers that are used to validate and sometimes diagnose disease, whereas differential correlation can reveal biological processes that are disregulated in disease and predict key players that ignite the signal transduction to affect the expression of biomarkers. There are many established packages and tools to analyze –omics with differential expression, such as limma, DiffSeq, DAVID, GSEA etc. (Anders and Huber, 2010; Huang et al., 2008b; Ritchie et al., 2015; Subramanian et al., 2007) but there are no standards for differential correlation. In our development of Discordant, we have introduced a method and R package that is a reliable and robust approach to perform differential correlation.

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Agilent Technologies (2007). Considerations for Selecting GC/MS or LC/MS for Metabolomics. Agil. Technol.

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression. PLoS Comput. Biol. *9*, e1002955.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anderson, K.C., and Carrasco, R.D. (2011). Pathogenesis of Myeloma. Annu. Rev. Pathol. Mech. Dis. *6*, 249–274.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Atchison, J., and Shen, S.M. (1980). Logistic-normal distributions:Some properties and uses. Biometrika *67*, 261–272.

Bahr, T.M., Hughes, G.J., Armstrong, M., Reisdorph, R., Coldren, C.D., Edwards, M.G., Schnell, C., Kedl, R., LaFlamme, D.J., Reisdorph, N., et al. (2013). Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease. Am. J. Respir. Cell Mol. Biol. *49*, 316–323.

Bairoch, A. (2004). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154–D159.

Bentz, E.-K., Pils, D., Bilban, M., Kaufmann, U., Hefler, L.A., Reinthaller, A., Singer, C.F., Huber, J.C., Horvat, R., and Tempfer, C.B. (2010). Gene expression signatures of breast tissue before and after cross-sex hormone therapy in female-to-male transsexuals. Fertil. Steril. *94*, 2688–2696.

Borrebaeck, C.A., and Wingren, C. (2007). High-throughput proteomics using antibody microarrays: an update. Expert Rev. Mol. Diagn. *7*, 673–686.

Bowler, R.P., Jacobson, S., Cruickshank, C., Hughes, G.J., Siska, C., Ory, D.S., Petrache, I., Schaffer, J.E., Reisdorph, N., and Kechris, K. (2015). Plasma Sphingolipids Associated with Chronic Obstructive Pulmonary Disease Phenotypes. Am. J. Respir. Crit. Care Med. *191*, 275–284.

Bradley, P.H., Brauer, M.J., Rabinowitz, J.D., and Troyanskaya, O.G. (2009). Coordinated Concentration Changes of Transcripts and Metabolites in Saccharomyces cerevisiae. PLoS Comput. Biol. *5*, e1000270.

Burnett, R.M., Craven, K.E., Krishnamurthy, P., Goswami, C.P., Badve, S., Crooks, P., Mathews, W.P., Bhat-Nakshatri, P., and Nakshatri, H. (2015). Organ-specific adaptive signaling pathway activation in metastatic breast cancer cells. Oncotarget *6*, 12682–12696.

Bussey, K.J. (2006). Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol. Cancer Ther. *5*, 853–867.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. *14*, 365–376.

Cannell, I.G., Kong, Y.W., and Bushell, M. (2008). How do microRNAs regulate gene expression? Biochem. Soc. Trans. *36*, 1224–1231.

Chatterjee, A., Stockwell, P.A., Rodger, E.J., and Morison, I.M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Res. *40*, e79–e79.

Choi, H., and Pavelka, N. (2012). When One and One Gives More than Two: Challenges and Opportunities of Integrative Omics. Front. Genet. *2*.

Choi, Y., and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. Bioinformatics *25*, 2780–2786.

Cohen, J. (1992). Statiscal Power Analysis. Curr. Dir. Psychol. Sci. *1*, 98–101.

Cornbleet, P., and Gochman, N. (1979). Incorrect least-squares regression coefficients in method-comparison analysis. Clin. Chem. *25*, 432–438.

Dawson, J.A., and Kendziorski, C. (2012). An Empirical Bayesian Approach for Identifying Differential Coexpression in High-Throughput Experiments. Biometrics *68*, 455–465.

Dawson, J.A., Ye, S., and Kendziorski, C. (2012a). R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression. Bioinformatics *28*, 1939–1940.

Dawson, J.A., Ye, S., and Kendziorski, C. (2012b). R/EBcoexpress: an empirical Bayesian framework for discovering differential co-expression. Bioinformatics *28*, 1939–1940.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977a). Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. *39*, 1–38.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). MESSENGER-RNA-BINDING PROTEINS AND THE MESSAGES THEY CARRY. Nat. Rev. Mol. Cell Biol. *3*, 195–205.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. (2003). Multiple Hypothesis Testing in Microarray Experiments. Stat. Sci. *18*, 71–103.

Dvorkin, D., Biehs, B., and Kechris, K. (2013). A graphical model method for integrating multiple sources of genome-scale data. Stat. Appl. Genet. Mol. Biol. *12*.

Eng, and Ruggeri, C. (2015). Inferring Active and Prognostic Ligand-Receptor Pairs with Interactions in Survival Regression Models. Cancer Inform. 67.

Fang, G., Kuang, R., Pandey, G., Steinbach, M., Myers, C.L., and Kumar, V. (2009). Subspace Differential Coexpression Analysis: Problem Definition and a General Approach. In Biocomputing 2010, (WORLD SCIENTIFIC), pp. 145–156.

Findley, D.F. (1991). Counterexamples to parsimony and BIC. Ann. Inst. Stat. Math. *43*, 505–514.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika *10*, 507–521.

Flint, J., and Eskin, E. (2012). Genome-wide association studies in mice. Nat. Rev. Genet. *13*, 807–817.

Fraley, C., and Raftery, A.E. (1999). MCLUST: Software for Model-Based Cluster Analysis. J. Classif. *16*, 297–306.

Friedman, J., and Alm, E.J. (2012). Inferring Correlation Networks from Genomic Survey Data. PLoS Comput. Biol. *8*, e1002687.

de la Fuente, A. (2010). From “differential expression” to “differential networking” – identification of dysfunctional regulatory networks in diseases. Trends Genet. *26*, 326–333.

Fukushima, A. (2013). DiffCorr: An R package to analyze and visualize differential correlations in biological networks. Gene *518*, 209–214.

Greely, H.T. (2001). Human Genomics Research: New Challenges for Research Ethics. Perspect. Biol. Med. *44*, 221–229.

Griffin, L.M., Cicchini, L., Xu, T., and Pyeon, D. (2013). Human Keratinocyte Cultures in the Investigation of Early Steps of Human Papillomavirus Infection. In Epidermal Cells, K. Turksen, ed. (New York, NY: Springer New York), pp. 219–238.

Grubbs, F.E. (1969a). Procedures for Detecting Outlying Observations in Samples. Technometrics *11*, 1–21.

Grubbs, F.E. (1969b). Procedures for Detecting Outlying Observations in Samples. Technometrics *11*, 1–21.

Hamza, M.A., and Gilbert, M. (2014). Targeted Therapy in Gliomas. Curr. Oncol. Rep. *16*.

Ho, J.W.K., Stefani, M., dos Remedios, C.G., and Charleston, M.A. (2008). Differential variability analysis of gene expression and its application to human diseases. Bioinformatics *24*, i390–i398.

Horak, E.R., Klenk, N., Leek, R., LeJeune, S., Smith, K., Stuart, N., Harris, A.L., Greenall, M., and Stepniewska, K. (1992). Angiogenesis, assessed by platelet/endothelial cell adhesion molecule antibodies, as indicator of node metastases and survival in breast cancer. The Lancet *340*, 1120–1124.

Hotelling, H. (1953). New Light on the Correlation Coefficient and its Transforms. J. R. Stat. Soc. *15*, 193–232.

Hu, R., Qiu, X., Glazko, G., Klebanov, L., and Yakovlev, A. (2009). Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. BMC Bioinformatics *10*, 20.

Hu, R., Qiu, X., and Glazko, G. (2010). A new gene selection procedure based on the covariance distance. Bioinformatics *26*, 348–354.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Hughes, G., Cruickshank-Quinn, C., Reisdorph, R., Lutz, S., Petrache, I., Reisdorph, N., Bowler, R., and Kechris, K. (2014). MSPrep--Summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. Bioinformatics *30*, 133–134.

Inoue, H., and Randazzo, P.A. (2007). Arf GAPs and Their Interacting Proteins. Traffic *8*, 1465–1475.

Jauhiainen, A., Nerman, O., Michailidis, G., and Jornsten, R. (2012). Transcriptional and metabolic data integration and modeling for identification of active pathways. Biostatistics *13*, 748–761.

Juretic, E., Gagro, A., Vukelic, V., and Petrovecki, M. (2004). Maternal and Neonatal Lymphocyte Subpopulations at Delivery and 3 Days Postpartum: Increased Coexpression of CD45 Isoforms. Am. J. Reprod. Immunol. *52*, 1–7.

Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008). Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. J. Proteome Res. *7*, 40–44.

Kayano, M., Takigawa, I., Shiga, M., Tsuda, K., and Mamitsuka, H. (2011). ROS-DET: robust detector of switching mechanisms in gene expression. Nucleic Acids Res. *39*, e74–e74.

Kayano, M., Shiga, M., and Mamitsuka, H. (2014). Detecting Differentially Coexpressed Genes from Labeled Expression Data: A Brief Review. IEEE/ACM Trans. Comput. Biol. Bioinform. *11*, 154–167.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. *111*, 6131–6138.

Kendziorski, C., Newton, M., and Sarkar, D. (2005). EBarrays: Empirical Bayes for microarrays. R Package Version.

Kitano, H. (2002). Systems Biology: A Brief Overview. Science *295*, 1662–1664.

Komsta, L. (2006). Processing data for outliers. R Nes *6/2*.

Kostka, D., and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. Bioinformatics *20*, i194–i199.

Kuska, B. (1998). Beer, Bethesda, and Biology: How “Genomics” Came Into Being. JNCI J. Natl. Cancer Inst. *90*, 93–93.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics *20*, 3146–3155.

Lai, Y., Adam, B. -l., Podolsky, R., and She, J.-X. (2007). A mixture model approach to the tests of concordance and discordance between two large-scale experiments with two-sample groups. Bioinformatics *23*, 1243–1250.

Lai, Y., Zhang, F., Nayak, T.K., Modarres, R., Lee, N.H., and McCaffrey, T.A. (2014). Concordant integrative gene set enrichment analysis of multiple large-scale two-sample expression data sets. BMC Genomics *15*, S6.

Langfelder, P., and Horvath, S. (2012). Fast *R* Functions for Robust Correlations and Hierarchical Clustering. J. Stat. Softw. *46*.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lay, J.O., Liyanage, R., Borgmann, S., and Wilkins, C.L. (2006). Problems with the "omics." TrAC Trends Anal. Chem. *25*, 1046–1056.

Lederberg, J. (2001). 'Ome Sweet 'Omics -- A Genealogical Treasury of Words. The Scientist.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. *49*, 764–766.

Li, D.-M., and Feng, Y.-M. (2011). Signaling mechanism of cell adhesion molecules in breast cancer metastasis: potential therapeutic targets. Breast Cancer Res. Treat. *128*, 7–21.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265–272.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liu, N., and Pan, T. (2015). RNA epigenetics. Transl. Res. *165*, 28–35.

Lodish, H., Berk, A., Kaiser, C.A., Krieger, M., Bretscher, A., Hidde, P. and Matsudaira, M. (2008). Molecular cell biology (New York: W.H. Freeman).

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: But which regression?: Linear regression for comparing methods. Clin. Exp. Pharmacol. Physiol. *37*, 692–699.

Lui, T.W., Tsui, N.B., Chan, L.W., Wong, C.S., Siu, P.M., and Yung, B.Y. (2015). DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. BMC Bioinformatics *16*.

Magwene, P.M., Willis, J.H., and Kelly, J.K. (2011). The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. PLoS Comput. Biol. *7*, e1002255.

Malone, J.H., and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. *9*, 34.

Marques, S.A. (2012). Paracoccidioidomycosis. Clin. Dermatol. *30*, 610–615.

McDermott, J.E., Wang, J., Mitchell, H., Webb-Robertson, B.-J., Hafen, R., Ramey, J., and Rodland, K.D. (2013). Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin. Med. Diagn. *7*, 37–51.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., M. Mastrogianakis, G., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Myers, J.L., and Well, A. (2003). Research design and statistical analysis (Mahwah, N.J: Lawrence Erlbaum Associates).

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *27*, 29–34.

Oshlack, A., Robinson, M.D., and Young, M.D. (2010). From RNA-seq reads to differential expression results. Genome Biol. *11*, 220.

Ottenheijm, C.A.C., Heunks, L.M.A., Li, Y.-P., Jin, B., Minnaard, R., van Hees, H.W.H., and Dekhuijzen, P.N.R. (2006). Activation of the Ubiquitin–Proteasome Pathway in the Diaphragm in Chronic Obstructive Pulmonary Disease. Am. J. Respir. Crit. Care Med. *174*, 997–1002.

Pasquinelli, A.E. (2012). MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat. Rev. Genet.

Patti, G.J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. Nat. Rev. Mol. Cell Biol. *13*, 263–269.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. *58*, 240–242.

Pearson, K. (1916). Mathematical Contributions to the Theory of Evolution. XIX. Second Supplement to a Memoir on Skew Variation. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. *216*, 429–457.

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics *30*, 360–368.

Rau, A., Gallopin, M., Celeux, G., and Jaffrezic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. Bioinformatics *29*, 2146–2152.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47–e47.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rovina, N., Koutsoukou, A., and Koulouris, N.G. (2013). Inflammation and Immune Response in COPD: Where Do We Stand? Mediators Inflamm. *2013*, 1–9.

Ru, Y., Kechris, K.J., Tabakoff, B., Hoffman, P., Radcliffe, R.A., Bowler, R., Mahaffey, S., Rossi, S., Calin, G.A., Bemis, L., et al. (2014). The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic Acids Res. *42*, e133–e133.

Sanli, T., Linher-Melville, K., Tsakiridis, T., and Singh, G. (2012). Sestrin2 Modulates AMPK Subunit Expression and Its Response to Ionizing Radiation in Breast Cancer Cells. PLoS ONE *7*, e32035.

Schnabel, D., Schröder, M., Fürst, W., Klein, A., Hurwitz, R., Zenk, T., Weber, J., Harzer, K., Paton, B.C., and Poulos, A. (1992). Simultaneous deficiency of sphingolipid activator proteins 1 and 2 is caused by a mutation in the initiation codon of their common gene. J. Biol. Chem. *267*, 3312–3315.

Shedden, K., and Taylor, J. (2005). Differential Correlation Detects Complex Associations Between Gene Expression and Clinical Outcomes in Lung Adenocarcinomas. In Methods of Microarray Data Analysis IV, (Springer), pp. 121–131.

Shrikhande, T. Hunziker, L. R. Braa, M. (2000). Increased Coexpression of Eotaxin and Interleukin 5 in Bullous Pemphigoid. Acta Derm. Venereol. *80*, 277–280.

Silva, C.L., Silva, M.F., Faccioli, L.H., Pietro, R.C.L., Cortez, S.A.E., and Foss, N.T. (1995). Differential correlation between interleukin patterns in disseminated and chronic human paracoccidioidomycosis. Clin. Exp. Immunol. *101*, 314–320.

de Siqueira Santos, S., Takahashi, D.Y., Nakata, A., and Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief. Bioinform. *15*, 906–918.

Siska, C., Bowler, R., and Kechris, K. (2015). The discordant method: a novel approach for differential correlation. Bioinformatics btv633.

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics *13*, 328.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. Am. J. Psychol. *15*, 72.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.

Tesson, B.M., Breitling, R., and Jansen, R.C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics *11*, 497.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Tsai, T., and Gill, J. (2013). Interactions in Generalized Linear Models: Theoretical Issues and an Application to Personal Vote-Earning Attributes. Soc. Sci. *2*, 91–113.

Ubhi, B.K., Cheng, K.K., Dong, J., Janowitz, T., Jodrell, D., Tal-Singer, R., MacNee, W., Lomas, D.A., Riley, J.H., Griffin, J.L., et al. (2012). Targeted metabolomics identifies perturbations in amino acid metabolism that sub-classify patients with COPD. Mol. Biosyst. *8*, 3125.

Venables, W.N., and Ripley, B.D. Modern Applied Statistics with S. Springer.

Voet, D., and Voet, J.G. (2009). Biochemistry (Wiley).

Wahlsten, D. (1999). SINGLE-GENE INFLUENCES ON BRAIN AND BEHAVIOR. Annu. Rev. Psychol. *50*, 599–624.

Walley, A.J., Jacobson, P., Falchi, M., Bottolo, L., Andersson, J.C., Petretto, E., Bonnefond, A., Vaillant, E., Lecoeur, C., Vatin, V., et al. (2012). Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. Int. J. Obes. *36*, 137–147.

Wang, X., Yan, Z., Fulciniti, M., Li, Y., Gkotzamanidou, M., Amin, S.B., Shah, P.K., Zhang, Y., Munshi, N.C., and Li, C. (2014). Transcription factor-pathway coexpression analysis reveals cooperation between SP1 and ESR1 on dysregulating cell cycle arrest in non-hyperdiploid multiple myeloma. Leukemia *28*, 894–903.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. BMC Bioinformatics *7*.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J.

Willis, A., Jung, E.J., Wakefield, T., and Chen, X. (2004). Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. Oncogene *23*, 2330–2338.

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics *29*, 638–644.

Xu, X., Su, S., Barnes, V.A., De Miguel, C., Pollock, J., Ownby, D., Shi, H., Zhu, H., Snieder, H., and Wang, X. (2013). A genome-wide methylation study on obesity: Differential variability and differential methylation. Epigenetics *8*, 522–533.

Zhu, J.-K. (2008). Epigenome Sequencing Comes of Age. Cell *133*, 395–397.

APPENDIX A

IDENTIFIERS AND LISTS OF VALIDATED FEATURES

A.1. TCGA GBM Sample IDs

TCGA ID	Type
TCGA-06-0192-01B-01R-0338-01	Tumor
TCGA-06-0216-01B-01R-0338-01	Tumor
TCGA-06-0649-01B-01R-0338-01	Tumor
TCGA-06-0673-11A-01R-0342-01	Control
TCGA-06-0675-11A-01R-0342-01	Control
TCGA-06-0676-11A-02R-0342-01	Control
TCGA-06-0678-11A-01R-0342-01	Control
TCGA-06-0680-11A-01R-0342-01	Control
TCGA-06-0681-11A-01R-0342-01	Control
TCGA-06-0686-01A-01R-0338-01	Tumor
TCGA-06-0743-01A-01R-0338-01	Tumor
TCGA-06-0744-01A-01R-0338-01	Tumor
TCGA-06-0745-01A-01R-0338-01	Tumor
TCGA-06-0747-01A-01R-0338-01	Tumor
TCGA-06-0749-01A-01R-0338-01	Tumor
TCGA-06-0750-01A-01R-0338-01	Tumor
TCGA-08-0623-11A-01R-0342-01	Control
TCGA-08-0625-11A-01R-0342-01	Control
TCGA-08-0626-11A-01R-0342-01	Control
TCGA-08-0627-11A-01R-0342-01	Control
TCGA-12-0654-01B-01R-0338-01	Tumor
TCGA-12-0656-01B-01R-0338-01	Tumor
TCGA-12-0657-01A-01R-0338-01	Tumor
TCGA-12-0688-01A-02R-0338-01	Tumor
TCGA-12-0692-01A-01R-0338-01	Tumor
TCGA-12-0703-01A-02R-0338-01	Tumor
TCGA-12-0707-01A-01R-0338-01	Tumor
TCGA-12-0772-01A-01R-0338-01	Tumor
TCGA-12-0773-01A-01R-0338-01	Tumor
TCGA-12-0775-01A-01R-0338-01	Tumor
TCGA-12-0776-01A-01R-0338-01	Tumor
TCGA-12-0778-01A-01R-0338-01	Tumor

TCGA-12-0780-01A-01R-0338-01	Tumor

A.2. GBM miRNAs

hsa-mir-124a
hsa-mir-137
hsa-mir-326
hsa-mir-92b

A.3. Sphingolipid-Related Features

Genes

Probe Name	Gene Name
1552632_a_at	ARSG
1553929_at	ACER1
1554030_at	ARSB
1554032_at	ARSB
1554252_a_at	CERS3
1554253_a_at	CERS3
1554460_at	ST8SIA4
1555041_a_at	NAGA
1558279_a_at	KDSR
1559776_at	NULL
1560086_at	NULL
1564274_at	C9orf47
1564333_a_at	PSAPL1
1567080_s_at	CLN6
200661_at	CTSA
200695_at	PPP2R1A
200866_s_at	PSAP
200871_s_at	PSAP
201289_at	CYR61
201576_s_at	NULL
201765_s_at	HEXA
201944_at	HEXB
202278_s_at	SPTLC1
202545_at	PRKCD
202549_at	VAPB

202550_s_at	VAPB
202944_at	NAGA
203089_s_at	HTRA2
203128_at	SPTLC2
203269_at	NSMAF
203608_at	ALDH5A1
203609_s_at	ALDH5A1
203768_s_at	STS
203769_s_at	STS
203770_s_at	STS
204417_at	GALC
204443_at	ARSA
204458_at	PLA2G15
204642_at	S1PR1
204691_x_at	PLA2G6
204881_s_at	UGCG
205051_s_at	KIT
205309_at	SMPDL3B
205622_at	SMPD2
205670_at	GAL3ST1
205894_at	ARSE
206129_s_at	ARSB
206258_at	ST8SIA5
206397_x_at	NULL
206435_at	B4GALNT1
206437_at	S1PR4
206831_s_at	ARSD
206925_at	ST8SIA4
206948_at	NEU3
207381_at	ALOX12B
207708_at	ALOXE3
207856_s_at	NULL
208065_at	ST8SIA3
208358_s_at	UGT8
208381_s_at	SGPL1
208478_s_at	BAX
208537_at	S1PR2
208780_x_at	VAPA
208926_at	NEU1

209093_s_at	NULL
209250_at	DEGS1
209275_s_at	CLN3
209355_s_at	PPAP2B
209529_at	PPAP2C
209727_at	GM2A
209799_at	PRKAA1
209810_at	SFTPB
209857_s_at	SPHK2
210073_at	ST8SIA1
210171_s_at	CREM
210401_at	P2RX1
210589_s_at	GBAP1
210647_x_at	PLA2G6
210764_s_at	CYR61
210859_x_at	CLN3
210946_at	PPAP2A
211152_s_at	HTRA2
211488_s_at	ITGB8
212226_s_at	PPAP2B
212321_at	SGPL1
212322_at	SGPL1
212442_s_at	CERS6
212737_at	GM2A
213508_at	SPTSSA
213936_x_at	SFTPB
214354_x_at	SFTPB
214490_at	ARSF
214655_at	GPR6
215471_s_at	MAP7
215543_s_at	LARGE
215891_s_at	GM2A
215938_s_at	PLA2G6
216230_x_at	SMPD1
218028_at	NULL
218099_at	TEX2
218161_s_at	CLN6
218421_at	CERK
218556_at	ORMDL2

219340_s_at	CLN8
219429_at	FA2H
219625_s_at	COL4A3BP
219695_at	SMPD3
219973_at	ARSJ
221285_at	ST8SIA2
221368_at	NEU2
221417_x_at	S1PR5
221765_at	UGCG
222212_s_at	CERS2
222383_s_at	ALOXE3
222571_at	ST6GALNAC6
222688_at	ACER3
222689_at	ACER3
222874_s_at	CLN8
222957_at	NEU4
223259_at	ORMDL3
223466_x_at	COL4A3BP
223695_s_at	ARSD
223696_at	ARSD
223912_s_at	CLN8
223921_s_at	GBA2
224627_at	GBA2
224951_at	CERS5
225095_at	SPTLC2
225280_x_at	ARSD
225286_at	ARSD
225923_at	VAPB
225950_at	SAMD8
225984_at	PRKAA1
225985_at	PRKAA1
226189_at	ITGB8
226277_at	COL4A3BP
226560_at	NULL
227038_at	SGMS2
227548_at	ORMDL1
227752_at	SPTLC3
227776_at	ACER3
228457_at	PPM1L

228480_at	VAPA
228801_at	ORMDL1
228956_at	UGT8
229448_at	NULL
229850_at	KDSR
229958_at	CLN8
230131_x_at	ARSD
230261_at	ST8SIA4
230262_at	ST8SIA3
230275_at	ARSI
230464_at	S1PR5
230482_at	ST6GALNAC5
230836_at	ST8SIA4
231286_at	NULL
231732_at	SMPD3
231741_at	S1PR3
231791_at	ASAH2B
232149_s_at	NSMAF
232197_x_at	ARSB
232423_at	ARSD
233743_x_at	S1PR5
234963_s_at	FA2H
235136_at	ORMDL3
235502_at	PPP2CA
235678_at	GM2A
236339_at	PPM1L
236496_at	DEGS2
238567_at	SGPP2
238702_at	SPTSSB
238719_at	PPP2CA
238945_at	ACER3
239147_at	ARSK
239401_at	NULL
239488_at	PPM1L
239750_x_at	VAPA
240180_at	NULL
242019_at	CERS6
242062_at	SAMD8
242943_at	ST8SIA4

242963_at	SGMS2
243141_at	SGMS2
244780_at	SGPP2
35820_at	GM2A
37004_at	SFTPB
40273_at	SPHK2

Metabolites

M/Z ratio with retention time	Annotation
311.2821_3.1555946	(4E,8E,9Me-d19:2)sphingosine
1619.9467_1.3947561	*1-3Galalpha1-3Galalpha1-3Galalpha1-4Galbeta1-4Glcbeta-Cer(d18:1/24:1(15Z))
607.5888_7.030736	Cer(d16:1/23:0)
619.5897_6.868578	Cer(d18:1/22:1(13Z))
635.6203_7.4143705	Cer(d18:1/23:0)
647.6227_7.266074	Cer(d18:1/24:1(15Z))
633.6054_7.086333	Cer(d18:2/23:0)
1493.1392_5.6300683	CerP(d18:1/24:1(15Z))
1215.8105_1.0137502	Fucalpha1-2Galalpha1-3Galbeta1-4Glcbeta-Cer(d18:1/18:0)
1521.1075_5.171591	GlcAbeta-Cer(d18:1/18:0)
633.5257_2.2732987	N-(tetradecanoyl)-sphing-4-enine-1-(2-aminoethylphosphonate)
1046.7281_1.9593751	NeuAcalpha2-3Galbeta-Cer(d18:1/20:0)
1643.9421_1.384165	NeuGcalpha2-3Galbeta1-4GlcNAcbeta1-3Galbeta1-4Glcbeta-Cer(d18:1/24:1(15Z))
671.6187_7.5962663	N-Lignoceroylsphingosine
537.5129_5.722461	N-Palmitoylsphingosine
688.5519_4.3623652	SM(d16:1/17:0)
702.5691_4.7004237	SM(d18:0/16:1(9Z))
730.598_5.387554	SM(d18:0/18:1(11Z))
852.6434_6.297378	SM(d18:0/24:1(15Z))
646.5014_3.3163707	SM(d18:1/12:0)
674.5371_4.01326	SM(d18:1/14:0)
703.5728_4.700575	SM(d18:1/16:0)
700.5533_4.1588397	SM(d18:1/16:1)
756.6135_5.5130873	SM(d18:1/20:1)
800.672_6.8920975	SM(d18:1/23:0)
812.6755_6.5607677	SM(d18:1/24:1(15Z))
672.518_3.4627964	SM(d18:2/14:0)

686.5328_3.8306012	SM(d18:2/15:0)
798.6601_6.4355063	SM(d18:2/23:0)
299.2832_2.9104302	Sphingosine
299.2832_2.9104302	Sphingosine
299.2832_2.9104302	Sphingosine

A.4. TCGA Breast Cancer Sample IDs

TCGA ID	Type
TCGA.AR.A1AH.01A	Tumor
TCGA.BH.A0BO.01A	Tumor
TCGA.BH.A0C1.01B	Tumor
TCGA.BH.A0DO.01B	Tumor
TCGA.BH.A0DO.11A	Tumor
TCGA.BH.A0DT.01A	Tumor
TCGA.BH.A0DT.11A	Control
TCGA.BH.A18F.01A	Control
TCGA.BH.A18G.01A	Tumor
TCGA.BH.A18H.01A	Tumor
TCGA.BH.A18I.01A	Tumor
TCGA.BH.A18J.01A	Tumor
TCGA.BH.A18J.11A	Tumor
TCGA.BH.A18K.01A	Control
TCGA.BH.A18K.11A	Tumor
TCGA.BH.A18L.01A	Tumor
TCGA.BH.A18L.11A	Tumor
TCGA.BH.A18M.01A	Tumor
TCGA.BH.A18M.11A	Tumor
TCGA.BH.A18N.01A	Tumor
TCGA.BH.A18N.11A	Control
TCGA.BH.A18P.01A	Control
TCGA.BH.A18P.11A	Tumor
TCGA.BH.A18Q.01A	Tumor
TCGA.BH.A18Q.11A	Tumor
TCGA.BH.A18R.01A	Tumor
TCGA.BH.A18R.11A	Tumor
TCGA.BH.A18S.01A	Tumor
TCGA.BH.A18S.11A	Control

TCGA.BH.A18T.01A	Control
TCGA.BH.A18U.01A	Control
TCGA.BH.A18U.11A	Tumor
TCGA.BH.A18V.01A	Tumor
TCGA.C8.A12Y.01A	Tumor
TCGA.C8.A133.01A	Tumor
TCGA.E2.A14P.01A	Tumor
TCGA.E2.A14Q.01A	Control
TCGA.E2.A14S.01A	Control
TCGA.E2.A14V.01A	Tumor
TCGA.E2.A14W.01A	Tumor
TCGA.E2.A14Y.01A	Tumor
TCGA.E2.A150.01A	Tumor
TCGA.E2.A152.01A	Tumor
TCGA.E2.A153.01A	Tumor
TCGA.E2.A153.11A	Tumor
TCGA.E2.A155.01A	Tumor
TCGA.E2.A158.01A	Tumor
TCGA.E2.A158.11A	Control
TCGA.E2.A15A.01A	Tumor
TCGA.E2.A15A.06A	Tumor
TCGA.E2.A15C.01A	Tumor
TCGA.E2.A15E.06A	Tumor
TCGA.E2.A15G.01A	Tumor
TCGA.E2.A15H.01A	Control
TCGA.E2.A15L.01A	Control
TCGA.E2.A15M.01A	Control

A.5. Breast Cancer miRNAs

hsa-mir-107
hsa-mir-150
hsa-mir-152
hsa-mir-191
hsa-mir-24-2
hsa-mir-374a
hsa-mir-574
hsa-mir-454

APPENDIX B

Tables of Model Assumptions

B.1. BIC of GBM and COPD data sets.

BIC of Normal vs. Pearson VII distributions with 1 to 5 components in GBM and COPD data. Pearson VII 1-component mixture model is NA because unable to get value.

	GBM		COPD	
	Normal	**Pearson VII**	**Normal**	**Pearson VII**
1	-26284924	NA	45507901	NA
2	-26294229	-26599590	45487152	44454561
3	-26280180	-26479171	45486345	44785415
4	-26280384	-26454494	45483803	44855883
5	-26278844	-26448430	45485535	44876453

B.2. BIC of Breast Cancer Datasets with Various Correlation Metrics

BIC of Normal vs. Pearson VII distributions with 1 to 5 components in breast cancer data. Pearson VII 1-component mixture model is NA because unable to get value.

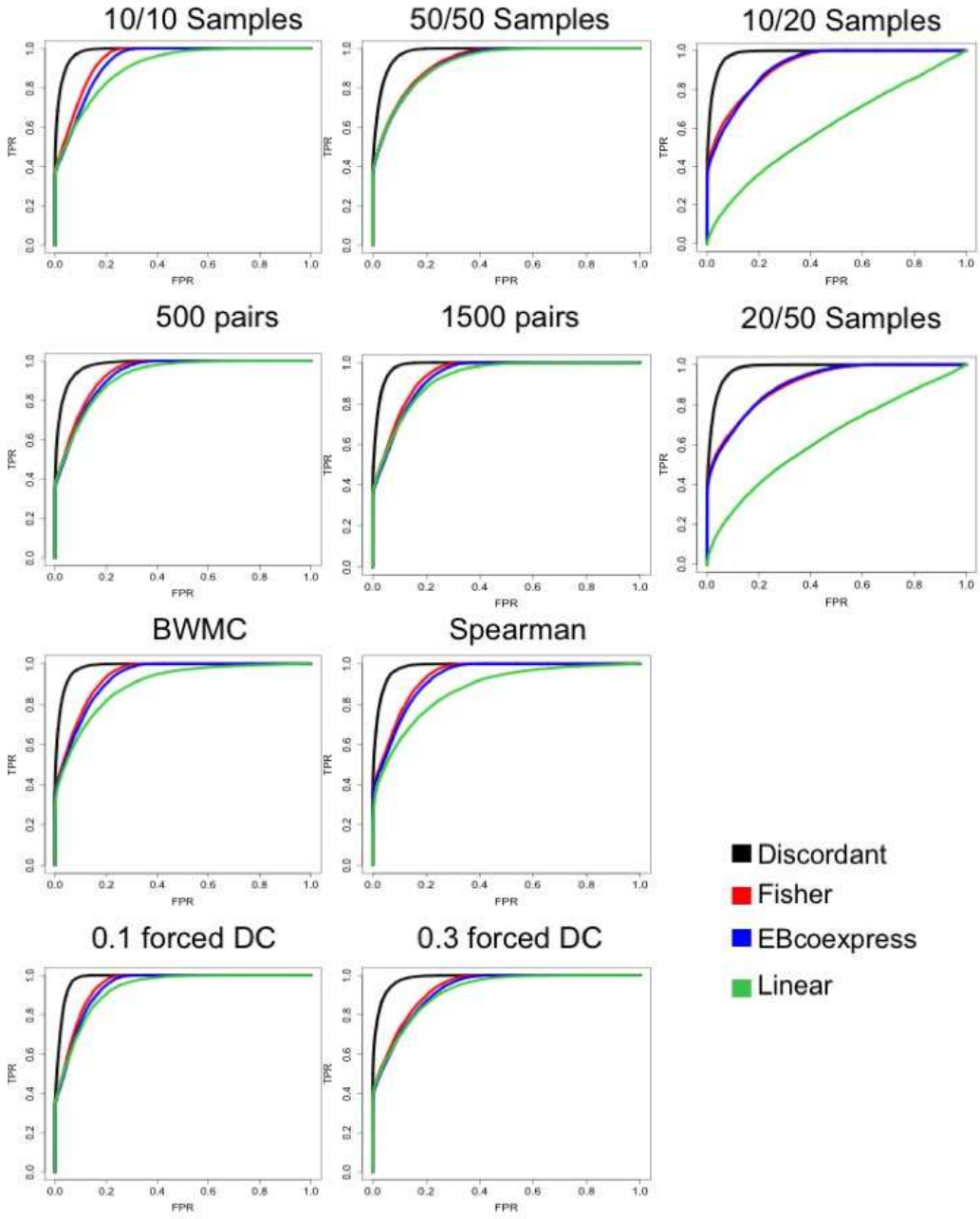
	Spearman		**Pearson**	
	Normal	**Pearson VII**	**Normal**	**Pearson VII**
1	-2508345	NA	-3534317	NA
2	-2490313	-2521102	-3522077	-3563371
3	-2491553	-2507935	-3522569	-3547161
4	-2491443	-2504859	-3522552	-3543296
5	-2491321	-2504034	-3522571	-3542188
	BWMC		**SparCC**	
	Normal	**Pearson VII**	**Normal**	**Pearson VII**
1	-3310034	NA	-3580273	NA
2	-3291900	-3332878	-3581509	-3610734
3	-3292002	-3315611	-3567261	-3591147
4	-3292158	-3310952	-3566941	-3587009
5	-3292426	-3309456	-3566512	-3586140

APPENDIX C

SIMULATIONS

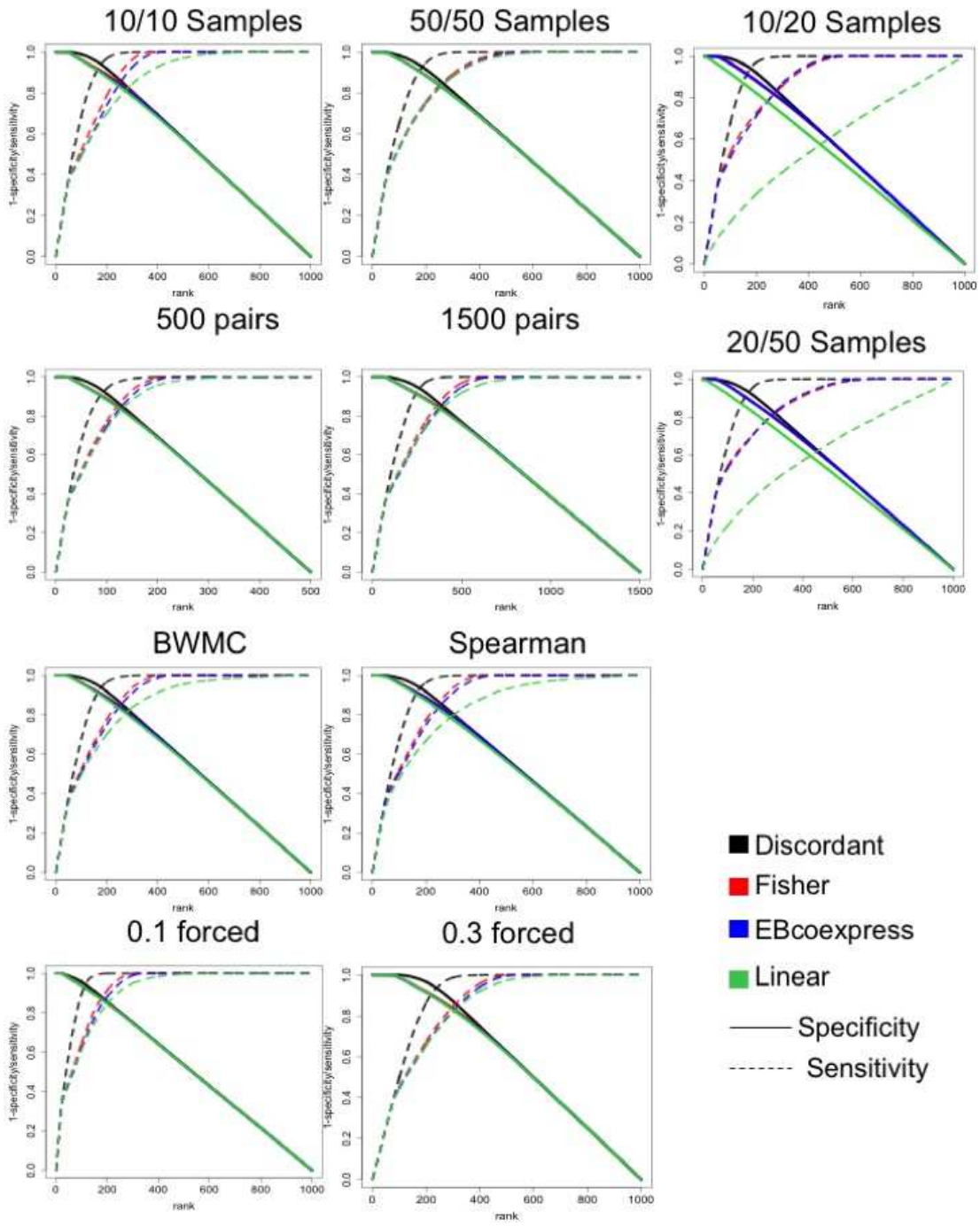
C.1. ROC Curves of Adjustments of Simulation Parameters

ROC curves with changes in simulation parameters. Title tells parameter changed from shaded red row in Table 1.

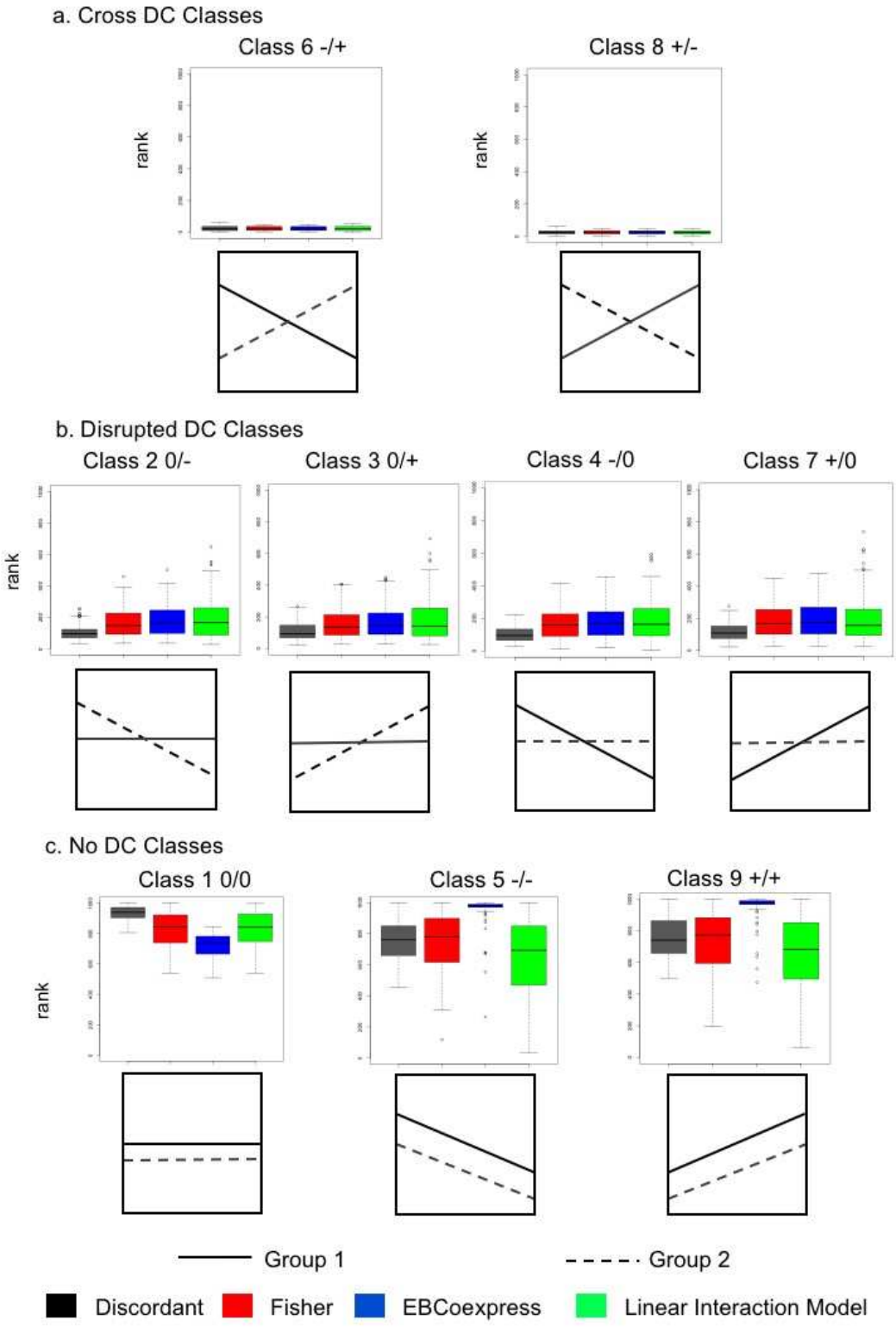


C.2. Sensitivity/Specificity of Adjustments of Simulation Parameters

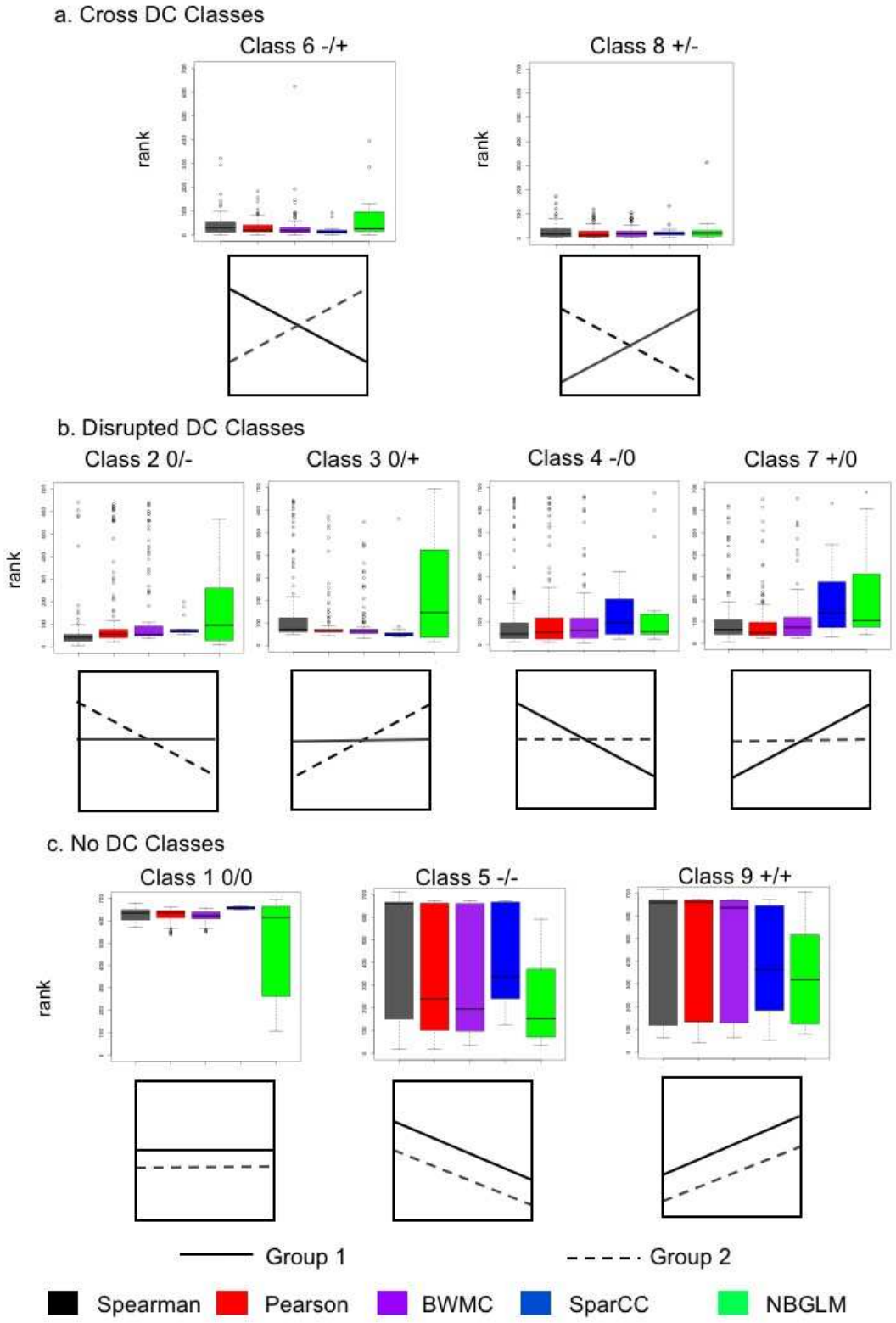
Sensitivity/1-Specificity curves with changes in simulation parameters. Title tells parameter changed from shaded red row in Table 1.



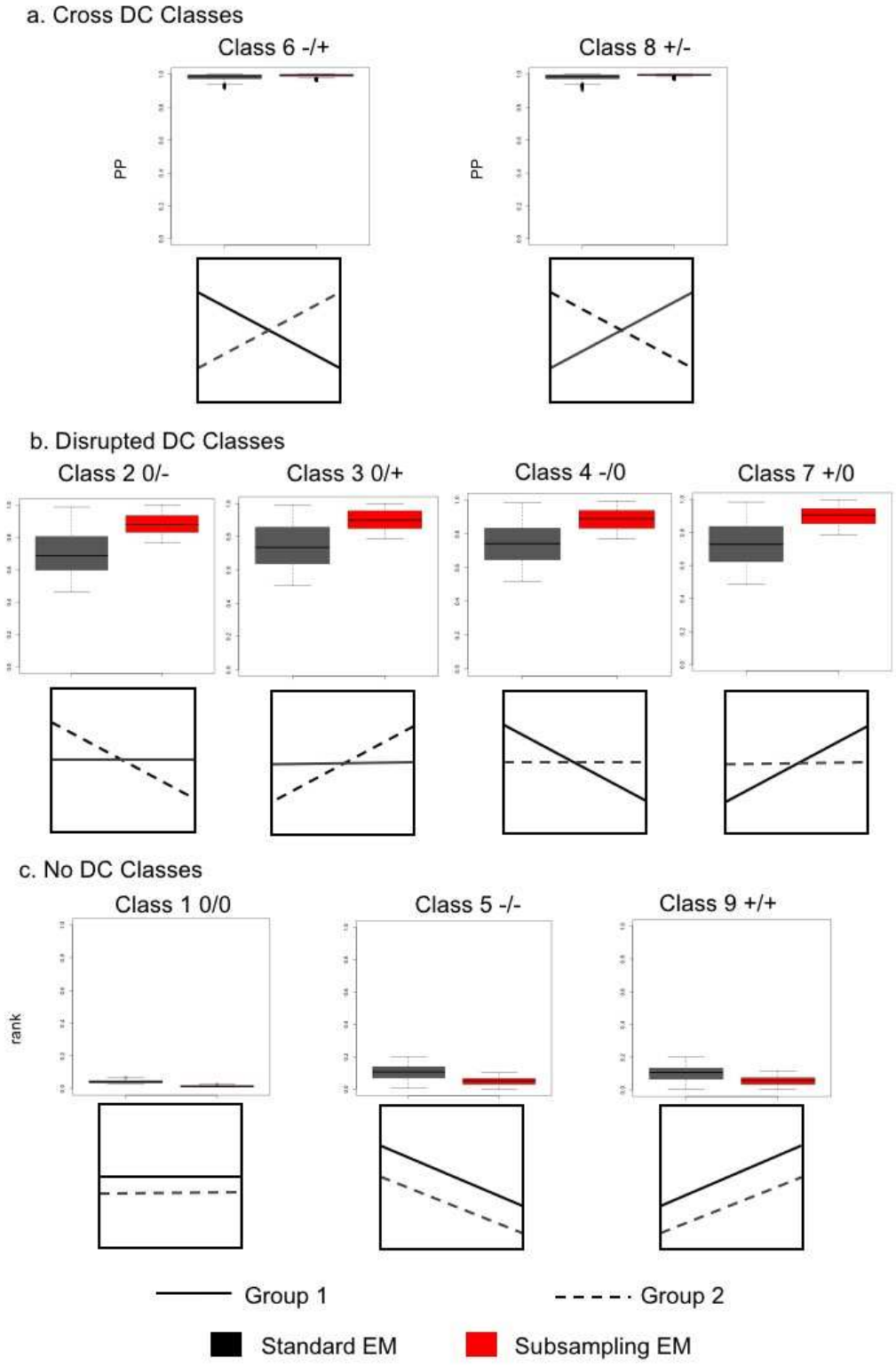
C.3. Boxplots of Rank Distributions of Each Class for Continuous Simulations to Compare Competing Methods



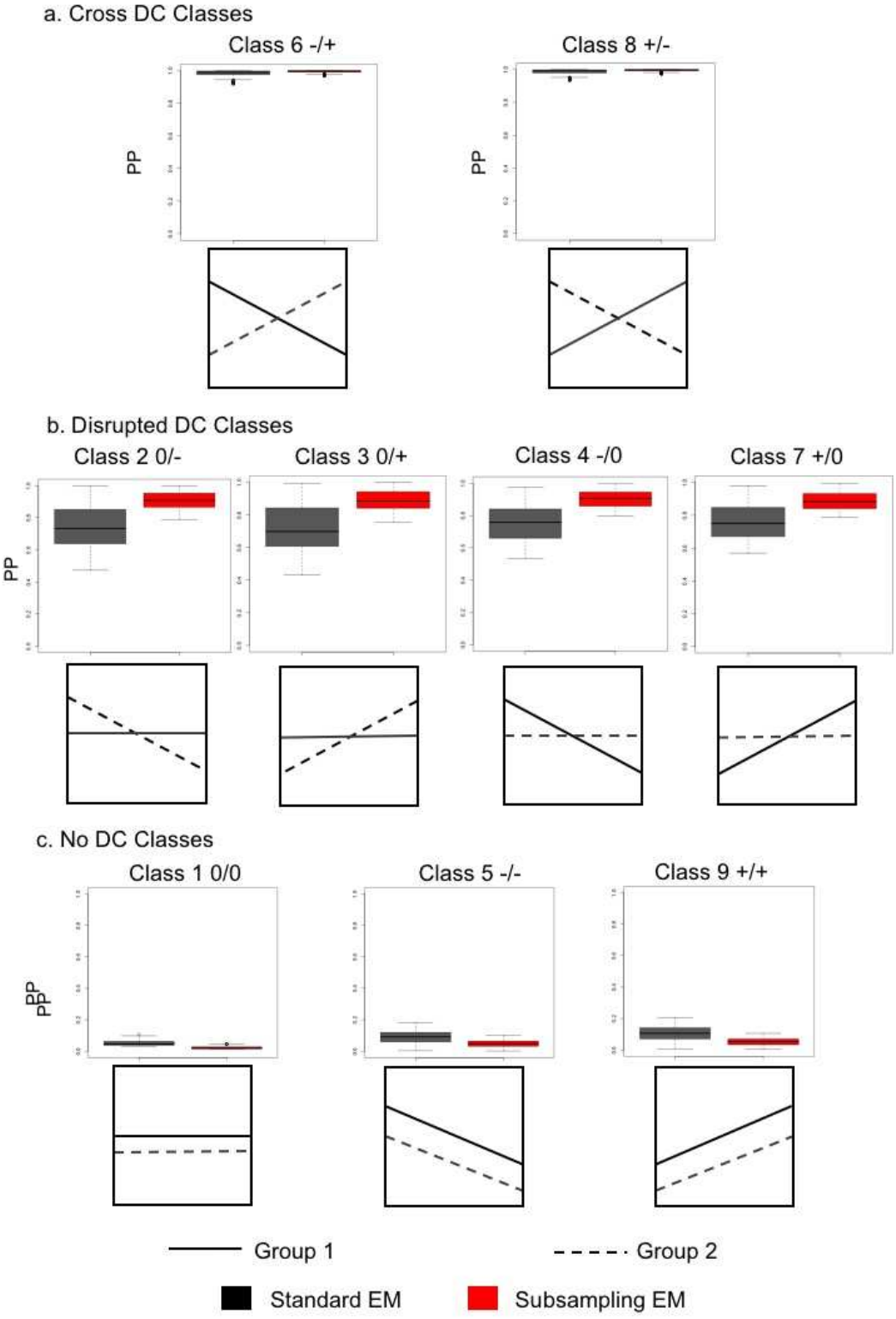
C.4. Boxplots of Rank Distributions of Each Class for Count Simulations Comparing Correlation Metrics



C.5. Boxplots of Posterior Probability Distributions of Each Class for Continuous Simulations Comparing Standard EM vs. Subsampling EM

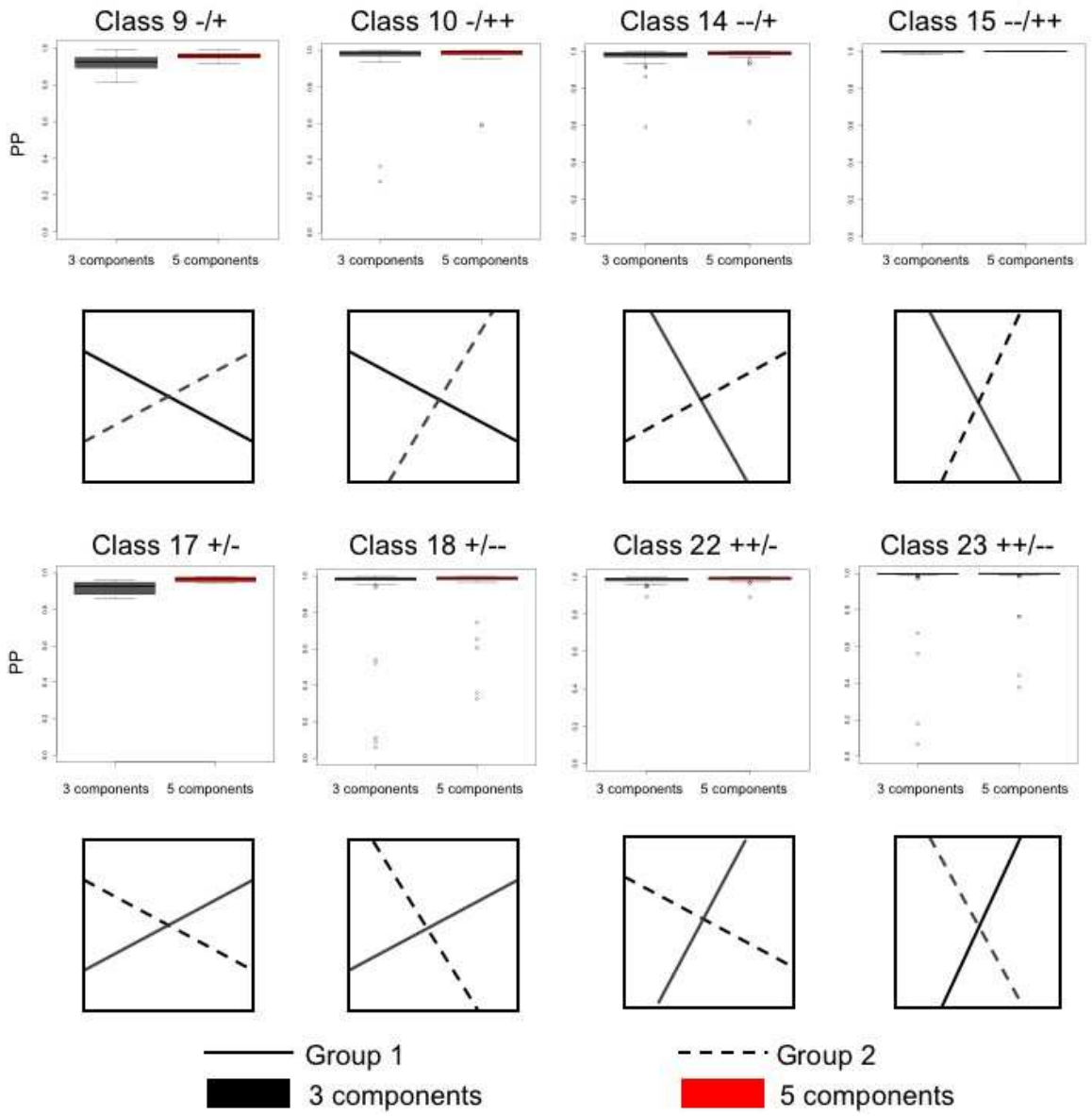


C.6. Boxplots of Posterior Probability Distributions of Each Class for Count Simulations Comparing Standard EM vs. Subsampling EM

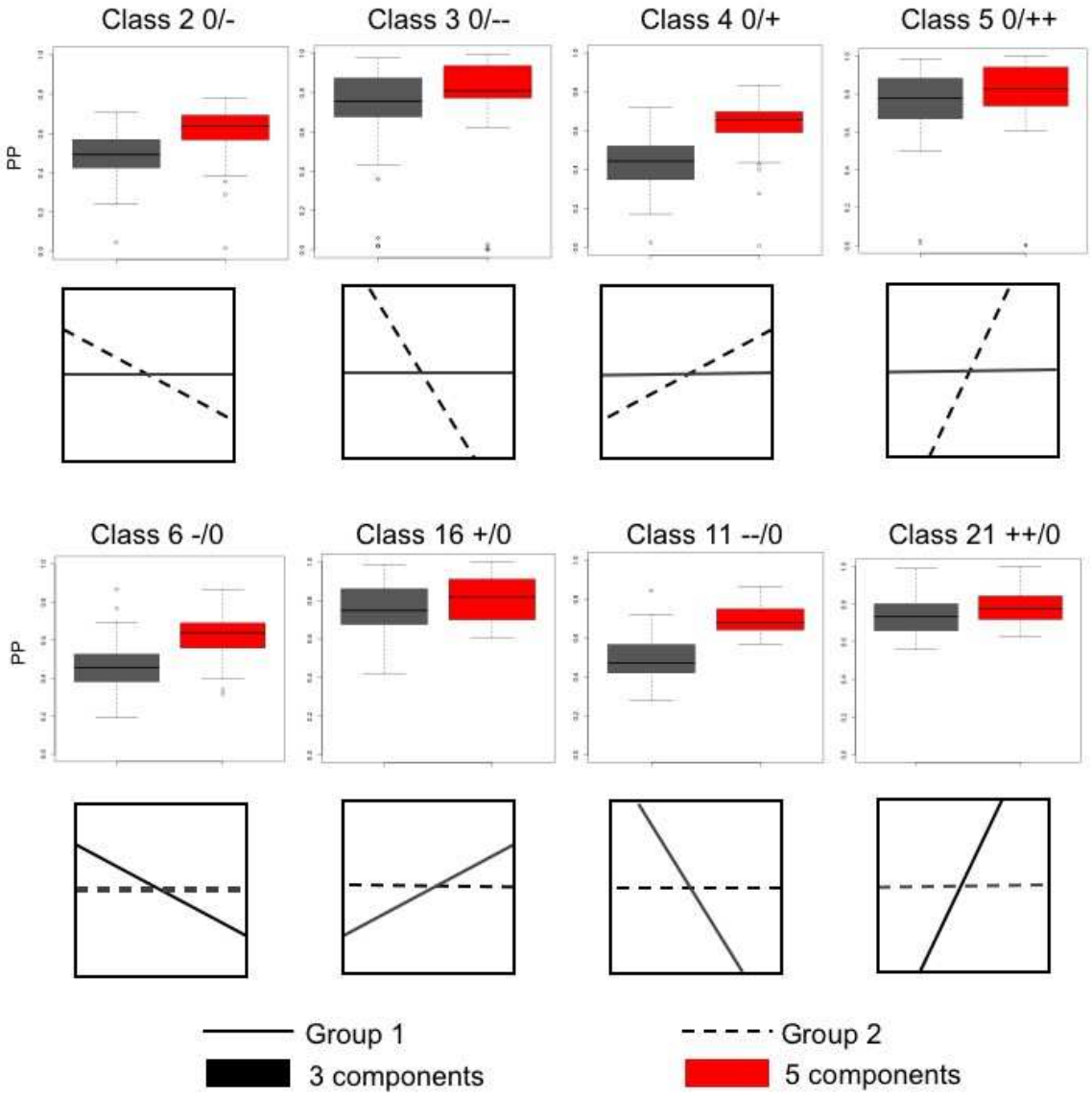


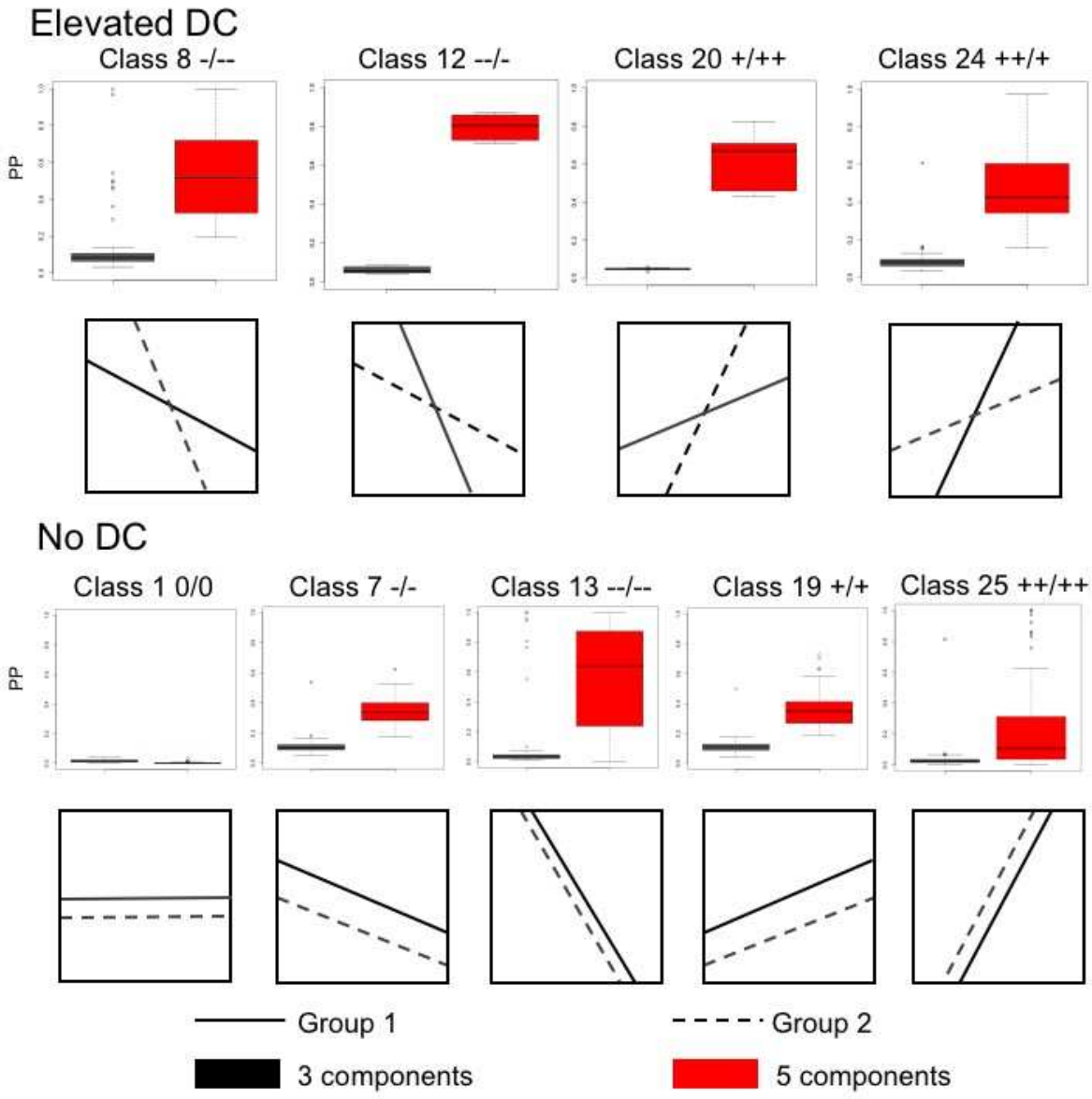
C.7. Boxplots of Posterior Probability Distributions of Each Class for Continuous Simulations Comparing Three Components vs. Five Components

Cross DC



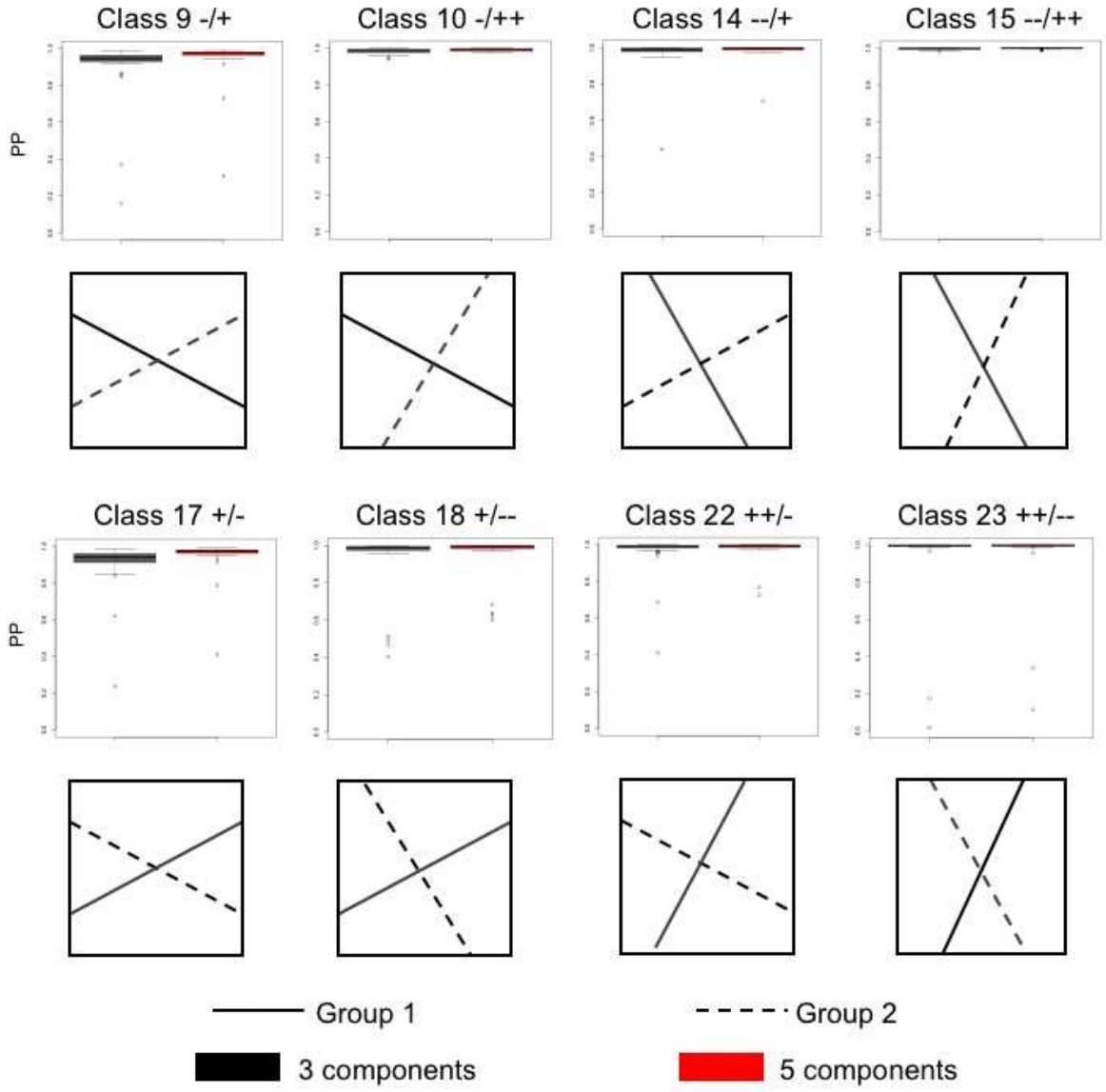
Disrupted DC



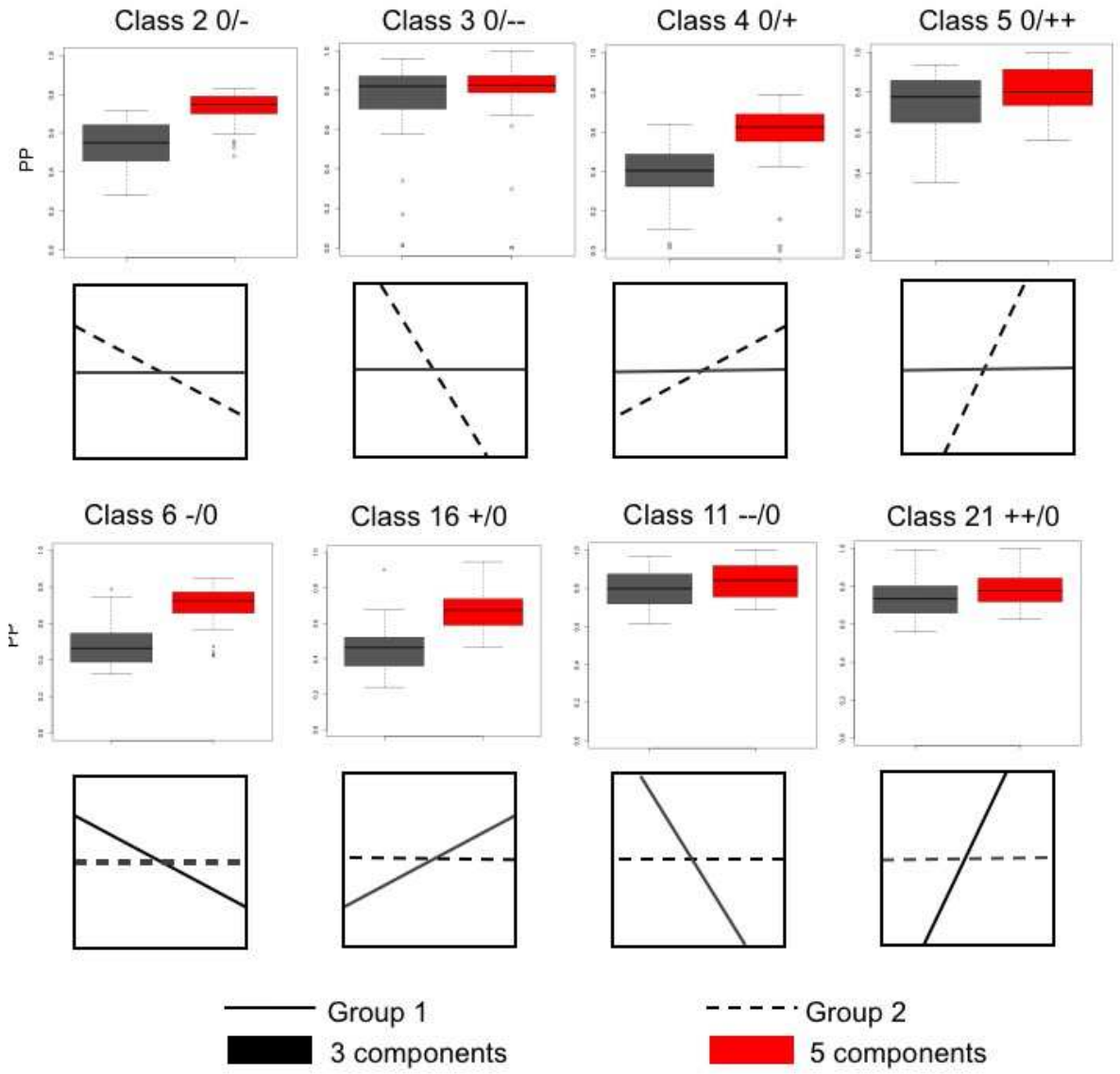

Elevated DC
Class 8 -/--
Class 12 --/-
Class 20 +/++
Class 24 ++/+
PP
No DC
Class 1 0/0
Class 7 -/-
Class 13 --/--
Class 19 +/+
Class 25 ++/++
PP
Group 1
Group 2
3 components
5 components

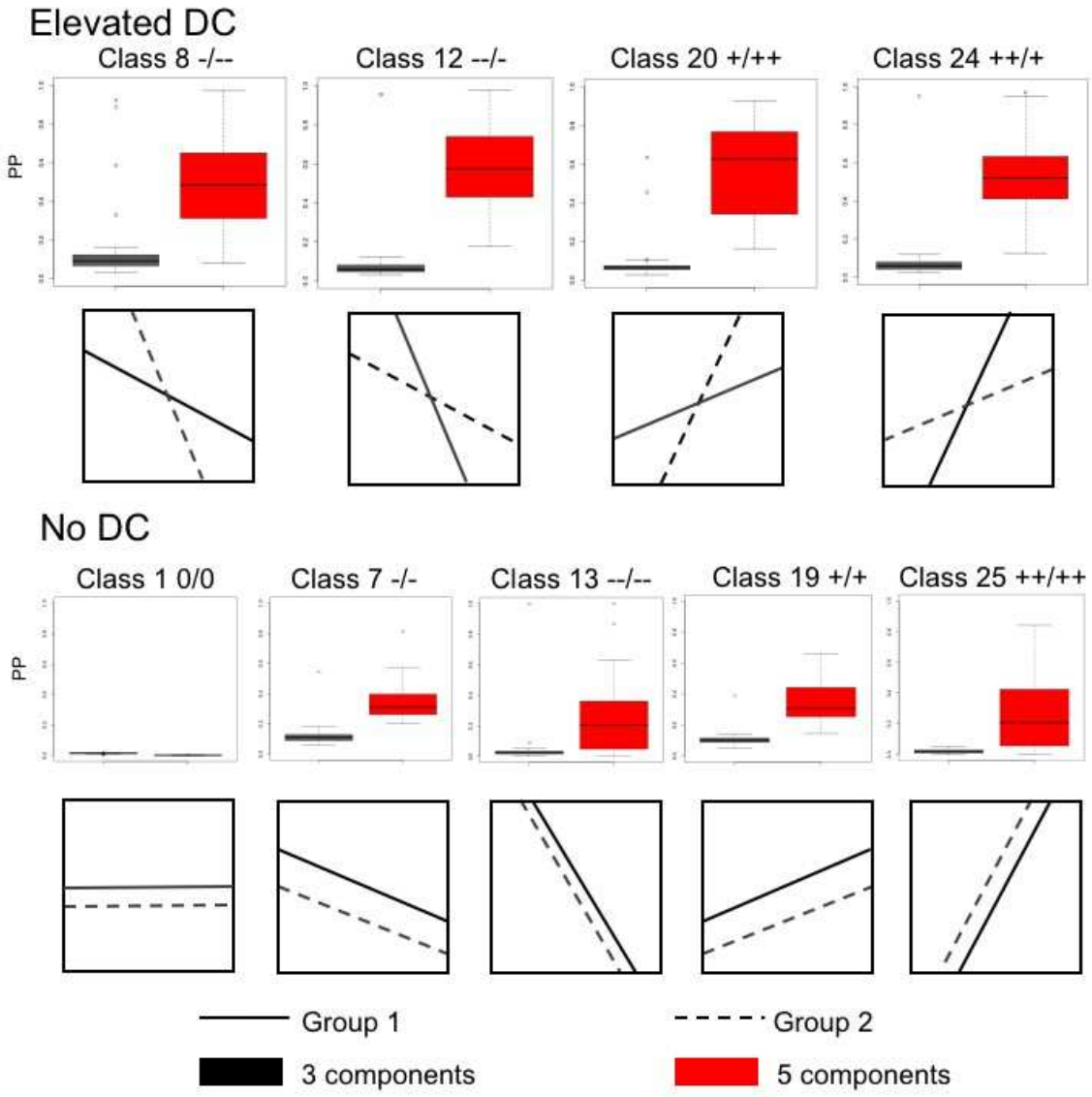
C.8. Boxplots of Posterior Probability Distributions of Each Class for Continuous Simulations Comparing Three Components vs. Five Components

Cross DC



Disrupted DC




Elevated DC
Class 8 -/--
Class 12 --/-
Class 20 +/++
Class 24 ++/+
PP
No DC
Class 1 0/0
Class 7 -/-
Class 13 --/--
Class 19 +/+
Class 25 ++/++
PP
Group 1
Group 2
3 components
5 components

APPENDIX D

Tables of Biological Validation

D.1. GBM miRNAs

Summary of unique GBM-related miRNAs top ranked pair with a transcript. Shown is rank, p-value/1 - posterior probability (1 - pp) and FDR/q-value for Discordant, EBcoexpress and Fisher.

Method	**statistic**	**hsa-miR-124a**	**hsa-miR-137**	**hsa-miR-326**	**hsa-miR-92b**
Discordant	Rank	223	1081	472	83
	1 - pp	2.51e-7	1.16e-6	5.04e-7	1.05e-7
	1 - q-value	4.98e-7	2.39e-6	1.02e-6	2.02e-7
EBcoexpress	Rank	585	1862	628	185
	1 - pp	3.54e-3	9.86e-3	3.84e-3	1.25e-3
	1 - q-value	6.66e-3	0.184	7.23e-3	2.44e-3
Fisher	Rank	799	1282	803	238
	p-value	3.61e-6	7.43e-6	3.63e-6	6.24e-7
	FDR	0.109	0.139	0.109	0.629
miRNA –Dep. Linear	Rank	3246	297	768	69
	p-value	5.65e-5	1.29e-6	6.34e-6	8.95e-8
	FDR	0.420	0.103	0.198	0.031
Transcript – Dep. Linear	Rank	467	8807	4	1108
	p-value	1.85e-5	4.72e-4	1.97e-7	4.78e-5
	FDR	0.950	1	0.578	1

D.2. Sphingolipid Metabolites

Summary of sphingolipid metabolite top ranked pair with a sphingolipid-related gene. Shown is rank, p-value/1 - posterior probability (1 - pp) and FDR/q-value for Discordant, EBcoexpress and Fisher.

	Discordant			EBcoexpress			Fisher		
metabolite	**Rank**	**1-pp**	**1 - q-value**	**Rank**	**1-pp**	**1 - q-value**	**Rank**	**p-value**	**FDR**
***1-3Galalpha1-3Galalpha1-3Galalpha1-4Galbeta1-4Glcbeta-Cer(d18:1/24:1(15Z))**	272483	6.87e-3	1.31e-2	312901	0.385	0.496	998686	0.216	1
3-O-Sulfogalactosylceramide (d18:1/18:1(9Z))	105638	2.91e-3	5.54e-3		0.253	0.345	99173	0.096	1
C18-OH Sulfatide	854304	1.96e-2	3.79e-2	554288	0.452	0.566	523347	0.173	1
Cer(d18:1/22:1(13Z))	158937	4.21e-3	8.02e-3	165520	0.314	0.417	91285	0.093	1
Cer(d18:1/23:0)	626399	1.47e-2	2.83e-2	372631	0.406	0.517	384322	0.155	1
Cer(d18:1/24:1(15Z))	70023	2.02e-3	3.82e-3	66774	0.226	0.311	53879	0.077	1
Cer(d18:2/23:0)	198078	5.14e-3	9.80e-3	372969	0.406	0.517	102053	0.097	1
Ceramide (d18:1/18:0)	161785	4.27e-3	8.14e-3	321358	0.388	0.499	70169	0.085	1
Ceramide (d18:1/22:0)	15721	5.49e-4	1.02e-3	15698	0.123	0.178	16451	0.050	1
Ceramide (d18:1/25:0)	134584	3.62e-3	6.90e-3	137213	0.295	0.394	78423	0.088	1
CerP(d18:1/24:1(15Z))	169281	4.45e-3	8.49e-3	255398	0.362	0.470	796240	0.200	1
Ganglioside GA1 (d18:1/16:0)	409392	9.96e-3	1.91e-2	441601	0.426	0.538	944945	0.212	1
Ganglioside GM1 (18:1/9Z-18:1)	168094	4.42e-3	8.44e-3	290467	0.377	0.486	457741	0.165	1
Ganglioside GM3 (d18:1/14:0)	663578	1.55e-2	2.99e-2	733865	0.486	0.599	423699	0.161	1
Ganglioside GM3 (d18:1/16:0)	1160391	2.60e-2	5.07e-2	633445	0.468	0.581	733809	0.195	1
GlcAbeta-Cer(d18:1/18:0)	927875	2.11e-2	4.10e-2	566984	0.455	0.568	1098109	0.224	1
Glucosylceramide (d18:1/25:0)	731165	1.69e-2	3.27e-2	527497	0.447	0.560	676768	0.189	1
Lactosyceramide (d18:1/18:1(9Z))	147435	3.93e-3	7.49e-3	217794	0.344	0.451	262888	0.136	1
N-(tetradecanoyl)-sphing-4-enine-1-(2-aminoethylphosphonate)	376192	9.22e-3	1.77e-2	432036	0.423	0.535	226664	0.129	1

NeuGcalpha2-3Galbeta1-4GlcNAcbeta1-3Galbeta1-4Glcbeta-Cer(d18:1/24:1(15Z))	472751	1.13e-2	2.18e-2	545833	0.451	0.564	1043493	0.220	1
N-Lignoceroylsphingosine	1871161	4.09e-2	8.05e-2	1334581	0.556	0.665	1877890	0.269	1
N-Palmitoylsphingosine	520229	1.24e-2	2.38e-2	319315	0.388	0.498	441743	0.163	1
SM(d16:1/17:0)	130542	3.52e-3	6.71e-3	107532	0.270	0.365	298258	0.142	1
SM(d18:0/24:1(15Z))	1207572	2.70e-2	5.27e-2	1803152	0.590	0.695	1106829	0.224	1
SM(d18:1/12:0)	12906	4.64e-4	8.58e-4	14048	0.117	0.169	95758	0.095	1
SM(d18:1/14:0)	3146	1.45e-4	2.61e-4	3449	0.060	0.090	23730	0.057	1
SM(d18:1/16:1)	1855826	4.06e-2	7.99e-2	2791220	0.637	0.735	2001704	0.276	1
SM(d18:1/20:1)	2624635	5.68e-2	1.12e-1	2090254	0.606	0.709	1590396	0.254	1
SM(d18:1/23:0)	140351	3.76e-3	7.17e-3	222316	0.453	0.453	188406	0.121	1
SM(d18:1/24:1(15Z))	198733	5.15e-3	9.83e-3	146886	0.402	0.402	334705	0.148	1
SM(d18:2/14:0)	493438	1.18e-2	2.27e-2	473547	0.547	0.547	523556	0.173	1
SM(d18:2/15:0)	214413	5.52e-3	1.05e-2	403837	0.527	0.527	74789	0.087	1
SM(d18:2/23:0)	144149	3.85e-3	7.34e-3	206117	0.338	0.444	75113	0.087	1
Sphingosine	187977	4.90e-3	9.34e-3	156783	0.308	0.410	451307	0.164	1
Sphingosine 1-phosphate	319191	7.94e-3	1.52e-2	394444	0.712	0.524	554183	0.176	1
Sphingosine-1-phosphocholine	99476	2.76e-3	5.24e-3	109241	0.272	0.367	461599	0.166	1
Trihexosylceramide (d18:1/16:0)	954799	2.17e-2	4.21e-2	536542	0.449	0.562	869371	0.206	1

D.3. Breast Cancer miRNAs

Summary of Breast cancer miRNAs top ranked pair with a transcript. Shown is rank, p-value/1 - posterior probability (1 - pp) for correlation metrics and NBGLM.

treatment	statistic	hsa-mir-107	hsa-mir-150	hsa-mir-152	hsa-mir-191	hsa-mir-24-2	hsa-mir-374a	hsa-mir-574	hsa-mir-454
Correlation method comparison									
Spearman	rank	4	1	72	135	85	7	368	40
	1-PP	2.7e-3	8.3e-4	1.2e-2	1.49e-2	1.3e-2	3.9e-3	2.5e-2	8.7e-3
SparCC	rank	133	3	473	1499	300	602	269	1068
	1-PP	1.2e-3	1.4e-3	2.3e-2	4e-2	1.8e-2	2.6e-2	2.7e-2	3.4e-2
NBGLM	rank	47	1329	220	1249	179	78	101	936
	p-value	7e-3	9e-4	1.8e-3	9.4e-3	1.5e-3	7.6e-4	9e-4	7.2e-3
Pearson	rank	578	445	627	315	77	1996	957	9
	1-PP	2.1e-2	1.9e-2	2.1e-2	1.7e-2	1.3e-2	2.9e-2	2.3e-2	8.6e-3
BWMC	Rank	364	2	207	408	89	342	864	79
	1-PP	4.5e-2	6.5e-3	0.037	4.9e-2	2.6e-2	4.6e-2	6.6e-2	2.5e-2

D.4. Standard EM vs. Subsampling EM Rank and 1-PP

Summary of Breast cancer miRNAs top ranked pair with a transcript for Standard EM and Subsampling EM. Shown is rank, p-value/1 - posterior probability (1 – pp).

GBM					
Method	statistic	hsa-mir-124a	hsa-mir-137	hsa-mir-326	hsa-mir-92b
Standard EM	Rank	223	1081	472	83
	1-PP	2.51e-7	1.16e-6	5.04e-7	1.05e-7
Subsampling EM	Rank	691	2833	648	91
	1-PP	5.76e-6	3.27e-5	5.41e-6	4.47e-7

Breast Cancer									
Method	statistic	hsa-mir-107	hsa-mir-150	hsa-mir-152	hsa-mir-191	hsa-mir-24-2	hsa-mir-374a	hsa-mir-574	hsa-mir-454
Standard EM	Rank	4	1	72	135	85	7	368	40
	1-PP	2.7e-3	8.3e-4	1.1e-2	1.5e-2	1.3e-3	3.9e-3	2.5e-2	8.7e-3
Subsampling EM	Rank	4	1	81	151	50	12	398	37
	1-PP	6.4e-4	1.5e-4	3.7e-3	5.0e-3	2.8e-3	1.2e-3	8.7e-3	2.4e-3

D.5. Three Component vs. Five Component Rank and 1-PP

Summary of Breast cancer miRNAs top ranked pair with a transcript for 3-component vs. 5-component mixture models. Shown is rank, p-value/1 - posterior probability (1 – pp).

GBM					
Method	statistic	hsa-mir-124a	hsa-mir-137	hsa-mir-326	hsa-mir-92b
3 Components	Rank	223	1081	472	83
	1-PP	2.51e-7	1.16e-6	5.04e-7	1.05e-7
5 Components	Rank	2101	1360	339	433
	1-PP	1.54e-8	7.30e-9	6.13e-10	9.43e-10

Breast Cancer									
Method	statistic	hsa-mir-107	hsa-mir-150	hsa-mir-152	hsa-mir-191	hsa-mir-24-2	hsa-mir-374a	hsa-mir-574	hsa-mir-454
3 Components	Rank	4	1	72	135	85	7	368	40
	1-PP	2.7e-3	8.3e-4	1.1e-2	1.5e-2	1.3e-3	3.9e-3	2.5e-2	8.7e-3
5 Components	Rank	2067	363	51	32	276	1811	444	421
	1-PP	9.5e-4	1.5e-4	1.8e-5	1.4e-5	1.1e-4	8.1e-4	1.8e-4	1.7e-4