Integrative statistical methods for the analysis of transcriptomic and metabolomic data

by

Laila M. Poisson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Jeremy M. G. Taylor, Co-Chair
Associate Professor Debashis Ghosh, Co-Chair, Pennsylvania State University
Professor Arul M. Chinnaiyan
Associate Professor Bhramar Mukherjee
Assistant Professor Arun Sreekumar, Medical College of Georgia

To Verity.
May you grow to be the woman that *you* want to be.

ACKNOWLEDGEMENTS

This dissertation marks the culmination of a long scholastic endeavor. Fresh out of high-school I once declared to my mother that I would not be so crazy as to be in school doing post-graduate studies at the age of thirty, but here I am, plus a couple of years, glad to have given it a go, glad to be seeing it come to an end. On this journey I have been grateful for the help and support of many people. I could not possibly list them all here but I would like to mention some key players.

Firstly, I thank my committee. To Debashis, thank you for sticking with me to the end. As my academic advisor, research assistantship advisor, thesis advisor and dissertation co-chair you have been there for every stage of this journey. I have appreciated that you have always valued my ideas even when they go a bit off track. To Jeremy, thank you for your continued support throughout the years, be it financially, academically, or personally. Your Cancer Biostatistics Training Program gave me the academic support I needed in the middle years of my studies as well as depth of training in cancer research. I also appreciate that you have always provided opportunities to meet and talk with other researchers be it through visiting speakers or introductions at conferences. To Arul, thank you for the opportunity to be a part of your incredible team. Transcriptomics, metabolomics, proteomics, it was these data that inspired the work contained in this thesis. To Arun, thank you for the last several years of collaboration as we tackled those omics datasets and wound up with some interesting results and great publications. I also thank you for being a role model of someone who stays true to family and beliefs while getting the work done. To Bhramar, thank you for providing an additional point of view to my research and for being available for questions or simply conversation.

Next, thank you to all of my friends, colleagues and instructors over the years here at U-M. Here is where I could not begin to list names, but if you are reading this then you are likely among them. Thanks to each of you for sharing your knowledge, your inspiration, and your time. Thank you also to the faculty and staff of the Genome Sciences Training Program for giving me the push that I needed to return to school full time, providing depth of training in genetics, and continuing to provide opportunities for professional growth as a program alumna.

Of course a special thank you is due to my friends at the Henry Ford Health System Department of Biostatistics and Research Epidemiology. Thank you for giving me such an enjoyable start to my biostatistical career that I came back to school so that I could do my job better. Thank you for the years of encouragement and moral support that you have given to me during my time at U-M. Words cannot express what your friendship has meant and I am thrilled to be able to continue my career with you.

Finally, to my family, near and far, living and deceased, thank you for being my cheerleaders – proud of my every accomplishment and encouraging through my failures. I am blessed to have each of you in my life. And, most importantly, an extra thank you goes to my husband. Without your suggestion, Chris, I may never have found the world of biostatistics. Without your love, support, and sacrifice I would have never made it through this journey. And now dear, it's your turn. Enjoy.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

In the last decade the high-throughput assessment of gene expression, i.e. transcriptomics, in a sample has grown to measure the whole genome with over 40,000 probes per microarray (e.g. Varambally et al. (2005) [68]). Microarray technology uses complementary DNA fragments (Agilent Technologies, `www.chem.agilent.com`) or oligonucleotide probes (Affymetrix, `www.affymetrix.com`) fixed on a slide to assess mRNA abundance via the complementary binding of single-stranded nucleotide sequences [49]. The reproducibility of the quantitative and qualitative assessment has been demonstrated within platforms and, to a lesser extent, between platforms [42]. Gene profiles resulting from gene expression analysis have even been translated to clinical application (e.g. Wigelt et al. (2009) [72]).

An area of high-throughput analysis that has recently emerged is metabolomics, the assessment of the small molecules in the sample [3, 25]. The data are generated by mass spectrometry (MS) preceded by either gas or liquid chromatography (GC or LC) [26, 69]. The initial chromatography step separates the molecules so that they can be identified by their mass spectrum. When this separation step is skipped the metabolic activity of the cell is measured as a metabolic fingerprint, but metabolites are not measured individually [18]. In this work we consider all molecules under 10,000

KDa. This includes such classes as amino acids, fatty acids, simple carbohydrates, and exogenous drugs within the cell. Metabolomic studies currently detect between one hundred and one thousand metabolites [21, 70, 60].

Small samples sizes and potentially high variance between samples can lead to low power in the analysis of metabolomic intensity changes between human populations. Specifically, the metabolome is affected by diurnal rhythms, diet, drugs and therapeutics, and other environmental factors beyond disease [38]. However, we know that the global snapshot of any one molecular component is an isolated view of a larger picture [31]. Thus we turn to data integration through systems biology to provide a broader view of the disease while enhancing power and providing new insights.

This work has been motivated by a study of metabolomics in prostate cancer by Sreekumar et al. (2009) [60]. Here 626 metabolites are detected across 42 prostate tissue samples of increasing cancer severity; 16 benign adjacent prostate, 12 localized prostate tumor, and 14 metastatic prostate tumors. Gene expression data were also measured for 40 of these samples (excluding two metastatic samples); data unpublished. Additionally, we consider another cancer progression dataset from the the same group [68]; 4 benign adjacent prostate, 5 localized prostate tumor, and 4 metastatic prostate tumors. Though data are available for metastatic disease we focus on the comparison of benign prostate and local prostate tumor tissues.

We integrate the data by mapping the two platforms to the metabolic pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG) [35, 36, 34]. These pathways represent the current literature for metabolomic biosynthesis and degredation. Genes are mapped to the pathways via the enzymes with which they are associated. For details on this mapping please refer to Appendix A.

In this dissertation we explore three analysis methods that are able to combine gene expression and metabolite intensity data. In Chapter II, we examine a classification method that allows us to utilize a differential list of elements from a prior study to make prognostic or diagnostic predictions about samples in a current study. A prognostic application is demonstrated using gene expression datasets of various cancer samples. The method is then applied to the Sreekumar et al. (2009) [60] metabolite data using a gene profile from the Sreekumar-matched gene expression data to build the classifier. In Chapter III, we explore the power gain available to per-metabolite tests of differential intensity when the tests are weighted according to a per-pathway ranking constructed from the gene expression data. We summarize the information in the gene expression data using set enrichment scores on the KEGG pathways. The methods are explored through two simulation models; one simplistic and one motivated by the Varabmally et al. (2005) [68] gene expression data and the Sreekumar et al. (2009) [60] metabolite data. Chapter IV considers the joint enrichment testing of differential results from the metabolomic and gene expression datasets. We extend two different set enchriment tests to allow multi-dimensional input. Through simulation we compare these methods to their univariate counterparts, the Fisher's exact test, and to two simple meta-analysis approaches. An additional statistical goal of this dissertation is to define the null hypothesis under consideration in each test so that appropriate testing measures can be defined. This will be discussed for each chapter in turn.

CHAPTER II

Statistical issues and analyses of in vivo and in vitro genomic data in order to identify clinically relevant profiles

Integration of in vitro studies, i.e. experimental studies, with human “in vivo” gene expression studies is an area that is being considered more frequently in the functional genomic analysis of cancer. Hypotheses about cancer development, progression, and risk factors are difficult to test directly in a patient population. However, in experimental studies on tissue cultures and model organisms, conditions can be specifically controlled to allow biological hypotheses to be tested. Integrating the results from such controlled experiments with in vivo cancer signatures holds the potential to both infer activity of specific oncogenic pathways in vivo and to identify relevant effectors of oncogenic pathways.

This biologically directed analysis of two gene expression data sets holds potential for the integration of omics data. Consider that transcriptomic studies are more readily available and currently more robust than metabolomic studies. Thus, we can use transcriptomic data in place of the controlled in vitro studies and metabolomic data in place of the in vivo data. We explore the ability to use this methodology to use gene expression data to make predictions about the metabolomic profiles of tumor tissues.

We begin with a description of the integration task at hand and a review of recent work in in vitro and in vivo integration. We outline an approach for quantifying the predictive ability of a gene expression profile determined from an in vitro experiment based on the tissue similarity approach of Sandberg and Ernberg (2005) [56]. We describe the application of the proposed methodology using in vitro data from a wound healing study conducted by Chang et al. (2004) [10] and in vivo data from Glinsky et al. (2004) [23], van't Veer et al. (2002) [67], and Beer et al. (2002) [4]. We follow-up with the integration of the Sreekumar et al. (2009) [60] metabolomic data and matched gene expression data (unpublished).

2.1 Integrating in vitro and in vivo studies

To understand the mechanisms by which oncogenes cause cancer, studies have used gene-expression profiling to identify downstream targets of oncogenic pathways in cell-culture systems. Conceptually, this involves manipulating a gene in an in vitro system, measuring the global profile using gene expression technology and then trying to relate the in vitro gene expression profile to an in vivo gene expression profile. Such an approach was taken by Lamb et al. (2003) [39] to determine the direct transcriptional effects of the oncogene Cyclin D1. In vitro experiments were performed in which the Cyclin D1 was both over and under expressed, and global gene expression profiles were determined. Lamb et al. (2003) [39] found that there was a significant correlation between the targets found in vitro and the ordered gene list in a human tumor dataset thus suggesting the role of Cyclin D1 regulation in tumorigenisis. Another example of in vitro/in vivo gene expression data integration appears in the study of Huang et al. (2003) [32]. They developed distinct in vitro oncogenic signatures for three transcription factors: Myc, Ras and E2F1-3. These signatures were able to predict the Myc and Ras state in mammary tumors that developed in transgenic mice expressing either Myc or Ras, suggesting that specific oncogenic events are encoded in global gene-expression profiles.

Additionally, studies have used gene-expression profiling of cancerous growths induced in model organisms to examine tumor development or progression. Though model organism studies have the added difficulty of mapping orthologous genes between organisms, a difficulty not shared with tissue and cell cultures of human origin, there have been promising applications. For example, Sweet-Cordero et al. (2005) [62] defined a KRAS induced lung cancer signature by comparing lung tumors generated from a spontaneous KRAS mutation mouse model to normal mouse lung tissue. They then correlated this KRAS lung cancer signature with gene expression profiles in human lung cancer studies and found that the mouse signature shared significant similarity with human lung adenocarcinoma but not with other lung cancer types. Next, Sweet-Cordero et al. (2005) [62] looked for evidence of the KRAS signature in human tumors carrying activating KRAS mutations relative to wild-type tumors. Although no individual genes were significantly associated with the KRAS mutation status in human tumors, the mouse KRAS signature was significantly enriched among genes rank-ordered by differential expression in human tumors with a KRAS mutation.

2.1.1 Background and Review

One class of methods that has been popular in the literature for in vitro/in vivo genomic data analysis is the following. First, one generates ordered lists of genes using the in vivo expression data. One then generates a differentially expressed gene list using the in vitro data and studies the overlap between the two lists. The seminal examples of this are in Mootha et al. (2003) [44] and Lamb et al. (2003) [39], which were then used as the basis of the Gene Set Enrichment Analysis (GSEA) method

[61]. We describe the GSEA methodology by briefly reviewing what was done in the Lamb et al. (2003) [39] study.

First, a list of differentially expressed genes was generated based on the comparison of Cyclin D1 overexpressing relative to wildtype (no Cyclin D1 manipulation) mammary epithelial cell lines. Next, each gene's expression in vivo, from 190 human tumor samples of various origins, was correlated to that of Cyclin D1 and the genes were ranked accordingly. Then, a Kolmogorov-Smirnov (KS) statistic was used to determine if the in vitro differential expression list clustered within the correlation-ordered in vivo list. Since there was significant evidence of clustering, Lamb et al. (2003) [39] determined that the in vitro-defined targets of Cyclin D1 were correlated with their respective levels in vivo. This suggests that the direct regulatory effects of Cyclin D1 may play an important role in tumorigenesis.

There are some desirable features of the GSEA method. First, it utilizes all the information available in the in vivo gene expression data; no thresholding is done in that dataset. Second, a Kolmogorov-Smirnov statistic is used for the analysis, which is a nonparametric method and thus provides some robustness. However, there are several disadvantages to GSEA as well. For instance, note that there is thresholding done in the in vitro gene expression dataset to select the differentially expressed gene set. A potential improvement to the GSEA method, to avoid this thresholding, would be the following. First, one determines the common genes in the in vivo and in vitro datasets. One then takes the scores of differential expression from the in vitro data, finds the corresponding correlation scores (correlation with Cyclin D1) in the in vivo data and examines a scatterplot of the two variables. If the association is linear, then one tests for association using the Pearson correlation coefficient between the two variables. If instead the association appears nonlinear, then one could use a smoothing-spline based test [40]. Such an approach would give a direct test of association between the correlations in vivo and the differential expression measurement in vitro without requiring thresholding of any datasets and would still allow for a nonlinear relationship between the two variables.

Before going further, let us consider the null hypothesis under consideration in the GSEA method, or the variants proposed above. Specifically, in the Lamb et al. (2003) [39] study they test:

> H_0: **There is no association between differential expression of Cyclin D1-overexpressed, relative to non-overexpressed, cell lines and correlation with Cyclin D1 in human tumors.**

The alternative hypothesis is that there is an association. In specifying the null hypothesis we uncover a more subtle disadvantage of the GSEA method - the determination of the distribution of the KS test statistic under the null hypothesis.

Two variants of permutation testing have been proposed by Subramanian et al. (2005) [61] to elucidate the distribution of the KS test statistic assuming the null hypothesis is true. In the first, the sample labels in the in vitro data are permuted, the differentially expressed gene signature is redefined, and the Kolmogorov-Smirnov statistic is recomputed based on this new signature; see Figure 2.1A, red. Here the implication is that the correlation between the two Cyclin D1 levels in the cell line experiment is removed by the permutation. However, this addresses the differential expression in the in vitro samples but does not address a null association with the in vivo samples. In the second version, the sample labels in the in vitro and in vivo datasets are permuted, both the in vitro differential expression signature and the in vivo correlations are redefined, and the Kolmogorov-Smirnov statistic is recomputed; see Figure 2.1A, blue. Again, the implication is to remove the association within the

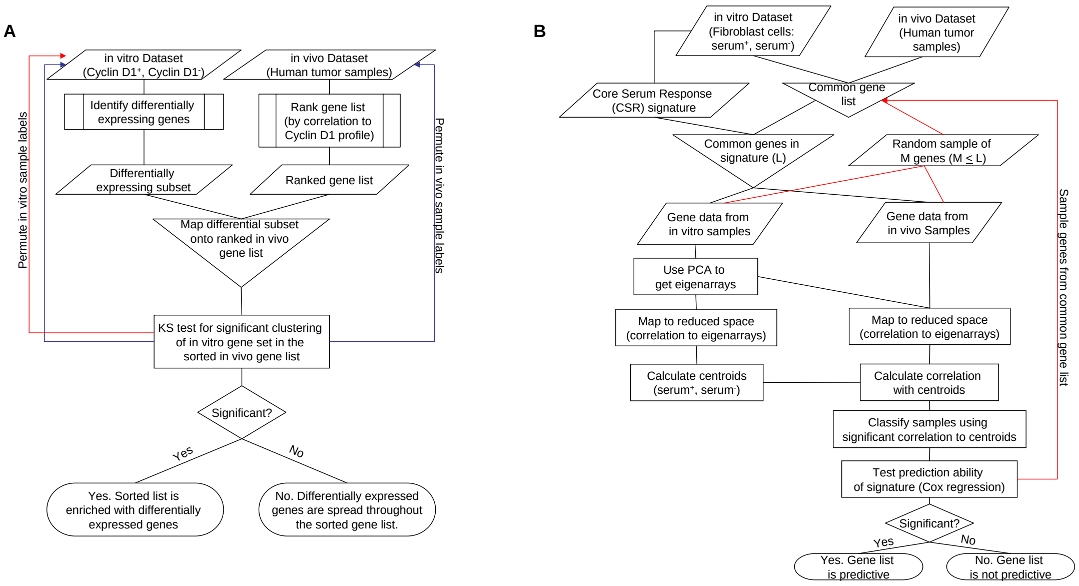

A
in vitro Dataset
(Cyclin D1+, Cyclin D1-)
in vivo Dataset
(Human tumor samples)
Identify differentially expressing genes
Rank gene list (by correlation to Cyclin D1 profile)
Differentially expressing subset
Ranked gene list
Map differential subset onto ranked in vivo gene list
KS test for significant clustering of in vitro gene set in the sorted in vivo gene list
Significant?
Yes
No
Yes. Sorted list is enriched with differentially expressed genes
No. Differentially expressed genes are spread throughout the sorted gene list.
Permute in vitro sample labels
Permute in vivo sample labels
B
in vitro Dataset
(Fibroblast cells: serum+, serum-)
in vivo Dataset
(Human tumor samples)
Core Serum Response (CSR) signature
Common gene list
Common genes in signature (L)
Random sample of M genes (M ≤ L)
Gene data from in vitro samples
Gene data from in vivo Samples
Use PCA to get eigenarrays
Map to reduced space (correlation to eigenarrays)
Map to reduced space (correlation to eigenarrays)
Calculate centroids (serum+, serum-)
Calculate correlation with centroids
Classify samples using significant correlation to centroids
Test prediction ability of signature (Cox regression)
Significant?
Yes
No
Yes. Gene list is predictive
No. Gene list is not predictive
Sample genes from common gene list

Figure 2.1: **Schematic representations of the GSEA-type and TSI-type algorithms** (A) The GSEA-type algorithm is depicted along with the two suggested permutation tests (red = permutation 1, blue = permutation 2). Details of the Lamb et al. (2003) [39] study are included for illustration. (B) The TSI-type algorithm is depicted along with the suggested permutation test (red). Details of the Chang et al. (2004) [10] Core Serum Response signature classification are included for illustration.

in vitro and in vivo experiments. Yet this permutation scheme still does not address the association between the in vitro differential expression and the in vivo correlation.

The role of permutation testing is to simulate the distribution of the test statistic assuming that H_0 is true; however, the two permutation schemes developed in the GSEA method do not do this. Permutation of the sample labels fails because the null hypothesis pertains to the population of genes in the two studies and not the relation of samples within a study. Additionally, Shedden (2004) [58] suggests that permuting the sample labels of both the in vitro and in vivo data sets is not appropriate. Simply, if the permutation does not correctly model the null hypothesis, then we are answering a different question than the one asked.

There is an alternative approach to the GSEA method for integrative analysis of in vitro and in vivo data, the implementation of which is the focus of this chapter.

It is based on ideas of classification and clustering since the goal in many genomic studies utilizing high-throughput expression technologies is to develop a signature that can discriminate between relevant classes or groups of samples. In general, demonstration of the predictive or prognostic ability of a classification signature on independent data sets is a crucial step in the validation of that signature [52]. Thus, differential expression signatures discovered in vitro are often "validated" on independent in vivo data sets, such that the in vitro data is the training dataset and the in vivo data is the testing dataset. In this validation setting, the null hypothesis that we wish to test is the following:

> H_0^{class}**: There exists no set of genes derived from the in vitro gene expression that can predict clinical outcome in the in vivo expression data.**

The alternative is that at least one set of genes derived from the in vitro data is predictive. Notice that this null hypothesis is different from the null hypothesis described for the GSEA method. For clarity, we will refer to H_0^{class} as the classification null hypothesis.

An advantage of the classification null hypothesis is that permutation testing becomes possible here. In particular if H_0^{class} is true, then any set of genes derived from the in vitro expression profile data will have no ability to separate samples in the in vivo expression dataset with regard to a clinical outcome. Thus, we can take random sets of genes from the in vitro data and apply the classification algorithm of interest. If the classification null hypothesis is true, then all sets of genes, including the derived signature, should provide equal prediction performance.

The classification null hypothesis has motivated the following algorithm that we have used in our previous work [68]. Here we are considering the genes common to the in vitro and in vivo expression datasets.

1. Derive a gene signature (i.e. interesting gene list) from the in vitro gene expression data;
2. Select those genes from the in vivo expression data that are included in the in vitro signature and cluster the samples from the in vivo expression data into two groups using hierarchical clustering with average linkage clustering and Euclidean distance;
3. Calculate the log-rank statistic for survival between the two groups of patients;
4. Let L denote the size of the gene list in 1. Randomly choose L genes from the in vitro data as the gene signature. Continue with steps 2 and 3 above.
5. Repeat steps 2-4 1000 times. Calculate the proportion of datasets in which the log-rank statistic is greater than the one calculated initially from the signature in step 1.

The proportion calculated in step 5 will be the permutation p-value under the classification null hypothesis. This permutation scheme will form the basis of assessing significance for our proposed analytical scheme described in the next section. We note that one could also modify the GSEA procedure in a similar way, as shown in Lamb et al. (2003) [39], such that we randomly draw the gene set from the in vitro data rather than assessing differential expression based on permuted sample labels. Unfortunately, Shedden (2004) [58] shows that when one does not account for gene-gene correlation, the resulting test statistic can be too liberal by as much as 10 times.

Notice that a limitation of the classification null hypothesis is that the alternative hypothesis states that there exists at least one signature from the in vitro expression data that is predictive in the in vivo expression data. In fact the experimentally derived gene list need not be a unique classifier. It has been recently noted that there are likely many gene signatures that have similar predictive power [14, 15]. It may be due in part to genetic redundancy or to the high correlation of genes within a pathway. Yet if the in vitro gene signature is able to predict prognosis better than a randomly selected set of genes we expect that there is biological significance to that signature. Thus permutation testing helps us to determine if the gene set derived from the in vitro experiments is of interest for further study of its biological relevance.

2.1.2 Proposed methodology for in vitro/in vivo analyses

The paper of Sandberg and Ernberg (2005) [56] considers the relationship between the gene expression of in vitro cell cultures and their respective in vivo tumor samples. To that end they developed an algorithm for comparing gene expression values across experiments that they call the tissue similarity index (TSI). We use that algorithm here to compare the in vivo tumor samples to the in vitro samples of a lab experiment.

The algorithm of Sandberg and Ernberg (2005) [56] is as follows; see Figure 2.1B. Principal component analysis is run on the covariance matrix of gene expression for genes in the in vitro dataset. Data are scaled across arrays so that each gene has a mean expression of zero and a unit standard deviation. The resulting Eigenarrays (Eigenvectors) are stored. To project the in vitro gene expression into the reduced dimensional space, created by the Eigenarrays, calculate the correlation between each Eigenarray and each in vitro sample array. The consensus signature for each experimental condition (serum induced and serum independent) is represented by its median centroid in the reduced space.

To integrate the in vivo data, first map the in vivo samples into the same reduced space of the in vitro samples by again calculating the correlation between each Eigenarray and in vivo sample array. To maintain scale in this correlation, the tumor samples are also standardized so that each gene has a mean expression of zero and a unit standard deviation. The distance between the in vivo tumor sample and each of the two consensus signatures, ie centroids, is calculated using Pearson correlation. Samples are classified with the experimental condition with whose centroid they correlate best.

There are several differences between their and our implementations of TSI. First, in contrast to Sandberg and Ernberg (2005) [56], we use positive statistical significance of the TSI to determine classification, thus allowing some samples to remain unclassified. In their paper they used an ad-hoc threshold value for TSI score, delineating moderate and high correlation groups. It is natural to believe that some of the in vivo samples will not correlate well with the in vitro conditions. These unclassified samples may actually be informative in that they define a subset of cases which do not meet our expectation as developed in the hypotheses tested in vitro. Second, the goal of the Sandberg and Ernberg (2005) [56] paper was qualitative assessment of cell line gene expression relative to in vivo tumor gene expression, thus they do not address the issue of statistical significance of their method. However, the classification provided by the TSI can be tested for prognostic or diagnostic value depending upon the study goal.

Since the gene signature on which the classification is based is determined from the in vitro data, and does not use the in vivo data, the statistical significance of any tests on the in vivo data can be accepted without bias. This is an example of using the in vitro data for the training dataset and the in vivo data for the testing dataset. Indeed, if this in vivo validation is not marginally significant it is not of interest to proceed further to test the classification null hypothesis.

Here the TSI method develops a classification scheme from the in vitro signature. The null hypothesis of interest is again the classification null hypothesis as presented above. We thus propose the use of a permutation test to determine the utility of the gene signature in its classification ability. In the following we slightly modify the permutation test procedure described in the previous section to account for gene-gene correlation within the in vitro gene signature. Specifically, as it is likely that genes within a pathway are correlated, it is reasonable to assume that the significantly differentially expressed genes that comprise the in vitro signature are correlated. Shedden (2004) [58] showed that this correlation can lead to liberal p-values. Additionally, in the classical genetics setting, Nyholt (2004) [50] shows that permutation tests that do not account for this correlation can be misleading and proposes a simple adjustment. In essence, rather than randomly selecting L genes in each cycle of the permutation test, only M ($M < L$) genes are selected, where M is calculated to be the effective number of independent genes in the gene signature [50]; see Figure 2.1B.

Finally, the permutation test for the TSI analysis has two interesting attributes against which the classification signature is compared. Specifically, in permuting the data, the TSI scores are recalculated using the randomly selected gene list and with each randomly selected set of genes there is a possibility of unclassified samples. Thus the classification is compared to: (1) the measure of association with predictive factors in vivo, and (2) the percentage of unclassified samples in vivo.

2.2 Gene expression data acquisition and preparation

For the purpose of demonstration we use, as the in vitro derived signature, the wound healing signature of Chang et al. (2004) [10]. Derived from cultured fibroblasts in the presence and absence of serum components, the wound healing signature is composed of 573 genes that are differentially expressed in response to serum. We consider the wound healing signature, or Core Serum Response (CSR), as the in vitro basis of classification of in vivo tumor samples - prostate tumor samples [23], breast tumor samples [67], and lung tumor samples [4] - into good and bad prognosis groups.

The fibroblast gene expression data [10] were downloaded from the Stanford Microarray Database (SMD, `http://smd.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=293`) (platform: cDNA microarray, 50 samples). The data were normalized using loess normalization by print block within array [74]. Inter-array variability was accounted for by scaling using the MAD (median absolute deviation). Missing data were imputed using KNN (K-nearest neighbors) imputation as implemented in the `pam.r` package [27, 65].

Localized prostate tumor probe-set level expression measures and recurrence free survival information [23] were obtained from the Sidney Kimmel Cancer Center website (no longer posted at time of submission) (platform: Affymetrix U95Av2, 295 samples). Lung adenocarcinoma probe-set level expression measures and overall survival information [4] were obtained from `http://dot.ped.med.umich.edu:2000/pub/Lung/index.html` (platform: Affymetrix HUgenFL, 86 samples). Sporadic breast cancer expression data and recurrence free survival information [67] were obtained from `http://www.rii.com/publications/2002/vantveer.html` (platform:

Table 2.1: **Genes are matched between gene expression array platforms using Unigene ID numbers.** For the permutation testing, all common genes between the in vitro experiment [10] and each in vivo experiment were considered (prostate [23], breast [67], lung [4], respectively). The classification of a set of in vivo samples was done based on only the CSR genes identified in that data set. Permutation sample size was determined based on the effective number of independent genes in the CSR signature.

Samples	Unigene maped genes per microarray	Genes common with in vitro samples	Unigene mapped CSR genes	Effective number of independent genes
In Vitro	20414	-	484	-
Prostate Cancer	11772	9753	367	345
Breast Cancer	17168	13600	421	399
Lung Cancer	4705	3891	158	136

Aglient Hu25K, 78 samples). Each of these experiments was normalized by global scaling per array. No imputation was done for missing data in the tumor sample data sets.

Unigene Cluster ID number was used to map genes between platforms. Annotation information was acquired from SOURCE [12]. If, for a given platform, multiple measurements were represented by the same Unigene Cluster ID, these expression values were averaged within array, thus allowing one-to-one mapping of genes between platforms. Genes were mapped to Unigene Cluster ID from GenBank Accession number if available [10, 23, 67] or from Unigene Symbol [4].

2.3 Application of the TSI based classifier

The classifier was built using the CSR in vitro signature and the TSI algorithm, described in the previous section and in Figure 2.1B. The classifier was built for each of the three in vivo experiments using only those genes in the CSR signature that were common to both the in vivo and in vitro experiments; see Table 2.1 and Figure 2.2. All 50 Eigenarrays were used for the TSI classification algorithm and classification is based on significant positive correlation with one of the two CSR

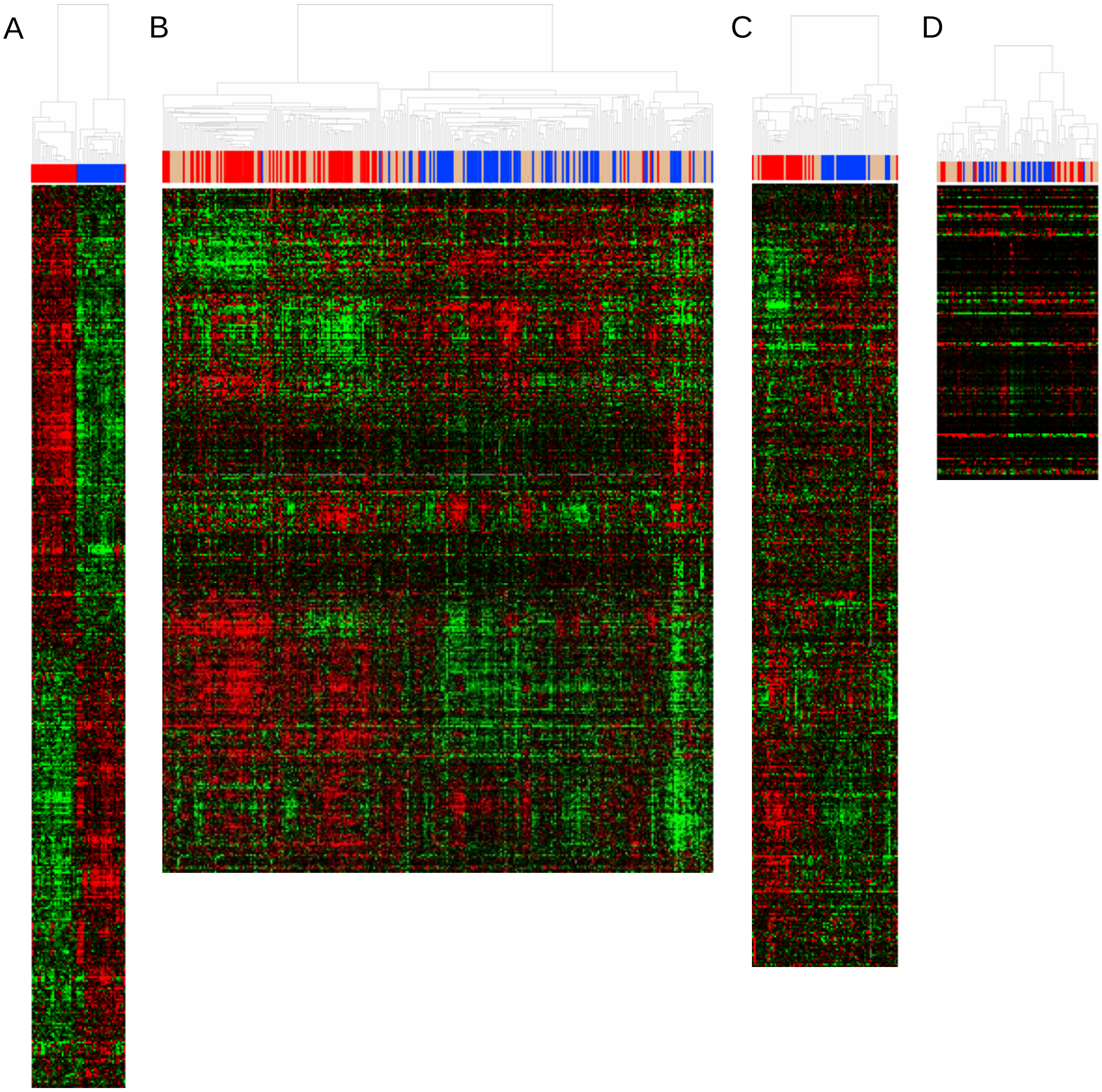


Figure 2.2: **Heatmap of core serum response genes in each of the four data sets considered.** Red are serum induced samples, blue are serum independent samples, and tan are unclassified samples. (A) Expression of 484 Unigene mapped core serum response genes in the 50 in vitro samples of the primary experiment [10]. (B) Expression of the 376 core serum response genes in the 295 prostate tumor samples [22] (C) Expression of the 421 core serum response genes in the 78 v'ant Veer samples [67] (D) Expression of the 158 core serum response genes in the 86 Beer samples [4]

group centroids. Figure 2.3 plots the first two dimensions of this reduced space for each of the three tumor types. The cell cultures that were grown in the presence of serum were considered to be serum induced, whereas those grown without serum components were serum independent. In vivo samples that correlate significantly ($p < 0.05$) with the composite serum induced signature, ie centroid, are classified as serum induced. Likewise, those in vivo samples correlating significantly with the centroid of the serum independent samples are labeled serum independent. In vivo samples that do not correlate significantly with either centroid remain unclassified. In Figure 2.3, the tumor samples are colored according their classification and the in vitro samples and centroids are included for reference.

According to H_0^{class}, we wish to see if the in vitro derived CSR signature has prognostic ability in vivo. Thus the prognostic ability of the CSR signature as a classifier was tested using univariate Cox regression; see Table 2.2. The TSI score was incorporated through its discrete classification of the in vivo samples, as described above. Figure 2.4 contains the Kaplan-Meier survival curves for this discrete classification. The red and blue curves represent the serum activated and serum independent classifications, respectively. Log rank statistics on the Kaplan-Meier estimates indicate that there is a significant separation between the curves for the prostate tumors ($p < 0.0001$), the breast tumors ($p = 0.0207$), and the lung tumors ($p = 0.0352$). The tan curve shows the survival of those samples that did not significantly correlate with either the serum activated or serum independent profiles and are thus left unclassified by the TSI algorithm. When this unclassified group was included in the Log-rank test of survival curve separation the prostate cancer and breast cancer samples remained significant ($p < 0.0001$ and $p = 0.0078$, respectively) whereas the lung cancer samples were marginally significant ($p = 0.0789$).

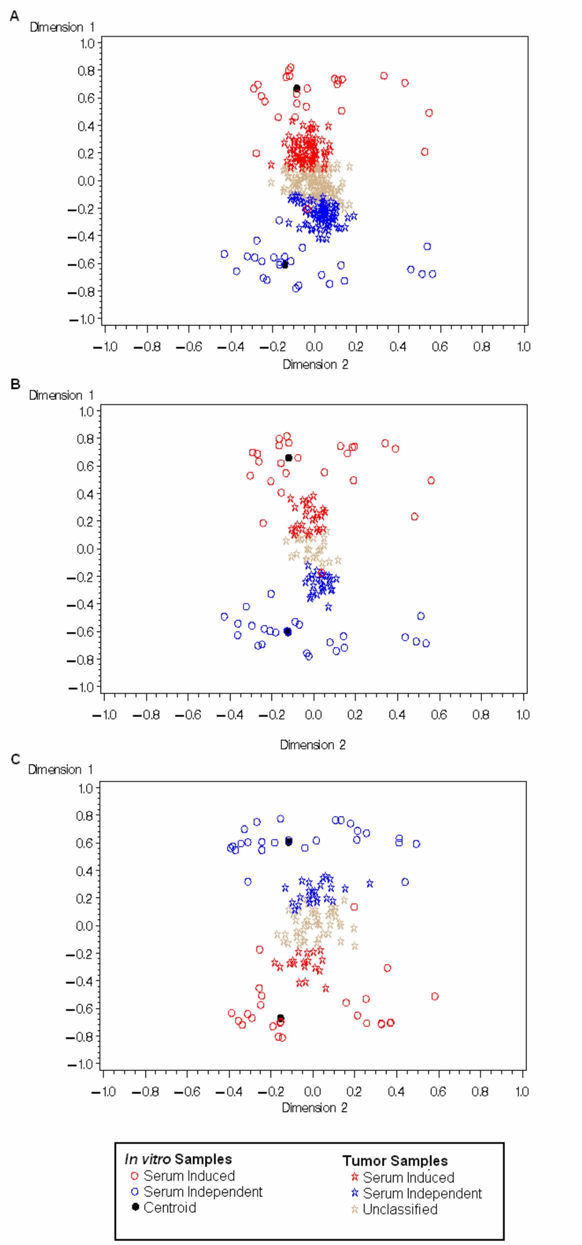


Figure 2.3: **In vitro samples and tumor samples are plotted for the first two dimensions of Eigenspace.** (A) 55% of the prostate tumor samples are classified: 78 as serum induced, 83 as serum independent (B) 69.3% of the breast tumor samples are classified: 27 as serum induced, 27 as serum independent. (C) 48.9% of the lung tumor samples are classified: 18 as serum induced, 24 as serum independent.

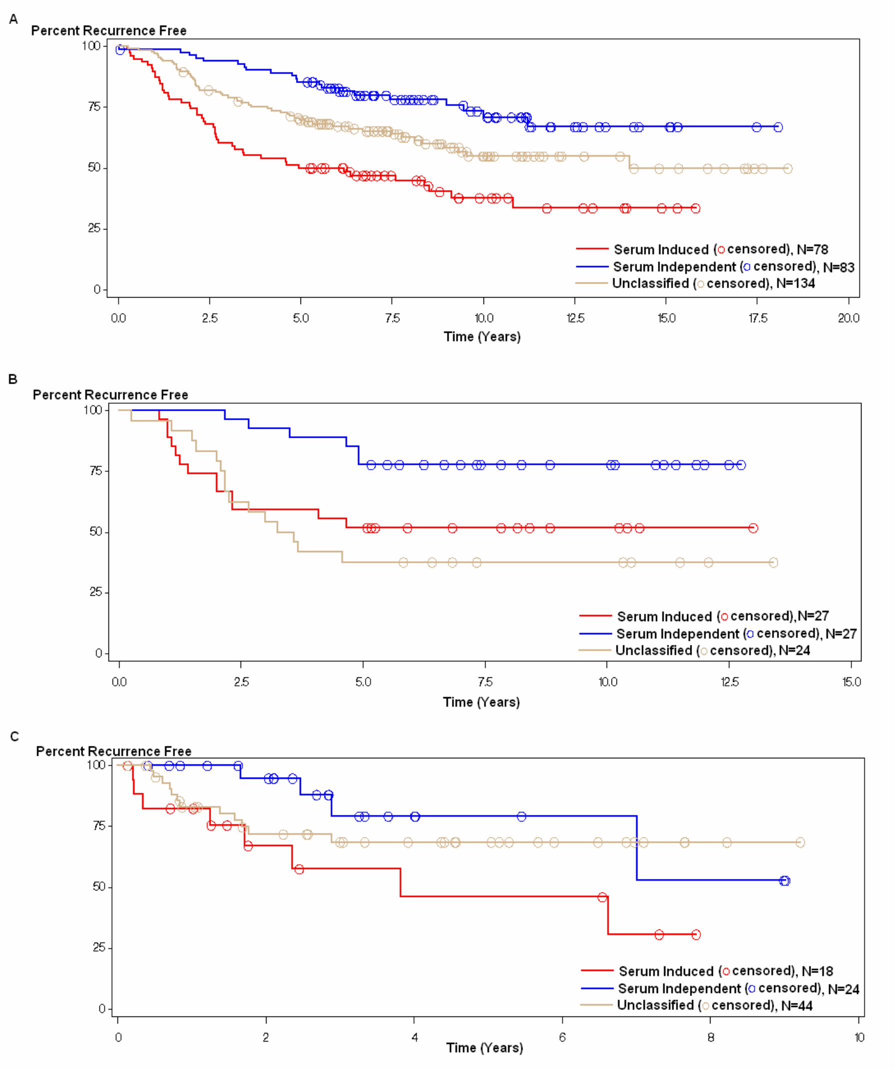


Figure 2.4: **Survival of tumor samples by classification.** These plots demonstrate that patients with samples in the serum induced class are likely to have a worse prognosis. Both the classified and unclassified samples are included in these plots. The log-rank statistic p-values from Kaplan Meier estimation are given for the separation of the classes without and with inclusion of the unclassified samples in the model (A) recurrence free survival in prostate cancer: $p < 0.0001$; $p < 0.0001$, (B) recurrence free survival in breast cancer: $p = 0.0207$; $p = 0.0078$, (C) overall survival in lung cancer: $p = 0.0352$; $p = 0.0789$

Table 2.2: **Cox regression results for the classified samples.** Cox regression was run on the samples that were classified as serum induced or serum independent by the CSR gene signature. Unclassified samples are excluded from this analysis. The hazard ratios are relative to the serum independent classification.

	Number Classified						
	Serum	Serum	Percent	Hazard	χ^2 Test		Empirical
Samples	Induced	Independent	Unclassified	Ratio	Statistic	p-value	p-value
Prostate	78	83	45.4%	3.35	20.9	<0.0001	0.0040
Breast	27	27	30.8%	2.96	4.80	0.0284	0.0783
Lung	18	24	51.2%	3.40	3.94	0.0471	0.0111

As depicted in Figure 2.2, the simple dichotomization of in vivo samples by hierarchical clustering is far from optimal. By the nature of hierarchical clustering, dichotomization can be achieved by splitting samples at the first node. In Figure 2.2 we have color coded the samples by their TSI predicted classification (red = serum activated, blue = serum independent, tan = unclassified) and we see that there is heterogeneity in the classification suggested by dichotomization at the first node of the dendrogram. This heterogeneity is apparent in the Kaplan Meier plots of Figure 2.4.

Notice that the prostate samples appear to be least heterogeneous, see Figure 2.2B, in that most of the serum activated samples are clustered on the left and most of the serum independent samples are clustered on the right with the unclassified samples interspersed among both branches. The Kaplan Meier plot in Figure 2.4A suggests that those samples which can be classified by their serum response have the best and worst recurrence free survival with the unclassified samples having intermediate recurrence free survival. The intermediate nature of the unclassified samples may be due to a third class of tumors with moderate serum response or it may be due to a blending of high risk and low risk samples that were not separated by the CSR signature.

The breast cancer samples appear to be have a more well defined subset of unclassified samples, see Figure 2.2C. The far right branch of the dendrogram (as split on the second node) contains a high percentage of unclassified samples. In the Figure 2.4B, the unclassified samples are associated with a recurrence free survival curve that is worse than for the serum activated samples. In the lung cancer data it is not clear that classification on any of the first three nodes of the dendrogram would result in homogeneous classification based on the CSR signature; see Figure 2.2D. However, using the TSI classification we are able to significantly split the samples into good and bad prognosis groups based on overall survival; see Figure 2.4C.

2.3.1 Permutation testing of H_0^{class}

Accepting the above significant separation of the Kaplan-Meier curves as validation of the CSR signature in vivo, we proceed to test the classification null hypothesis using 1000 random samples from the genes in common between the in vivo and in vitro samples, see Table 2.1. The size of the randomly drawn set of genes was determined by the correlation in the original CSR genes, such that the randomly drawn sets contained an equivalent number of effectively independent genes as the CSR set. The TSI score was recalculated on each of these 1000 random gene sets. It was then used to classify the in vivo samples and predict survival.

Figure 2.5 depicts the classification and prediction ability of the 1000 random sets for each of the three in vivo data sets. The CSR gene predictor is colored red in these plots. The vertical axis plots the predictive ability of the gene set as the chi-squared test statistic associated with univariate Cox regression on the classifier. If we look at the vertical margin we arrive at the permutation p-value as depicted by the marginal histogram. However, we have additional information about the utility of the CSR signature as a classifier. The horizontal axis provides the percentage of the samples that remained unclassified in each of the 1000 random sets. In each case, the classifier based on the CSR genes has a lower percentage of unclassified samples than any of the randomly drawn gene sets. Finally, note that for some of the randomly drawn gene sets, see Figures 2.5B and 2.5C, the samples were classified into only one group and thus the chi-squared test statistic could not be calculated. This occurred when the percentage of unclassified samples was high.

The three plots of Figure 2.5 carry a lot of interesting information regarding the utility of the CSR gene signature as a predictor of survival among the three tumor types. First, consider the horizontal axes of Figures 2.5A-C. It is intriguing that for all three tumor types the CSR signature has the lowest percentage of unclassified samples. Yet we see that percentage of classified samples is not the sole predictor of significant separation in the survival curves since there are randomly selected gene sets that have higher percentages of unclassified samples but also have higher test statistics.

Next, consider the empirical p-value for testing H_0^{class}. In the prostate samples, Figure 2.5A, the empirical p-value is 0.0040, whereas the p-value obtained from a simple training/testing strategy is very small (chi-squared test statistic = 20.89, $p < 0.0001$). In fact from the scale on the vertical axis we see that most of the random permutation samples were able to predict a significant separation in the survival of the prostate cancer patients. Thus had we relied only on the training/testing strategy we could not distinguish that the CSR signature is superior to 99.6% of the randomly selected signatures. The range of scale of the test statistics for the breast cancer and lung cancer samples are less dramatic. In fact the empirical p-value for the lung cancer dataset behaves we would normally expect, showing that a minimally significant test statistic in the training/testing setting ($p = 0.0352$) is indeed superior to test statistics generated under the classification null hypothesis

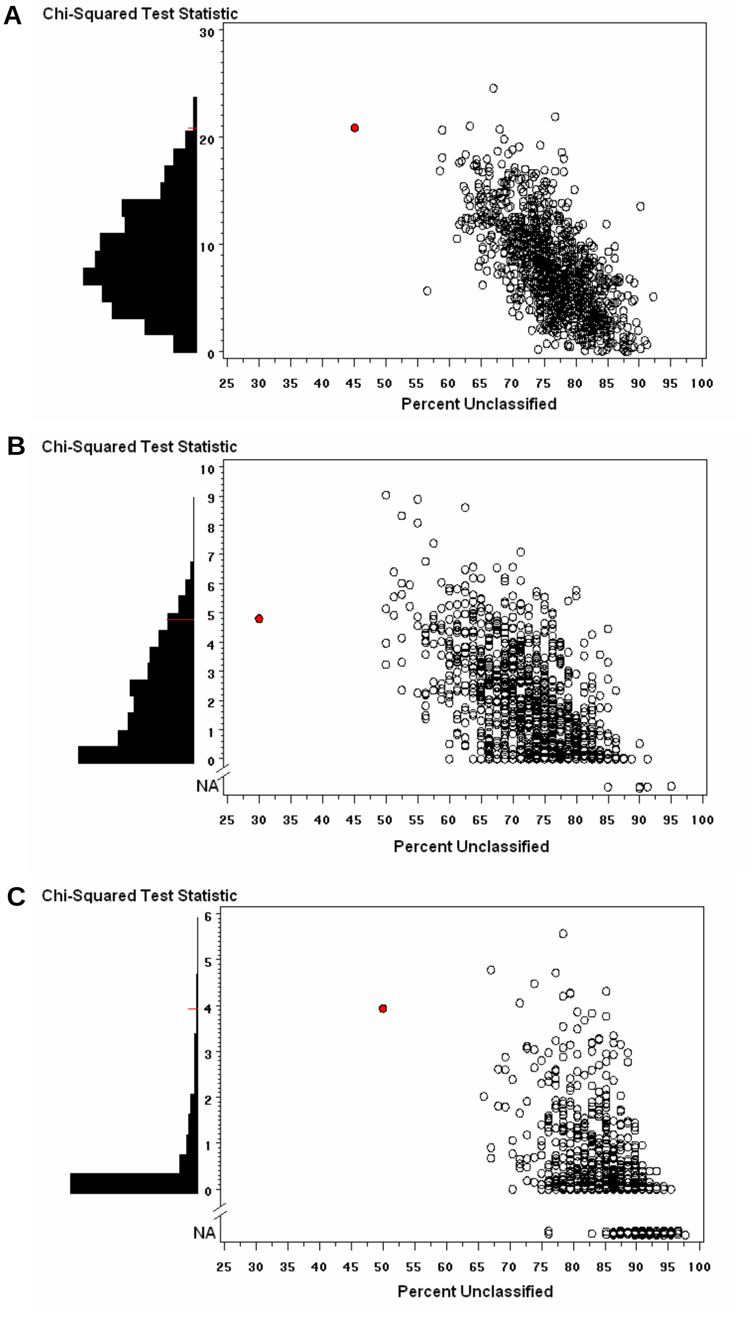


Figure 2.5: **The Cox regression test permutation results by percentage unclassified.** The Cox regression statistic is plotted against the percentage of unclassified samples for each of the 1000 permutations. The circle filled in red denotes the original classification. A test statistic listed as 'NA' indicates that there samples were classified in only one class and thus no test statistic could be calculated. The empirical p-value from the chi-squared statistics is depicted as a histogram in the left margin. (A) Glinsky et al. (2004) [23] prostate tumor samples. $P < 0.0001$; (B) van't Veer et al. (2002) [67] breast tumor samples, $p = 0.0783$; (C) Beer et al. (2002) [4] lung tumor samples, $p = 0.0111$.

(empirical $p = 0.0111$).

2.3.2 Differences in array configurations may reduce utility of the in vitro signature

One problem encountered in this analysis was the integration of gene expression data across microarray platforms. We attempted to compensate for this numerically by global standardization that centered the array-wise median values at zero. Furthermore, in the TSI algorithm genes were standardized to zero mean and unit standard deviation before being mapped into the reduced space. An additional complication, beyond numerical scaling, is that the differing array configurations between the in vitro and in vivo experiments mean that only those genes with Unigene ID numbers common to both data sets can be considered. This initially excludes ESTs from the in vitro signature as well as other features that do not have Unigene ID numbers. The signature is further reduced by focusing on only the common genes between data sets as determined by Unigene ID. We expect that there is correlation between the genes within the CSR signature and thus the loss of some genes from this signature will be tolerable.

The most dramatic decrease in CSR genes available for the analysis was for the Beer et al. (2002) [4] lung samples which measured only 32.6% of the 484 Unigene mapped CSR genes; see Table 2.1. It is possible that the high observed percentage of unclassified samples, 51.2%, is related to this diminished in vitro signature. Also, notice that in Figure 2.3C, that the mapping of the in vitro samples into the reduced space appears to have flipped about horizontal axis from what we saw for the other two in vivo data sets. Since the reduced space is determined by the in vitro data we expect that this inversion is a result of the diminished in vitro signature. However, this inversion does not affect the association of the classification with prognosis. As shown in Figure 2.4C the serum induced class has worse overall survival than the serum independent class, as expected. This change in the reduced space mapping highlights the necessity to calculate the TSI classifier independently for each in vivo dataset, or particularly for each different array platform and configuration used by the in vivo experiments.

2.4 Application of the TSI based classifier to integrate gene expression and metabolomic data

In the interest of data integration between transcriptomics and metabolomics we applied this classification method to these data. Specifically we used the Sreekumar-matched gene expression data to build the TSI classifier for benign versus cancer samples. We then used this classifier to predict the sample diagnosis of the Sreekumar et al. (2009) [60] metabolomics data.

Integrating these data sources for this purpose requires a means of mapping the genes and metabolites. We used enzymes as the common variable for genes and metabolites, obtaining this information from the Kyoto Encyclopedia of Genes and Genomes (KEGG, version 50) [35, 36, 34]. That is the genes whose products compose an enzyme were matched to the metabolites on which the enzyme acted. This resulted in between 1 and 205 gene probes from the Agilent Whole Human Genome microarray mapped per metabolite. There were 133 metabolites that mapped to at least one enzyme. We reduced the list of genes to a 1:m mapping by selecting the gene probe with the maximum variation across the samples for each metabolite. It is not a 1:1 mapping because multiple metabolites are associated with only a single gene. The duplication was as much as 6 for one gene but the mapping was 1:1 for a majority of the metabolites. Duplications in the data may cause stability issues in the principal components analysis but did not appear to cause problems in this analysis.

We were also concerned that the metabolomic data is measured by mass spectrometry, and thus will have a dynamic range different from data collected by gene expression microarray. As for the differing microarray platforms used in the above work, we standardized each gene and metabolite to have a mean expression of zero and a unit standard deviation.

In the above work we used the CSR gene list to build the classifier. Here we used the set of genes that were differential ($p \leq 0.05$) by a two-sided two-sample Welch's t-test between benign and local cancer samples (n=16 and 12, repectively). This set has 39 genes mapping to 43 metabolites. Interestingly when the set size is corrected for correlation the effective sample size is estimated at 38 genes.

The TSI classifier is built using the 39 differential genes and then applied to the metabolite data. We see in Figure 2.6 that the first two principal components do well separating the two diagnostic classes. For the classification of the metabolites we use the first three principal components. The choice to use three components was made *a priori* but *post hoc* analysis finds that this may be an optimal choice. Classification was made by the absolute TSI score, that is the sample is classified with the centroid with which it correlates best. Requiring a significant correlation did not result in any samples being classified.

The classification of the metabolomic samples resulted in 91.7% sensitivity and 81.3% sensitivity; see Table 2.3. The one undetected cancer sample has a Gleason grade of 3+4 so it is of moderate severity, however, this patient also contributed an adjacent benign sample. Of the three misdiagnosed benign samples, one was contributed by a patient who also contributed a moderate (4+3) cancer sample and a high grade (4+4) cancer sample. No information was available regarding the tumor grade of patients contributing the other two misclassified bengin samples.

A feature of the TSI analysis presented above is the ability to test the classification

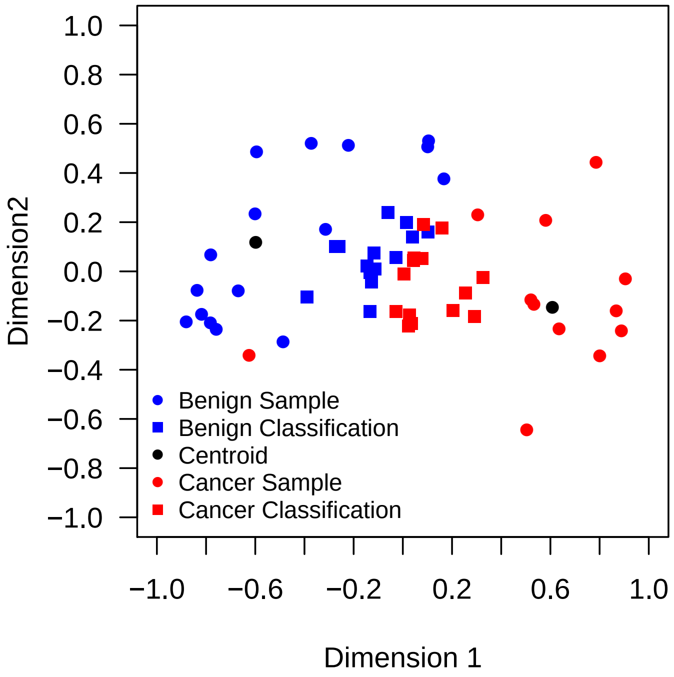


Figure 2.6: **Metabolite prediction by gene classifier** Samples are plotted by their mapping to the first two Eigenarrays either by their gene expression measures (circles) or metabolite intensities (squares).

Table 2.3: **Classification of metabolomic samples** The first three prinicpal components were used to generate a classifier from the differentially expressed genes. This classifier predicted the metabolite sample dianosis with 91.7% sensitivity and 81.3% sensitivity.

	Classification		
Diagnosis	Benign	Cancer	Total
Benign	13	3	16
Cancer	1	11	12
Total	14	14	28

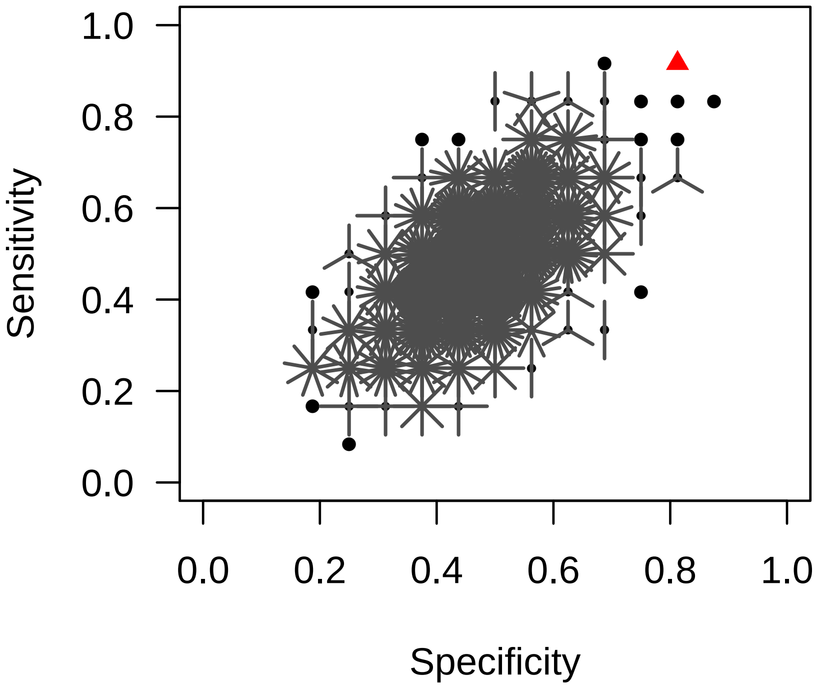


Figure 2.7: **Comparison of differential gene classifier to random sets** The sunflower plot adds one 'petal' for every observation and shows highest petal density around 50% specificity and 50% sensitivity as expected for these 1000 randomly drawn sets. The observed value is plotted with a red triangle.

null hypothesis. Here we are interested in the sensitivity and specificity obtained by the classifier. For this integrative analysis we can write H_0^{class}: There exists no set of genes derived from the in gene expression data that can predict diagnosis in the metabolite intensity data. To test this null hypothesis we generate 1000 random sets of 43 genes from the 133 gene set and compare their classification ability to that of our differential gene list. We did not adjust for the correlated data by drawing the effective sample size of 38 genes since there is likely to be correlation in any randomly drawn set given that we are selecting approximately one-third of the data with each draw.

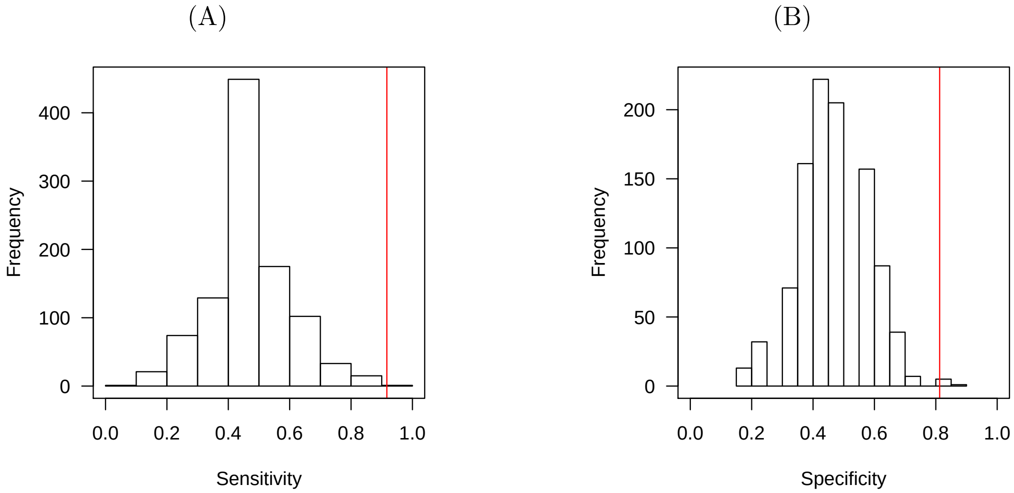


Figure 2.8: **Histograms of (A) sensitivity and (B) specificty** as achieved by random sets of genes. P-values can be computed by a count of the random sets that are exceed the observed value plotted in red. The observed sensitivity, 91.7%, has $p = 0.001$ and the observed specificity, 81.3%, has $p = 0.006$.

Figure 2.7 shows that a majority of the 1000 randomly drawn sets have sensitivity and specificity around 50% as expected. We can marginally assess how likely it is that a random set would meet or exceed the sensitivity and specificity seen by the original differential gene set. One set achieves the same sensitivity; a p-value of 0.001. There are 5 sets that achieve and one set that exceeds the observed specificty; a p-value of 0.006. No random sets achieve the same levels of both sensitivity and specificity. The marginal results can also be seen in the histograms in Figures 2.8 i and ii, respectively.

The low p-values recommend that we reject the null hypothesis H_0^{class}. This means that there is at least one set of genes that can predict the diagnosis of the metabolite samples. Caution should be exercised however since the gene expression and metabolite intensity data were measured from the same set of 40 samples. This is an ideal situation and classification may be more difficult for data from different samples or from different labs. However, these results show give evidence that gene expression data can be used to make predictions in metabolite data.

CHAPTER III

Pathway-directed weighted testing procedures for the integrative analysis of gene expression and metabolomic data

3.1 Introduction

Currently we are experiencing an explosion of high-throughput technology for assessing global snap-shots of the molecular behavior of cells. Global assays exist for measuring DNA sequence and copy number, mRNA transcript levels, protein presence and abundance, as well as metabolite abundance in biological samples of healthy and diseased tissues [43].

Transcriptomics is the high-throughput study of the transcriptome, the cellular complement of gene transcripts. Gene expression microarrays use complementary DNA fragments or oligonucleotide probes fixed on a slide to assess the mRNA abundance in a sample using the complementary binding of single-stranded nucleotide sequences. A gene may be represented by one or more probes targeting varying regions of the gene. Commercially available gene expression microarrays, such as Affymetrix (www.affymetrix.com) and Agilent Technologies (www.chem.agilent.com) can measure the full complement of known and estimated genes and gene elements with over 40,000 elements per array. Moving away from fixed array technology, Next Generation Sequencing Digital Gene Expression is able to measure gene expression at a much greater dynamic range than available by microarrays [63], and it detects all transcripts without the need for *a priori* probe development.

Metabolomics is the high-throughput study of the metabolome, the cellular complement of small molecules. In this work we consider all molecules under 10,000 KDa. This includes such classes as amino acids, fatty acids, simple carbohydrates, and exogenous drugs within the cell. Metabolomic data are generated by mass spectrometry (MS) preceded by either gas or liquid chromatography (GC or LC) [26, 69]. The initial chromatography step separates the molecules so that they can be identified by their mass spectrum. When this separation step is skipped the metabolic activity of the cell is measured as a metabolic fingerprint, but metabolites are not measured individually [18].

Metabolomic studies currently detect between one hundred and one thousand metabolites [21, 70, 60], compared to the tens of thousands of genes probed on a microarray [68]. Additionally, unlike a gene expression array, there is no pre-defined set of metabolites measured in each experimental run. This allows for metabolite discovery but adds another dimension of missing data in that it is not clear if the metabolite was not present or simply not detected. Similarly to the estimated sequence tags (ESTs) on a gene expression array not all metabolites detected will be identified [19].

In the classic dogma of biology DNA gives rise to mRNA transcripts as genes are expressed which direct the construction of proteins. From this view the metabolites are the functional elements upon which the proteins act. Proteins construct, degrade or alter metabolites in predictable patterns for energy transfer or other functions vital to the cell. These reactions can be classified into metabolic pathways and provide a means for connecting genes to metabolites. Though we know now that information transfer is not strictly passed from DNA to RNA to proteins this gives us an introductory view of the relationship between mRNA and metabolites.

Integration of gene expression and metabolomic data has been used in the study of model organisms for gene function discovery as well as sample class differentiation. Unsupervised classification methods are predominantly used and even recommended by Weckwerth and Morgenthal (2005) [71] who suggest that supervised analyses may be particularly biased by choices in data preprocessing. Hirai et al. (2004 [29], 2005 [28]) use principal components analysis and self organizing maps (SOM) to predict gene function by correlation with metabolite profiles in time series experiments on Arabidopsis. Unsupervised methods are also commonly used for discrimination of classes with results of concatenated gene expression and metabolomic datasets often performing better than either individually [66, 45, 59].

Metabolomic studies in human populations that include gene expression data are fewer [33, 71]. Most recently, Spicker et al. (2008) [59] looked separately at data reduction models for genes and metabolites and then jointly from a concatenated list, finding that the joint model is more interpretable. They discuss the risks of concatenating such distinct data sets and recommend either block scaling or hierarchical modelling where the results of the first model (e.g. principal components analysis) are used to construct a second model. Additionlly, Ferrara et al. (2008) [17] combined metabolomic and gene expression data with genomic markers to construct hypotheses regarding causal relationships between genes and metabolites.

Most of the current methods are based on correlation matrices [45, 59, 9]. As expected, correlation within a platform, such as between metabolites, is higher than correlations between platforms [17, 9]. Carrari et al. (2006) [9] states that correlation is low between gene and metabolite and which is supported by Urbanczyk-Wochniak et al. (2003) [66] who found only 2% of the pairs to have correlation estimates significantly different from zero. However, the magnitude and prevelance of the correlation varies between studies and may be dependent upon data pre-processing steps chosen. For instance, Camacho et al. (2004) [8] found that most metabolite-metabolite pairs had low correlation whereas Ferrara et al. (2008) [17] found a high percentage of metabolite pairs with correlation over 0.5.

Our motivation is to discover biomarkers in case-control studies. Thus, supervised methods are of primary interest. We are encouraged by the previously described unsupervised results that demonstrate better separation of classes when gene expression and metabolite data are considered jointly. Though, a simple approach to supervised integration is to analyze the data from each platform independently and then assess the concordance of the results by looking at unions and intersections of the elements highlighted in each experiment, we posit that it is more powerful to use a higher level of data in the integration. We propose that p-value weighting, the method of adjusting the p-value or threshold of a test according to some *a priori* importance measure, is a viable method for the integration of differential analysis. Under the proper conditions p-values can be weighted without an increase in type I error [20] and even minimally informative weights can provide power enhancement [54].

In this work we extend the use of p-value weights to a systems biology data integration. We find that gene expression data can be used to construct weights for per-metabolite tests providing a boost in power to detect differential metabolites. We begin with a brief description of the motivating data sets (Section 3.2). In Section 3.2.2 we discuss p-value weighting and its use in genomics. We follow with a description of the weighting methods that we consider (Section 3.3). Two simulations are employed to study these weights in Sections 3.4.1 and 3.4.2. In Section 3.5 we use the Sreekumar-matched gene expression data to construct weights for the Sreekumar et al. (2009)[60] metabolite data. We conclude in Section 3.6 with some discussion and recommendations.

3.2 Background

3.2.1 Gene expression and metabolic profiling data used

For this work, we use metabolomic data and gene expression data. Metabolomic data can be derived from mass spectra techniques or other means of high throughput measurement. Unlike metabolic fingerprinting, which uses the spectral intensities directly, metabolomics requires an identification step that matches the spectral intensities to distinct compounds, though not all compound will be known. Thus the data can be represented in an $M \times N_M$ matrix with M metabolites and N_M samples. Gene expression data is derived from microarrays or other high-througput assessment. The gene expression data can be represented in a $G \times N_G$ matrix with G genes and N_G samples. For motivation we consider the metabolic profiling data of Sreekumar et al. (2009 [60]; $N_M = 28$, $M = 518$) the gene expression data of the matched samples (GEO number GSE8511, unpublished, $N_{G_S} = 28$, $G_S > 40,000$), and the gene expression data of Varambally et al. (2005 [68]; GEO number GSE3325, $N_{G_V} = 9$, $G_V > 40,000$).

Each of these motivating examples compares prostate cancer tumor tissues samples to adjacent benign tissue samples. Tissues are collected from prostate glands extracted by prostatectomy. Benign adjacent tissue is removed from histopathologically determined non-tumor sections of the extracted prostate. In the Sreekumar metabolic data and matched gene expression data, seven men contributed both a tumor sample and a benign sample. One man contributed three samples; a high grade tumor sample (Gleason Major: 4, Gleason Minor: 4), a moderate grade tumor sample (Gleason Major: 4, Gleason Minor: 3), and a benign sample. As not all tumor samples have a matched benign sample, we chose to ignore this matching in our analysis and we treat the samples as independent between case and control.

The metabolomic data set is pre-processed as described in Sreekumar et al. (2009) [60] and two-sample testing per-metabolite is conducted using Wilcoxon rank-sum tests with an emprically calculated p-value to assess differences between tumor and benign tissues [60]. The Varambally dataset was preprocessed using LOESS models for standardization, as described in Varambally et al. (2005) [68]. The Sreekumar-matched gene expression was globally standardized by subtracting the median and dividing by the interquartile range of each sample. The gene expression datasets were assessed for differential expression using per-gene Welch t-tests. In the following we consider only the p-value per metabolite, $P^M = (P_1^M, \ldots, P_M^M)$, and we use either the test statistic, $T^G = (T_1^G, \ldots, T_G^G)$, or p-value, $P^G = (P_1^G, \ldots, P_G^G)$, per gene.

For integration, the gene expression and metabolomic datasets are mapped to the pathways defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG [34, 36, 35]. The mapping between gene and metabolite is not one-to-one, i.e. a single metabolite may be associated with several genes and likewise a single gene may be associated with several metabolites (see Appendix A for details). To build per-metabolite weights, we must summarize the gene information related to each metabolite. In this chapter we use gene-set enrichment tests to capture the information in the gene expression data regarding the differential characteristics of the genes in a pathway. Specifically, we use either the p-value or test statistic from the gene set enrichment tests to construct p-value weights for the per-metabolite tests.

3.2.2 Weighted multiple testing and applications to genomic analysis

P-value weighting was suggested by Holm (1979) [30] as a method for controlling error while retaining power in multiple testing situations. Consider a set of

m tests in which a positive constant weight, w_i, is applied to each p-value, P_i, for test $i = 1, \dots, m$, according to its perceived importance. Holm then showed that the sequentially rejective Bonferroni test (now referred to as Holm's test) could be generalized to use the new p-value, $P_i^* = P_i/w_i$, to assess significance. Specifically, if the ordered weighted p-values are written as $P_{(1)}^* \leq \dots \leq P_{(m)}^*$ then we reject $H_{(i)}^*$ when

$$P_{(j)}^* \leq \frac{\alpha}{\sum_{k=j}^{m} w_{(k)}^*}, j = 1, \dots, i \tag{3.1}$$

where $H_{(i)}^*$ and $w_{(i)}^*$ are the hypothesis and weight associated with weighted p-value $P_{(i)}^*$. The weighted Holm's test (wHT) is designed to control the family-wise error rate (FWER), i.e. the probability that at least one null hypothesis is falsely rejected, and does not require that the multiple tests be independent. Holm's only requirement for wHT is that $w_i \geq 0$.

More recently Genovese et al. (2006) [20] proposed p-value weighting with application to genomics studies. Instead of Holm's test, they use the Benjamini-Hochberg (BHT) step-down test Benjamini and Hochberg (1995) [5] as the basis of their method. With the BHT, $H_{(1)}, \dots, H_{(i)}$ are rejected for $\max_{i=1,\dots,m}\{P_{(i)} : P_{(i)} \leq i\alpha/m\}$ where $H_{(i)}$ is the hypothesis associated with ordered p-value $P_{(i)}$. The BHT controls the false discovery rate (FDR), i.e. the expected rate of incorrectly rejected null hypotheses among all rejected null hypotheses. For testing hundreds, or thousands, of hypotheses controlling the FDR is less conservative than controlling the FWER. Yet, when all null hypotheses are true the FDR and FWER are equivalent [6].

As with the work of Holm (1979)[30], Genovese et al. (2006) [20] consider the weight $w_i \geq 0$ for test i resulting in the weighted p-value, $P_i^* = P_i/w_i$. Again, there is no requirement on the independence of the tests. However, they allow that the set of weights $W = \{w_1, \ldots, w_m\}$ be random variables and they additionally require that $\bar{w} = m^{-1} \sum_{i=1}^{m} w_i = 1$ to maintain control of the FDR. Then the weighted-BHT (wBHT) rejects the null hypotheses, $H^*_{(1)}, \ldots, H^*_{(i)}$ for $\max_{i=1,\ldots,m}\{P^*_{(i)} : P^*_{(i)} \leq i\alpha/m\}$, where $H^*_{(i)}$ is the hypothesis associated with ordered weighted p-value $P^*_{(i)}$, while controlling the FDR at level α.

With only two requirements on the weights, Genovese et al. (2006) [20] note that p-value weighting is quite flexible. In fact they show that for binary weighting schemes, where all hypotheses are either up-weighted by w_1 or down-weighted by w_0, the power is improved for informative weight choices and is only minimally reduced for completely non-informative weights [53]. The concern for power loss stems from the required balancing of the weights across all tests such that, $\bar{w} = m^{-1} \sum_{i=1}^{m} w_i = 1$. Here all up-weighting must be balanced by equal down-weighting. If the up-weighting is sparse, the weight can be strong as the down-weight is spread over many tests. In contrast, broadly applied up-weights must be more moderate to reduce the effect of the down-weights. More recently, Roeder et al. (2009) [55] showed this to be true and found that the power can be greatly increased when sparse weights are well assigned, yet, they also show that the power loss is small for poorly assigned weights. For broad coverage of up-weighting they show that there is only small power loss for poorly specified weights and that in most cases the weighted tests have more power than the un-weighted tests.

Roeder et al. (2006) [53] applied the wBHT to large-scale genomic studies where thousands of tests are performed and controlling the error rates leads to a loss of power. Multiple testing adjustments are particularly problematic for the identification of subtle changes that are of most interest. In their work, Roeder used linkage studies to construct p-value weights to improve the power to identify disease variants in genome wide association studies (GWAS).

Consider the linkage test statistic z_i at locus i where $Z_i \sim N(\mu, 1)$ with $\mu = 0$ for unlinked loci and $\mu > 0$ for linked loci. Here the information from the linkage test is contained in the parameter μ. They define two functional relations between μ and the test statistic z_i to use for weight construction. First they consider the posterior odds that a locus is linked which is proportional to $v_i = exp(z_i\mu)$. They also consider the standard normal cumulative distribution function (CDF), $\Phi(\cdot)$, such that $v_i = \Phi(z_i - \mu)$. In both cases the values v_i must be standardized to form weights $w_i = v_i/\bar{v}$ that meet the requirement that $\bar{w} = m^{-1}\sum_{i=1}^{m} w_i = 1$. The strength of the weights is determined by the choice of $\tilde{\mu}$ to estimate μ.

By simulation, Roeder et al. (2006) [53] found that both continuous weighting schemes were near to one when no linkage signal was simulated. With stronger linkage signals the weights increased. The posterior odds weights increased more dramatically and had high variance resulting in spikes in the informative regions. The CDF weights showed less intensity and lower variance. They had broader peaks but that resulted in deeper down weights due to the average weight constraint. For highly informative data the posterior odds weighting is ideal since they provide sparse weights. The cumulative weights are preferred when there is less certainty in the prior data, however, the up-weighting must be moderate to avoid strong down-weighting. There are no restrictions on the weighting functions and others may be used [51]. However, the choice of weighting function should have a meaninful interpretation and Roeder et al. (2006) [53] wisely advise that they should be chosen prior to analysis of the data to be weighted.

In the context of metabolomic profiling we have hundreds of tests, not thousands, and we are concerned with improving the power of the per-metabolite tests regardless of multiple test correction. Low sample numbers and high inter- and intra-person variation due to diurnal rhythm, diet, and other environmental factors contribute to low power to detect differential metabolites. Yet, metabolites do not act in isolation within the cell. Here we use p-value weights to add information about the behavior of other molecular components such as gene transcripts in an effort to add power to nominate metabolites of interest. Through simulations described in Sections 3.4.1 and 3.4.2 we find that the power can be improved without raising the Type I error rates relative to the unweighted tests.

3.3 Proposed weight functions

Let us denote the metabolomic data by y_{lm}, for samples $l = 1, \ldots, N_M$ and metabolites $m = 1, \ldots, M$. Likewise denote the gene expression data by x_{jg} for samples $j = 1, \ldots, N_G$ and genes $g = 1, \ldots, G$. We integrate these two datasets using the KEGG pathway maps $k = 1, \ldots, K$. As the genes and metabolites do not have a one-to-one mapping we must summarize the information in the gene expression data to construct the weights. Here we use enrichment testing to capture the information about the differential gene expression data per pathway. That is, each gene is assessed for its ability to differentiate diagnostic classes, say by a two-sample t-test, as captured by the statistic T_g^G, $g = 1, \ldots, G$. An enrichment test assesses the level of differential ability within a pathway k.

As listed in Table 3.1 we consider four different enrichment test types. First we consider tests based on either binary or continuous differential expression results. Binary tests require that the per-gene tests are thresholded to categorize each gene as "differential" (e.g. $|T_g^G| \geq \tau$) or "non-differential" (e.g. $T_g^G < \tau$), for a given threshold τ. Continuous tests use the per-gene test statistic, T_g^G, in its continuous form. The benefit of the continuous tests is that does not rely on an arbitrarily defined threshold.

Second, we consider the null hypothesis style of the test, namely competitive versus self-contained [24]. A competitive test compares the genes in the set of interest, say set ξ, to all other genes, say set ξ^c. The null hypothesis is that ξ contains the same proportion of differential genes, say π, as ξ^c. A self-contained test considers only the genes within the set of interest, ξ, and ignores the genes in ξ^c. The hypothesis is that there are no more differential genes than expected where the expected value is determined *a priori*, i.e. 5% based on an $\alpha = 0.05$ error rate, or by sample permutation. Competitive tests allow selection of a "best" set, that is one that is enriched above the rest, but they are limited such that a given set ξ' with π percent differential genes will receive a different test statistic depending upon the proportion of differential genes, say π^c, in the set $(\xi')^c$. Self-contained tests will always give the same result for the same set of data since the test of ξ does not depend on ξ^c. However, if differential genes are uniformly distributed across all pathways, that is $\pi_k = \pi$ for all $k = 1, \ldots, K$, then all pathways will be called enriched by a self-contained test if π is great enough.

Each of the four enrichment tests listed in Table 3.1 will be described in Section 3.3.1. The enrichment tests capture the gene expression information in each pathway k, $k = 1, \ldots, K$, by a test statistic S_k or its corresponding p-value P_k^E. To utilize this information as a weight we first transform it using one of five weighting functions described in Section 3.3.2.

Table 3.1: **Four gene-set enrichment tests are considered.** The competitive hypergeometric and weighted Kolmogorov-Smirnov tests compare the level of differentiation in the set of interest to all other sets. The self-contained binomial and sum of squared statistics tests are global tests of differentiation within the set. Thresholding of per-gene tests prior to enrichment testing is required for the hypergeometric and binomial tests.

	Competitive	Self-contained
Binary	Hypergeometric	Binomial
Continuous	Weighted Kolmogorov-Smirnov	Sum of Squared Statistics

3.3.1 Enrichment test methods

Directional Hypergeometric Test

For a given pathway ξ_k, each gene tested (T_g^G, $g = 1, \ldots, G$) is categorized by its inclusion in the pathway ($g \in \xi_k$) and whether it is differentially expressed (e.g. $|T_g^G| \geq \tau$, for a given τ). This categorization is depicted in Table 3.2. In the following we use two-sample t-tests for assessing per-gene differential expression. We then threshold the test statistics at $|T_g^G| \geq \tau_\alpha$ where τ_α is chosen according to a t-distribution with $N_G - 2$ degrees of freedom and $\alpha = 0.05$. We use a directional test of enrichment, $Pr(X \geq x | G_k = g_k, G_0 = g_0, D = d)$, assuming $X \sim Hyper(G_k, G_0, D)$, where X, G_k and G_0 are defined as in Table 3.2. Assuming a hypergeometric distribution for this 2x2 table, we can get an exact p-value without permutation testing and we use $S_k = X_k$ as the statistic of interest. Because the hypergeometric test uses the genes of ξ_k^c to define the null proportion of differential genes, this is a competitive test. This test differs from the Fisher's Exact test in that it does not consider the depletion of differential genes in a pathway ξ_k as an interesting case.

Table 3.2: **The classification of genes that underlies the gene set enrichment testing.** Competitive tests consider the entire table whereas self-contained tests focus on the first row.

	Differential	Not Differential	Total
In Pathway ξ_k	X	$G_k - X$	G_k
In Pathway Complement ξ_k^c	$D - X$	$G_0 - (D - X)$	G_0
Total	D	$G - D$	G

Binomial Test of Proportions

Also called Tukey's Higher Criticism [24], the binomial test of proportions is self-contained such that only genes contained in pathway ξ_k are considered for the test of that pathway (i.e. the top row of Table 3.2). Specifically we test the null hypothesis that $X \sim Bin(G_k, \alpha)$ and we reject if $X/G_k \geq x$ where α is set *a priori* and X and G_k are defined as in Table 3.2. In the following we set $\alpha = 0.05$ as this is the assumed error rate for each of the per-gene differential tests, $\underline{T}^G = (T_1^G, \ldots, T_G^G)$. Significance is determined using permutation sampling of the N_G sample labels to construct the null distribution or, in the simulation, draws from the null distribution of test statistics. The test statistic of interest is $S_k = X_k/G_k$ for each pathway k.

Weighted Kolmogorov-Smirnov Test

A weighted Kolmogorov-Smirnov (K-S) test is used to compare the test statistics of those genes in the pathway k, i.e. $T_g^G : g \in \xi_k$, against the statistics of those not in the pathway and thus is a competitive test. The K-S test compares the two groups of test statistics in a single ranked list, testing if they arise from the same distribution by assessing the spread of the two sets throughout the ranked list. Specifically, begin by ranking the vector of t-statistics $\underline{T}^G$ as $(T_{(1)}^G, T_{(2)}^G, \ldots, T_{(G)}^G)$. Construct a corresponding pathway inclusion indicator vector $\underline{\gamma} = (\gamma_{(1)}, \gamma_{(2)}, \ldots, \gamma_{(G)})$ where $\gamma_{(g)} = 1$ if $g \in \xi_k$ and 0 otherwise. The statistic S_k is then maximum deviation of the empirical distributions

$$S_k = max_h |P(\xi_k, h) - P(\xi_k^c, h)| \tag{3.2}$$

where

$$P(\xi_k, h) = \sum_{g \leq h} \frac{\nu_g \gamma_{(g)}}{\sum_{g=1}^{G} \nu_g \gamma_{(g)}} \text{ and } P(\xi_k^c, h) = \sum_{g \leq h} \frac{1 - \gamma_{(g)}}{G - G_k} \tag{3.3}$$

Here $\nu_g \in [0,1]$ is the weight for gene g and an unweighted K-S test would have $\nu_g = 1$ for all $g = 1, \ldots, G$. Additionally, G_k is the number of genes in ξ_k; see Table 3.2.

The weighted K-S test as proposed by Subramanian et al. (2005) [61] uses $\nu_g = |corr_j(x_{jg}, \psi_j)|$. That is, each gene g is weighted by the correlation across samples, j between the expression value x_{jg} and the case status ψ_j, where $\psi_j = 1$ for cases and 0 for controls. In this way tests that cluster in the tails of the ranked list are given higher weight. In our simulations we define ν_k based on a function of the simulated test statistic $\nu_g = |Z_g|/(1 + |Z_g|)$. We choose a Z-score based weight for conveience in our simulation but the relationship between the Pearson's correlation coefficient in the Subramanian ν_k and Z score is monotonic so only the magnitude of the test statistic will be affected. Significance is determined using permutation sampling of the sample sample labels ψ_j to construct the null distribution or, in the simulation, draws from the null distribution of test statistics.

Sum of Squared Test Statistics

The test statistic from the the sum of squared test statistics method is simply the sum of the squared per-gene test statistics in the set ξ_k [1]. That is

$$S_k = \sum_{g \in \xi_k} T_g^G \tag{3.4}$$

Significance is determined using permutation sampling of the N_G sample labels to construct the null distribution or, in the simulation, draws from the null distribution of test statistics. This is a self-contained test since S_k and its null distribtution consider only the genes in ξ_k.

3.3.2 Weight functions

The above enrichment tests provide a test statistic, S_k, and corresponding p-value, P_k^E, that summarize the level of differential expression in each pathway $k = 1, \ldots, N_k$. To utilize these data as p-value weights we must transform them to ensure that they are positive and that they increase with increasing levels of differential expression, that is they must be positively correlated with increasing importance. We may also want to adjust the scale or distribution of the enrichment information for better performance under the $\bar{w} = m^{-1} \sum_{i=1}^{m} w_i = 1$ restriction.

We consider five functions of the enrichment test information, ω_k. These are (A) $\omega_k = -\log_{10}(P_k^E)$, (B) $\omega_k = |S_k|$, (C) $\omega_k = |\tilde{S}_k|$, (D) $\omega_k = \Phi(\tilde{S}_k - \tilde{\mu})$, and (E) $\omega_k = exp(\tilde{S}_k \tilde{\mu})$ where S_k is the test statistic and P_k^E is the p-value from the enrichment test of gene expression values for pathway k. $\tilde{S}_k = (S_k - E(S_0))/\sqrt{Var(S_0)}$ is a standardized test statistic using the null distribution of the test statistic to determine the mean, $E(S_0)$, and variance, $Var(S_0)$, of S_0. Here $\Phi(\cdot)$ is the cumulative distribution function (CDF) for the standard normal distribution. We set the parameter $\tilde{\mu}$ to 2 according to Genovese et al. (2006) [20]. Higher $\tilde{\mu}$ results in more conservative weights for the CDF function (D) by shifting the distribution of S_k. Higher $\tilde{\mu}$ results in stronger weights for the exponential function (E).

We chose to explore the weight functions suggested by Roeder et al. (2006) [53], termed weights D and E here, as they present both bounded (D) and unbounded (E) options for weight construction. We expect that the bounded weights of the CDF function (D) will out-perform the strong peaks of the exponential function (E) in this application to metabolomic data. Specifically, applying the same weight to all members of a pathway is contrary to the sparse nature of the exponential function weighting and strong downweighting is likely to arise. We also considered some more simplistic functions based directly on the enrichment p-value and statistic (weights A, B, and C).

To apply these pathway level weight to the per-metabolite p-values, $P^M_m = 1, \ldots,$ P^M_M, we first associate the component ω_k with each metabolite in the pathway, say $v_i = \omega_k$ for metabolite i in pathway ξ_k. The weights must then be standardized such that $w_i = v_i/\bar{v}$ where $\bar{v} = \sum_{i=1}^{m} v_i$ to arrive at w_i. Finally, we define $P^{M*}_i =$ P^M_i/w_i and P^{M*}_i can be assessed for significance in the usual way. As many of the metabolites in the KEGG pathways are associated with more than one pathway, additional summarization of the pathway weights ω_k must be done such that $v_i =$ $f(\omega_k I(i \in k))$ where $I(i \in \xi_k)$ is an indicator function for metabolite i in pathway k and $f(\cdot)$ is a summary function. In Section 3.4.2 we explore median and seventy-fifth percentile summaries.

3.4 Numerical examples

The simulations were programmed using the R statistical software v 2.7 and greater. SAS v. 9.1 was used for preparation of the real datasets. The gene and metabolite information for each human pathway map in KEGG was acquired from KEGG version 50 (April 2009) using perl scripts and the KEGG API [37]. The Varambally et al. (2005) [68] gene expression measures are from an Affymetrix HU133Av2 genechip and the gene symbols were obtained from GEO (Gene Expression Omnibus, `www.ncbi.nlm.nih.gov/geo`, GSE3325, August 2009) using the GPL570 platform information file. The Sreekumar-matched gene expression measures are from an Agilent Whole Human Genome Oligo Array (data: GSE8511 (private), platform: GPL1708, March 2008). We use these gene expression data sets in combination with the Sreekumar et al. (2009) [60] metabolomic data to construct two simulation situations. We begin with a simplistic model with disjoint pathways such that every metabolite contributes to only one pathway. Following is a more complex dataset with pathways modelled from the KEGG pathway mapping of the real data. The simplistic model allows us to explore properties of the different weight functions in an easily interpretable setting. The complex model provides a more realistic scenario with which to test our methods where the truth is still known.

3.4.1 Simulation I: Disjoint Pathways

Simulation Model I

Z-scores are simulated from a standard multivariate normal distribution to represent the per-gene test statistics of differential expression and per-metabolite test statistics of differential intensity. A constant correlation between like elements, i.e. gene-gene (ρ_{GG}) and metabolite-metabolite (ρ_{MM}), and a constant but lesser correlation between gene and metabolite (ρ_{GM}) within a pathway are assumed. For simplicity we assume that pathways are disjoint, that is no element appears in multiple pathways and there is no correlation between elements in different pathways. The case of non-disjoint pathways will be considered in Simulation II.

We model each pathway to have N_k^G genes and N_k^M metabolites. We draw a vector of z-scores $(\underline{Z}^G, \underline{Z}^M)$, where $\underline{Z}^G = (z_1^G, \ldots, z_{N_k^G}^G)$ and $\underline{Z}^M = (z_1^M, \ldots, z_{N_k^M}^M)$, from

$$(\underline{Z}^G, \underline{Z}^M) \sim MVN((\underline{\beta}, \underline{\phi}), \underline{\underline{\Sigma}}).$$

The variance covariance matrix is defined per pathway as

$$\underline{\underline{\Sigma}} = \begin{bmatrix} 1 & \rho_{GG} & \cdots & \rho_{GM} & \rho_{GM} \\ \rho_{GG} & 1 & & & \rho_{GM} \\ \vdots & & \ddots & & \vdots \\ \rho_{GM} & & & 1 & \rho_{MM} \\ \rho_{GM} & \rho_{GM} & \cdots & \rho_{MM} & 1 \end{bmatrix}_{(N_k^G+N_k^M)\times(N_k^G+N_k^M)}$$

where $\rho_{GM} < \min(\rho_{GG}, \rho_{MM})$. Under the null model $(\underline{\beta}, \underline{\phi})$ is a vector of zeros.

Under the alternative model, z-scores are simulated from a multivarite normal distribution with the same variance-covariance matrix, $(\underline{\underline{\Sigma}})$, of the null model but with shifted means ($\beta_1 = \cdots = \beta_{N_k^G} = \beta > 0$ and $\phi_1 = \ldots = \phi_{N_k^M} = \phi > 0$). Genes and metabolites are drawn from this alternative model according to a Bernoulli($\pi_k^{\cdot}$) distribution thereby assigning some elements to be truly differential. The probability of differential elements can differ for genes, $\pi_k^G \in [0,1]$, and metabolites, $\pi_k^M \in [0,1]$. We also allow π_k^G and π_k^M to differ by pathway (k) thereby defining some pathways to be enriched. We retain the simulated state of differential intensity for each metabolite, $i \in (i, \ldots, M)$ in the vector $\underline{H}$ where $H_i = 0$ for the null case and $H_i = 1$ for the differential case. P-values are calculated from the simulated z-scores using the standard normal distribution, i.e. $p = 2Pr(|Z| \geq z_{\alpha/2})$.

We consider a scenario with 50 pathways and allow the following parameters to vary:

- Alternative means (β, ϕ): (1.5, 2), (1.5, 3), (2,3)
- Pathway size (N_k^M, N_k^G): (3, 20), (5, 40)
- Percentage of enriched pathways: 10%, 20%
- Correlation between like elements (ρ_{MM}, ρ_{GG}): 0.2, 0.4, 0.6

- Correlation between gene and metabolite (ρ_{GM}): 0.0, 0.10, 0.15, 0.25, 0.50 where $\rho_{GM} < min(\rho_{GG}, \rho_{MM})$

To attain the desired level of enrichment, we set $(\pi_k^M, \pi_k^G) = (0.75, 0.50)$ for enriched pathways and $(\pi_k^M, \pi_k^G) = (0, 0)$ for non-enriched pathways when $(N_k^M, N_k^G) = (3, 20)$. For the larger pathways, $(N_k^M, N_k^G) = (5, 40)$, we set $(\pi_k^M, \pi_k^G) = (0.50, 0.25)$ for enriched pathways and $(\pi_k^M, \pi_k^G) = (0, 0)$ otherwise.

Each of the four enrichment tests described in Section 3.3.1 was applied to the gene expression z-scores for each of the 50 pathways. For tests requiring it, the null distribution was simulated by the generation of 1000 null vectors of z-scores ($\pi_k^G = \pi_k^M = 0.05$ for all k). The enrichment test statistic (S_k) and p-value (P_k^E) for each pathway ($k = 1, \ldots, 50$) was retained. Each of the five weight functions of Section 3.3.2 was then applied: (A) $\omega_k = -\log_{10}(P_k^E)$, (B) $\omega_k = |S_k|$, (C) $\omega_k = |\tilde{S_k}|$, (D) $\omega_k = \Phi(\tilde{S_k} - \tilde{\mu})$, and (E) $\omega_k = exp(\tilde{S_k}\tilde{\mu})$. The standardized test statistic, $\tilde{S_k} = (S_k - E(S_0))/\sqrt{Var(S_0)}$, uses the 1000 null vectors to determine the null mean, $E(S_0)$, and null standard deviation, $Var(S_0)$. Estimates of the mean and variance were determined from the hypergeometric distribution for the directional hypergeometric test. Again, $\Phi(\cdot)$ is the cumulative distribution function (CDF) for the standard normal distribution and we set $\tilde{\mu} = 2$ unless otherwise noted.

With four enrichment scores and five weight functions we have twenty weights constructed for each pathway, ω_{kj} where $j = 1, \ldots, 20$. The pathway level weights can be applied to the simulated metabolite p-values for each pathway such that $v_{ij} = \omega_{kj}$ for metabolite i in pathway k and weight option j. The weights are then standardized, $w_{ij} = v_{ij}/\bar{v}_{\cdot j}$ where $\bar{v}_{\cdot j} = \sum_{i=1}^{m} v_{ij}$. The per-metabolite p-values are determined from the z-score vector $\underline{Z}^M$ by comparing the z-scores to a standard normal distribution, i.e. $P_i^M = 2Pr(|Z_i^M| \geq z_{\alpha/2})$. The weighted per-metabolite p-value, P_{ij}^{M*}, is calculated by $P_{ij}^{M*} = P_i^M / w_{ij}$. This results twenty weighted p-values for each metabolite, i.e. an $N_M \times 20$ matrix.

To assess the Type I error rate for each method, with respect to metabolites, we simulated the situation of completely null data by generating Z-scores under a model where $\pi_k^G = \pi_k^M = 0$ for all k. We also simulated a second null setting in which we assume that there are differentially expressed elements but that they are not associated with the pathways. Here we set $\pi_k^G > 0$ and $\pi_k^M > 0$ to be constant non-zero rates for all pathways, $k \in (1, \ldots, K)$, to generate differential elements uniformly across all pathways. The second null model helps us to determine error rates and to assess any power loss from the marginal weighting of the null pathways. The power, or the probability of correctly identifying a differential result, is assessed using the true state of metabolite differential intensity, H_i, as simulated by the Bernoulli(π_k^M) draws. We use receiver operating characteristic (ROC) curves, varying the significance threshold for P_i^{M*}, and the associated area under the curve (AUC) to compare the properties of the different methods.

Results for Simulation Model I

Graphical representation of the results are presented in a 2 × 2 grid mimicking Table 3.1 with each enrichment test occupying a quadrant of the figure. Each of the five weight functions is labelled A–E and color coded . This coloring is consistent throughout the paper where the unweighted (raw) p-values are black, and the five weight functions are (A), $\omega_k = -log_{10}(p)$, green; (B), $\omega_k = |S|$, blue; (C), $\omega_k = |\tilde{S}|$, purple; (D), $\omega_k = \Phi(\tilde{S} - \tilde{\mu})$, orange; and (E), $\omega_k = exp(\tilde{S}\tilde{\mu})$, red. Average ROC curves, across the simulation runs, are used to compare the sensitivity and specificity of correctly identifying differential metabolites. Boxplots are used to demonstrate differences in error rates under the null models.

The null model in which no genes and no metabolites are selected to be differential ($\pi_k^G = \pi_k^M = 0$ for all k) was simulated for varying correlations. Figure 3.1 shows a representative plot of the error rates for 1000 replications of the simulation where $N_k^G = 20$, $N_k^M = 3$, $\rho_{GG} = \rho_{MM} = 0.20$ and the correlation $\rho_{GM} \in (0, 0.1, 0.15)$. We see that the Type I error is near the nominal level when there is no correlation between genes and metabolites. In some cases the error rate may increase slightly as the correlation increases. However, the boxplots each overlap the nominal 0.05 error rate (black horizontal line) except for the exponential weight function (E, red) which is conservative. Additionally, we are simulating uniform correlation within pathways. In reality we find from assessment of the Sreekumar metabolite and Sreekumar-matched gene expression data that pairwise correlation of genes and metabolites across matched samples are highly correlated for only about 8% of the pairs by pathway. This result is supported by Urbanczyk-Wochniak et al. (2003) [66] who find only 2% of the total pairs to be significantly correlated. Thus this Type I error estimate is likely conservative and actual error may be lower.

Under the second null model we assume that there are differential metabolites and genes in the dataset but that they are uniformly distributed across the dataset, that is simulated without pathway enrichment. Figure 3.2 shows the Type I error rates for 1000 replicates of the simulation with $\pi_k^G = \pi_k^M = 0.1$ for all k, $N_k^G = 20$, $N_k^M = 3$, $\rho_{GG} = \rho_{MM} = 0.20$ and the correlation $\rho_{GM} \in (0, 0.1, 0.15)$. Figure 3.3 shows the Type I error for the same settings except that $\rho_{GG} = \rho_{MM} = 0.60$ and the correlation is explored up to $\rho_{GM} = 0.50$. We see more Type I inflation in these high correlation cases. We do not expect such high correlations to exist in our data but high correlations have been observed by others [9].

As expected, there is some power lost when there is no enrichment of the pathways,

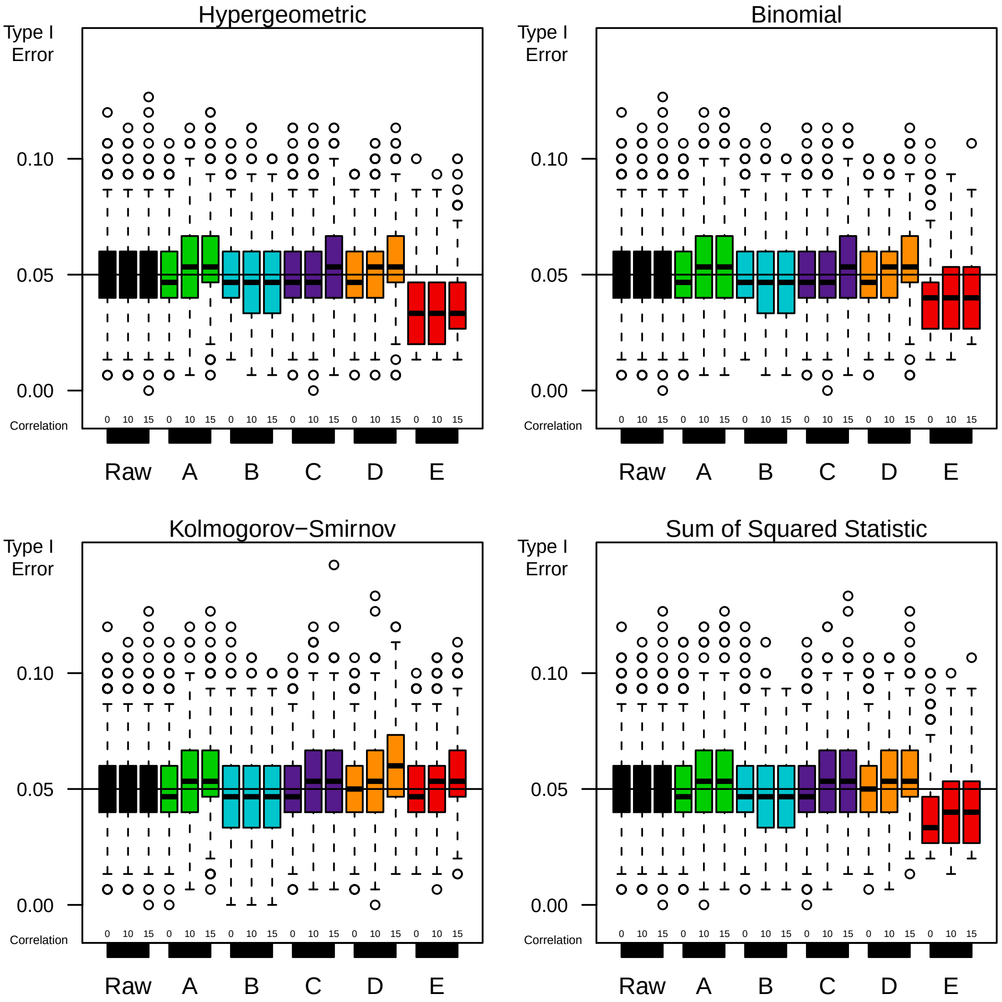


Figure 3.1: **Type 1 error** for per-metabolite tests using a significance threshold of $\alpha = 0.05$ without multiple testing adjustments. 1000 datasets were simulated assuming within pathway correlation of 0.2 for each metabolites and genes. Unweighted (Raw) p-values and the five weight functions (A, $-log_{10}(p)$; B, $|S|$; C, $|\tilde{S}|$; D, $\Phi(\tilde{S}-\tilde{\mu})$; E, $exp(\tilde{S}\tilde{\mu})$) are depicted with increasing between element correlation, $\rho_{GM} \in (0, 0.1, 0.15)$.

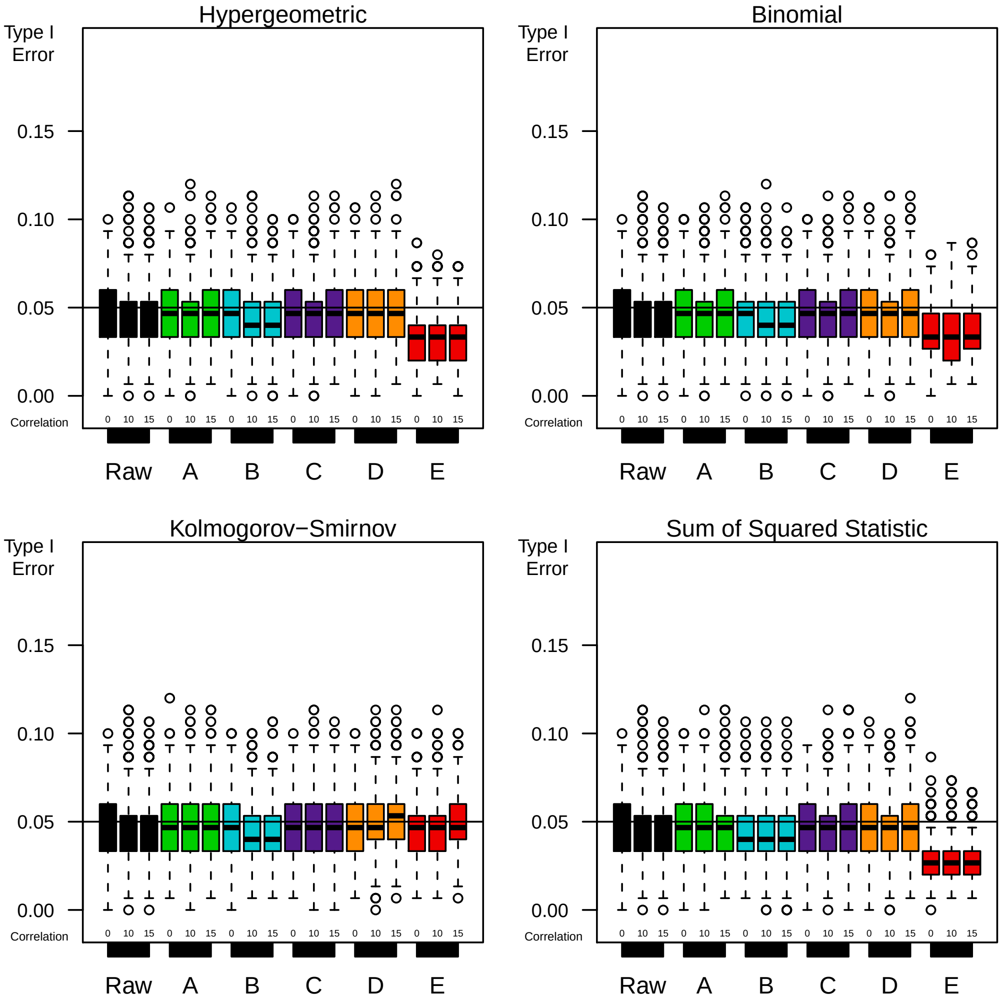


Figure 3.2: **Type 1 error, under uniformly distributed differential elements**, for per-metabolite tests using a significance threshold of $\alpha = 0.05$ without multiple testing adjustments. 1000 datasets were simulated assuming within pathway correlation of 0.2 for each metabolites and genes and $\pi_k^G = \pi_k^M = 0.1$ for all k. Unweighted (Raw) p-values and the five weight functions (A, $-log_{10}(p)$; B, $|S|$; C, $|\tilde{S}|$; D, $\Phi(\tilde{S} - \tilde{\mu})$; E, $e^{(\tilde{S}\tilde{\mu})}$) are depicted with increasing between element correlation, $\rho_{GM} \in (0, 0.1, 0.15)$.

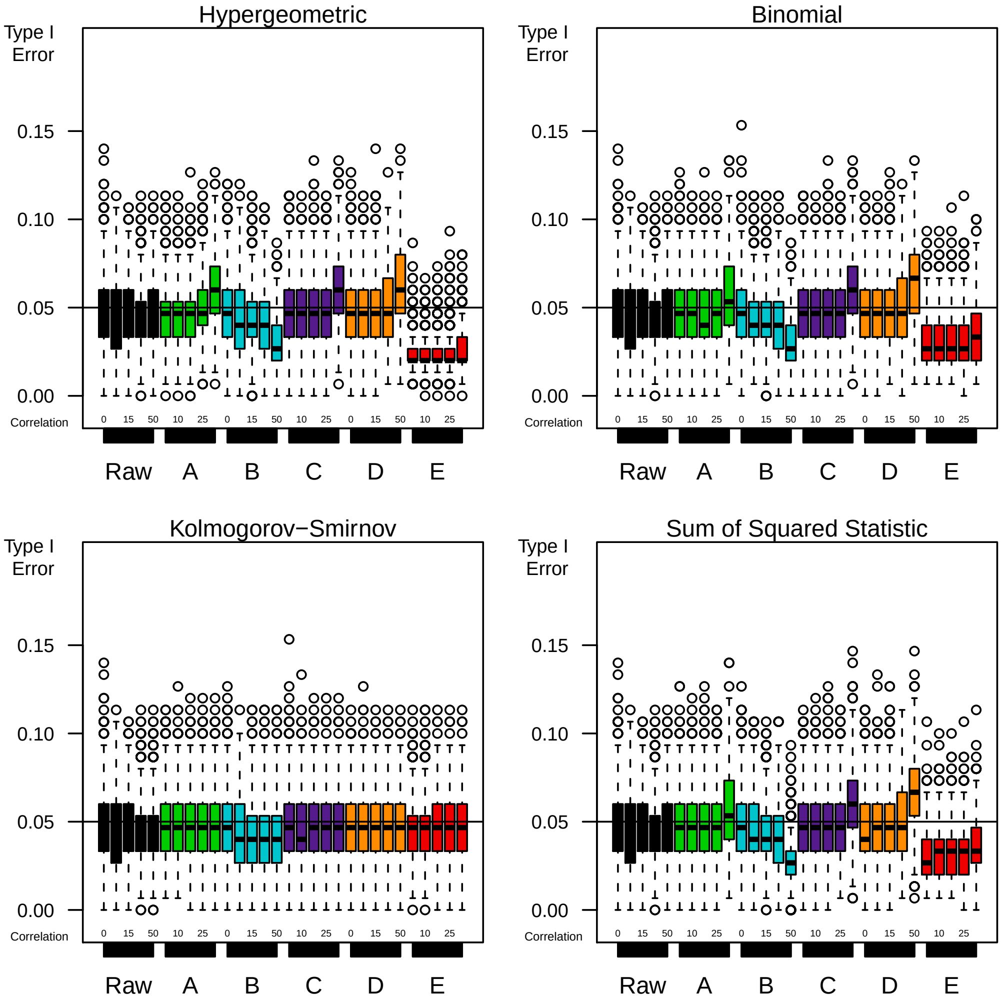
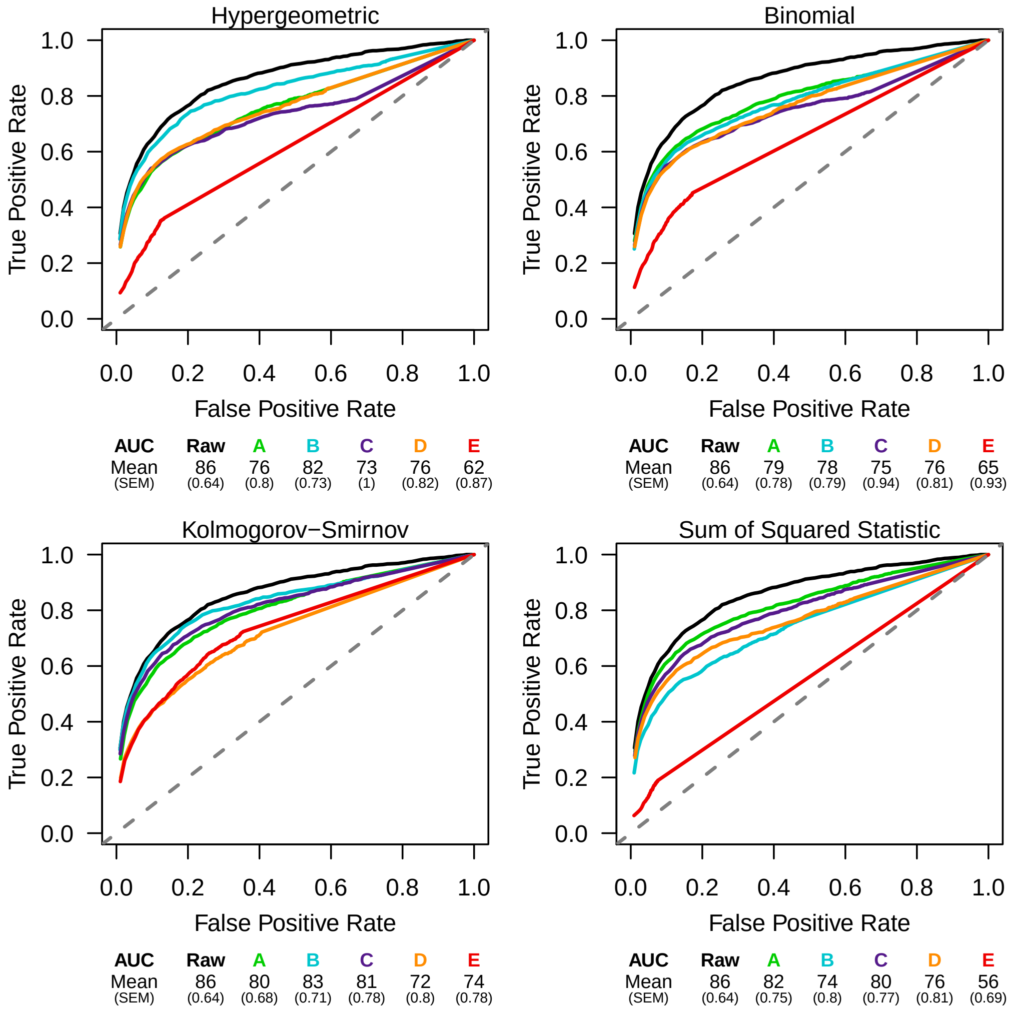


Figure 3.3: **Type 1 error, under uniformly distributed differential elements**, for per-metabolite tests using a significance threshold of $\alpha = 0.05$ without multiple testing adjustments. 1000 datasets were simulated assuming within pathway correlation of 0.6 for each metabolites and genes and $\pi_k^G = \pi_k^M = 0.1$ for all k. Unweighted (Raw) p-values and the five weight functions (A, $-log_{10}(p)$; B, $|S|$; C, $|\tilde{S}|$; D, $\Phi(\tilde{S}-\tilde{\mu})$; E, $e^{(\tilde{S}\tilde{\mu})}$) are depicted with increasing between element correlation, $\rho_{GM} \in (0, 0.1, 0.15, 0.25, 0.5)$.

i.e. the gene expression data is not informative by pathway. On average the loss is between 5–15 points on the AUC scale, except for the exponential weight function where the conservative error rates seen above are reflected in the poor power. Figure 3.4 shows the ROC curves and AUC levels for each method under this null model with $\rho_{GM} = 0.1$. The loss is similar for other correlation values and for $\pi_k^G = \pi_k^M = 0.05$ (data not shown).



Hypergeometric

AUC	Raw	A	B	C	D	E
Mean (SEM)	86 (0.64)	76 (0.8)	82 (0.73)	73 (1)	76 (0.82)	62 (0.87)

Binomial

AUC	Raw	A	B	C	D	E
Mean (SEM)	86 (0.64)	79 (0.78)	78 (0.79)	75 (0.94)	76 (0.81)	65 (0.93)

Kolmogorov−Smirnov

AUC	Raw	A	B	C	D	E
Mean (SEM)	86 (0.64)	80 (0.68)	83 (0.71)	81 (0.78)	72 (0.8)	74 (0.78)

Sum of Squared Statistic

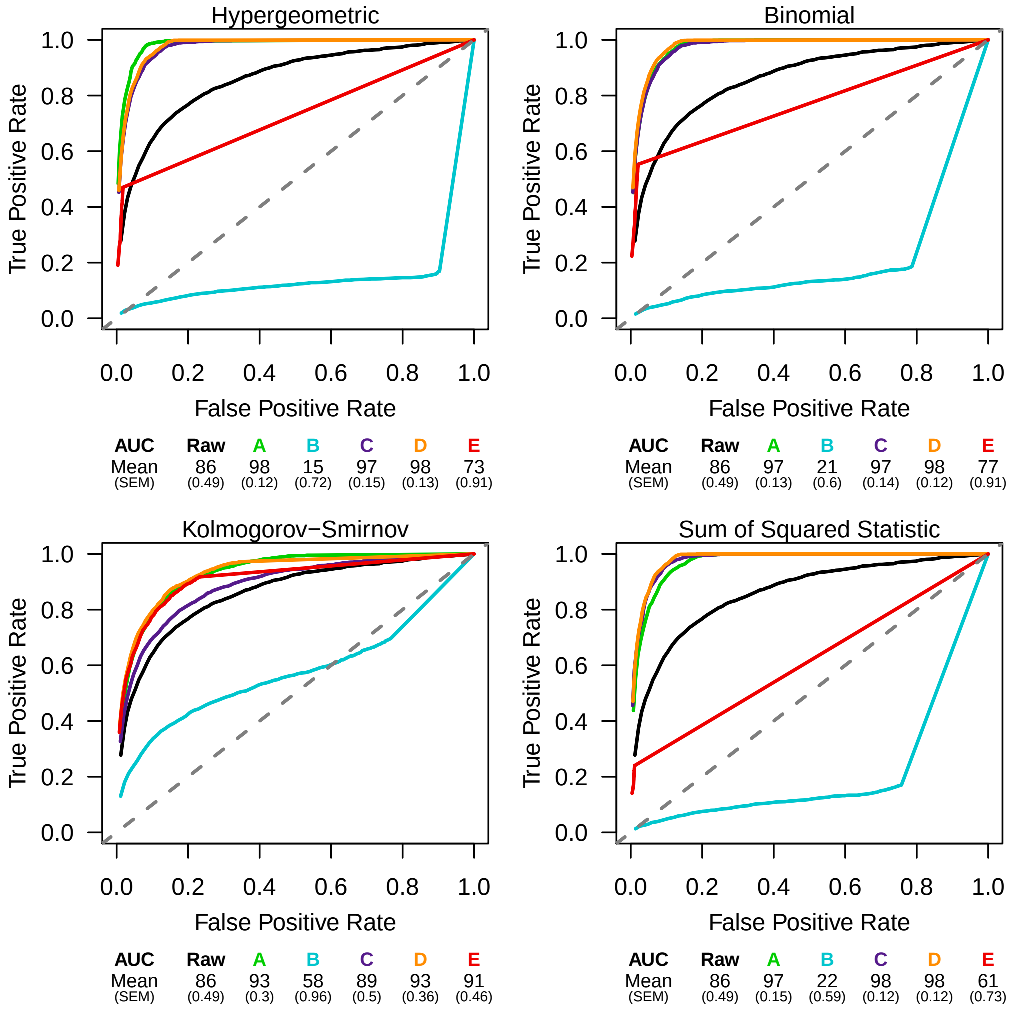
AUC	Raw	A	B	C	D	E
Mean (SEM)	86 (0.64)	82 (0.75)	74 (0.8)	80 (0.77)	76 (0.81)	56 (0.69)

Figure 3.4: **Power loss under uniformly distributed differential elements** is depicted by the average ROC curve for 1000 simulated datasets with $\pi_k^G = \pi_k^M = 0.1$ for all k, $\rho_{GG} = \rho_{MM} = 0.2$, and $\rho_{GM} = 0.1$.

When there is enrichment in a subset of the pathways three of the five weight functions show robust increases in power over the unweighted case (raw, black); the p-value weight (A, green), the standardized test statistic (C, purple), and the CDF transformation (D, orange). Figure 3.5 shows the average receiver operating characteristic (ROC) curves from 100 simulated datasets where the alternative means are $(\underline{\beta}, \underline{\phi}) = (2, 3)$. This representative plot is of data with pathway sizes $(N_k^M, N_k^G) = (3, 20)$ and correlations $\rho_{GG} = \rho_{MM} = 0.2$ and $\rho_{GM} = 0.10$. Ten of 50 pathways (20%) are simulated to be enriched with at $(\pi_k^M, \pi_k^G) = (0.75, 0.50)$ for $k \in (1, \ldots, 10)$ and $(\pi_k^M, \pi_k^G) = (0, 0)$ otherwise. Increasing the correlation to $\rho_{GG} = \rho_{MM} = 0.6$ and $\rho_{GM} = 0.15$ provides only a marginal increase in the AUC for the absolute statistic (B, blue) and exponential (E, red) weight functions, see Figure 3.6. The other weight functions appear to have a minimal loss of power, e.g. AUC=98 versus AUC=96 for the hypergeometric test p-value weight (A, green) in this higher correlation model.

When we reduce the effect size of the differential elements to have alternative means $(\underline{\beta}, \underline{\phi}) = (1.5, 2)$ there is still a substantial increase in power for the p-value weight (A, green), the standardized test statistic (C, purple), and the CDF transformation (D, orange) with AUC values of 90 or greater (see Figure 3.7).

As expected from the recommendations of Roeder et al. (2006) [53], the exponential weight function (Figure 3.5 – 3.7, E, red) is poorly suited for this application. The exponential function, $\omega_k = e^{(\tilde{S}\tilde{\mu})}$, provides strong weights and is best suited for defining sparse up-weights. The balancing down-weights are then spread across the remaining tests. In the simulations presented thus far we have looked at situations in which 20% of the pathways are enriched leading to approximately 20% up-weighting. Given the small number of metabolites tested, compared to genomic studies, the corresponding down-weighting is only shared across a couple hundred metabolites. The strength of the exponential weight is also amplified by the choice $\tilde{\mu}$ which has been



Hypergeometric

AUC	Raw	A	B	C	D	E
Mean	86	98	15	97	98	73
(SEM)	(0.49)	(0.12)	(0.72)	(0.15)	(0.13)	(0.91)

Binomial

AUC	Raw	A	B	C	D	E
Mean	86	97	21	97	98	77
(SEM)	(0.49)	(0.13)	(0.6)	(0.14)	(0.12)	(0.91)

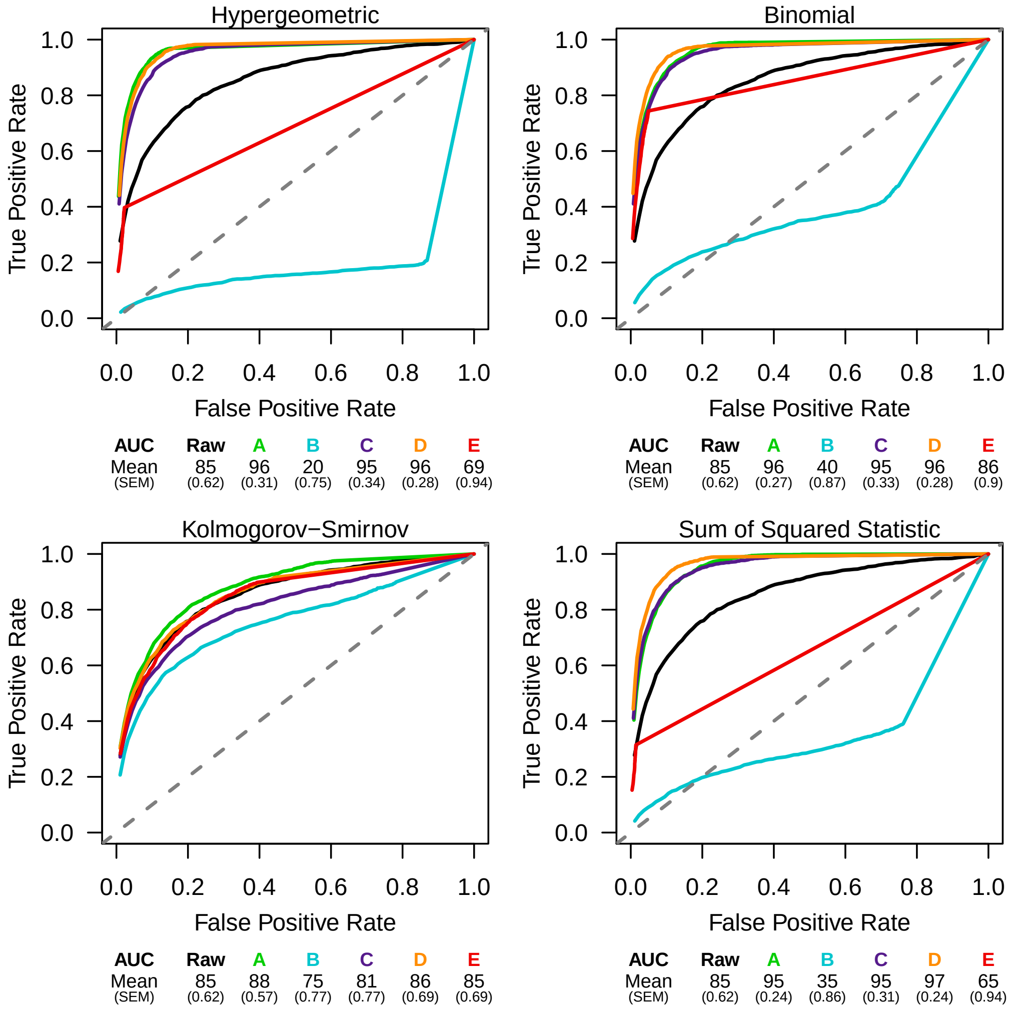
Kolmogorov−Smirnov

AUC	Raw	A	B	C	D	E
Mean	86	93	58	89	93	91
(SEM)	(0.49)	(0.3)	(0.96)	(0.5)	(0.36)	(0.46)

Sum of Squared Statistic

AUC	Raw	A	B	C	D	E
Mean	86	97	22	98	98	61
(SEM)	(0.49)	(0.15)	(0.59)	(0.12)	(0.12)	(0.73)

Figure 3.5: **Average receiver operating characteristic (ROC) curves** (n=100) depict the sensitivity and specificity for each test method and weight function when applied to per-metabolite tests. Data are simulated assuming within pathway correlation of 0.2 for each metabolites and genes and between element correlation of 0.1. Ten of fifty pathways were simulated as enriched where differential test statistics have mean of two and three for metabolites and genes, respectively. The mean area under the curve (AUC) estimate and associated standard error are provided in the table below each plot.



Hypergeometric

AUC	Raw	A	B	C	D	E
Mean (SEM)	85 (0.62)	96 (0.31)	20 (0.75)	95 (0.34)	96 (0.28)	69 (0.94)

Binomial

AUC	Raw	A	B	C	D	E
Mean (SEM)	85 (0.62)	96 (0.27)	40 (0.87)	95 (0.33)	96 (0.28)	86 (0.9)

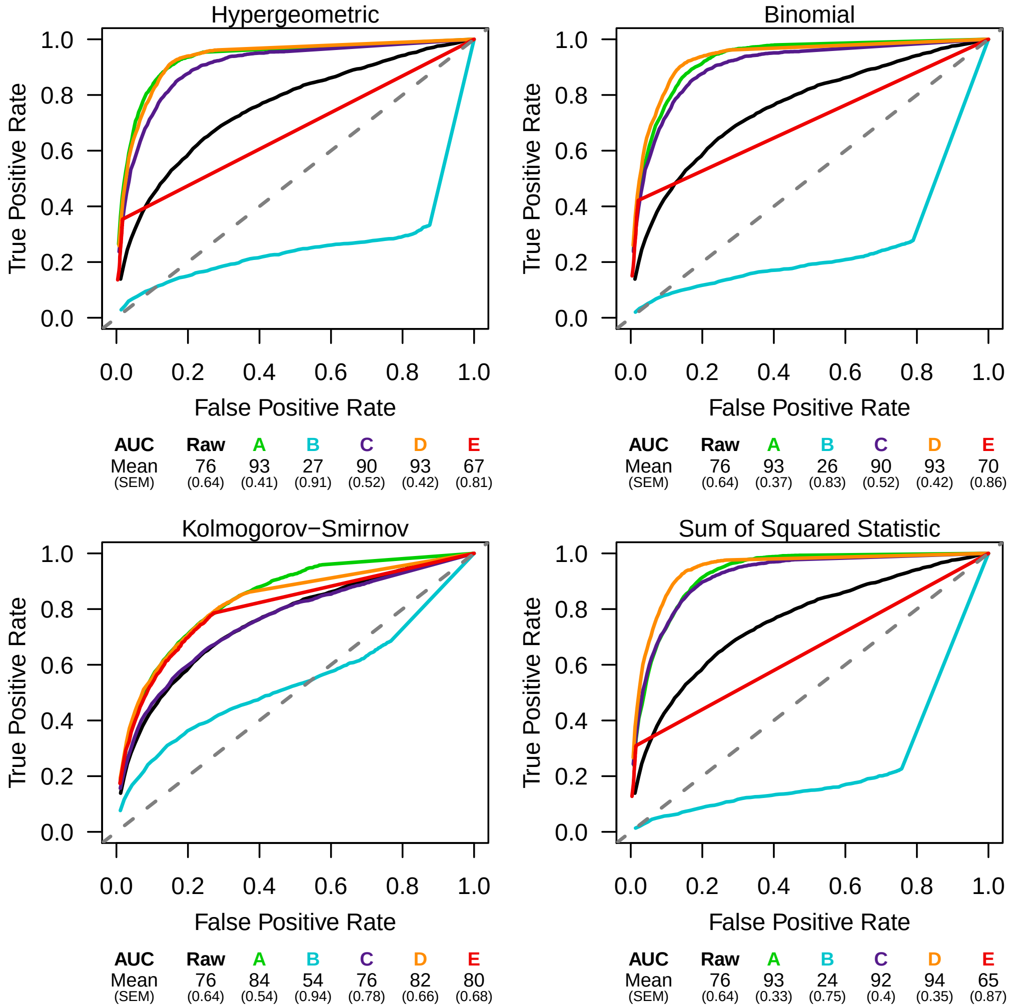
Kolmogorov−Smirnov

AUC	Raw	A	B	C	D	E
Mean (SEM)	85 (0.62)	88 (0.57)	75 (0.77)	81 (0.77)	86 (0.69)	85 (0.69)

Sum of Squared Statistic

AUC	Raw	A	B	C	D	E
Mean (SEM)	85 (0.62)	95 (0.24)	35 (0.86)	95 (0.31)	97 (0.24)	65 (0.94)

Figure 3.6: **Average receiver operating characteristic (ROC) curves** (n=100) depict the sensitivity and specificity for each test method and weight function when applied to per-metabolite tests. Data are simulated assuming within pathway correlation of 0.6 for each metabolites and genes and between element correlation of 0.15. Ten of fifty pathways were simulated as enriched where differential test statistics have mean of two and three for metabolites and genes, respectively. The mean area under the curve (AUC) estimate and associated standard error are provided in the table below each plot.



Hypergeometric

AUC	Raw	A	B	C	D	E
Mean	76	93	27	90	93	67
(SEM)	(0.64)	(0.41)	(0.91)	(0.52)	(0.42)	(0.81)

Binomial

AUC	Raw	A	B	C	D	E
Mean	76	93	26	90	93	70
(SEM)	(0.64)	(0.37)	(0.83)	(0.52)	(0.42)	(0.86)

Kolmogorov−Smirnov

AUC	Raw	A	B	C	D	E
Mean	76	84	54	76	82	80
(SEM)	(0.64)	(0.54)	(0.94)	(0.78)	(0.66)	(0.68)

Sum of Squared Statistic

AUC	Raw	A	B	C	D	E
Mean	76	93	24	92	94	65
(SEM)	(0.64)	(0.33)	(0.75)	(0.4)	(0.35)	(0.87)

Figure 3.7: **Average receiver operating characteristic (ROC) curves** (n=100) under the same simulation conditions as in Figure 3.5 except that differential test statistics have mean of 1.5 and 2 for metabolites and genes, respectively.

set $\tilde{\mu} = 2$ thus far. When we consider reducing the parameter to $\tilde{\mu} = 1$ we see that the exponential weight method (E, red) is improved but not to the level of the CDF transformation weights (D, orange; see Figure 3.8).

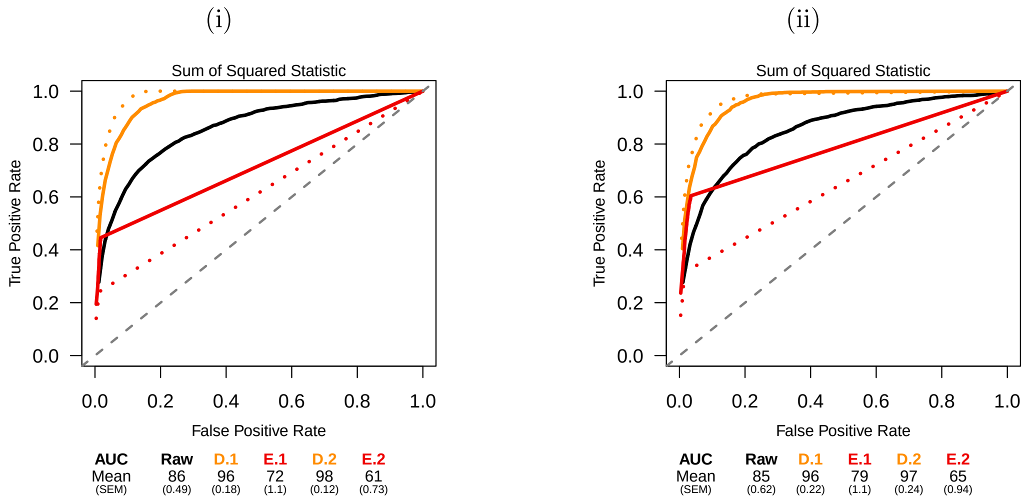
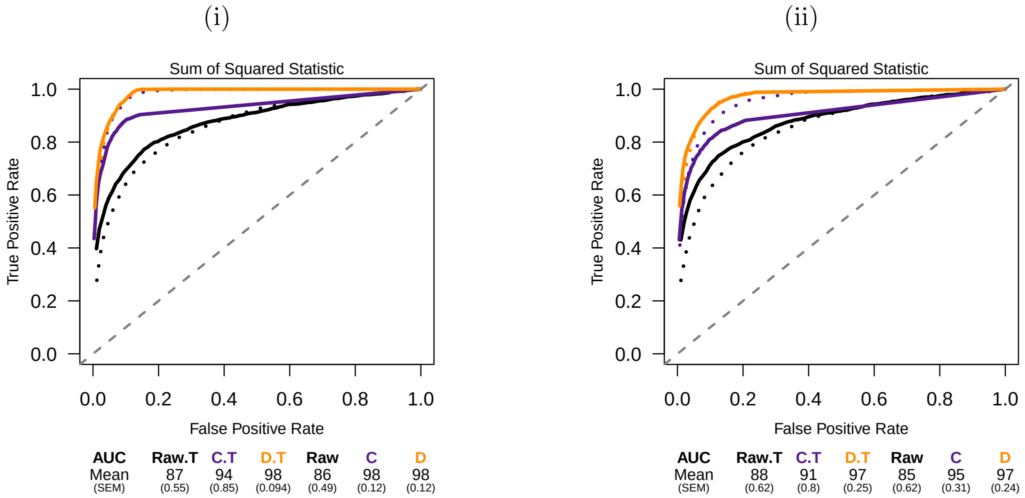


Figure 3.8: **Average receiver operator characteristic (ROC) curves (n=100) assuming $\tilde{\mu} = 1$.** The CDF transformation (D, orange) and exponential (E, red) weight functions are shown for the sum of squared statistic enrichment test. For comparison, both $\tilde{\mu} = 1$ (solid lines) and $\tilde{\mu} = 2$ (dotted lines) are plotted (i) under low correlation conditions as in Figure 3.5 and (ii) under high correlation conditions as in Figure 3.6.

The absolute value of the test statistic (Figure 3.5, B, blue) has poor AUC in most scenarios. We expect that this is primarily because the unstandardized enrichment test statistics vary so dramatically that the standardized weights must be very extreme in order to sum to one. It is interesting to note that the absolute statistic (B, blue) has appropriate error control (see Figures 3.1 and 3.2) and minimal power loss (see Figure 3.4) in the null cases. Perhaps under the null the variability of the pathway enrichment scores is low since no pathways are modelled as enriched resulting in more stable weight values.

In contrast, the p-value weight (A, green), the standardized test statistic (C, purple), and the cdf of the test statistic (D, orange) are all consistently more powerful than the unweighted p-value (Figures 3.5 – 3.7). We consider the potential weaknesses of each in turn. First, like the exponential or absolute test statistic weights, the p-value weight can become quite large when tests of pathway enrichment are significant. Having one strongly enriched pathway could produce down-weighting in the other pathways resulting in power loss. However, notice that for three of the four set enrichment tests the p-value is determined by permutation test. Here we use 1000 permutations so the precision of the p-value cannot be lower than 1/1000 and thus the negative log_{10} value will not be greater than 3. The hypergeometric test is an exact test and the p-value precision is thus limited by the sample size. These constraints put a ceiling on the range of the negative log_{10} p-values under which restriction they appear to behave well.

The absolute standardized statistic (C, purple) and the CDF function (D, orange) behave similarly. The standardized statistic reduces the magnitude of the statistic and reigns in the extreme values that occur prior to standardization. The CDF weight function is based on the standard normal density and thus it works well with the standardized test statistic. The benefit to using the CDF function is that it smooths out the test statistic thus reducing the effect of extreme test statistics. We ran a set of simulations using a multivariate t-distribution for drawing test statistics under the alternative hypothesis. The heavy tails of the t-distribution resulted in reduced power with the standardized statistic weight (C, purple) but not the CDF weight (D, orange) in the non-thresholded tests (see Figure 3.9). However, the power was still above that of the unweighted test and continued to show comparable power gains for both weight functions.and continued to show comparable power gains for both weight functions. The thresholded tests were not affected as they do not use the per-gene test statistics directly.



(i)

AUC	Raw.T	C.T	D.T	Raw	C	D
Mean	87	94	98	86	98	98
(SEM)	(0.55)	(0.85)	(0.094)	(0.49)	(0.12)	(0.12)

(ii)

AUC	Raw.T	C.T	D.T	Raw	C	D
Mean	88	91	97	85	95	97
(SEM)	(0.62)	(0.8)	(0.25)	(0.62)	(0.31)	(0.24)

Figure 3.9: **Average receiver operator characteristic (ROC) curves (n=100) under a multivariate t-distribution alternative.** The standardized statistic (C, purple) and the CDF transformation (D, orange) weight functions are shown for the sum of squared statistic enrichment function. Results from per-gene test statistic simulation models using both the t-distribution (solid lines) and the Normal distribution (dotted lines) alternatives are plotted (i) under low correlation conditions as in Figure 3.5 and (ii) under high correlation conditions as in Figure 3.6.

3.4.2 Simulation II: KEGG Based Pathways

Simulation Model II

This simulation makes use of the data structure of the KEGG pathways between genes and metabolites to define the pathways. This introduces overlapping pathways and pathways of varying sizes into the simulation. Rather than drawing the data from a multivariate normal distribution we use bootstrap resampling of published gene expression data to populate our vector of per-gene test statistics. The metabolite data is modeled from Sreekumar et al. (2009) [60]. Here we have 12 prostate cancer tissues from localized tumors of varying grade and 16 benign prostate tissues taken from resected prostate tissues. The gene expression experiment was performed by the same group (GEO:GSE3325 [68]) so we expect that the tissue diagnosis is consistent between the two experiments. There are five localized prostate tumor

Table 3.3: **Fifteen metabolites were chosen to be differential.** These metabolites are associated with up to five pathways, of which up to two pathways are simulated as enriched.

Metabolite	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Pathways	5	5	5	4	4	3	3	3	2	2	2	1	1	1	1
Number Enriched	2	1	0	2	0	2	1	0	2	1	0	1	1	0	0

samples and four benign prostate samples in the gene expression study. Further sample information and mapping details can be found in Appendix A.

For the simulation we selected eight of 76 pathways that have between 10 and 100 genes measured to be enriched. We then selected 15 (10.2%) of the metabolites to be differential. These metabolites were selected in such a way that they are members of between one and five pathways with up to two of the pathways being enriched (see Table 3.3). By examining metabolites in multiple pathways we can assess the affect of weight summaries across pathways. By examining metabolites that are not in enriched pathways we can examine potential power loss due to down-weighting.

The Varambally gene expression data were analyzed per gene using a two-sample t-test with pooled variance. All 126 permutations of the samples were run and the t-test recalculated to form the permutation null distribution. To prevent overcounting, the t-statistics were averaged across probes by gene symbol prior to gene set enrichment testing. All four enrichment tests, as described for Simulation I (Section 3.4.1) were run on the gene expression data. Additionally, all five weight functions were calculated (Section 3.3.2). To accomodate metabolites that belong to multiple pathways we summarized the ω_k values across pathways within metabolite, $v_{ij} = f(\omega_{jk} I_{(i \in k)})$. The weights were then standardized to average to one, $w_{ij} = v_{ij}/\bar{v_{\cdot j}}$. We consider both the median and the 75th percentile as the summary statistics, $f(\cdot)$.

The per-gene test statistic data are simulated by randomly sampling the 2375 mapped genes, with replacement, from the t-statistic matrix (t-statistic and 126 permutation statistics). To induce enrichment, the genes in the enriched pathways were randomly selected according to a Bernoulli(π) distribution from the subset of genes that were differential (n=177 at $\alpha = 0.05$), where $\pi \in (0.2, 0.5)$. That is on average 100π% of the genes in an enriched pathway were selected to be differential with an effect size seen in the Varambally data.

The 147 metabolite p-values were drawn from a uniform distribution on [0,1]. Those 15 metabolites that were selected to be differential are chosen from a Beta(3, 37) distribution. The shape parameters were chosen for a mean of 0.075 and a relatively narrow variance to provide the marginal p-values of interest. Specifically, the probability of selecting a p-value less than 0.05 is approximately 31% whereas the probability of selecting a p-value greater than 0.2 is less than 1%. The gene-set enrichment tests, weight functions, and per-metabolite weights are calculated for each of 1000 generated gene expression datasets and metabolite p-value vectors.

Results for Simulation Model II

Figure 3.10 shows the frequency of significant p-values (at $\alpha = 0.05$) for each of the 15 differential metabolites across the 1000 simulated data sets under $\pi = 0.2$. The color scheme and quadrant style figure is retained from the first simulation (Section 3.4.1). Additionally we denote the median summary weights by circles and the 75th percentile weights by triangles. Notice that these weights are similar but not exact for the four metabolites in a single pathway. Even though their v_{ij} measures will be identical they are standardized against $\bar{v}_{\cdot j}$ which will vary depending on the other $v_{i'j}$ values ($i' \neq i$). The black squares represent the unweighted test p-value and, as expected by the Beta(3,37) distribution used in simulation, they are significant in about 30% of the datasets.

We quickly see in Figure 3.10 that there are a few metabolites that are weighted well by the pathway data and show great power increases; namely metabolite 5(2) and the first 1(1) metabolite. Additionally, the CDF method (D, orange) appears to increase success of identifying these metabolites in most situations. It does show power loss for those metabolites that have no associated gene expression pathway enrichment such as 3(0). As expected, the exponential weights (E, red) perform poorly in most cases; performing well for only the two strongest metabolites and severely downweighting the others. However the unstandardized absolute test statistic (B, blue) does better than expected in some situations such as with the hypergeometric test.

The 132 metabolites that were not chosen to be differential can be used to determine false positive rates by considering the percentage of significant calls per metabolite across the 1000 simulated data sets. In Figure 3.11 these percentages are plotted as boxplots per weight function. The nominal 5% error rate is noted with a black horizontal line. For most weight methods the majority of the metabolites have error near the nominal line, i.e. the box contains 5%. The unweighted tests (black) are tightly centered at 0.05 as expected by the simulation design. There are some methods, however, that have high error rates for a handful of metabolites. Specifically, the exponential function (red) behaves poorly with all four enrichment tests with error rates reaching 50% for some metabolites. The p-value weighting method (green) has higher error for the hypergeometric test and the Kolmogorov-Smirnov test. Surprisingly, the absolute test statistic (blue) has low error in all cases except for the binomial test. The CDF function (orange) has low error in the self-contained tests (binomial and sum of squared statistic). There is no obvious difference between the median summary and the 75th percentile summary with respect to the error rates (see Figure 3.11). The 75th percentile may has a slight advantage in the rate

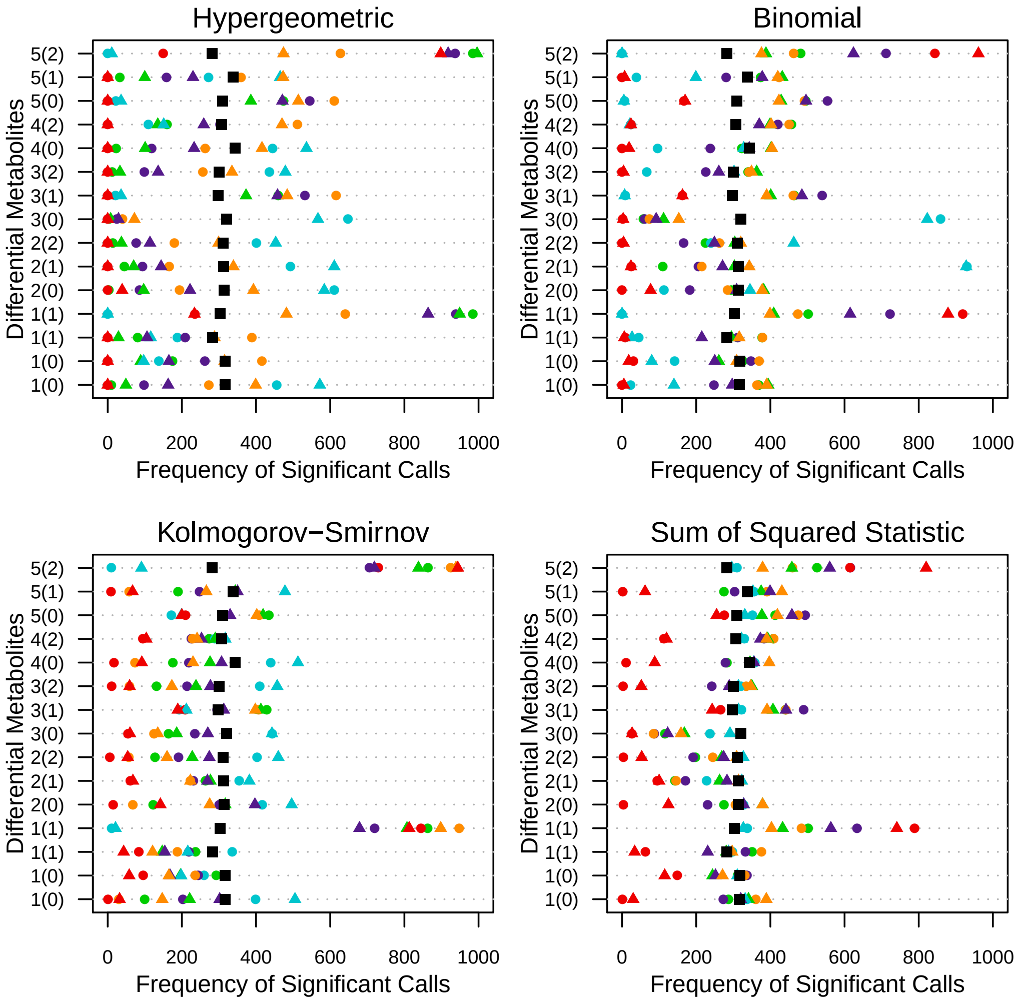


Figure 3.10: **The frequency of a significant call for each of the fifteen differential metabolites.** The metabolites are noted on the vertical axis by the number of associated pathways (enriched) for that metabolite. The weight functions are color coded as before (A, green; B, blue; C, purple; D, orange; E, red) with the median summary denoted by circles and the 75th percentile summary noted by triangles. The unweighted result is noted by black squares.

of true significant calls (see Figure 3.11).

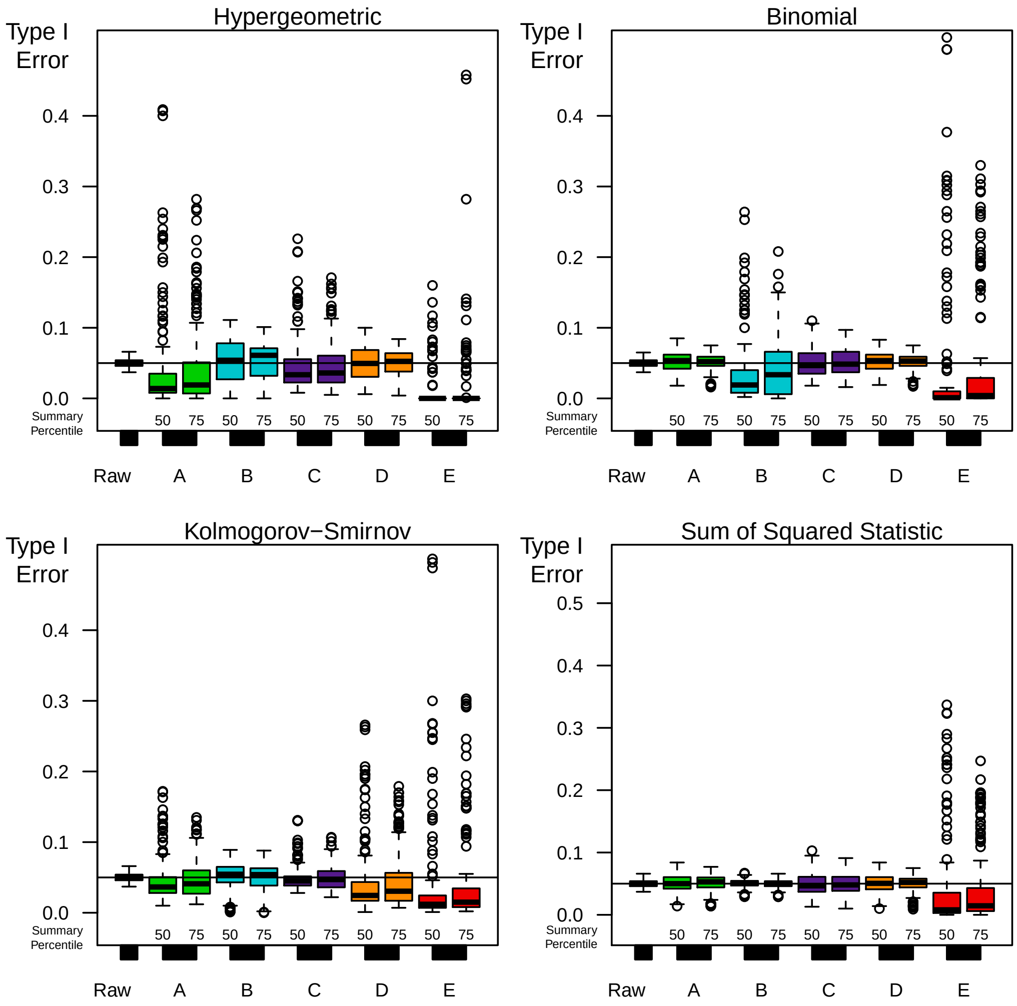


Figure 3.11: **Boxplots of the frequency of falsely finding a null metabolite to be significant.** Each method is presented by a boxplot and each point represents metabolite. For instance, the exponential weights (E) tend to have low error rates overall, but a handful of metabolites are called significant up to 500 times. The weight functions are color coded as before with the median summary in the left box and the 75th percentile summary in the right box.

3.5 Application

We wish to apply the method of p-value weighting to the motivating data example of the Sreekumar et al. (2009) [60] metabolite data and matched gene expression data. We begin by assessing the pathway enrichment of the gene expression data. Two-sample t-tests were used to assess the difference of gene expression between localized prostate cancer and adjacent benign tissues, per gene. All four enrichment methods (see Section 3.3.1) were then applied. When required, permutation p-values were calculated from 1000 permutations of the sample labels.

Permutation p-values are limited in precision by the number of permutations used. Here this resulted in little discrimination of pathways for the two self-contained tests; binomial and sum of squared statistics. The statistic S_k, however, showed potential for discriminating between the k pathways.

The results of Simulation I (Section 3.4.1) lead us to favor the CDF weighting function (D, orange). The results of Simulation II (Secion 3.4.2) show the CDF weighting function performing well for all but the K-S test. The hypergeometric test appears to perform best here but it is not robust when the parameter π is reduced from 0.5 to 0.2, unlike the two self-contained tests, data not shown. Interestingly in this application $Cor(S_k^{hyper}, S_k^{SST}) = 0.98$ across k thus we chose to proceed with the sum of squared test of enrichment.

In the simulations we set $\tilde{\mu} = 2$ for the CDF function as the presumed differential effect under the alternative distribution. Yet, the distribution of $\tilde{S}_k$ across k shows that this choice may not be optimal; see Figure 3.12. By assigning $\tilde{\mu} = 2$ we may be severly tempering the upper range of test statistics and thereby reducing the ability of the weights to differentiate pathway contributions. To assess this we look back at the data of Simulation II under the sum of squared test. The range of the $\tilde{S}_k$ values is less in the simulated data but still suggests that $\tilde{\mu} = 2$ may not be optimal.

Using the first 500 simulated data sets from Simulation II we assessed how varying $\tilde{\mu}$ affected the frequency of detection for each of the 15 simulated differential metabolites; see Figure 3.13. In the upper panels we see that the detection rate increases

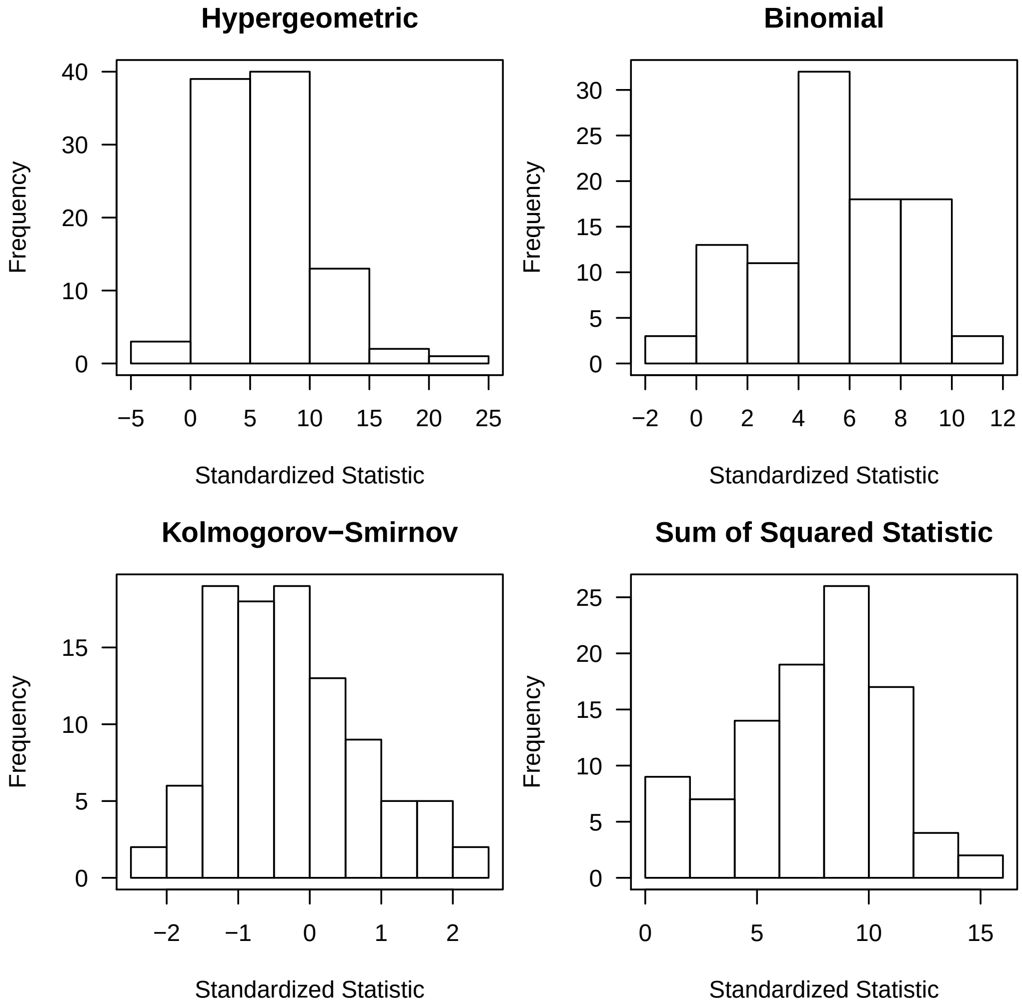


Figure 3.12: **Distribution of** $\tilde{S}_k$ across $k = 1, \ldots, 98$ pathways in the Sreekumar-matched gene expression data. Each of the four pathway enrichment tests is shown.

at first and then decreases as $\tilde{\mu}$ becomes large. In contrast the error rates shown in the bottom panels decrease overall (boxes) but a handful of metabolites are falsely discovered at increasing rates. It is interesting to note that when a median summary is used for combining weights for a metabolite in multiple pathways (Figure 3.13 left) the frequency of detecting truly differential metabolites quickly decreases for a majority of the metabolites. In contrast, if the seventy-fifth percentile is used to combine the weights across pathways (Figure 3.13 right) then the power loss is less and the error rates do not increase as quickly.

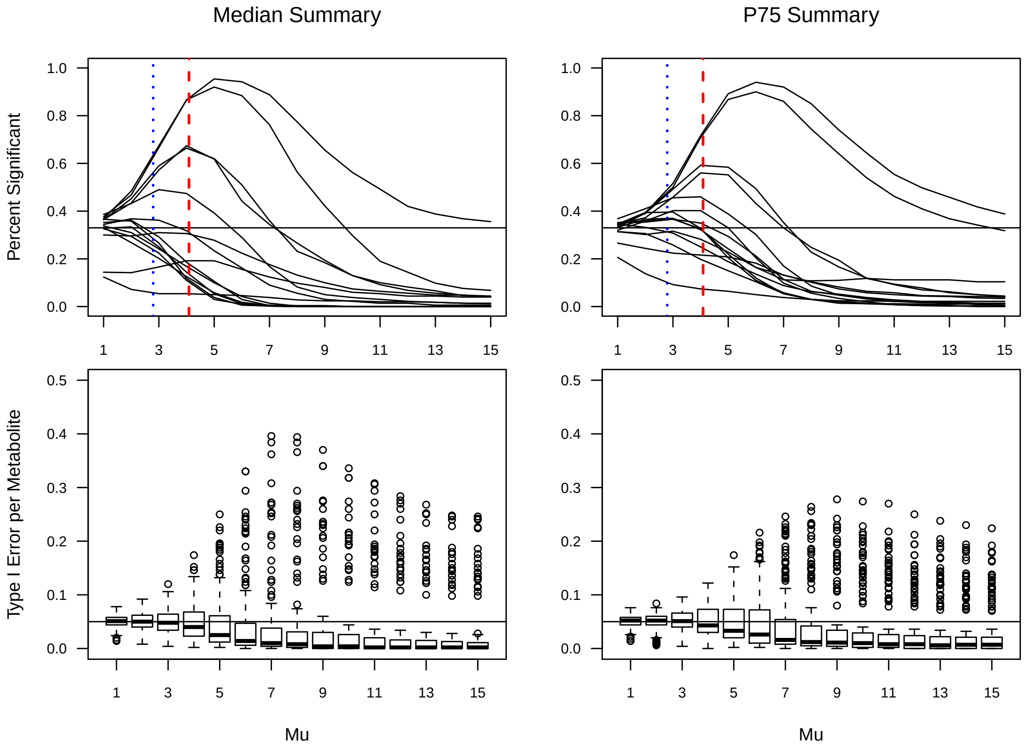


Figure 3.13: **The effect of adaptive estimation** with respect to power in the 15 differential metabolites (top) and Type I error in the remaining 132 metabolties (bottom). Both median (left) and seventy-fifth percentile (right) summaries for metabolites involved in multiple pathways are considered.

From these results we consider assigning $\tilde{\mu} = P_{75}(\underline{\tilde{S}})$, i.e., the 75-th percentile of the S_k values accross all k. The median 75th percentile, across all 500 simulation datasets, is shown with a red dashed vertical line in Figure 3.13. The median 50th percentile (median) is shown by a blue dotted vertical line. The seventy-fifth percentile appears to be the better choice according to Figure 3.13.

To test this adaptive $\tilde{\mu} = P_{75}(\underline{\tilde{S}})$ we analyzed the remaining 500 Simulation II datasets. Weight construction used the sum of squares statistic, the CDF weighting function with $\tilde{\mu} = P_{75}(S_k)$ across all k and the seventy-fifth percentile statistic was used to summarize the pathway weight components (ω_k) when a metabolite was represented in multiple pathways.

Table 3.4: **Percentage of times the metabolites are detected to be differential** in the remaining 500 datasets from Simulation II. The fifteen differential metabolites are listed by the number of pathways (enriched) in which each is included. An adaptive and a fixed estimation of $\tilde{\mu}$ are compared to the unweighted results. Median and maximum Type I error rates for the 132 non-differential metabolites are given in the bottom rows.

Metabolite	$\tilde{\mu} = P_{75}$	$\tilde{\mu} = 2$	Unwt.
5(2)	0.422	0.362	0.3
5(1)	0.568	0.394	0.32
5(0)	0.746	0.408	0.312
4(2)	0.424	0.4	0.312
4(0)	0.066	0.172	0.358
3(2)	0.356	0.37	0.312
3(1)	0.278	0.334	0.288
3(0)	0.23	0.314	0.304
2(2)	0.332	0.372	0.296
2(1)	0.538	0.392	0.292
2(0)	0.236	0.292	0.298
1(1)	0.328	0.368	0.308
1(1)	0.252	0.284	0.34
1(0)	0.694	0.398	0.302
1(0)	0.206	0.298	0.294
Median Error	0.041	0.050	0.050
Max Error	0.104	0.078	0.074

The fixed estimate of $\tilde{\mu} = 2$ shows more consistent, though marginal, increases in detection rates among the fifteen differential metabolites. The adaptive estimate of $\tilde{\mu} = P_{75}(\underline{\tilde{S}})$ shows stronger gains but they are balanced by stronger losses. Error rates are near to the nominal 0.05 rate, however, the adaptive estimate gives a wider spread of errors with one metabolite being falsely detected in up to 10.4% of the datasets.

For the application to the Sreekumar et al. (2009) [60] metabolite data we accept the potential losses of the adaptive method in favor of the potential for strong gains. Thus we use the sum of squares enrichment test for gene expression with the CDF weight function using $\tilde{\mu} = P_{75}(\underline{\tilde{S}})$. Additionally we use the seventy-fifth percentile summary for metabolites that participate in multiple pathways. The resulting shift in p-values can be seen in Figure 3.14.

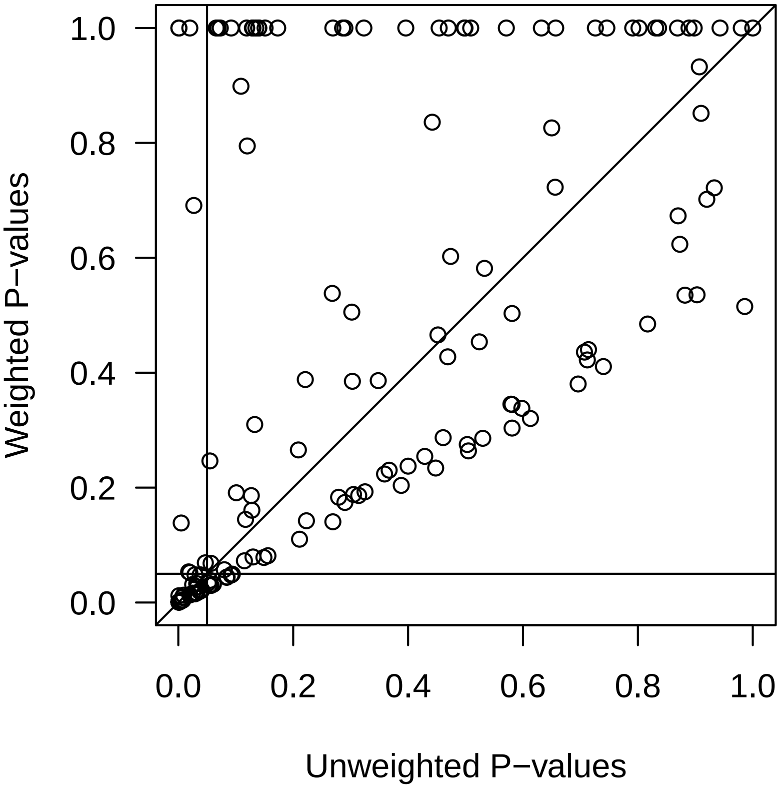


Figure 3.14: **Scatterplot comparing weighted and unweighted p-values** for each of the 147 mapped metabolties in the Sreekumar et al. (2009) [60] metabolite dataset. The matched gene expression data were used to construct the weights.

Twenty-five metabolites were found to be significant at $p < 0.05$ by both the weighted and unweighted p-values. There was a loss of eight metabolites by the weighted method; homocysteine, asparagine, bradykinin, cysteine, leucine, malate, N-acetylaspartate, and oxalate. However, there was a gain of ten metabolites resulting in a net gain of two metabolites; N-acetylneuraminate, adenine, argininosuccinate, aspartate, glycerol, guanosine, hypoxanthine, orotidine-5'-phosphate, spermine and xanthosine. In the original publication [60] leucine was listed as a metabolite upregulated from benign to metastatic disease so this loss is notable. Additionally there was an enrichment of amino acids detected in the differential metabolites originally and some are lost with the weighted analysis (e.g. leucine, cysteine) but aspartate is gained. Finally, sarcosine, which was of primary interest originally, had a decreased p-value in the weighted analysis (0.029 unweighted; 0.016 weighted) suggesting that this finding is supported by gene-set enrichment results.

3.6 Discussion

Here we have explored the utility of p-value weighting to enhance the power to detect differential metabolites. Gene set enrichment scores were used to summarize the gene expression data. Four enrichment tests of varying style and five weight functions were considered. As expected, the CDF function (D, orange) is better suited to the integration of gene expression and metabolite data by pathways than the exponential function (E, red) which is better suited for strong and sparse regions of upregulation. The standardized enrichment test statistic (C, purple) also performed well. Standardization of the enrichment test statistic makes the distribution of the test statistic better behaved than if the absolute statistic (B, blue) were used directly. However, the CDF function of the standardized test statistic is more robust when the tails of the per-gene test statistics are long. The p-value weight (A, green) performed well in Simulation I (Section 3.4.1) but had mixed results in the more complex second simulation (Section 3.4.2). In contrast, the absolute test statistic (B, blue) performed very poorly in Simulation I but performed moderately well in the second simulation.

Considering both simulations we recommend using the CDF function ($\omega_k = \Phi(\tilde{S} - \tilde{\mu})$, with $\tilde{S} = \frac{S-E(S_0)}{\sqrt{Var(S_0)}}$ and $\tilde{\mu} = 2$) when the distribution of the per-gene test statistics has long tails. Alternatively, a thresholded enrichment test can be used which ignores the magnitude of the per-gene test statistics by classifying each test as differential or not. In application we discovered that the use of $\tilde{\mu} = 2$ may not be optimal and an adaptive method for estimating $\tilde{\mu}$ was explored. This adaptive method produced more consistent gains for some differential metabolites but also resulted in more severe losses. The choice of a fixed or adaptive $\tilde{\mu}$ should be made in consideration of the study goals.

When a continuous enrichment test is desired, we prefer the self-contained sum of squared statistic test to the weighted Kolmogorov-Smirnov (K-S) test. This may be the difference between a self-contained and a competitive test for the application of summarizing gene expression information for constructing weights. Alternatively, the weighted K-S test has been receiving poor reviews in the recent literature (e.g. Ackerman and Strimmer, 2009) so perhaps a different continuous-measure competitive test would have better performance.

Another appealing feature of the p-value weighting method is that all metabolites can be considered in the analysis. Currently, as few as one third of the metabolites measured are identified (Sreekumar, 2009). As mass spectrometry libraries are expanded this number will increase but those metabolites that are unknown or are not mapped to gene expression are simply awarded a weight of 1. This does not adjust the p-value nor does it affect the requirement that the weights average to 1. Thus the unidentified metabolites are tested as they would have been had weighting not been considered.

Moving forward, we plan to explore other methods of summarizing the gene expression data to construct weights for the metabolites. Combining the data on a pathway level can lose power as the number of pathways to which the metabolite belongs increases. We are thus working on a method to summarize the gene expression information on a per-metabolite basis.

We additionally considered two weighting functions that include the metabolite data in the weight. The appeal of these weight functions is that the gene expression information would not be allowed to dominate the metabolite data. Expanding on the grouped weight method of Roeder et al. (2006) [53] we define groups using the metabolic pathways and the weight as a mixture of the estimates of gene set enrichment and metabolite enrichment. We hoped that inclusion of external gene expression information would reduce the bias of an internal weight since the number of metabolites per group is often smaller than the twenty recommended by Roeder et al. (2006) [53] to engage the sieve principle. We used either the percentage of metabolites measured per group or the correlation between the gene expression and the metabolite data to determine the mixing parameter. However, in simulation this method produced inflated type I error rates for both weight functions and the method was abandonded. For clarity details are omitted here but interested readers can see Appendix B.

CHAPTER IV

Integrative pathway enrichment testing methods for joint assessment of multiple omics platforms

4.1 Introduction

In case-control studies we are interested in comparing two groups of samples on a collection of measured variables possibly associated with case status. When the variables of interest are measurements from a high-throughput molecular assessment, such as from a gene expression microarray, the result is thousands of comparisons. The resulting list of differential genes can be unwieldy with hundreds of entries. Given this, researchers are often interested in grouping these lists into sets of genes with common functionality. The area of enrichment testing looks at an *a priori* defined gene set, such as from KEGG (Kyoto Encyclopedia of Genes and Genomes) or GO (Gene Ontology), and asks if the number of differential genes in the set is remarkable; either more or less than expected.

When the high-throughput assessment is metabolomics the number of molecules measured is reduced by at least an order of magnitude compared to gene expression assays [60, 70]. But the list of differential molecules can still be lengthy with respect to the number of leads feasibly followed. Thus, the desire to elucidate common functional patterns remains. The pitfall with a smaller list is that the sets of interest may not be represented well for testing. For instance, if three molecules are measured from a given set and two are differential this is 67% enrichment but it may not be statistically significant and it is not clear if it is biologically interesting.

Enrichment tests work by assessing the overall evidence of differential behavior of the elements (e.g. genes, metabolites) within the set. Tests that do not depend on dichotomizing the elements, into differential and not, are able to incorporate even moderate changes [64]. When we have measured many molecular aspects, e.g. gene expression, metabolites, proteins, it seems beneficial to assess their differential tendencies jointly across platforms. Integration of omics technologies has proved to be beneficial in other areas resulting in more interpretable results [45] and more meaningful associations [17] than when the platforms are assessed separately. In an effort to translate this success to the area of set enrichment we explore two set enrichment tests and describe methods for their multivariate application.

We begin with an overview of enrichment testing theory in Section 4.2. The methods of interest are described in Section 4.3 and the multivariate extensions are detailed. We explore the properties of these methods by simulation as detailed in Section 4.4. A metabolomics dataset [60] and related gene expression dataset [68] are used in Section 4.6 for an application of the top methods to existing data. We conclude with discussion and recommendations in Section 4.7.

4.2 Background on Enrichment Testing

Recently enrichment testing methods have been classified into two general flavors; competitive and self-contained [24, 48, 64]. Additionally two resampling methods are commonly used to assess the null hypotheses of these models [48]. In this section we introduce these testing styles and their underlying null hypotheses. We describe the resampling methods used for estimating the null distribution of the test statistic

Table 4.1: **The general scheme for a hypergeometric test of differential genes** A set of G genes is divided by the criteria of inclusion in the set of interest (S) and inclusion in the set of differential genes (D).

	Differential gene (D)	Non-differential gene (D')	Total
In the set (S)	G_{SD}	$G_{SD'}$	G_S
Not in the set (S')	$G_{S'D}$	$G_{S'D'}$	$G_{S'}$
Total	G_D	$G_{D'}$	G

and discuss the pros and cons of each method. For clarity the following discussion will use gene expression as the omics platform of interest, however these methods are applicable to any omics platform whose elements can be classified into *a priori* defined sets.

For reference we define the 2×2 classification depicted in Table 4.1. Here G genes have been individually tested for differential expression, perhaps using two-sample t-tests per gene. An interesting set of genes S has been defined. We can classify each of the G genes by whether they are differential (D) and whether they are in the set of interest (S).

4.2.1 Competitive Tests

For a set of genes, S, a competitive test assesses whether the amount of differential expression differs from that of its complement S'. The competitive null hypothesis, H_0^{comp}, then assumes that

H_0^{comp}**: genes within the set S show the same amount of association with the phenotype as those in set S'**

[64, 24]. In this way each gene set *competes* against its complementary set of measured genes.

A popular competitive set enrichment test is the Fisher's Exact test run on Table 4.1. The Fisher's Exact test assesses independence of the columns and rows and a

(i)

	D	D'	Total
In set A	4	4	8
In set B	6	2	8
Total	10	6	16

(ii)

	D	D'	Total
In set A	4	4	8
In set C	2	6	8
Total	6	10	16

Table 4.2: **An example of the relative estimation problem in competitive tests** Consider testing a set A in which half of the genes are differential (D). Table (i) uses set B as the reference set and table (ii) uses set C as the reference group, where sets B and C switch labels D and D'.

significant result suggests that the rate of differentially expressed genes is associated with pathway status. Specifically, we test that $Pr(G_{SD} \neq g_{SD} | G_S = g_s, G_{S'} = g_{S'}, G_D = g_d)$, assuming $G_{SD} \sim Hyper(G_S, G_{S'}, G_D)$, where G_{SD}, G_S and $G_{S'}$ are defined as in Table 4.1. As it is a two-sided test, a detected association may be due to enrichment or depletion of differential genes. The Fisher's Exact test is a competitive test in that it uses information on the differential expression of genes not in the set.

The chief complaint against enrichment tests based on relative enrichment estimation is that the significance of a set S can differ depending upon the gene sets used for the reference set S' [2]. Consider the gene sets A, B, and C as depicted in Table 4.2.1. Each of the three sets has eight genes with 50%, 75%, and 25% enrichment, respectively. If we compare set A to set B, as in Table 4.2.1i, then the one-sided hypergeometric test is $Pr(G_{AD} > 4 | G = 16, G_D = 10, G_A = 8) = 0.69$. However, if we compare set A to set C, as in Table 4.2.1ii, then the one-sided hypergeometric test is $P(G_{AD} > 4 | G = 16, G_D = 6, G_A = 8) = 0.06$. Clearly the significance of the enrichment in set A is affected by the reference set though the number of genes called differential in set A does not change.

Though viewed as a limitation, critics concede that relative estimation is useful when there is a large number of genes that are differential, as when comparing cancer versus normal tissue [24, 48]. In essence, competitive test results can be used to rank a series of gene sets of interest thereby distinguishing potentially interesting results from those arising by chance.

An additional argument against competitive testing is that most rely on gene-resampling to generate the null distribution [48, 24]. Essentially we can rewrite H_0^{comp} to say that S has as many differential genes as if they were drawn by chance from the set of all genes, $S \cup S'$. Empirical estimates of the null hypothesis are then generated by randomly sampling G_S genes from $S \cup S'$ and repeating the test on the random set S^*.

Arguments against gene-resampling methods are three-fold: lack of independence between genes, improper p-value interpretation, and sample size inflation. First, under gene-resampling we assume that the genes are independent with no distinction other than the label of differential (D) or not, (D'). However, we know that it is not accurate to assume that the expression levels of the full complement of genes are independent. Thus, it is possible that evidence against the null hypothesis actually reflects the underlying correlation structure of the genes and does not reflect a true enrichment of differential expression [64]. One may argue that identification of strongly correlated genes may also be valuable but it is important to understand that this may arise.

The second criticism of gene-resampling methods, as pointed out by Goeman and Bühlmann (2007) [24], is that the generated p-value must be interpreted in the context of gene sampling. We define a p-value to be the probability that a test statistic as extreme, or more, is observed when the experiment is repeated hundreds of times and the null hypothesis is true. Under a gene sampling scheme the experimental unit becomes the gene and the p-value indicates confidence that a new set of genes would show similar association with the set of interest. No claims can be made, however, about the association should a new set of biological samples be procured.

Finally, when constructing a test based on gene sampling, the "sample size" becomes the number of genes, G, which is often larger than the number of biological samples by a few orders of magnitude. This leads to sample size inflation and an inflation of power [24]. In an extreme example, Breitling et al. (2004) [7] claim that the 2x2 table can be used to produce meaningful statistical results with only a single pair of biological samples — one case and one control — used to determine differential genes, say through a thresholded ranked list of ratios.

4.2.2 Self-contained Tests

In contrast to competitive tests, self-contained tests do not utilize S' in the assessment of S. Specifically, we consider only the first row of Table 4.1 and ignore the second row. The self-contained null hypothesis, H_0^{sc}, assumes that

H_0^{sc}: the gene set S does not contain any genes whose expression levels are associated with the phenotype of interest

[64, 24].

In a simple example, we can test the cell G_{SD} in Table 4.1 using a binomial test of proportions based on $G_{SD} \sim binomial(G_S, \alpha)$, where α is an expected rate of differential genes which can be set to zero or the expected error rate, say $\alpha = 0.05$. Notice that, contrary to the relative estimation of the competitive hypergeometric test, this self contained binomial test gives the same p-value for both Table 4.2.1i and Table 4.2.1ii. Specifically, the one sided test of $Pr(G_{AD} > 4|G_A = 8, \alpha = 0.05) =$

1.5×10^{-5}.

An additional benefit of testing H_0^{sc} is that it reduces well so that gene sets of size one are treated as a differential expression test. At the other extreme it can be used as a global test of differential expression using the entire microarray as the set of interest [24]. Arguments against self-contained tests focus on the strong null hypothesis in relation to its biological interpretation. In particular, a single differential gene may be able to give enough evidence to reject the null hypothesis depending upon the differential threshold used and the number of genes in the set, however, this may not represent a biologically interesting result [24].

In their favor, self-contained tests primarily utilize subject-resampling methods rather gene-resampling to determine the null distribution of the test statistic [48]. Subject-resampling assumes that the subjects are independent and that under the null hypothesis the sample labels are meaningless and could have been assigned randomly. Permutation testing is a common method of subject-resampling in which the diagnostic class labels are shuffled between subjects and the test is repeated under the new assignment.

In contrast to gene-resampling, subject-resampling follows the experimental design of most studies by assuming that the subjects are independent realizations of the study population. By sampling the subjects the between-gene correlation structure is maintained. Additionally, the sample size is reflective of the number of subjects included and the p-value is generalizable to experiments with new subjects under study.

Subject-resampling tests are not without objection though. First, while gene sampling methods exaggerate the sample size, subject-resampling tests are limited by their often small sample size. In this way, p-values derived from permutation testing may be coarse and the level of discrimination desired may not be available [2, 48]. Second the null hypothesis being tested by subject sampling may be more difficult to state, especially when it is used to assess competitive tests (see Section 4.2.1).

4.2.3 Other Tests and Recommendations

Nam and Kim (2008) [48] suggest that there is a third test style which uses a null hypothesis intermediate to H_0^{sc} and H_0^{comp}. This mixed null hypothesis, H_0^{mixed}, states that

H_0^{mixed}**: none of the gene sets considered is associated with the phenotype**

[44]. In essence it is a simultaneous test of all gene sets. H_0^{mixed} is used by the GSEA test [44, 61] and the GSA test [13]. However, rejection of H_0^{mixed} only implies that there is at least one set that is associated with the phenotype. Further steps are required to make set-wise assessments.

Goeman and Bühlmann (2007) [24] note that tests of H_0^{sc} are more sensitive than H_0^{comp} for detecting changes within a set. However, Nam and Kim (2008) [48] show by simulation that tests of H_0^{sc} are not specific. The recommendation of Nam and Kim (2008) [48] is vague such that a test should be chosen according to the interest of the study, relying on H_0^{mixed} as a moderate choice, but preferably testing all three hypotheses simultaneously when possible as they each address slightly different questions.

With respect to sampling methods, Nam and Kim (2008) [48] suggest that gene-resampling methods be used when sample sizes are small. In contrast, Allison et al. (2006) [2] suggests that both subject sampling and gene sampling methods be considered as they are testing different hypotheses. Finally, some authors suggest a compromise that incorporates features of both gene sampling and subject sampling such as in Erfon and Tibshirani (2007) [13]. In this work we consider both competitive and self-contained methods. Gene-resampling and subject-resampling are used according to the null hypothesis of the test.

4.3 Joint Assessment of Enrichment

We consider conducting tests of enrichment that incorporate both the gene expression and metabolite information. We begin with per-gene and per-metabolite assessments of differential ability. Thus we have two vectors of test statistics, $\underline{T}^G = (T_1^G, T_2^G, \ldots, T_g^G)$ and $\underline{T}^M = (T_1^M, T_2^M, \ldots, T_m^M)$, and their corresponding p-values, $\underline{P}^G = (P_1^G, P_2^G, \ldots, P_j^G)$ and $\underline{P}^M = (P_1^M, P_2^M, \ldots, P_j^M)$, for the g genes and m metabolites, respectively.

The univariate tests can be employed as a means of joint assessment by simply concatenating the per-gene and per-metabolite test statistics to form a single vector of data, e.g. $\underline{T} = (\underline{T}^G, \underline{T}^M)$. However, concatenation of the lists may lead to bias favoring the larger dataset [59]. Additionally, the joined vectors must be made comparable before concatenation. For instance, if the different platforms have different sample sizes, $N_G \neq N_M$, then two-sample t-tests per-element will not be comparable as they have different degrees of freedom. Concatenating lists of p-values will resolve this problem as they are comparable by design. However, p-values may not be as precise if they were determined by empirical distributions and they lose directionality that may be of interest.

Enrichment can be assessed separately for each of $\underline{T}^G$ and $\underline{T}^M$ and then combined by considering intersections and unions of the individual enrichment test results. This method may help with ranking sets of interest since sets that are enriched independently by both platforms are likely to be of interest. However, it does not benefit by sharing of strength between the platforms. To arrive at a single statistic we can join the two results using a p-value combining method such as Fisher's method [46]. Here we assume that $-2 \times (log_e P_S^M + log_e P_S^G) \sim \chi_4^2$ for the enrichment p-values for metabolites and genes in set S. However, the χ^2 distribution is not correct when the tests being summed are dependent [73] which is likely the case here.

In the following we propose multivariate extensions of two univariate tests that use the information from each platform to form a single test of enrichment. As described below, we consider extending the competitive logistic regression test of Sartor et al. (2009) [57] and the self-contained sum of squared statistics [1]. We compare our multivariate extension to the univariate tests of each platform, the univariate test of concatenated datasets, and the sum of p-values in Section 4.4.

4.3.1 Competitive test: Logistic regression analysis with 2-df Wald test

Logistic regression was introduced by Sartor et al. (2009) [57] as an alternative test to the Fisher's Exact test which does not require dichotomization of the genes into differential and not. The logistic model proposed by Sartor et al. (2009) [57] is $logit(Pr(G_j \in S)) = \gamma_0 + \gamma(-log_{10}(p_j^G))$ where p_j^G is the p-value from the per-gene test of differential expression for gene j. The test of $H_0^{LR} : \gamma = 0$ can be obtained from standard statistical software using a 1-degree of freedom Wald test where rejection of H_0^{LR} indicates enrichment or depletion of the set. This is a competitive test because it takes an indicator of the genes inclusion in the set of interest, S, as the dependent variable, thereby comparing genes in S to all other genes, i.e. S'.

For the joint assessment we begin by modelling the genes and metabolites separately using the absolute value of the per-element t-statistic as the measure of differential ability. We chose to use the absolute t-statistic instead of $-log_{10}(p)$ because it appeared to be more stable in bootstrap resampling described below. Thus for set S we fit the following two models:

$$logit(Pr(G_j \in S)) = \gamma_0 + \gamma(|T_j^G|) \tag{4.1}$$

$$logit(Pr(M_k \in S)) = \mu_0 + \mu(|T_k^M|). \tag{4.2}$$

We construct a joint test of $H_{02}^{LR} : \gamma = 0, \mu = 0$ using a two degree-of-freedom Wald test. Specifically, $EV^{-1}E^T$, where $E = [\hat{\gamma}, \hat{\mu}]$, and V is an estimated variance-covariance matrix for γ and μ. We can obtain estimates of the variance of γ and μ from the univariate model estimates. However we do not have a convenient estimate of the correlation between the two parameters.

To estimate the correlation between the esimates $\hat{\gamma}_S$ and $\hat{\mu}_S$, for pathway S, we construct the bootstrap distributions of the two parameters, say $\underline{\tilde{\gamma}} = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_B)$ and $\underline{\tilde{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_B)$. We use row-resampling which reflects the use of genes and metabolites, not subjects, as input into the model. Through simulation we found that subject-resampling, which does not reflect the use of genes and metabolites as input into in the logistic regression model, dramatically underestimates the variance of the parameters compared to the variance estimates obtained from the univariate models of Equations 4.1 and 4.2. Additionally, when resampling the elements, we stratify the sample by inclusion in S to retain a fixed number of elements in S in each bootstrap sample. That is, for the set S we can split $\underline{T}^G$ into $\underline{T}^{GS}$ and $\underline{T}^{GS'}$. We sample $n(S_G)$ genes from $\underline{T}^{GS}$ with replacement, where $n(S_G)$ is the number of genes in S. The remainder of the genes are sampled with replacement from $\underline{T}^{GS'}$. The metabolite test statistics are then resampled in the same fashion. Stratification is especially important for platforms that tend to have small set counts.

Upon generation of B bootstrap estimates of γ and μ, $\underline{\tilde{\gamma}} = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_B)$ and $\underline{\tilde{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_B)$, we can compute an estimate of the correlation $\rho_{\gamma\mu}$ =corr($\underline{\tilde{\gamma}}, \underline{\tilde{\mu}}$). To reduce convergence error associated with small samples sizes and logistic regression bootstrapping we use a one-step bootstrap procedure as described by Moulton and Zeger (1991) [47]. Essentially the iterated weighted least-squares (IWLS) estimation algorithm is seeded with the observed parameters values of $(\hat{\gamma}_0, \hat{\gamma})$ for the gene expression. Then, given the b^{th} bootstrap resampled vector $\underline{T}^{Gb}$, one step is taken in the IWLS algorithm and the new estimate $(\tilde{\gamma_0}^b, \tilde{\gamma}^b)$ is reported. The IWLS algorithm is stopped here and not allowed to continue to convergence. One-step estimation is particularly important for models on small sets where the full IWLS can have problems of separation resulting in estimates nearing positive or negative infinity. Such non-convergent estimates inflate the variance and in simulation the variance estimates tended to be large compared to their model based counterparts with some variances on the order of 10^3 and greater when $n(S) = 4$.

Through simulation we find that there is not strong correlation between the parameters $\hat{\gamma}$ and $\hat{\mu}$ even in sets where the genes and metabolites were simulated to be correlated. Figure 4.1 shows a histogram of N correlation estimates, each from $B = 500$ bootstrap resamples, for a set S in which there is correlation between genes, $\rho_{GG} = 0.6$, between metabolties, $\rho_{MM} = 0.6$, and between genes and metabolites, $\rho_{MG} = 0.25$. The correlation in S' is not homogeneous and most genes and metabolites are simulated to be independent. These correlations are defined further in Section 4.4.2 where we have described our model for simulating data.

We see in Figure 4.1, i that the distribution of the estimates are fairly symmetric about zero. If we perform Fisher's transformation on the correlations to arrive at standardized z-scores we see that the estimates are underdispersed compared to the

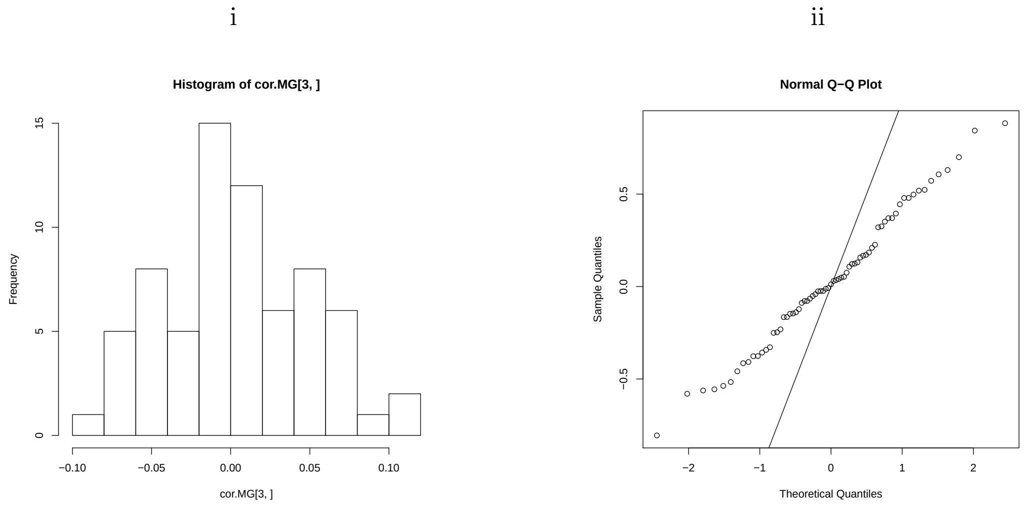


Figure 4.1: **Correlation of γ and μ** (i) Correlation of the estimates of γ and μ do not differ much from zero. (ii) They are under-dispersed compared to the expected normal distribution under the Fisher's z-score transformation.

quantiles of a standard normal distribution; see Figure 4.1, ii. The loss of correlation is likely due to the row-resampling that is used for this bootstrap.

Given these findings, we assume that the correlation is zero between $\hat{\gamma}$, and $\hat{\mu}$ for all sets. Thus V can be estimated as a diagonal vector with $var(\hat{\gamma})$ and $var(\hat{\mu})$ estimated from the univarite models. This reduces our test statistic to the sum of the two one-degree-of-freedom tests. Specifically, the test statistic for set S can be written as $U_S^{LR} = EV^{-1}E^T = \hat{\gamma}^2{\sigma_\gamma}^{-2} + \hat{\mu}^2{\sigma_\mu}^{-2}$, where $E = [\hat{\gamma}, \hat{\mu}]$, and $V = diag(\sigma_\gamma^2, \sigma_\mu^2)$. We assume that $U_S^{LR} \sim \chi_2^2$ under the null hypothesis $H_{02}^{LR} : \gamma = 0, \mu = 0$.

4.3.2 Self-contained test: Sum of squared statistics with 2-dimensional permutation test

We begin with the test statistics measuring the differential ability $\underline{T}^G = (T_1^G, T_2^G, \ldots, T_g^G)$ and $\underline{T}^M = (T_1^M, T_2^M, \ldots, T_m^M)$, for each gene and metabolite, respectively, separately. This enrichment test is simply the sum of all squared test statistics within set S, $W_S^G = \sum_{i=j}^{g} (T_j^G)^2 I_{(j \in S)}$, for $j = 1, \ldots, g$ genes and $I_{(a)}$ is an indicator function where $I_{(a)} = 1$ if a is true and 0 otherwise [1]. The statistic W_S^M for metabolites can be defined equivalently. Significance of W_S^G and W_S^M is determined separately by generating the null distribution for each using permutation of sample labels to form null datasets. Notice that this is a self-contained test as it only depends on the elements in the S but not S'.

To obtain a two-dimensional test we must decide on a way to assess how extreme the observed statistics (W_S^G, W_S^M) are when compared to each pair of null estimates $(\tilde{W}_S^G, \tilde{W}_S^M)_h$, $h = 1, \ldots, H$. Marginally, we can simply estimate $P_S^G = H^{-1}\sum_{h=1}^{H} I((\tilde{W}_S^G)_h \geq \hat{W}_S^G)$ for genes and $P_S^M = H^{-1}\sum_{h=1}^{H} I((\tilde{W}_S^M)_h \geq \hat{W}_S^M)$ for metabolites. For a multivariate application we calculate the Mahalanobis distance from the observed statistics $(\hat{W}_S^G, \hat{W}_S^M)$ to the centroid of the cloud of permutation statistic pairs.

The Mahalanobis distance on the H permutation statistics for set S can be written as

$$D_{SH} = \sqrt{((\underline{\tilde{W}}_S^G, \underline{\tilde{W}}_S^M) - \underline{1}(\psi_H^G, \psi_H^M))^T V_H^{-1}((\underline{\tilde{W}}_S^G, \underline{\tilde{W}}_S^M) - \underline{1}(\psi_H^G, \psi_H^M))} \tag{4.3}$$

where (ψ_H^G, ψ_H^M) are central measures of the H pairs $(\tilde{W}_S^G, \tilde{W}_S^M)$ and V_H is their variance-covariance matrix [11]. The vector $\underline{1}$ is an $H \times 1$ vector of 1s. Given (ψ_H^G, ψ_H^M) and V_H, the Mahalanobis distance can be calculated for any pair (W_S^G, W_S^M)

$$D_{SH}(W_S^G, W_S^M) = \sqrt{((W_S^G, W_S^M) - (\psi_H^G, \psi_H^M))^T V_H^{-1}((W_S^G, W_S^M) - (\psi_H^G, \psi_H^M))} \tag{4.4}$$

including the observed pair $(\hat{W}_S^G, \hat{W}_S^M)$. Thus to calculate the joint permutation p-value for $(\hat{W}_S^G, \hat{W}_S^M)$ we calculate

$$P_S^{GM} = \frac{1}{H}\sum_{h=1}^{H} I(D_{SH}(\tilde{W}_S^G, \tilde{W}_S^M)_h \geq D_{SH}(\hat{W}_S^G, \hat{W}_S^M)). \tag{4.5}$$

Figure 4.2 gives an example of the application of the Mahalanobis distance. Here a cloud of points representing the null distribution is drawn from the bivariate normal

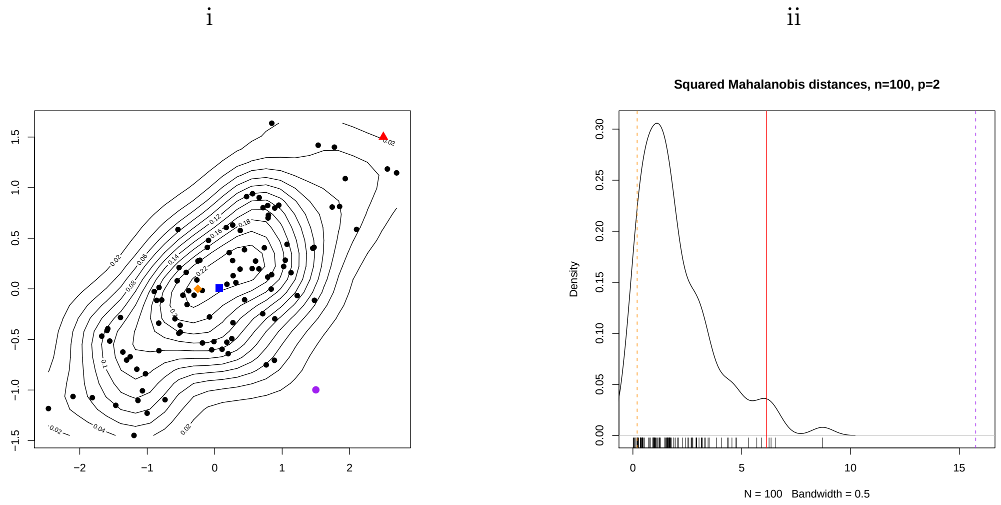


Figure 4.2: **Mahalanobis Distance** (i) A contour plot overlaying the scatterplot of 100 random draws from a bivariate normal distribution with mean zero, unit variance, and 50% correlation. The centroid defined by the marginal means is noted by a blue square. Three points of interest are added as the red triangle, orange diamond and purple circle (ii) The distribution of the Mahalanobis distance from the centroid for each point in the scatter with the three new points highlighted. Their rank can be used to determine a p-value.

distribution, $N_2([0,0], \sigma_x^2 = \sigma_y^2 = 1, \sigma_{xy} = 0.5)$. This hypothetical null distribution is plotted with black circles in Figure 4.2i. The centroid (ψ_x, ψ_y), determined by marginal means, is highlighted by a blue square. Though, in practice, each observed pair of statistics would have a unique null distribution, we suggest three possible observed pairs plotted here by a red triangle, purple circle, and orange diamond.

We chose to use the Mahalanobis distance metric since it accounts for the shape or spread of the null distribution [11]. The utility of this become apparant when we calculate the permutation p-values for each of the three hypothetical observations. Figure 4.2ii provides a distribution for the distances of the null values (black circles) from the centroid (blue square). The distance of each of the three hypothetical observations to the centroid is marked by a like-colored vertical line. The orange diamond, being near to the centroid (blue) is surpassed by most null values resulting in a p-value of 0.94. The red triangle is on the edge of the null distribution but is still surpassed by a few null points so it is given a p-value of 0.04. The purple circle is outside of the cloud of null points and thus results in a p-value $< 1/H$, where H is the number of permutation datasets used.

Points such as the hypothetical purple observation in Figure 4.2 are of most interest because they would be missed marginally. Additionally, had we not accounted for the non-spherical shape of the null distribution, such as by using an Euclidean distance measure, $E(x, y) = (((x, y) - (\psi_x, \psi_y))^T((x, y) - (\psi_x, \psi_y)))^{1/2}$, the purple point would not have been identified as extreme.

4.4 Simulation Models

We use simulation to assess the properties of our multivariate enrichment tests and to compare them to various univariate methods. Two simulation models are used. Each will be described in turn and results will be presented in the following section. The pathway maps of the Kyoto Encyclopedia of Genes and Genomes (KEGG, [35, 36, 34]) were used as motivating examples; see also Appendix A. Thus in the following we will use the term "pathway" instead of "set" to indicate an *a priori* defined collection of genes and metabolites.

4.4.1 Disjoint Pathway Simulation

In this simulation model we assume that the genes and metabolites can be separated into fifty disjoint pathways. That is any gene or metabolite is included in only one pathway. The correlation structure is the same for each pathway but no correlation is assumed between pathways. Additionally, ten pathways are simulated to have association with disease and the level of enrichment is consistent across these pathways. This simple model with homogeneous pathways allows us to explore very specific hypotheses about the properties of our methods.

Let Y_{ij} be the gene expression measurement for sample i and gene j. Likewise let $Z_{i'k}$ be the metabolite intensity measure for sample i' and metabolite k. The gene expression measures and metabolite measures need not arise from the same subjects, but it is possible for some or all samples to be matched by subject, i.e. $i = i'$. Define W_i and $W_{i'}$ as the case-control status for samples i and i', respectively, where

$$W_i, W_{i'} = \begin{cases} 1 & \text{case} \\ 0 & \text{control} \end{cases}$$

We assume the data, Y and Z, can be sorted into $\mathscr{P}$ pathways. Let I_p be an indicator for association of pathway p with case status,

$$I_p = \begin{cases} 1 & \text{Associated pathway} \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, define the indicator variables g_{jp} and m_{kp} for the inclusion of gene j and metabolite k, respectively, in pathway p,

$$g_{jp} = \begin{cases} 1 & \text{Gene } j \text{ in pathway } p \\ 0 & \text{otherwise} \end{cases} ; \; m_{kp} = \begin{cases} 1 & \text{Metabolite } k \text{ in pathway } p \\ 0 & \text{otherwise} \end{cases}$$

Define D_j and C_k as indicator variables of differential gene expression and differential metabolite intensity, respectively, between cases and controls. We can use pathway association to define Bernoulli distributions for D_j and C_k such that a gene or metabolite has a chance d_1 or c_1 of being differentially expressed provided that it is in at least one associated pathway, i.e. $max_p(I_p g_{jp}) = 1$ and $max_p(I_p m_{kp}) = 1$, respectively for genes and metabolites.

$$D_j \sim \begin{cases} Bern(d_1), & max_p(I_p g_{jp}) = 1 \\ Bern(d_0), & \text{otherwise} \end{cases} ;\ C_k \sim \begin{cases} Bern(c_1), & max_p(I_p m_{kp}) = 1 \\ Bern(c_0), & \text{otherwise} \end{cases}$$

Here d_1, d_0, c_1, and c_0 are fixed values and can be adjusted in simulation. To simulate pathway enrichment we assign $d_1 > d_0$ and $c_1 > c_0$. Interestingly, as $d_1 \rightarrow d_0$ (or as $c_1 \rightarrow c_0$) the effect of being in the pathway diminishes under the competitive definition of enrichment because no pathway will be enriched above its complement. However, this is not a concern for tests of the self-contained null hypothesis provided that d_1 and c_1 are still sufficiently large, since they do not consider the elements in the complement of the pathway

Let us then write the simulation model as:

$$\begin{aligned} Y_{ij} &= \alpha + \beta_j + \omega_j D(I_p g_{jp})_j W_i + e_{Y_{ij}} \\ Z_{i'k} &= \theta + \phi_k + \eta_k C(I_p m_{kp})_k W_{i'} + e_{Z_{i'k}} \end{aligned} \tag{4.6}$$

This additive model allows for a non-zero global mean expression (intensity) level through α (θ). It assumes a mean expression (intensity) level per gene (metabolite) as defined by β_j (ϕ_k) which is modified for case samples by ω_j (η_k) according to the distributions $D(I_p g_{jp})_j$ ($C(I_p m_{kp})_k$). We allow $e_{Y_{ij}}$ and $e_{Z_{i'k}}$ to be correlated, ρ_{mg}, in order to simulate matched samples. We also allow correlation between genes $\rho_{gg} = Corr(e_{Y_{ij}}, e_{Y_{ij'}})$ and between metabolites $\rho_{mm} = Corr(e_{Z_{i'k}}, e_{Z_{i'k'}})$. In this simulation model these correlations are limited to genes and metabolites within the same pathway, thereby reducing the complexity of the simulated data structure.

The simulation of the data is then as follows:

1. Determine sizes N_{gene}, $N_{metabolite}$, N_{sample}, $N_{pathway}$

2. Determine case status W_i, $i = 1, \ldots, N_{sample}$, fixing N_{case} and $N_{control}$

3. Determine pathway associations I_p, $p = 1, \ldots, N_{pathway}$, fixing N_{assoc} and pathway memberships g_{jp} and m_{kp}.

4. Determine correlation terms $\rho_{g,g}$, $\rho_{m,m}$, and $\rho_{m,g}$.

5. Jointly simulate the gene expression data matrix (size $N_{gene} \times N_{sample}$) and the metabolite data matrix (size $N_{metabolite} \times N_{sample}$)

    (a) Assume data is globally standardized per sample, $\alpha = 0$ and $\theta = 0$.

    (b) Draw variance terms $\sigma^2_{Y_j} \sim \chi_4^{-2}$, for $j = 1, \ldots, N_{gene}$ and $\sigma^2_{Z_k} \sim \chi_4^{-2}$, for $k = 1, \ldots, N_{metabolite}$.

    (c) Draw element-wise mean terms $\beta_j \sim N(0, 4\sigma^2_{Y_j})$, for $j = 1, \ldots, N_{gene}$ and $\phi_k \sim N(0, 4\sigma^2_{Z_k})$, for $k = 1, \ldots, N_{metabolite}$.

    (d) Jointly draw $y^{'}_{ij}$ and $z^{'}_{i'k}$ from a multivariate normal distribution. Specifically,

    $$(\underline{y}^{'}_{i}, \underline{z}^{'}_{i'}) \sim MVN((\underline{\beta}, \underline{\phi}), \underline{\underline{\Sigma}}_{YZ}),$$

    where

    $$\underline{\underline{\Sigma}}_{YZ} = \begin{bmatrix} \sigma^2_{Y_1} & \sigma_{YY} & \cdots & \sigma_{YZ} & \sigma_{YZ} \\ \sigma_{YY} & \sigma^2_{Y_2} & & & \sigma_{YZ} \\ \vdots & & \ddots & & \vdots \\ \sigma_{YZ} & & & \sigma^2_{Z_{N_m-1}} & \sigma_{ZZ} \\ \sigma_{YZ} & \sigma_{YZ} & \cdots & \sigma_{ZZ} & \sigma^2_{Z_{N_m}} \end{bmatrix}$$

    and $\sigma_{\cdot\cdot}$ is chosen to retain the desired correlation $\rho_{\cdot\cdot}$.

6. Apply differential effects

    (a) Draw mean effect sizes $\omega_j \sim Unif([-2.5, -0.5] \cup [0.5, 2.5])$, for $j = 1, \ldots, N_{gene}$ and $\eta_k \sim Unif([-1.5, -0.5] \cup [0.5, 1.5])$, for $k = 1, \ldots, N_{metabolite}$.

(b) For fixed d_1, d_0, c_1, and c_0 draw $D_j \sim I_p g_{jp} Bern(d_1) + (1 - I_p g_{jp}) Bern(d_0)$ and $C_k \sim I_p m_{kp} Bern(c_1) + (1 - I_p m_{kp}) Bern(c_0)$

(c) Set $y_{ij} = y^{'}_{ij} + \omega_j D_j W_i$ and $z_{i'k} = z^{'}_{i'k} + \eta_k C_k W_{i'}$

For this simulation, we assume that the number of samples is the same for cases and controls, with $N_{sample} \in (30, 100)$. We allow the correlations to vary: $\rho_{YY} = \rho_{ZZ} \in (0.2, 0.6)$ and $\rho_{YZ} \in (0.10, 0.25)$ where $\rho_{YY} = \rho_{ZZ} > \rho_{YZ}$. We consider gene pathways with 20 measurements, i.e. $N_{G_p} = 20$, and metabolite pathways with $N_{M_p} \in (4, 20)$. The enrichment levels (d_1, d_0) and (c_1, c_0) are allowed to vary with $(d_1, d_0) = (c_1, c_0) \in [(0.5, 0), (0.25, 0), (0.10, 0), (0.25, 0.05), (0.05, 0.05), (0.10, 0.10), (0, 0)]$ with the last three pairs representing null models.

4.4.2 Varying Pathway Simulation

This simulation model generates the same number of genes and metabolites in total and per-pathway as for the disjoint simulation above. The simulation model of Equation 4.6 is used as the basis of the data generation. However, m_{kp}, g_{jp}, D_j, and C_k are fixed to construct clusters of genes and metabolites with varying levels of association with disease. We also allow the pathways to overlap and to be non-homogeneous in correlation structure and enrichment. This style of simulation was used by Ackermann and Strimmer (2009) [1] in their review of various single-platform enrichment tests. Here we can assess how well the methods are able to detect various pathway types in a non-homogeneous setting. Null pathways are also included providing a reference set for competitive tests and allowing us to estimate false discovery.

The $N_{gene} \times N_{sample}$ gene expression matrix and $N_{metabolite} \times N_{sample}$ metabolite intensity matrix are drawn in blocks of $N_{G_p} \times N_{sample}$ genes and $N_{M_p} \times N_{sample}$

Table 4.3: **The data are generated from 10 multivariate distributions** with the following correlation structures and differential patterns. Twenty genes and four metabolites are drawn from each distribution ($h = 1, \ldots, 9$). Background genes (n=820) and metabolites (m=164) are simulated to be non-differential and without correlation, see distribution $h = 0$.

Distribution (h)	0	1	2	3	4	5	6	7	8	9
% Differential Genes	0	100	100	100	100	100	100	0	0	0
% Differential Metabolites	0	100	100	100	0	0	0	100	100	100
ρ_{YY}, ρ_{ZZ}	0	0	r_1	r_1	0	r_1	r_1	0	r_1	r_1
ρ_{YZ}	0	0	0	r_2	0	0	r_2	0	0	r_2

metabolites according to multivariate normal distributions $h = 1, \ldots, 9$; see Table 4.3. These 9 distributions have varying levels of correlation and enrichment between genes and metabolites. The remaining genes and metabolites required to reach size N_{gene} and $N_{metabolite}$, respectively, are drawn from distribution $h = 0$ to represent the null elements.

The data are simulated as follows:

1. Set N_{gene}, $N_{metabolite}$, N_{sample}, and N_{case}, where $N_{control} = N_{sample} - N_{case}$.

2. Assume data is globally standardized per sample, i.e. $\alpha = 0$ and $\theta = 0$.

3. For each gene j, $j = 1, \ldots, N_{gene}$, and each metabolite k, $k = 1, \ldots, N_{metabolite}$

   (a) draw variance terms $\sigma^2_{Y_j} \sim \chi_4^{-2}$ and $\sigma^2_{Z_k} \sim \chi_4^{-2}$,

   (b) draw element-wise mean terms $\beta_j \sim N(0, 4\sigma^2_{Y_j})$ and $\phi_k \sim N(0, 4\sigma^2_{Z_k})$, and

   (c) draw mean effect sizes $\omega_j \sim Unif([-2.5, -0.5] \cup [0.5, 2.5])$ and $\eta_k \sim Unif([-1.5, -0.5] \cup [0.5, 1.5])$.

4. Jointly draw y_{ij} and z_{ik} from each of the ten multivariate normal distributions ($h = 0, 1, \ldots, 9$, see Table 4.3). Specifically,

   (a) for distribution $h \in (1, 2, 3)$ containing N_{Gp} genes and N_{Mp} metabolites, draw $N_{control}$ control samples from $(\underline{y}_i, \underline{z}_i)_h \sim MVN((\underline{\beta}, \underline{\phi})_h, \underline{\underline{\Sigma_h}}_{YZ})$, and

N_{case} case samples from $(\underline{y}_i, \underline{z}_i)_h \sim MVN((\underline{\beta}+\underline{\omega}, \underline{\phi}+\underline{\eta})_h, \underline{\underline{\Sigma_h}}_{YZ})$, where

$$\underline{\underline{\Sigma_h}}_{YZ} = \begin{bmatrix} \sigma^2_{Y_1} & \rho_{gg}\sigma_{Y_1}\sigma_{Y_2} & \cdots & \rho_{mg}\sigma_{Y_1}\sigma_{Z_3} & \rho_{mg}\sigma_{Y_1}\sigma_{Z_4} \\ \rho_{gg}\sigma_{Y_1}\sigma_{Y_2} & \sigma^2_{Y_2} & & & \rho_{mg}\sigma_{Y_2}\sigma_{Z_4} \\ \vdots & & \ddots & & \vdots \\ \rho_{mg}\sigma_{Y_1}\sigma_{Z_3} & & & \sigma^2_{Z_3} & \rho_{mm}\sigma_{Z_3}\sigma_{Z_4} \\ \rho_{mg}\sigma_{Y_1}\sigma_{Z_4} & \rho_{mg}\sigma_{Y_2}\sigma_{Z_4} & \cdots & \rho_{mm}\sigma_{Z_3}\sigma_{Z_4} & \sigma^2_{Z_4} \end{bmatrix}$$

and $(\rho_{gg}, \rho_{mm}, \rho_{mg})$ are defined in Table 4.3.

(b) for distribution $h \in (4, 5, 6)$ containing N_{Gp} genes and N_{Mp} metabolites, draw $N_{control}$ control samples from $(\underline{y}_i, \underline{z}_i)_h \sim MVN((\underline{\beta}, \underline{\phi}), \underline{\underline{\Sigma_h}}_{YZ})$, and N_{case} case samples from $(\underline{y}_i, \underline{z}_i) \sim MVN((\underline{\beta}+\underline{\omega}, \underline{\phi})_h, \underline{\underline{\Sigma_h}}_{YZ})$, where $\underline{\underline{\Sigma_h}}_{YZ}$ is as defined above.

(c) for distribution $h \in (7, 8, 9)$ containing N_{Gp} genes and N_{Mp} metabolites, draw $N_{control}$ control samples from $(\underline{y}_i, \underline{z}_i) \sim MVN((\underline{\beta}, \underline{\phi})_h, \underline{\underline{\Sigma_h}}_{YZ})$, and N_{case} case samples from $(\underline{y}_i, \underline{z}_i) \sim MVN((\underline{\beta}, \underline{\phi}+\underline{\eta})_h, \underline{\underline{\Sigma_h}}_{YZ})$ where $\underline{\underline{\Sigma_h}}_{YZ}$ is as defined above.

(d) for distribution $h = 0$ containing $N_{gene} - 9 \times N_{Gp}$ genes and $N_{metabolite} - 9 \times N_{Mp}$ metabolites, draw N_{sample} samples, that is for n_{cases} and $n_{controls}$, from $(\underline{y}_i, \underline{z}_i) \sim MVN((\underline{\beta}, \underline{\phi}), \underline{\underline{\Sigma_h}}_{YZ})$, where $\underline{\underline{\Sigma_h}}_{YZ} = diag(\underline{\sigma}^2_Y, \underline{\sigma}^2_Z)_h$.

As in the disjoint simulation, Section 4.4.1, we set $N_{G_p} = 20$ and N_{M_p} to be either 4 or 20. This results in 1000 genes and either 200 or 1000 metabolites, according to the simulation parameters. The data are correlated for some genes and some metabolites, though not all are correlated, see Table 4.3. The overall rate of differential expression is 12% for the genes and 12% for the metabolites in each dataset.

To assess various pathway structures we subset these data into various "pathways"

Table 4.4: **Simulated pathways to be tested for enrichment.** Each pathway p contains N_{G_p} genes and N_{M_p} metabolites drawn such that π- percent of the elements are from a differential distribution $h \in (1, 2 \ldots, 9)$ and the remainder are from the null distribution $h = 0$.Pathways 25–29 are constructed by random draws across all 10 distributions, $h \in (0, 1, \ldots, 9)$. Pathways $30 - 70$ are a disjoint partition of the null set, $h = 0$, so that each element in this set contributes to at least one pathway.

π	Distribution (h)	1	2	3	4	5	6	7	8	9
1		1	2	3	4	5	6	7	8	9
0.5		10	11	12	13	14	15	16	17	18
0.25		19	20	21	22	23	24	-	-	-

of interest. In particular, we are interested to know how well each test can find the 24 pathways described in Table 4.4. We also construct five random pathways, i.e. pathways 25-29, where the pathway membership is a random draw from all simulated genes or metabolites. These five pathways allow us to consider the rate of non-specific pathway identification since genes and metabolites are selected across all distributions $h = (1, 2, \ldots, 9)$. Finally, the remaining 41 pathways, i.e. 30–70, are a partitioning of the null elements, $h = 0$, so that each element participates in at least one pathway. The null pathways allow us to look at false discovery error rates. The pathway participation indicators, g_{jp} and m_{kp} for pathway p, gene j and metabolite k, respectively, are determined at the start of the analysis. Random draws as required for pathways 10 - 29 are done once so that pathway membership is not changing throughout the analysis since we would not expect this in application to non-simulated data.

For clarity, an example of the indicator matrix, m_{kp}, for the inclusion of metabolite k in pathway p, is given in Table 4.5. This matrix represents a dataset of 200 metabolites assigned to 70 pathways of size $N_{M_k} = 4$. Notice that accoring to Table 4.4, 100% of the metabolites of pathway $p = 1$ are drawn from distribution $h = 1$. Likewise 50% of the metabolites in pathway $p = 10$ are drawn from distritubion

$h = 1$ and 50% are drawn from $h = 0$. The randomly drawn pathways $p = 25$ and $p = 29$ are shown. Notice that pathway $p = 25$ does not include any of the differential metabolites, $k \in (1, \ldots, 36)$, where as one differential metabolite, $k = 4$, was selected to be included in pathway $p = 29$. Pathways $p \in (30, \ldots, 70)$ are formed by a partitioning of the null metabolites, $h = 0$ and $k \in (37, \ldots, 200)$.

4.5 Simulation Results

Using the simulation models above we explored several enrichment testing methods. We are particularly interested in the two multidimensional methods that we devised; the 2-df Wald test for logistic regression and the 2-dimensional permutation test for the sum of squared statistics. For comparison we also considered the univariate counterparts for these methods testing enrichment based on the genes alone and on the metabolites alone. We also consider joining the data via concatenation and joining the tests via Fisher's method of combining p-values, abbreviated in figures a "p-sum". Each of these methods is described in Section 4.3. Finally, given its continued popularity, we also consider the Fisher's Exact test, though no multi-dimensional extension was devised.

4.5.1 Varying pathway simulation results

Let us first consider some results from the variable pathway simulation of Section 4.4.2. These simulations provide an overview of the behavior of the methods. For each simulation scenario, 100 datasets were generated and tested. In Figures 4.4 and 4.5 we depict the frequency with with each pathway, from $1 - 29$, was determined to be significant at $\alpha = 0.05$ for each test considered. The symbol key for these plots can be found in Figure 4.3. The average rate of false positives is computed across each of the 41 null pathways per test. Boxplots of these error rates across the 100

k	h	1	2	⋯	10	11	⋯	25	⋯	29	30	⋯	45	46	47	⋯	70
1	1	1	0	⋯	1	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
2	1	1	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
3	1	1	0	⋯	0	0	⋯	0	⋯	1	0	⋯	0	0	0	⋯	0
4	1	1	0	⋯	1	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
5	2	0	1	⋯	0	1	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
6	2	0	1	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
7	2	0	1	⋯	0	1	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
8	2	0	1	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
36	9	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	0	⋯	0
37	0	0	0	⋯	0	0	⋯	1	⋯	0	1	⋯	0	0	0	⋯	0
38	0	0	0	⋯	0	0	⋯	0	⋯	0	1	⋯	0	0	0	⋯	0
39	0	0	0	⋯	0	0	⋯	0	⋯	0	1	⋯	0	0	0	⋯	0
40	0	0	0	⋯	0	0	⋯	0	⋯	1	1	⋯	0	0	0	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
99	0	0	0	⋯	0	0	⋯	1	⋯	0	0	⋯	1	0	0	⋯	0
100	0	0	0	⋯	0	0	⋯	1	⋯	0	0	⋯	1	0	0	⋯	0
101	0	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	1	0	⋯	0
102	0	0	0	⋯	1	0	⋯	0	⋯	0	0	⋯	0	1	0	⋯	0
103	0	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	1	0	⋯	0
104	0	0	0	⋯	0	0	⋯	0	⋯	1	0	⋯	0	1	0	⋯	0
105	0	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	1	⋯	0
106	0	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	1	⋯	0
107	0	0	0	⋯	0	0	⋯	0	⋯	0	0	⋯	0	0	1	⋯	0
108	0	0	0	⋯	0	1	⋯	0	⋯	0	0	⋯	0	0	1	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
152	0	0	0	⋯	1	0	⋯	1	⋯	0	0	⋯	0	0	0	⋯	0
153	0	0	0	⋯	0	0	⋯	0	⋯	1	0	⋯	0	0	0	⋯	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
200	0	0	0	⋯	0	1	⋯	0	⋯	0	0	⋯	0	0	0	⋯	1
N_{M_k}		4	4	⋯	4	4	⋯	4	⋯	4	4	⋯	4	4	4	⋯	4

Table 4.5: **This matrix represents an example m_{kp} matrix** for 200 metabolites (rows) and 70 pathways (columns) with pathway size $N_{M_k} = 4$. The metabolite number, k, and the distribution from which it was drawn, $h \in (0, 1, \ldots, 9)$, are listed to the left of the indicator matrix.

simulated data sets are presented in Figures 4.6 and 4.7. That is, one point on the boxplot represents the average error rate for that test in one simulated dataset.

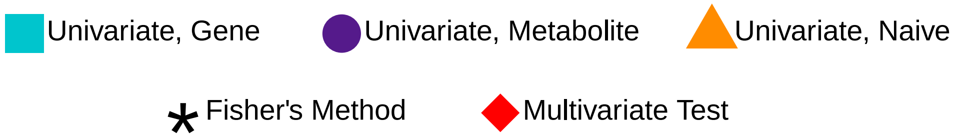


Figure 4.3: **The symbol key** for Figures 4.4 – 4.7.

Looking at the univariate tests in Figure 4.4 (univariate gene, blue square; univariate metabolite, purple circle) we see the behavior that we would expect for the different test styles. The tests are able to detect pathways 1–9 according to their enrichment. Genes but not metabolites are detected for pathways 4–6. Metabolites but not genes are detected for pathways 7–9 but not perfectly in the competitive Fisher's exact and logistic regression tests (panels A and B). This is likely due to the small pathway size, $N_{M_p} = 4$ in this simulation. In fact the Fisher's exact test loses all ability to detect metabolite enrichment beyond 100% enrichment. The logistic regression model has about 50% detection of the metabolite pathways when they are 50% enriched, pathways 10–12 and 16–18. This distinction is likely due to the loss of information in the Fisher's exact test due to dichotomization of genes and metabolites as differential or not. The logistic regression model does not require this and can make use of even marginal effect sizes. The self-contained sum-of-squared statistic (panel C) test only begins to have trouble detecting the metabolite enrichment at 25% enrichment, pathways 19-21.

When we turn our attention to the multivariate methods we see that the multivariate tests (red diamonds) show a similar pattern to the results from p-values combined by p-value sum (i.e., Fisher's method; black stars). For the competitive

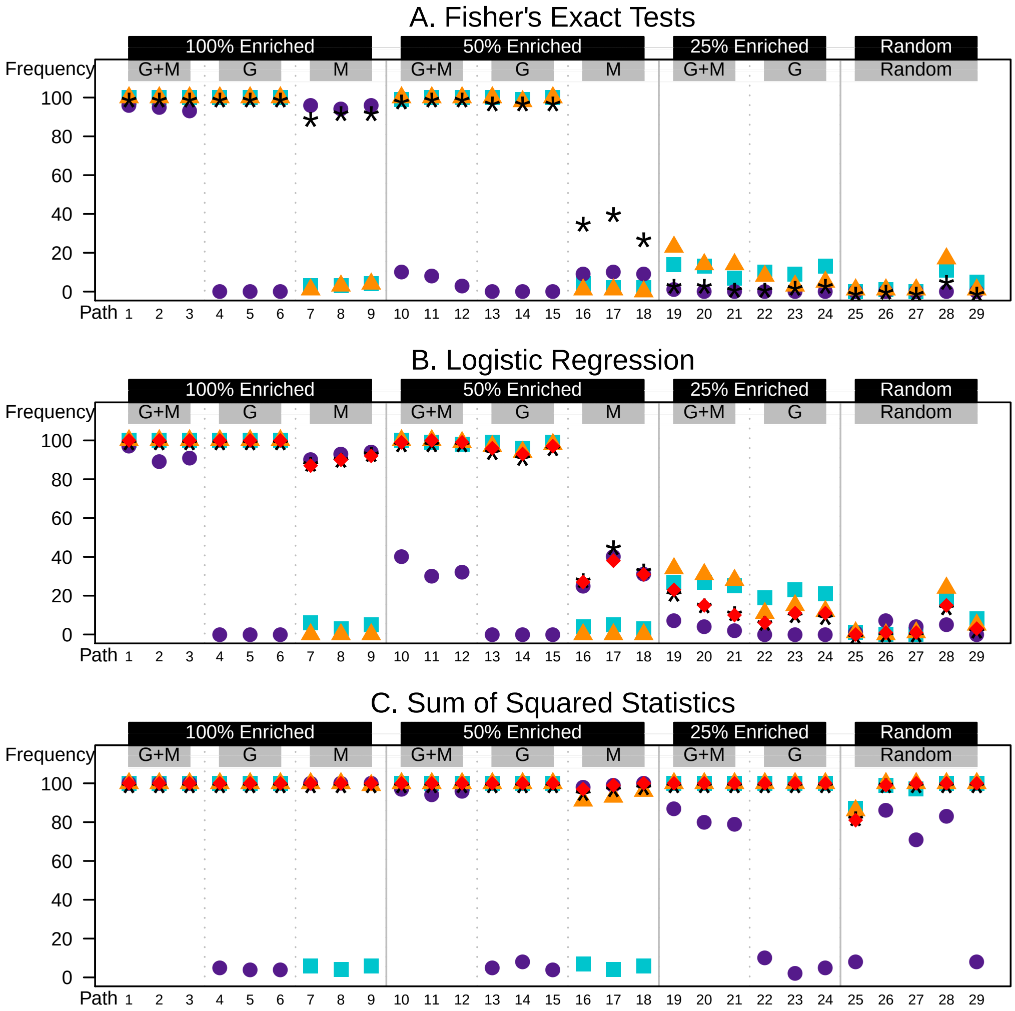


Figure 4.4: **Simulation results for 1000 genes and 200 metabolites generated for 30 samples.** Seventy pathways are assumed with pathways 31–70 representing null pathways. The correlation values are $\rho_{GG} = \rho_{MM} = 0.20$, and $\rho_{MG} = 0.10$. Four metabolites and 20 genes are included in each pathway. The symbols represent the frequency of rejecting the null hypothesis in 100 simulated datasets. The symbol key can be found in Figure 4.3.

tests these two methods tend to follow the gene expression data results. There is some improvement in the metabolite only pathways number 16–18. For the logistic regression test they also show a moderate effect, between that of the gene and metabolite only tests for pathways 19-24 (panel B). All methods appear to perform maximally in the self-contained sum-of-squared statistics test. Increased power is even provided to pathways 19-21 (panel C).

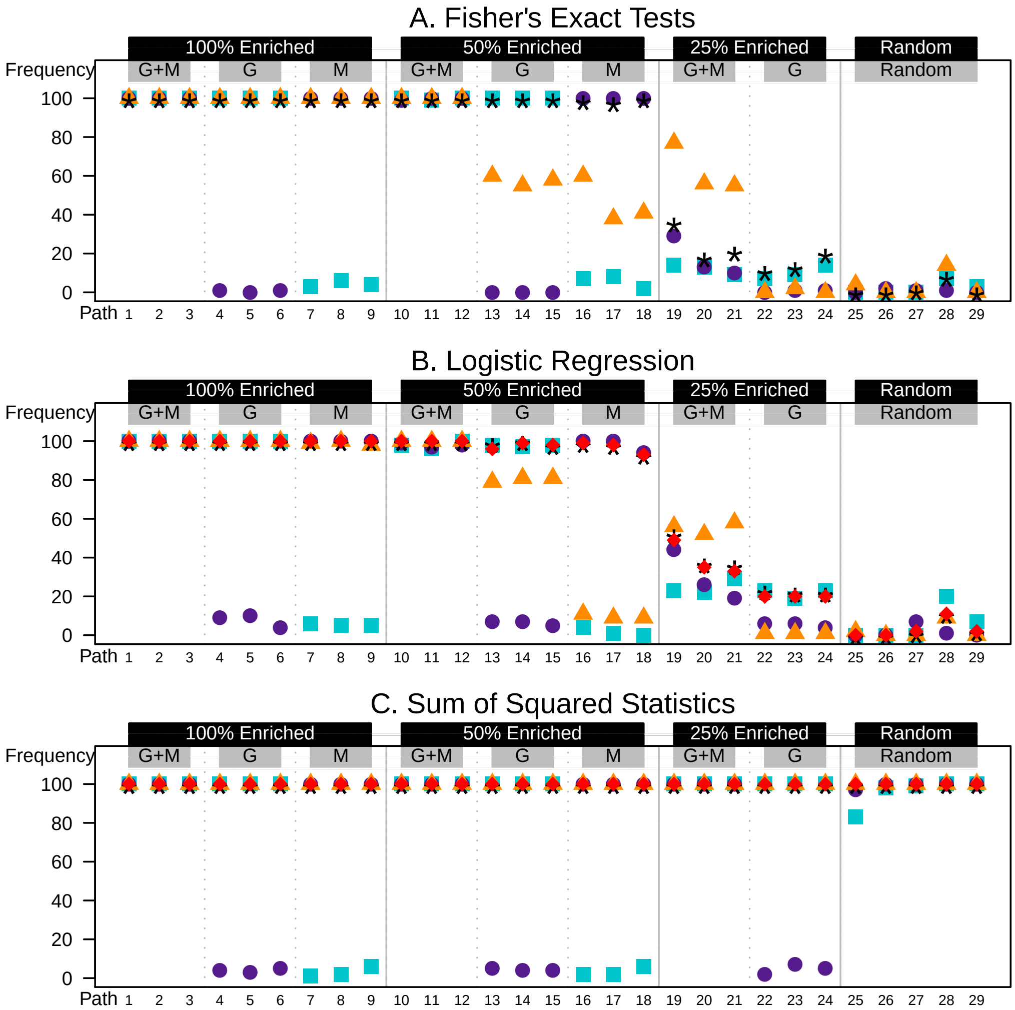


Figure 4.5: **Simulation results for 1000 genes and 1000 metabolites generated for 30 samples.** Seventy pathways are assumed with pathways 31–70 representing null pathways. The correlation values are $\rho_{GG} = \rho_{MM} = 0.20$, and $\rho_{MG} = 0.10$. Each pathway includes 20 metabolites and 20 genes. The symbols represent the frequency of rejecting the null hypothesis in 100 simulated datasets. The symbol key can be found in Figure 4.3.

In Figure 4.5 the metabolites now have large pathway memberships, $N_{M_p} = 20$, to be more comparable to the gene pahtways of size $N_{G_p} = 20$. The metabolite intensities are still generated with lower effect size than the gene expression values. What stands out in this figure are the orange triangles representing the concatenated data. In Figure 4.4 the $N_{gene} = 1000$ dataset dominated the $N_{metabolite} = 200$ dataset in the concatenated list. This is noticable by the orange triangles closely following the pattern of the blue squares of the gene-only analysis. However, in Figure 4.5, $N_{gene} = N_{metabolite} = 1000$ so the concatenated list is showing more mixed results. For the Fisher's exact test (panel A) the strength of a single enriched platform gets muddied by the non-enriched platform as in the gene-only and metabolite-only pathways numbered 13–18. Additionally, the pathway detection frequency is improved for the 25% enrichment pathways numbered 19-21, showing rates exceeding either single platform method. For the logistic regression tests (panel B) the concatenated data is still more closely related to the gene expression data. This is likely due to the higher effect sizes of the gene expression data compared to the metabolite data. The other combined p-values (black star) and 2-df Wald test (red diamond, panel B) appear to be improved for the low enrichment case of 25% enrichment for genes and metabolites, pathways 19-21. This shows that the joint enrichment methods are useful in these marginal cases.

The sum-of-squared statistics test (panel C) continues to perform maximally for all tests. One other distinction between this self-contained method and the two competitive tests (panels A and B) can be seen in pathways 25–29. These five pathways were determined by random selection. Given that 12% of the genes and metabolites are simulated to be differential in the full dataset these five pathways will have 12% enrichment on average. These pathways are not detected by the competitive tests because the null hypothesis for the competitive test can be written as H_0^{comp}: S has as many differential genes as if they were drawn by chance from the set of all genes, $S \cup S'$.

To ensure that our power gains are real we must also look at the error rate of these tests. Figures 4.6 and 4.7 are the corresponding error plots for Figures 4.4 and 4.5. The boxplots reflect 100 estimates of the average error rate across all 41 null pathways. The striking feature of both plots is the high error rates for the naïve concatenation of the data when using the Fisher's exact test (panel A). This error comes from a high rate of depletion calls. In essence these pathways are detected as having too few differential elements. This problem is amplified when the pathway size increases as in Figure 4.7. Given that there are now 40 elements per pathway in the concatenated list, zero differential elements is significantly smaller than the 12% expected by random selection.

The combined p-values also show inflated error rates for the competitive tests (panels A and B) when the larger pathway size is considered, $N_{M_p} = 20$. This may also be a symptom of detecting depleted pathways. Recall that in these competitive tests the sample size for the test is based on the number of elements. The larger pathway size may offer stronger depletion results that are then amplified by the joining of the two tests.

We do not observe error inflation in the sum-of-squares statistic methods (panels C). Firstly, the p-value is calculated for a one-sided test. Thus, as currently defined, the sum of squares test cannot detect depletion. Second, the p-values are determined by permutation so there is a limit on the level of precision for the p-values which thus limits the precision of the p-value as combined by summation in the Fisher's method.

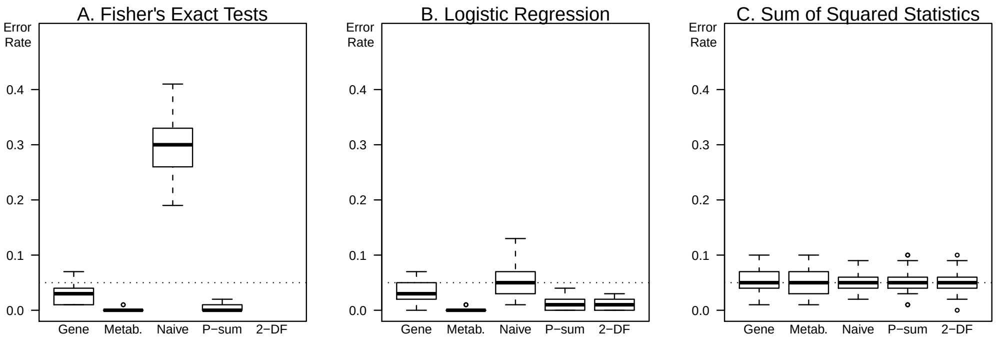


Figure 4.6: **Type I error associated with Figure 4.4.** Each boxplot represents 100 measurements of the error rate across the 41 null pathways.

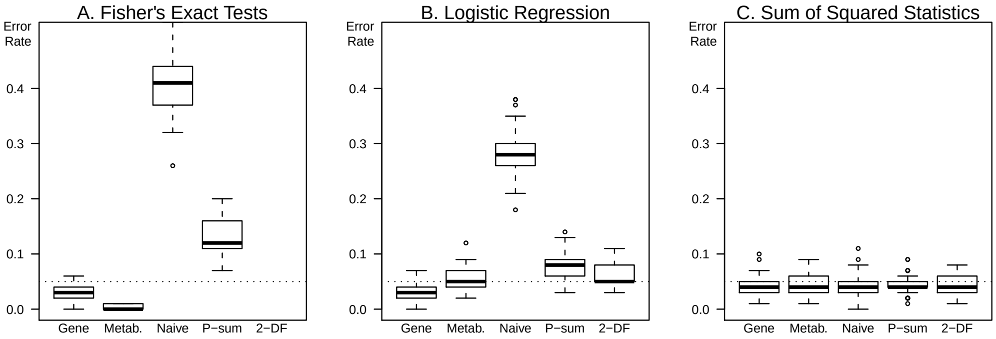


Figure 4.7: **Type I error associated with Figure 4.5.** Each boxplot represents 100 measurements of the error rate across the 41 null pathways.

4.5.2 Disjoint pathway simulation results

Now that we have the general pattern of operation for each of these methods let us explore some specific hypotheses using the disjoint simulations of Section 4.4.1. Recall that in these simulations we generate data for 50 disjoint pathways, of which 10 are designed to be enriched. The correlation structure is homogeneous in that each pathway has the same structure. However, there is no correlation simulated between pathways.

Here we use a different metric to assess the results of the methods. Specifically, we ask, if we were to choose the top ten pathways by ranking p-values, would we select the 10 associated pathways? Instead of looking at frequencies of being in the top 10 we consider the sum of the ranks for the 10 associated pathways. When the 10 associated pathways form the top 10 pathways selected the sum of the ranks is $R = \sum_{x=1}^{10} x = 55$. When there is no association between the pathway and disease then the 10 pathways of interest should have a sum of the ranks distributed as $R \sim Unif(55, 455)$ and we would thus expect R to fall near $E(R) = 255$.

Under the null model of no enrichment, that is $d_1 = d_0 = c_1 = c_0 = 0$, the rank sum of the associated pathways fall nicely around $E(R) = 255$; see Figure 4.8. Under the null model of uniform enrichment, that is $d_1 = d_0 = c_1 = c_0 = \delta$ we also see that the rank sum of the associated pathways matches $E(R) = 255$ when $\delta = 0.05$ as in Figure 4.9 and when $\delta = 0.10$ (data not shown).

To get a better understanding of the methods under specific scenarios we now consider some non-null simulations. For reference we begin with Figure 4.10. This simulation assumes that on average 25% of the elements in the associated pathways are differential, that is $d_1 = c_1 = 0.25$ and $d_0 = c_0 = 0$. This results in a 5% rate of differential elements within the datasets. We see that the sum-of-squares statistic,

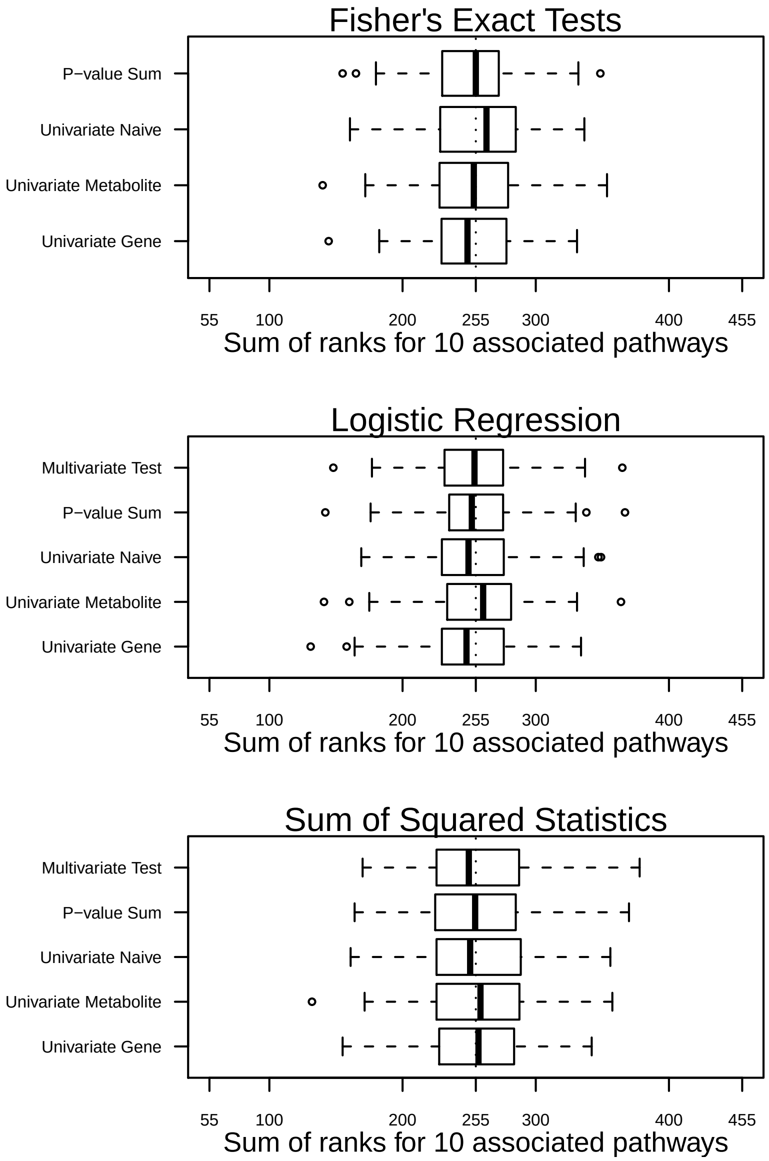


Figure 4.8: **The sum of the ranks of the 10 associated pathways assuming no differential elements** were measured, that is $d_1 = d_0 = c_1 = c_0 = 0$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.20$, $\rho_{MG} = 0.1$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample} = 30$.

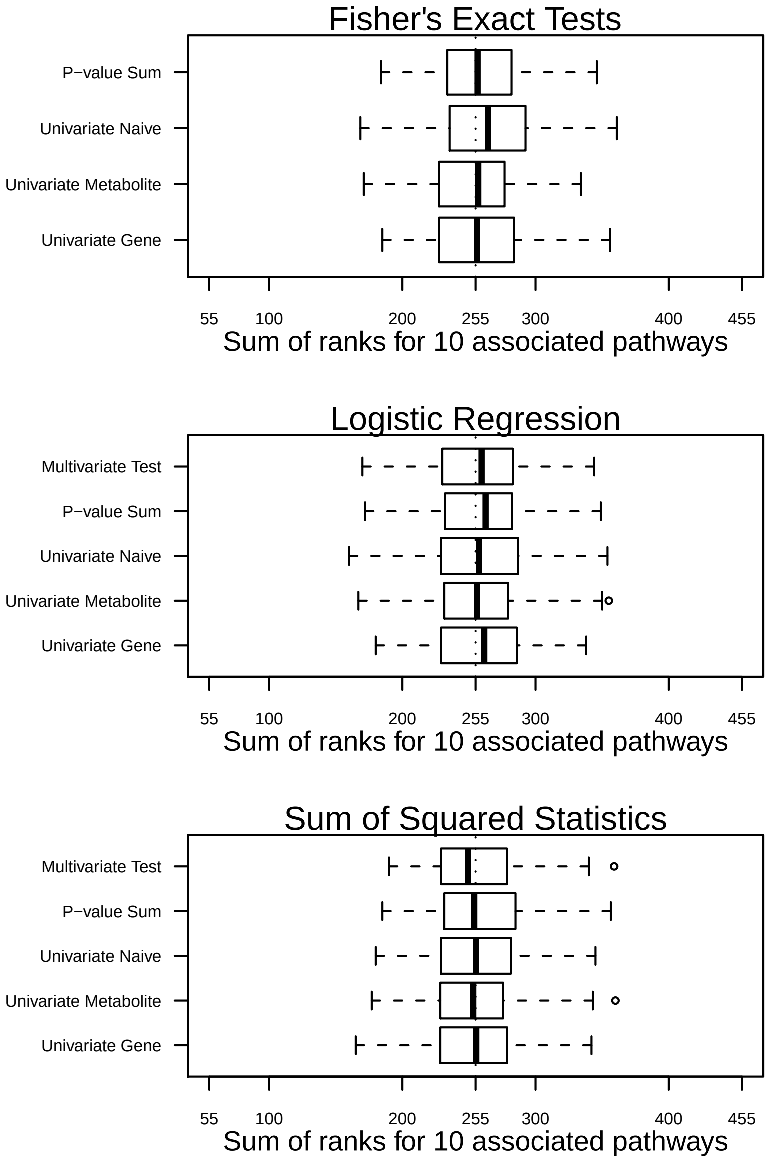


Figure 4.9: **The sum of the ranks of the 10 associated pathways assuming a constant probability of differential elements** across the pathways measured, that is $d_1 = d_0 = c_1 = c_0 = 0.05$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.20$, $\rho_{MG} = 0.1$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample} = 30$.

R, achieves nearly perfect rank sums for all tests. In the competitive tests we see improvement in R when any of the joint tests are used compared to the univariate gene or metabolite tests.

Under the higher correlation model shown in Figure 4.11, R is larger in the competitive tests compared to the lower correlation model of Figure 4.10. The loss of power is possibly due to loss of information attributed to the dependent measurements. However, it is not likely that such high correlations will be homogenously present in real applications [9, 66].

We next consider the behavior of the test in a scenario of few differential elements; see Figure 4.12. Here we set $d_1 = c_1 = 0.1$ and $d_0 = c_0 = 0$. The overall enrichment is 2% on average so the competitive tests still perform better than if the pathways were randomly assigned. Since d_1 and c_1 are probabilities we expect that on average 10% of the elements of the associated pathways are differential. It may be the case that one, or none of the elements are simulated to be differential. These low counts likely contribute to the increase in R for the sum-of-squared statistics in Figure 4.12.

Finally, we consider the model where $d_1 = c_1 = 0.25$ but $d_0 = c_0 = 0.05$, that is we simulate noise in the null pathways; see figure 4.13. It is in this scenario that the sum-of-squared statistic begins to falter. In fact we see that, beyond an increase in R, under this scenario the joint enrichment test performs more poorly than the univariate tests of the gene or metabolites alone. It is not surprising that this self-contained test performs poorly as this non-specific behavior is a criticism of self-contained method. It is surprising, however, that the joint methods appear to fare worse in this situation. However, notice that R is still less than $E(R) = 255$.

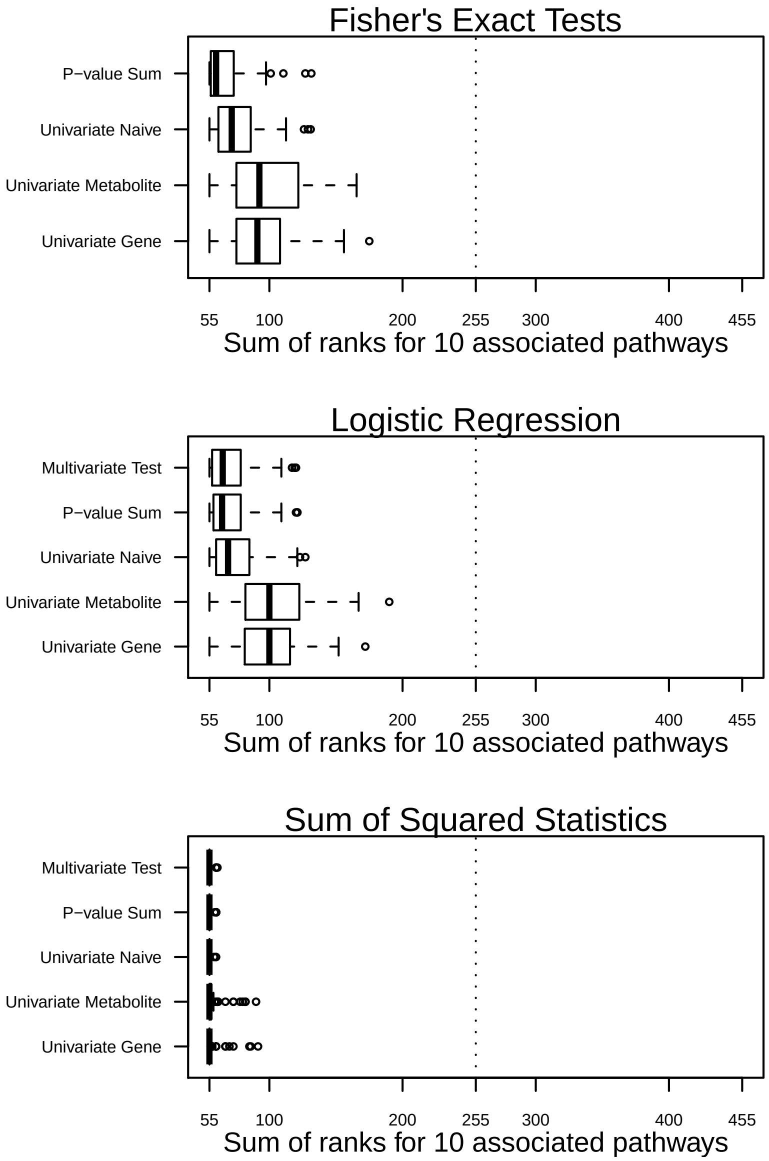


Figure 4.10: **The sum of the ranks of the 10 associated pathways assuming 25% of the elements are differential** in these pathways, that is $d_1 = c_1 = 0.25$ and $d_0 = c_0 = 0$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.20$, $\rho_{MG} = 0.1$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample} = 30$.

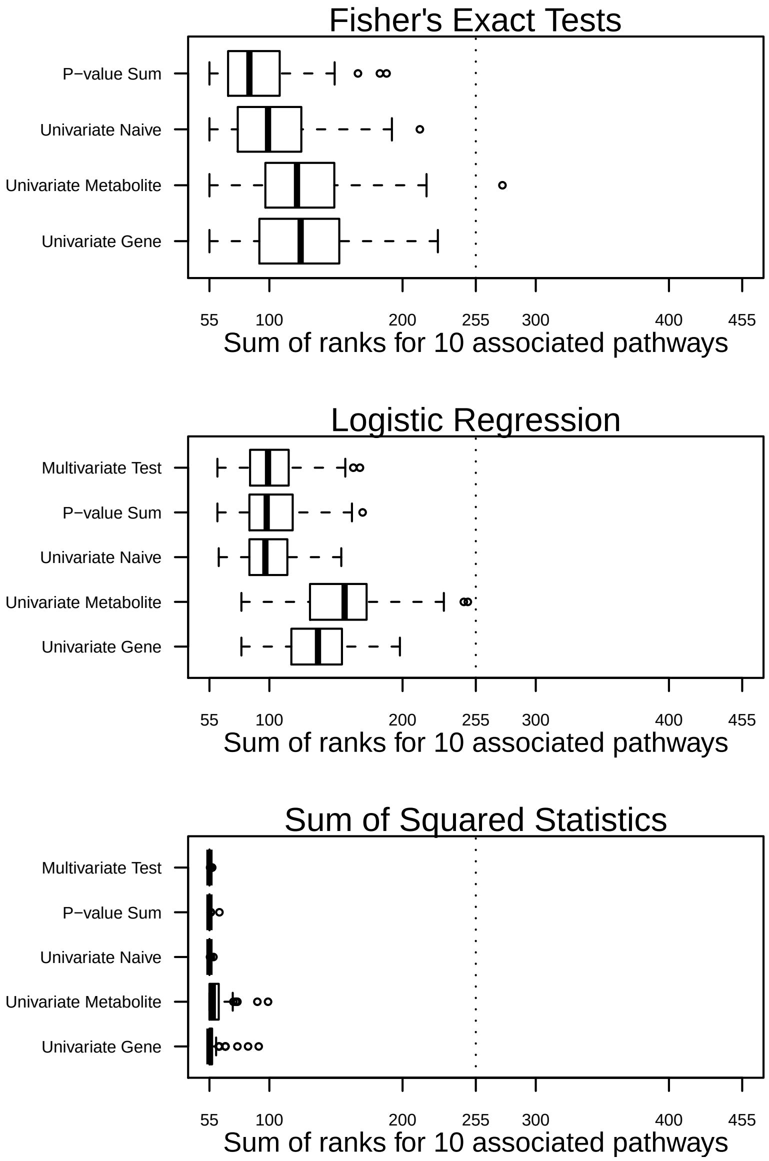


Figure 4.11: **The sum of the ranks of the 10 associated pathways assuming 25% of the elements are differential** in these pathways, that is $d_1 = c_1 = 0.25$ and $d_0 = c_0 = 0$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.60$, $\rho_{MG} = 0.25$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample=30}$.

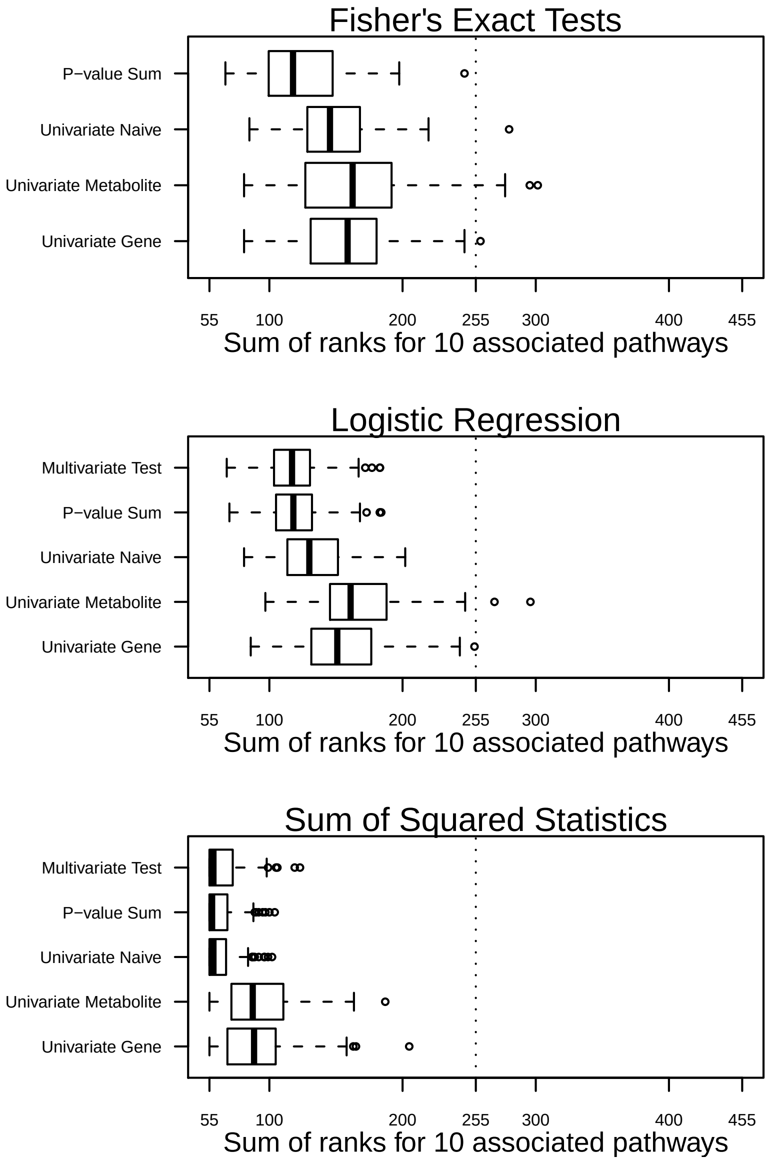


Figure 4.12: **The sum of the ranks of the 10 associated pathways assuming 10% of the elements are differential** in these pathways, that is $d_1 = c_1 = 0.10$ and $d_0 = c_0 = 0$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.20$, $\rho_{MG} = 0.1$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample} = 30$.

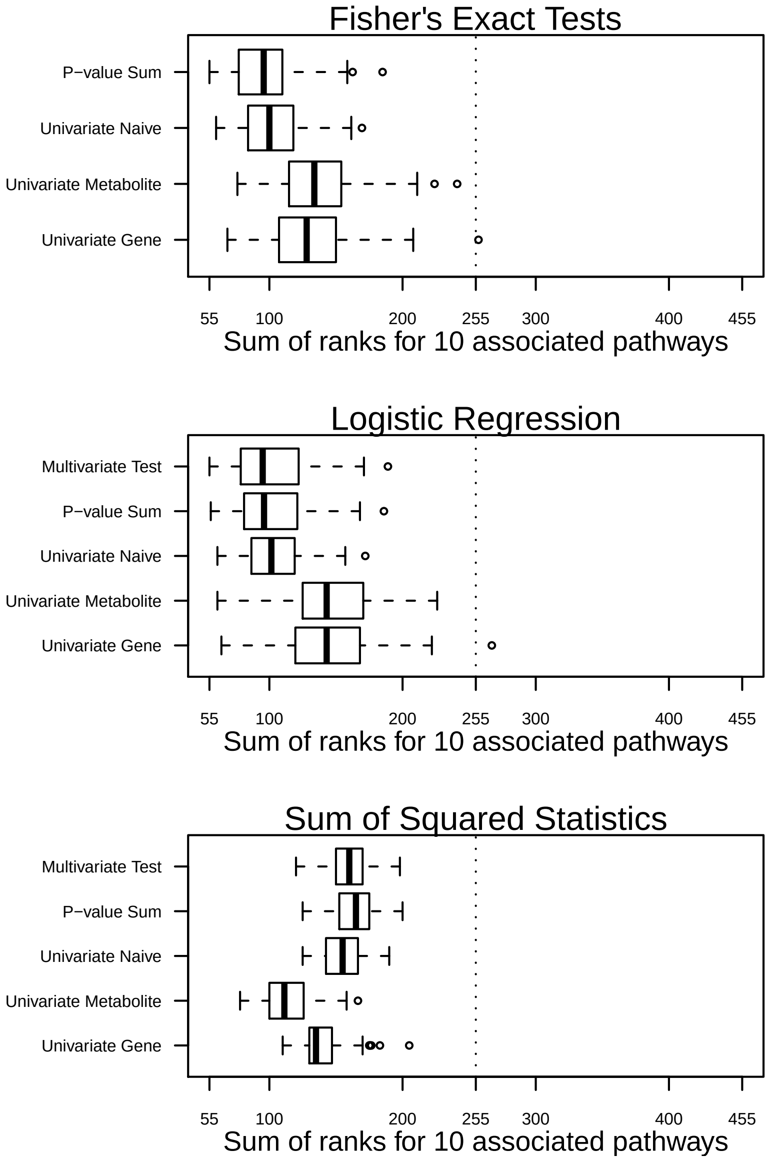


Figure 4.13: **The sum of the ranks of the 10 associated pathways assuming 25% of the elements are differential** in these pathways, that is $d_1 = c_1 = 0.25$ and 5% of the elements are differential in the remaining pathways, $d_0 = c_0 = 0.05$. The correlation structure $\rho_{GG} = \rho_{MM} = 0.20$, $\rho_{MG} = 0.1$ is assumed. Each of the 50 disjoint pathway were simulated to contain 20 metabolites and 20 genes. $N_{sample} = 30$.

4.6 Application to prostate metabolomics and transcriptomic data

We apply these enrichment methods to the metabolomic data of Sreekumar et al. (2009) [60] and the gene expression data from the same samples (unpublished). There are 40 samples in this dataset; 16 adjacent benign prostate tissue, 12 localized prostate tumors, and 12 metastatic prostate tumors. We consider the comparison between localized tumor and benign tissues. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG, version 50, April 2009) to determine the pathway mapping. Of 518 well measured metabolites, there are 147 metabolites that are named and can be mapped to the KEGG pathways. Of the over 40,000 gene probes measured on the Agilent Whole Human Genome microarray, there are 2169 genes that can be mapped to KEGG. To prevent overcounting, probes representing the same gene are averaged so that each gene is represented only once. There are 98 pathways in which at least one gene and one metabolite are measured; see Appendix A.

Each of the enrichment methods is run on this data. As this is experimental data, we do not know the true association of the genes and metabolites with the KEGG pathways. To assess our results we compare the findings of each method. Figure 4.14 shows a selection of these comparisons. Here we consider the number of pathways detected to be enriched at $p < 0.05$ using the logistic regression model. Additionally the 15 pathways selected by at least one enrichment test are listed in Table 4.6. Besides significance of the enrichment test (Y/N) the rank of the pathway by each test is given in parentheses.

Considering the Venn diagrams of Figure 4.14 we see that only combining p-values via p-value summation (i.e., Fisher's method; panel ii) detects a pathway not already detected by one or more of the univariate methods (panels ii – iii). However,

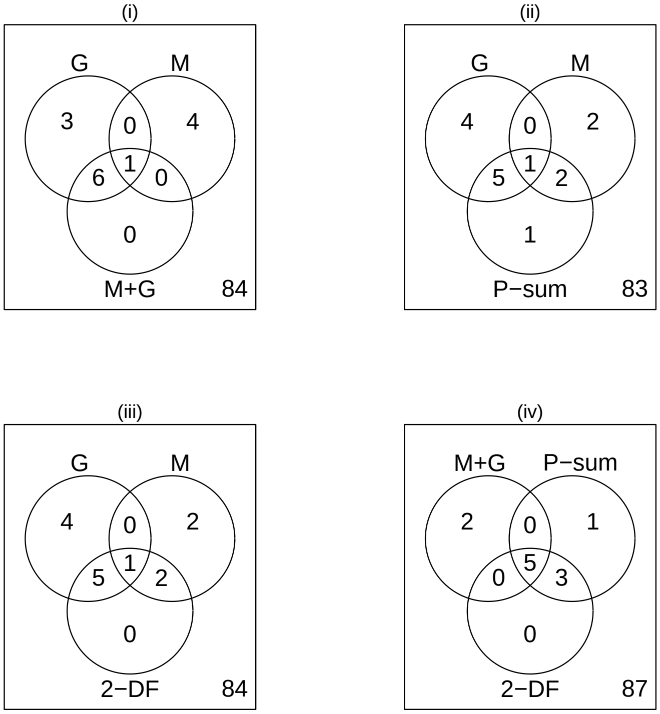


Figure 4.14: **Venn diagrams comparing enrichment methods** Panels i, ii, and iii compare the univariate methods to each of the joint enrichment tests for the logistic regression model; (i) Univariate naïve, (ii) p-value sum, (iii) 2-df Wald test. Panel iv compares the three joint tests.

Table 4.6: **The pathways identified by enrichment testing** These 15 pathways were identified by at least one of the five enrichment tests run on the logistic regression analysis with the threshold $\alpha \leq 0.05$. The pathway ranks are presented in parentheses. The number of metabolites N_{M_p} and genes N_{G_p} measured per pathway are provided for reference.

Pathway	N_{M_p}	N_{G_p}	M	G	$M+G$	P-sum	2-DF
ABC transporters	4	60	Y (1)	N (15)	N (67)	Y (4)	Y (4)
Neuroactive ligand-receptor interaction	2	4	Y (2)	Y (1)	Y (2)	Y (1)	Y (1)
Nitrogen metabolism	12	140	Y (3)	N (85)	N (43)	N (13)	N (11)
Aminoacyl-tRNA biosynthesis	6	9	Y (4)	N (67)	N (36)	N (12)	N (13)
Arginine and proline metabolism	5	22	Y (5)	N (19)	N (8)	Y (8)	Y (8)
Autoimmune thyroid disease	1	40	N (32)	Y (2)	Y (1)	Y (2)	Y (2)
Asthma	1	26	N (73)	Y (3)	Y (3)	Y (5)	Y (5)
Biotin metabolism	26	38	N (46)	Y (4)	N (9)	Y (7)	Y (6)
Taste transduction	2	43	N (26)	Y (6)	Y (6)	Y (6)	Y (7)
Purine metabolism	9	205	N (6)	Y (5)	Y (4)	Y (3)	Y (3)
Fc epsilon RI signaling pathway	2	71	N (90)	Y (7)	Y (5)	N (14)	N (12)
Renal cell carcinoma	2	67	N (85)	Y (8)	Y (7)	N (18)	N (15)
Valine, leucine and isoleucine biosynthesis	11	31	N (98)	Y (9)	N (14)	N (20)	N (16)
Glycerophospholipid metabolism	17	31	N (23)	Y (10)	N (10)	N (10)	N (10)
Fatty acid biosynthesis	6	5	N (9)	N (11)	N (86)	Y (9)	N (9)

the joint models provide a more refined list of pathways compared to using the union of the results of the two univariate methods. It may be preferable to consider those pathways with a significant joint association as preferred candidates for follow-up. Panel (iv) of Figure 4.14 compares the results of these three joint enrichment methods. We see that in this situation the sum of the p-values by Fisher's method selects nearly the same pathways as the 2-df Wald test. Since we are not assuming correlation between γ and μ the 2-df Wald test is simply a sum of the univariate Wald statistics so its behavior should be similar to the sum of $-log(p_\gamma)$ and $-log(p_\mu)$ as in the Fisher's method. The similarity of these methods is also seen in their similar pathway rankings given in Table 4.6.

Though we do not know the true result enrichment state in these observational data it is interesting to consider the pathways listed in Table 4.6 according to current knowledge of prostate cancer. First, in Sreekumar et al. (2009) [60] pathways of amino acid metabolism and nitrogen metabolism were identified as enriched. This is supported by our analysis here with both the nitrogen metabolism and the arginine and proline metabolism pathways detected by the metabolites alone. Valine, leucine and isoleucine biosynthesis, another amino acid metabolism pathway, is detected by the genes only but not by the joint tests.

The Fisher's exact test methods behaved similarly to the logistic regression tests shown in Figure 4.14. The sum-of-squared statistics tests were overly liberal identifying over 90% of the pathways as enriched. This implies that there was a high rate of differential elements throughout the datasets similar to the scenario of Figure 4.13. Such background noise makes a competitive test the preferred choice of enrichment test. Additionally this may suggest that the KEGG pathway maps, as applied, may not accurately capture the co-regulation in the data.

4.7 Conclusion

In this work we have considered the application of two-dimensional set enrichment testing methods for the joint analysis of transcriptomic and metabolomic datasets. We consider two novel methods: the logistic regression 2-degree of freedom Wald test and the 2-dimensional permutation p-value for the sum-of-squared statistics test. Through simulation we explored the properties of these tests in relation to their univariate counterparts and two simplistic joining methods, namely data concatenation and the Fisher's method for combining p-values. We find that the joint tests can improve our ability to detect results that are marginal univariately; see Figures 4.4 and 4.5. We also find that joint tests improve the ranking of associated pathways compared to their univariate counterparts; see Figures 4.10 and 4.11.

The various joint methods performed similarly for most simulations. The concatenation of datasets and the Fisher's method of combining p-values had inflated error in the competitive test; see Figure 4.7. For the logistic regression test, the 2-df wald test currently peforms similarly to the Fisher's method for combining the two p-values. This is likely due to the assumption that $\rho_{\gamma\mu} = 0$ in the 2-df test. Non-zero correlation would have provided a weighted sum in the 2-df test. Though we were not estimating $\rho_{\gamma\mu}$ to be non-zero, the slightly inflated error rate of the Fisher's method test, see Figure 4.7, suggests that the independence assumption may not always hold. In future work we will continue to explore if and when correlation may be a contributing factor or if there are other methods for combining the tests in a weighted fashion either at the level of the test statistic or p-value. Though perhaps the most commonly known, Fisher's method for combining p-values is only one of many methods available [41].

One of the more attractive features of the 2-df Wald test and 2-dimensional permutation test is that they can easily be extended to n-dimensions. This will allow for the incorporation of multiple omics platforms such as proteomics, genomics, or gene copy number. The data concatenation and sum of p-values methods can also be extended, but this may compound their potential error.

CHAPTER V

Conclusions and future work

In this work we have explored three different avenues of omics integration important to the area of cancer research; (1) the classification of samples, (2) biomarker discovery, and (3) systems biology.

In Chapter II we utilized a classification method that allowed us to utilize a differential list of elements from a prior study to make prognostic or diagnostic predictions about samples in a current study. We extended the classification method by providing a testing scenario for the classifier. Though originally motivated by the integration of in vitro and in vivo gene expression datasets we showed that it can be applicable across omics platforms as well. We demonstrated our result on the metabolomic and matched gene expression data of Sreekumar et al. (2009) [60]. We demonstrated that the gene expression profile could be used to distinguish tissue diagnosis using metabolomic intensities. Though the diagnosis of cancer in prostatectomy tissues is not of clinical importance this same method could be used to derive classifiers for biofluids in which classifiers are difficult to build because the true diagnostic state of the patient is not fully realized.

In Chapter III we explored the use of p-value weighting to improve the power of per-metabolite tests of differential intensity. Metabolites have the potential to be good biomarkers for screening, tracking disease progression, or for drug targeting [38]. However, metabolite levels can be affected by diurnal rhythms, diet, medications, and other illnesses leading to noisy data and reduced effect sizes when comparing populations [38]. We used the gene expression information per-metabolic pathway to devise pathway-based weights. In this way, metabolites that are involved in a pathway that is disregulated in its gene expression are given higher importance. With many publicly available gene expression studies and the robustness of microarray results we felt that this was an appropriate source of prior information. However, the simulations were not platform-specific, and the recommendations in the chapter can be applied to other omics platforms. In the future we would also like to consider reaction-based models for the weights. A reaction-based approach will allow us to derive a single weight per metabolite thereby removing the summarization step currently needed when a metabolite participates in multiple pathways.

Finally, in Chapter IV we adopted a systems biology perspective to search for sets of genes and metabolites that are coordinately differential. We extended two univariate set enrichment tests to jointly test the gene expression and metabolite data results. These tests readily expand to N-dimensions and may provide a means for simultaneously testing a series of omics platforms. Further work will need to be done to assess correlations between multiple omics pathways for the logistic regression model as additional platforms are likely to contribute some redundant information. Use of the score functions for the logistic regression model may prove helpful here. Additionally, the simple meta-anaysis approach of the Fisher's method for combining p-values showed promise but any anti-conservative tendencies due to correlated tests will be increased as more tests are added. Thus, going forward, we will consider other meta-analysis techniques that either account for, or are robust to, correlations between tests.

Expanding on the work of this dissertation we see the potential for employing a Bayesian approach. Hierarchical modelling may be used to assess importance of the per-metabolite hypotheses from corresponding gene expression results as a Bayesian corollary to the p-value weighting of Chapter III, lend themselves to hierarchical modelling. In fact, Genovese et al. (2006) [20] allude to this extension in the future work section of their paper. Concern arises with the minimal level of correlation expected in a majority of the gene-metabolite pairs. However, the classificatory ability of the gene signature on metabolite intensities in Chapter II strengthens the idea that gene expression data could be used to develop a prior distribution for the metabolomic data. Additionally, empirical Bayes approaches could possibly be extended [16] in lieu of a fully Bayesian approach.

Additionally, an interesting experimental design question is whether or not to assess the various omics on matched samples. Matching samples in a case-control type comparative analysis certainly has its benefits since we can compare the case to a measure of itself in a non-diseased state. However, the potential benefits of matching samples are less obvious in omics integration. Some methods such as correlative analysis are not possible without matched samples. However, methods such as the enrichment tests of Chapter IV are less likely to be affected by sample matching since summary measures of sample differentiation per element, e.g. two-sample t-statistics, are used at the point of data integration. The variety of simulation models developed in this thesis lend themselves to an exploration of this topic.

APPENDICES

APPENDIX A

Mapping genes and metabolites in the Kyoto Encyclopedia of Genes and Genomes (KEGG)

To integrate the datasets we use the pathways maps of the Kyoto Encyclopedia of Genes and Genomes (KEGG, www.genome.jp/kegg), [34, 36, 35]. These include metabolic pathways, signalling pathways, and disease associated pathways. KEGG has 200 pathways that are attributed to *Homo sapiens* (HSA). The pathway maps are drawn to reflect the current literature at the time of the build (v. 50, April 2009). Of these, 165 pathways contain at least one compound ID. Gene information can also be extracted from the pathways. All 200 HSA pathways include at least one gene.

A.1 Gene mapping

There are 4686 unique geneIDs associated with the HSA pathways. Each gene ID was associated with a single gene name except for seven that were associated with two gene names (e.g., MYL10 and alias MYLC2PL). The gene IDs were mapped onto the gene expression data using gene symbol as the index variable. The gene expression measures of Varambally et al. (2005) [68] were from an Affymetrix HU133Av2 genechip and the gene symbols were obtained from GEO (Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo, August 2009) using the GPL570 platform information file. Of the 54675 probes on this array, 9491 probes mapped to a KEGG gene ID. This translates to 4240 unique KEGG gene IDs represented on the array. Of the gene IDs represented, 39.8% were represented uniquely (see Table A.1). The remainder were measured by multiple probes. The matched gene expression measures for Sreekumar et al. (2009) [60] (unpublished data) were from an Agilent Whole Human Genome Oligo Microarray and the gene symbols were obtained from GEO platform information file number GPL1708 (March 2009). Of the 41675 non-control probes on this array, 6566 probes mapped to a KEGG gene ID. This translates to 4161 unique KEGG gene IDs represented on the array. Of the gene IDs represented, 59.9% were represented uniquely (see Table A.1). The remainder were measured by multiple probes.

Each of the 200 HSA pathways is represented on the Affymetrix array. The Aglient array represents 198 of the pathways, excluding only “Fluorobenzoate degradation” (pathway hsa:00364) and “1,4-Dichlorobenzene degradation” (hsa:00627). In Table A.2 we consider the number of pathways to which a gene ID contributes to determine the ammount of overlap between pathways. On each array approximately 82% of the genes are represented in three pathways at most. However, as almost half (approximately 45%) of the genes are found in multiple pathways and this may need attention in pathway based methods.

A.2 Metabolite mapping

There were 3076 unique compound IDs found among the HSA pathways. All possible naming conventions for the molecule associated with the compound ID were extracted from its description in KEGG (e.g. cpd:C00001 is named “H2O” or “Water”). The compound name was used to map the KEGG compound ID numbers onto the metabolomics dataset [60]. The data were merged directly by compound name and then additional curration was done by hand to resolve inconsistencies in nomenclature. One hundred and eighty-seven (of 626) compounds were named in the Sreekumar dataset; the remainder were labeled as unknown. Of these 187 named compounds, 13 were not found in KEGG and 3 were found only in non-HSA pathways. Eleven metabolites had a KEGG compound ID that was not associated with any pathway. Approximately 15 metabolites could be mapped only if either an isomeric form was chosen (e.g. cpd:C00303, Glutamine, is not mapped whereas cpd:C00064, L-Glutamine, is associated with eight pathways) or an alternate KEGG ID was used (e.g. cpd:C15571, Catechol (generic), is not mapped whereas cpd:C00090, Catechol, is mapped). In total 160 named metabolites can be mapped to at least one HSA pathway in KEGG. For analysis we consider only 147 of these 160 mapped metabolites, excluding 13 because they were poorly measured (see Sreekumar et al, 2009).

The 147 compound IDs map to 100 pathways. Table A.3 shows the number of pathways in which a compound ID is found. Only 38.8% of the compounds are represented in a single pathway. As metabolites for one pathway likely feed into other pathways this should not be surprising. Again, this overlap may need attention in any pathway based methods.

A.3 Integrative pathway mapping

There are up to 100 pathways for which both gene and metabolite measures are available. Since we are interested in the integration of gene and metabolite data we focus our attention on these 100 pathways. As at least 100 pathways are removed from consideration from the gene list we recalculate the pathway overlap for genes

(Table A.2) in Table A.4. As expected, the amount of overlap is reduced with a reduced pathway list.

Considering the 100 metabolite mapped pathways, Table A.5 shows the number of elements measured in each pathway. Not surprisingly the pathway population of Table A.5 is driven by the gene expression measures; see Table A.6. There are multiple metabolite measures made for a majority (75%) of the pathways as well, see Table A.7. However, this leaves 25 pathways in which only a single metabolite is measured. Pathway based methods may need to take this into consideration.

(i)

Number of Probes	Frequency	Percent
1	1168	39.8%
2	1180	27.8%
3	703	16.6%
4	344	8.1%
5	167	3.9%
> 5	159	3.8%

(ii)

Number of Probes	Frequency	Percent
1	2492	59.9%
2	1223	29.4%
3	318	7.64%
4	86	2.1%
5	12	0.3%
> 5	30	0.7%

Table A.1: **Frequency of mutliple probe measures.** (i) There are 4240 unique KEGG gene ID numbers represented by 9491 probes on the Affymetrix HU133Av2 array. (ii) There are 4161 unique gene ID numbers represented by 6566 probes. One geneID (hsa:6144, RPL21) is represented by 14 probes.

(i)

Number of Pathways	Frequency	Percent
1	2312	54.5%
2	739	17.4%
3	436	10.2%
4	279	6.6%
5	131	3.1%
6	82	1.9%
7	62	1.5%
8	48	1.1%
9	39	0.9%
10	25	0.6%
> 10	87	2.2%

(ii)

Number of Pathways	Frequency	Percent
1	2289	55.0%
2	717	17.2%
3	417	10.0%
4	273	6.6%
5	129	3.1%
6	81	2.0%
7	57	1.4%
8	46	1.1%
9	43	1.0%
10	25	0.6%
> 10	84	2.0%

Table A.2: **Pathway overlap for genes.** (i) Over half (54.5%) of the 4020 unique KEGG gene ID values on the Affymetrix array contribute to a single pathway. (ii) Over half (55.0%) of the 4161 unique KEGG gene D values on the Agilent array are associated with a single pathway. On both platforms, two genes (hsa:5594, MAPK1; hsa:5595, MAPK3) contribute to 33 pathways.

Number of Pathways	Frequency	Percent
1	57	38.8%
2	37	25.2%
3	14	9.6%
4	15	10.3%
5	5	3.4%
6	6	4.1%
7	4	2.7%
> 7	9	6.2%

Table A.3: **Pathway overlap for metabolites.** There are 147 compound ID numbers measured that can be associated with a KEGG HSA pathway. Less than half (38.8%)of these are associated with a single pathway, whereas, one metabolite (cpd:C00025, Glutamate) is associated with 18 pathways.

(i)

Number of Pathways	Frequency	Percent
1	1396	58.8%
2	472	19.9%
3	265	11.2%
4	109	4.6%
5	60	2.5%
6	19	0.8%
7	23	0.5%
> 7	31	1.3%

(ii)

Number of Pathways	Frequency	Percent
1	1346	58.6%
2	458	20.0%
3	255	11.1%
4	105	4.6%
5	64	2.8%
6	16	0.7%
7	22	1.0%
> 7	30	1.3%

Table A.4: **Overlap for genes in the pathways shared with metabolites.** (i) There are 100 pathways, represented by 2375 genes, shared by the metabolites and genes on the Affymetrix array. (ii) There are 99 pathways, represented by 2296 genes, shared by the metabolties and the genes on the Agilent array (pathway hsa:00364, “Fluorobenzoate degradation” is not represented). A slightly higher percentage of the genes are now associated with a single pathway (see Table A.2). On both platforms there is now one gene (hsa:218, ALDH3A1) associated with 16 pathways. The two MAPK genes, previously associated with 33 pathways, are reduced to only 11 pathways each.

(i)

Number of Elements	Frequency (of 100)
2 − 9	11
10 − 19	19
20 − 29	10
30 − 39	14
40 − 49	10
50 − 75	21
75 − 99	5
≥ 100	10

(ii)

Number of Elements	Frequency (of 99)
2 − 9	13
10 − 19	15
20 − 29	11
31 − 39	15
40 − 49	14
50 − 74	18
75 − 99	4
≥ 100	9

Table A.5: **Number of elements measured per pathway.** (i) Of the 100 pathways for which both metabolite and Affymetrix gene expression information are available 89 of have at least 10 elements. Two pathways contain only two measured elements – a gene and a metabolite. (ii) Of the 99 pathways for which both metabolite and Agilent gene expression information are available 86 have at least 10 elements and one pathway contains only two measured elements. On both platforms there are three pathways with over 200 elements in each (path:hsa04810, Regulation of actin cytoskeleton ((i) 206, (ii) 205), path:hsa04080, Neuroactive ligand receptor ((i) 263, (ii) 250), and path:hsa05200, Pathways in cancer ((i) 330, (ii) 323)

(i)

Number of Genes	Frequency (of 100)
1 − 9	17
10 − 19	16
20 − 29	14
30 − 39	12
40 − 49	15
50 − 74	14
75 − 99	4
≥ 100	8

(ii)

Number of Genes	Frequency (of 99)
1 − 9	18
10 − 19	14
20 − 29	13
30 − 19	14
40 − 49	14
50 − 74	14
75 − 99	3
≥ 100	9

Table A.6: **Number of genes measured per pathway.** (i) Two pathways contains only a single measured gene each from the Affymetrix data. (ii) One pathway contains only a single measured gene from the Agilent data. On both platforms the highly poplulated (> 200 genes) pathways include “Pathways in cancer” (path:hsa05200, (i) 327, (ii) 320), “Neuroactive ligand-receptor interaction” (path:hsa04080, (i) 254, (ii) 241), and “Regulation of actin cytoskeleton” (path:hsa04810, (i) 205, (ii) 204).

Number of Metabolites	Frequency (of 100)
1	25
2	20
3	9
4	7
5	12
6	10
7	3
8	3
9	2
10	1
> 10	8

Table A.7: **Number of metabolites measured per pathway.** Seventy-five percent of the pathways are represented by at least two metabolites with as many as 26 metabolites in one pathway (path:hsa02010, ABC transporters - General - Homo sapiens).

APPENDIX B

P-value weighting incorporating gene expression and metabolite information

The following is an extension of pathway-based weighting that we abandoned because the type I error rate was substantially inflated compared to the methods presented in Chapter III. Some results and discussion of this method are provided here for interested readers. It is assumed that readers are familiar with the material in Chapter III.

B.1 Introduction

An intriguing idea was presented by Roeder et al. (2007) [54] that uniform p-value weights be assigned to per-element tests according to an *a priori* defined grouping. The twist was that the elements being tested in a group, say k, would be used to define the weight for the group, say ω_k.

Consider a vector of elements $\mathscr{E}$ and a series of prior studies $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_K$ each of which results in a subset of $\mathscr{E}$ being selected, say $\mathscr{A}_i(\mathscr{E})$ for study i. We can then use results of these prior studies to subset $\mathscr{E}$ into disjoint sets, $\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_{K+1}$. To construct disjoint subsets of $\mathscr{E}$ Roeder et al. (2007) [54] suggest assigning set membership based upon a hierarchy of the prior studies. Thus if we assume that study $\mathscr{A}_1$ is most related to the current study then $\mathscr{E}_1 = \mathscr{A}_1(\mathscr{E})$. If study $\mathscr{A}_2$ is the second most interesting study then $\mathscr{E}_2 = \mathscr{A}_2(\mathscr{E}) \cap \mathscr{A}_1(\mathscr{E})^c$, where the set S^c denotes the complement of the set S. This subsetting is repeated for each of the prior studies and the remaining elements of $\mathscr{E}$ form the subset $\mathscr{E}_{K+1}$.

Since the prior studies are used only to define the grouping, the weights are derived from information from the elements of the set. That is for set k the pre-standardized weight $u_k = f(T_{\mathscr{E}_k})$, where $\underline{T}_{\mathscr{E}_k}$ is some vector of statistics from set $\mathscr{E}_k$ and $f(\cdot) > 0$ is a function of that information. This weight component u_k is assigned to every element in $\mathscr{E}_k$, $v_i = u_k$ for all $e_i \in \mathscr{E}_k$. Finally the weights are standardized, $w_i = v_i/\bar{v}$, where $\bar{v} = n(\mathscr{E})^{-1} \sum_{i=1}^{n(\mathscr{E})} v_i$ so that $n(\mathscr{E})^{-1} \sum_{i=1}^{n(\mathscr{E})} w_i = 1$.

A potential difficulty for using gene expression results to construct per-metabolite p-value weights is that the gene expression result could dominate the analysis. That is, the metabolite p-value may be upweighted or downweighed so heavily that the significance of the test is determined solely from the weight, e.g. $p_i^* = p_i/w_i > 1$. A weighting method in which the metabolites contribute information in the construction of the weights is therefore appealing.

Using a single gene expression study we could rank the metabolites by gene set enrichment of the KEGG pathways. That is $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_k$ would be a ranked list of differential gene enriched pathways. Alternatively, we can simply group the metabolites by pathway and use an average weight across overlapping pathways in the definition of the per metabolite weight as in Chapter III. For either grouping scheme, the drawback to using the Roeder et al. (2007) [54] grouped weighting method is that the number of metabolites in the resulting sets, $\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_{K+1}$, is likely to be low; see Appendix A for the number of measured metabolites per KEGG pathway in the Sreekumar et al. (2009) [60] metabolite data. The grouped weight of Roeder et al. (2007) [54] relies on the sieve principle to maintain error control and thus must have relatively large set membership; Roeder et al. recommend at least 20 elements per group. This group size requirement reduces that chance that a groups weight is dominated by a single element. Thus to utilize grouped weights in the context of metabolic pathways we propose an estimate for the weights that utilizes both gene and metabolite information.

B.2 The two-component weight

Prior work by Genovese et al. (2006) [20] used per-element weighting. Essentially this is a group size of one where the weight is determined solely by the prior data. The work of Roeder et al. (2007) [54] defines grouped weights using only the current data to define the weights. Thus we propose a two-component weight. Borrowing information from related metabolites and prior knowledge from gene expression data we consider a weight ω_k for metabolic pathway k defined as

$$\omega_k = \theta_k \omega_{1k} + (1 - \theta_k)\omega_{2k} \tag{B.1}$$

with ω_{1k} representing the metabolic component for pathway k ($\omega_1 \geq 0$) and ω_{2k} representing the component based on the gene expression analysis ($\omega_{2k} \geq 0$).

We consider two estimates of the mixing parameter $\theta_k \in [0, 1]$. First we estimate $\hat{\theta}_k = (\tilde{\eta}_k - 1)/\eta_k$ for pathway k. Here η_k is the number of metabolites with $\tilde{\eta}_k$ measured. This has the nice property that $\hat{\theta}_k = 0$ when $\tilde{\eta}_k = 1$ so that $\omega_k = \omega_{2k}$ which is the weight of Chapter III. However, the disadvantage to this estimate is that due to highly transient metabolites, η_k may never achieve $\tilde{\eta}_k$ and thus $\hat{\theta}_k$ will never be 1. This means that even large pathways that would satisfy the 20 element minimum of Roeder et al. (2007) [54] will contain gene information in the weights.

The alternative θ_k considered is $\hat{\theta}_k = 1 - |\rho_k|$ where $|\rho_k|$ is the absolute correlation between genes and metabolites within pathway k. The advantage of this estimate is that the gene component, ω_{2k}, will have little effect if the genes are not correlated with the metabolites. The disadvantage is that ρ_k is difficult to estimate. Additionally, pathways with minimal correlation will rely heavily on the metabolic component ω_{1k} regardless of the size η_k.

Finally, we define the components ω_{1k} and ω_{2k} of Equation B.1 as

$$\begin{aligned} \hat{\omega}_{1k} &= -\log_{10}(P_k^m) \\ \hat{\omega}_{2k} &= -\log_{10}(P_k^g) \end{aligned}$$

where P_k^m and P_k^g are the p-values from tests of pathway enrichment for metabolites and gene expression, respectively, in pathway k. In this way, $\omega_{1k} \geq 0$ captures the metabolite information in the pathway k and $\omega_{2k} \geq 0$ captures the gene expression information in that pathway. A separate ω_k is constructed for every pathway $k = 1, 2, \ldots, K+1$ for which there is gene and metabolite data available. Notice that the weight function A used in Chapter III can be written as $\omega_k = \omega_{2k}$.

To translate the pathway weight ω_k to the per-metabolite weight, say w_m, we first assign an unstandardized weight $v_m = \omega_k$ for each metabolite m within the pathway k. When a metabolite is present in more than one pathway we can use an average of the pathway weights such that $v_m = \sum_{k=1}^{K} \hat{\omega}_k I(m \in k) / \sum_{k=1}^{K} I(m \in k)$ where $I(m \in k)$ is an indicator of association for metabolite m with pathway k. Additionally we can use a rank-based summary such as the median or upper percentile, as in Chapter III, to determine v_m in cases of overlapping pathways. The standardized weight w_m can be obtained by standardizing the v_m terms by their average, $\bar{v} = \frac{1}{M} \sum_{m=1}^{M} v_m$. The condition that $w_m \geq 0$ for all $m = 1, \ldots, M$ is satisfied by $\omega_k \geq 0$. Finally, unmapped metabolites are given the weight $w_m = 1$ which gives no adjustment to the p-value. Thus the per-metabolite weight w_m, for $m = 1, \ldots, M$, can be written as

$$w_m = \begin{cases} v_m/\bar{v} & \text{mapped metabolites} \\ 1 & \text{otherwise} \end{cases} \tag{B.2}$$

B.3 Numerical results

This two-component weighting method was testing using the simulation model of Section 3.4.1 (Chapter III). Here we simulate disjoint pathways under various enrichment and correlation conditions. The same four enrichment test are also considered: a directional hypergeometric test, a binomial test of proportions, a weighted Kolmogorov-Smirnov test, and a sum of squared test statistics test. Only a subset of the simulation scenarios presented in Chapter III were run for the two-component weight models.

Some results are presented here using the same four-panel graphic style of Chapter III. The two-component weight using the coverage based estimate $\hat{\theta}_k = (\tilde{\eta}_k - 1)/\eta_k$ will be labelled weight function “F” and will be colored brown. The two-component weight using the correlation-based estimate $\hat{\theta}_k = |\rho_k|$ will be labelled weight function “G” and will be colored yellow (notice that the definition used in simulation is reversed from that described above). Since weight function “G” reduces to weight function “A”, $\omega_k = -\log_{10}(P^g_k) = \omega_{2k}$ when $\rho = 0$, we also include this function, colored green, for comparison.

The primary concern with including a metabolic component in the weighting of the metabolites is that the test will be biased. Under a null model of no differentially expressed genes or differential metabolites we find that the coverage-based model (F, brown) has error rates is above the nominal level a majority of the time; see Figure B.1. This is for the scenario of $\eta_k = 3$ and $\tilde{\eta}_k = 10$ for all pathways $k = 1, \ldots, K+1$. The correlation-based model (G, yellow) is equivalent to the gene expression p-value model (A, green) when the correlation is zero. However, it too shows increased error when the correlation is non-zero. The weight function A (green) has error rates that contain the nominal level in the box of the boxplot for each of these low correlation scenarios.

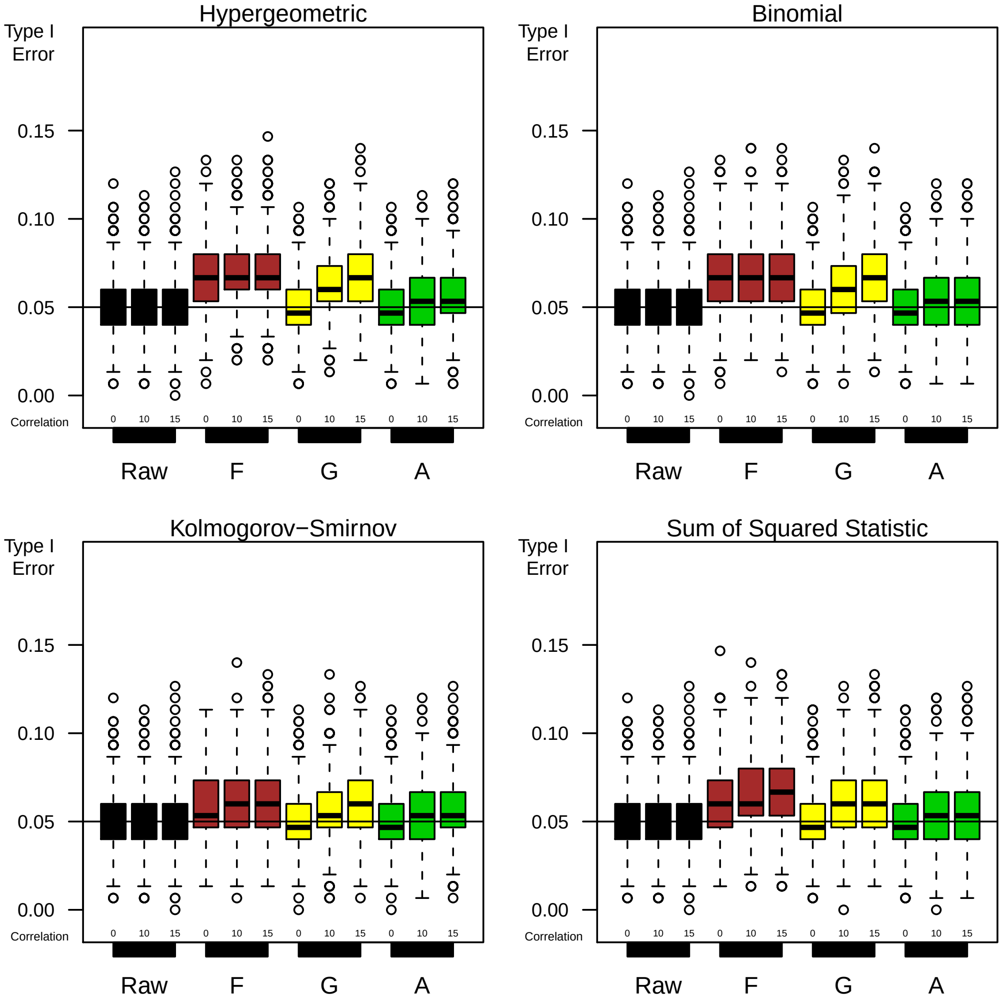


Figure B.1: **Type 1 error** for per-metabolite tests using a significance threshold of $\alpha = 0.05$ without multiple testing adjustments. 1000 datasets were simulated assuming within pathway correlation of 0.2 for each metabolites and genes. Unweighted (Raw) p-values and the three weight functions (F, $\hat{\theta}_k = (\tilde{\eta}_k - 1)/\eta_k$, brown; G, $\hat{\theta}_k = |\rho|$, yellow; A, $\hat{\theta}_k = 0$, green) are depicted with increasing between element correlation, $\rho \in (0, 0.10, 0.15)$.

When we look at a high correlation scenario, as in Figure B.2, we see that the error rates continue to increase with increasing ρ. Again, the single-component weight (A, green) maintains lower error rates except for the highest correlation scenario, $\rho = 0.5$. We happened to define the correlation-based θ estimate counter-intuitively for the simulation, $\hat{\theta}_k = |\rho_k|$, such that the gene expression component ω_{2k} is dominant when the genes and metabolites are not correlated. However, we see here, that as the correlation increases the amount of contribution of the metabolite component ω_{1k} increases and so does the bias. Thus this correlation-based θ is clearly not the best choice for this model.

If we turn our attention to a non-null case we quickly see that the single-component model (A, green) peforms almost equivalently to the two-component models in these scenarios; see Figures B.3 and B.4. There is a power gain but this is without correcting for the inflated type I error rates. Should error controlling measures, such as a Bonferroni correction, be made this power would likely be reduced.

Given these issues we chose to abandon the two-component weights in favor of a single-component weight. In our simulations we found stable single-component pathway-based weights based on gene expression that did not dominate the metabolic data; see Chapter III Section 3.6 for recommendations.

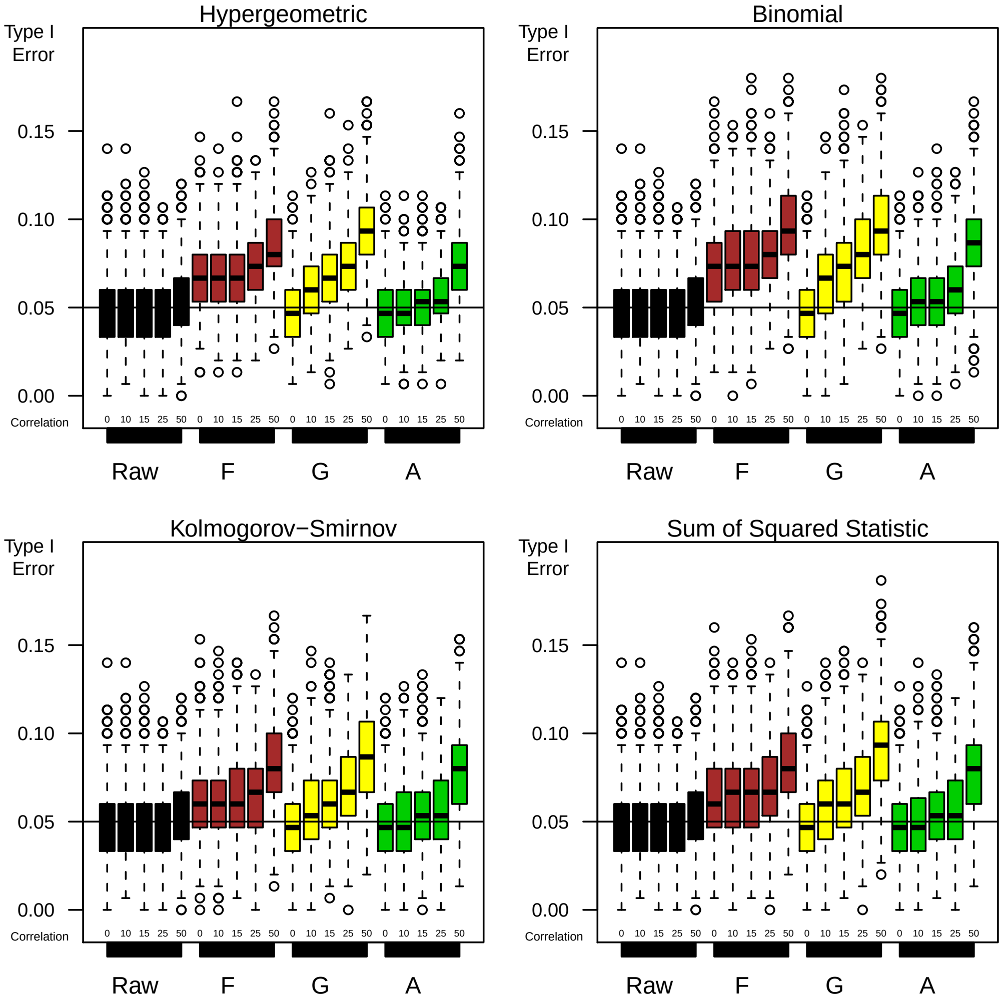
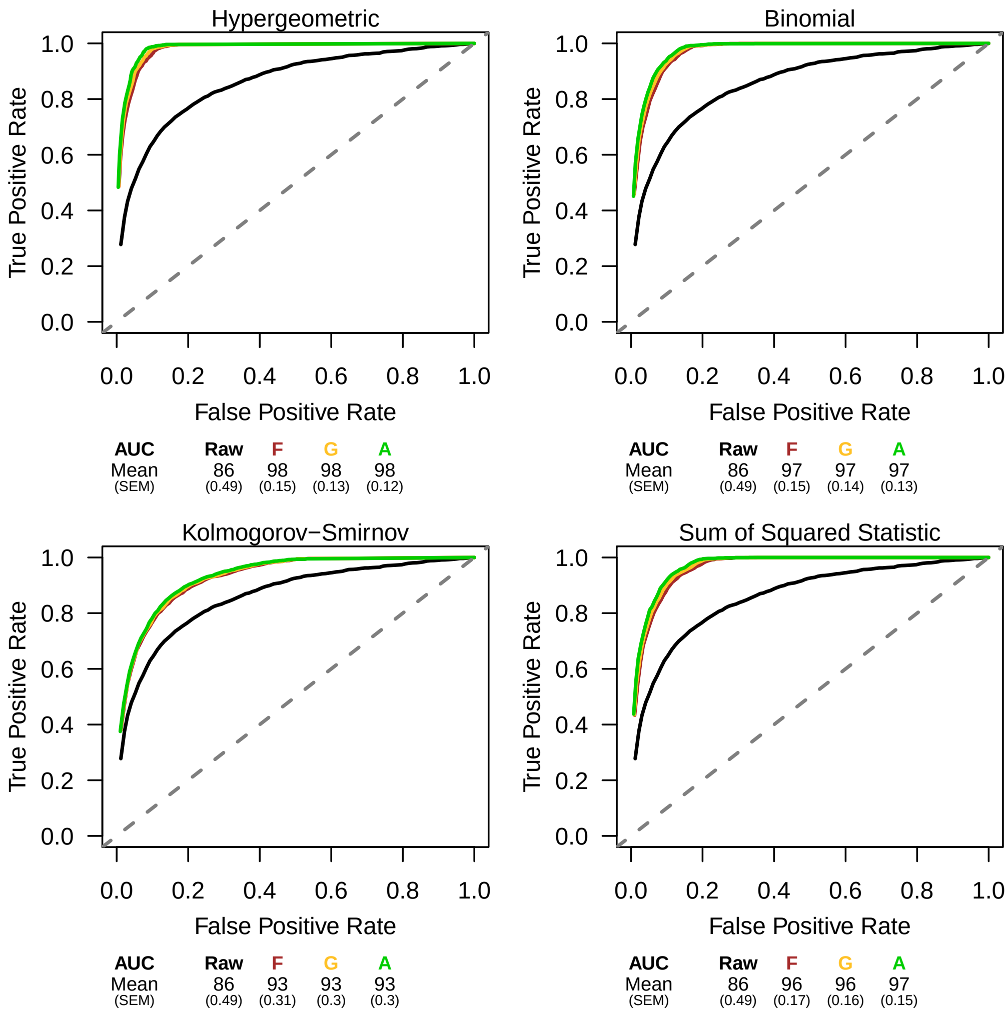


Figure B.2: **Type 1 error** for per-metabolite tests as in Figure B.1 except that the within pathway correlation is 0.6 for each metabolites and genes and the between element correlation is increased as high as 50%; $\rho \in (0, 0.10, 0.15, 0.25, 0.50)$. Unweighted (Raw) p-values and the three weight functions (F, $\hat{\theta}_k = (\tilde{\eta}_k - 1)/\eta_k$, brown; G, $\hat{\theta}_k = |\rho|$, yellow; A, $\hat{\theta}_k = 0$, green) are depicted.



Hypergeometric

AUC	Raw	F	G	A
Mean	86	98	98	98
(SEM)	(0.49)	(0.15)	(0.13)	(0.12)

Binomial

AUC	Raw	F	G	A
Mean	86	97	97	97
(SEM)	(0.49)	(0.15)	(0.14)	(0.13)

Kolmogorov−Smirnov

AUC	Raw	F	G	A
Mean	86	93	93	93
(SEM)	(0.49)	(0.31)	(0.3)	(0.3)

Sum of Squared Statistic

AUC	Raw	F	G	A
Mean	86	96	96	97
(SEM)	(0.49)	(0.17)	(0.16)	(0.15)

Figure B.3: **Average receiver operating characteristic (ROC) curves** (n=100) depict the sensitivity and specificity for each test method and weight function when applied to per-metabolite tests. Data are simulated assuming within pathway correlation of 0.2 for each metabolites and genes and between element correlation of 0.1. Ten of fifty pathways were simulated as enriched where differential test statistics have mean of two and three for metabolites and genes, respectively. The mean area under the curve (AUC) estimate and associated standard error are provided in the table below each plot. Here $\eta_k = 3$ and $\tilde{\eta}_k = 10$.

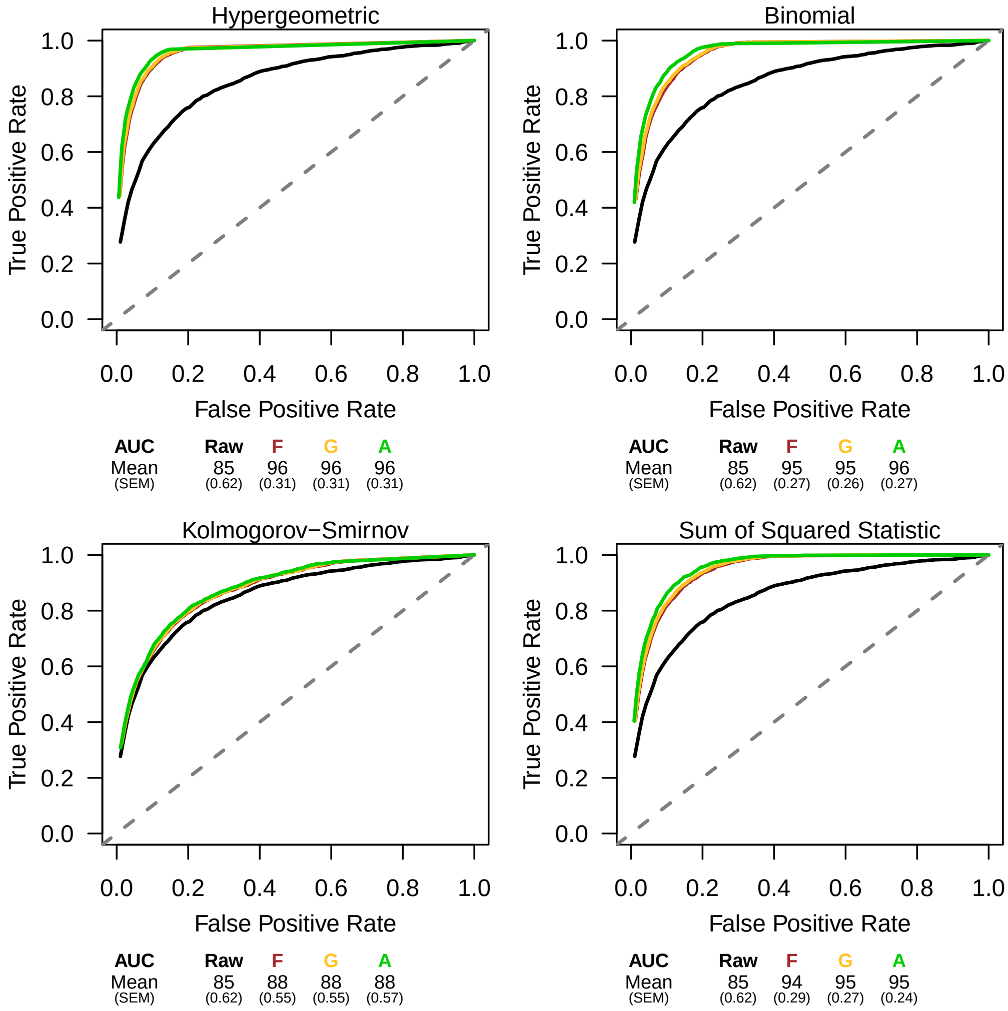


Figure B.4: **Average receiver operating characteristic (ROC) curves** (n=100) as in Figure B.3 except that data are simulated assuming within pathway correlation of 0.6 for each metabolites and genes and between element correlation of 0.15.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47, 2009.

[2] D. Allison, X. Cui, G. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and concensus. *Nature Reviews: Genetics*, 7:55–65, 2006.

[3] C. Beecher. Metabolomics: The newest of the "omics" sciences. *Innovations in Pharmaceutical Technology*, June:57–64, 2002.

[4] D. Beer, S. Kardia, C. Huang, T. Giordano, A. Levin, D. Misek, L. Lin, G. Chen, T. Gharib, D. Thomas, M. Lizyness, R. Kuick, S. Hayasaka, J. Taylor, M. Iannettoni, M. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8):816–824, 2002.

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

[6] Y. Benjamini and Y. Hochberg. Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics*, 24:407–418, 1997.

[7] R. Breitling, A. Amtmann, and P. Herzyk. Iterative group analysis (iga): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5:34, 2004.

[8] D. Camacho, A. de la Fuente, and P. Mendes. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, 2004.

[9] F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L. Sweetlove, and A. Fernie. Integrated analysis of metabolite and transcript levels reveal the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiology*, 142:1380–1396, 2006.

[10] H. Chang, J. Sneddon, A. Alizadeh, R. Sood, R. West, K. Montgomery, J. Chi, M. van de Rijn, D. Botstein, and P. Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLOS Biology*, 2(2):206–214, 2004.

[11] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[12] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J. Matese, T. Hernandez-Boussard, C. Rees, J. Cherry, D. Botstein, P. Brown, and A. Alizadeh. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research*, 31(1):219–223, 2003.

[13] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1):107–129, 2007.

[14] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21:171178, 2005.

[15] C. Fan, D. Oh, L. Wessels, B. Weigelt, D. Nuyten, A. Nobel, L. van't Veer, and C. Perou. Concordance among gene-expression-based predictors for breast cancer. *New England Journal of Medicine*, 355(6):560–569, 2006.

[16] E. Ferkingstad, A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong. Unsupervised empirical bayesian multiple testing with external covariates. *Annals of Applied Statistics*, 2(2):714–735, 2008.

[17] C. Ferrara, P. Wang, E. Neto, R. Stevens, J. Bain, B. Wenner, O. Ilkayeva, M. Keller, D. Blasiole, C. Kendziorski, B. Yandell, C. Newgard, and A. Attie. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. *PLoS Genetics*, 4(3):e1000034, 2008.

[18] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.

[19] O. Fiehn and W. Weckwerth. Deciphering metabolic networks. *European Journal of Biochemistry*, 270(4):579–588, 2003.

[20] C. R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[21] Y. Gibon, B. Usadel, O. Blaesing, B. Kamlage, M. Hoehne, R. Trethewey, and M. Stitt. Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in arabidopsis rosettes. *Genome Biology*, 7(8):R76, 2006.

[22] G. Glinsky, O. Berezovska, and A. Glinskii. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple type of cancer. *The Journal of Clinical Investigation*, 115(6):1503–1521, 2005.

[23] G. Glinsky, A. Glinskii, A. Stephenson, R. Hoffman, and W. Gerald. Gene expression profiling predicts clinical outcome of prostate cancer. *The Journal of Clinical Investigation*, 113(6):913–923, 2004.

[24] J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[25] R. Goodacre. Metabolomics the way forward. *Metabolomics*, 1(1):1–2, 2005.

[26] R. Goodacre, S. Vaidyanathan, W. Dunn, G. Harrigan, and D. Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5):245–252, 2004.

[27] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein. Imputing missing data for gene expression arrays. Technical report, Stanford University Statistics Department, http://www-stat.stanford.edu/ hastie/Papers/missing.pdf, 1999.

[28] M. Hirai, M. Klein, Y. Fujikawa, M. Yano, D. B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, and K. Saito. Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *The Journal of Biological Chemistry*, 280(27):2559095, 2005.

[29] M. Hirai, M. Yano, D. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings of the National Academy of Science of the United States of America*, 101(27):10205–10210, 2004.

[30] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[31] L. Hood. Systems biology: integrating technology, biology, and computation. *Mechanisms of Ageing and Development*, 124:9–16, 2003.

[32] E. Huang, S. Ishida, J. Pittman, H. Dressman, A. Bild, M. Kloos, M. D'Amico, R. G. Pestell, M. West, and J. R. Nevins. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics*, 34:226–230, 2003.

[33] J. Ippolito, J. Xu, S. Jian, K. Moulder, S. Mennerick, J. Crowlery, R. Townsend, and J. Gordon. An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 102(28):9901–9906, 2005.

[34] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36:480–484, 2008.

[35] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[36] M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:354–357, 2006.

[37] S. Kawashima, T. Katayama, Y. Sato, and M. Kanehisa. KEGG API: A web service using SOAP/WSDL to access the KEGG system. *Genome Informatics*, 14:673–674, 2003.

[38] Y. S. Kim and P. Maruvada. Frontiers in metabolomics for cancer research: Proceedings of a National Cancer Institute workshop. *Metabolomics*, 4:105113, 2008.

[39] J. Lamb, S. Ramaswamy, H. L. Ford, B. Contreras, R. V. Martinez, F. S. Kittrell, C. A. Zahnow, N. Patterson, T. R. Golub, and M. E. Ewen. A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell*, 114:323–334, 2003.

[40] X. Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84:309326, 1997.

[41] T. Loughin. A systematic comparison of methods for combining p-values. *Computational Statistics and Data Analysis*, 47:467–485, 2004.

[42] MAQC Consortium. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

[43] J. Mathew, B. Taylor, G. Bader, S. Pyarajan, M. Antoniotti, A. Chinnaiyan, C. Sander, S. Burakoff, and B. Mishra. From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Computational Biology*, 3(2):e12, 2007.

[44] V. Mootha, C. Lindgren, K. Eriksson, S. A., S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golumb, P. Tamayo, and B. Spiegelman. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

[45] K. Morgenthal, S. Wienkoop, M. Scholz, J. Selbig, and W. Weckwerth. Correlative gc-tof-ms-based metabolite profiling and lc-ms-based protein profiling reveal time-related systemic regulation of metaboliteprotein networks and improve pattern recognition for multiple biomarker selection. *Metabolomics*, 1(2):109–121, 2005.

[46] F. Mosteller and R. Fisher. Questions and answers. *The American Statistician*, 2(5):30–31, 1948.

[47] L. Moulton and S. Zeger. Bootstrapping generalized linear models. *Computational Statistics and Data Analysis*, 11:53–63, 1991.

[48] D. Nam and S. Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[49] D. Nguyen, A. Arpat, N. Wang, and R. Carroll. Dna microarray experiments: Biological and technological aspects. *Biometrics*, 58:701–717, 2002.

[50] D. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, 74:765–769, 2004.

[51] W. Pan. Network-based model weighting to detect multiple loci influencing complex diseases. *Human Genetics*, 124(3):225–234, 2008.

[52] D. Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer*, 4(4):309–314, 2004.

[53] K. Roeder, S. Bacanu, L. Wasserman, and B. Devlin. Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics*, 78:243–252, 2006.

[54] K. Roeder, B. Devlin, and L. Wasserman. Improving power in genome-wide association studies: weights tip the scale. *Genetic Epidemiology*, 31(7):741–747, 2007.

[55] K. Roeder and L. Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 2009.

[56] R. Sandberg and I. Ernberg. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proceedings of the National Academy of Science of the United States of America*, 102(6):2052–2057, 2005.

[57] M. Sartor, G. Leikauf, and M. Medvedovic. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

[58] K. Shedden. Confidence levels for the comparison of microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):32, 2004.

[59] J. Spicker, S. Brunak, K. Frederiksen, and H. Toft. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. *Toxicological Sciences*, 102(2):444–454, 2008.

[60] A. Sreekumar, L. Poisson, T. Rajendiran, A. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. Lonigro, Y. Li, M. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. Omenn, D. Ghosh, S. Pennathur, D. Alexander, A. Berger, J. Shuster, J. Wei, S. Varambally, C. Beecher, and A. Chinnaiyan. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457(7231):910–914, 2009.

[61] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Science of the United States of America*, 102:15545–15550, 2005.

[62] A. Sweet-Cordero, S. Mukherjee, A. Subramanian, H. You, J. J. Roix, C. Ladd-Acosta, J. Mesirov, T. R. Golub, and J. T. An oncogenic kras2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, 37:48–55, 2005.

[63] P. 't Hoen, Y. Ariyurek, H. Thygesen, E. Vreugdenhil, R. Vossen, R. de Menezes, J. Boer, G. van Ommen, and den Dunnen J.T. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36(21):e141, 2008.

[64] L. Tian, S. Greenberg, S. Kong, J. Altschuler, I. Kohane, and P. Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Science of the United States of America*, 102:13544–13549, 2005.

[65] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[66] E. Urbanczyk-Wochniak, A. Luedemann, J. Kopka, J. Selbig, U. Roessner-Tunali, L. Willmitzer, and A. Fernie. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports*, 4(10):989–993, 2003.

[67] L. van't Veer, H. Dai, M. van de Vivjer, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[68] S. Varambally, J. Yu, B. Laxman, D. Rhodes, R. Mehra, S. Tomlins, R. Shah, U. Chandran, F. Monzon, M. Becich, J. Wei, K. Pienta, D. Ghosh, M. Rubin, and A. Chinnaiyan. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, 8:393–406, 2005.

[69] W. Weckwerth. Metabolomics in systems biology. *Annual Review of Plant Biology*, 54:669–689, 2003.

[70] W. Weckwerth. Integration of metabolomics and proteomics in molecular plant physiology coping with the complexity by data-dimensionality reduction. *Physiologia Plantarum*, 132:176189, 2008.

[71] W. Weckwerth and K. Morgenthal. Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today*, 10(22):1551–1558, 2005.

[72] B. Weigelt, F. Baehner, and J. Reis-Filho. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *Journal of Pathology*, 10.1002/path.2648, 2009.

[73] J. Yang. Distribution of fisher's combination statistic when the tests are dependent. *Journal of Statistical Computation and Simulation*, 2009.

[74] Y. Yang, S. Dudiot, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.