ABSTRACT

HU, JUN. Integrative Analysis of Multi-Platform Genomic Data. (Under the direction of Dr. Jung-Ying Tzeng).

Rapid advancement of high-throughput technology has revolutionized the biomedical research in last two decades. Huge amount of data have been generated from different high-throughput platforms. To analyze these data appropriately, numerous statistical methods have been developed. Nevertheless, data from different high-throughput platforms often have their unique statistical properties, e.g., NGS data vs. microarray data. Methods developed for one platform usually cannot be applied to other high-throughput platform directly. There are few algorithms or packages that can jointly analyze datasets with multiple platforms. It often becomes a challenge to properly compare or integrate outcomes from different high-throughput platforms. And this dissertation is devoted to the integrative statistical analysis across different high-throughput platforms.

We start the dissertation with a comprehensive background introduction in chapter one. Microarray and NGS, two major high-throughput platforms, are briefly reviewed. We then discuss the availability of multi-platform genomic data and emphasis the importance to develop novel statistical methods for multi-platform data. We introduce gene set analysis (GSA), its application in biomedical data analysis, and also provide a systematic review of major statistical methods developed for single platform GSA. In addition, we reveal the concept of sample heterogeneity and its impact to data analysis. Single platform statistical methods to deal with sample heterogeneity are then summarized, including both parametric and nonparametric approaches. We end the chapter with a brief review on popular batch

effect removal (BER) methods that are used to correct batch effects and platform biases, which are commonly encountered during multi-platform data analysis.

In Chapter 2, we discuss the latest research progresses of GSA on multi-platform data. We compare the performances of major algorithms that were developed for multi-platform GSA at different simulated scenarios. We show that multi-platform methods outperform single platform methods. Among all the evaluated multi-platform methods, INT (integrative method developed by Tyekucheva *et al.*) have the best overall performance.

In Chapter 3, we demonstrate that existing multi-platform GSA methods perform poorly when sample heterogeneity exists. We develop three new methods to account for sample heterogeneity in multi-platform GSA, i.e., *MPMWS* (Multi-Platform Mann-Whitney Statistics), *MPORT* (Multi-Platform Outlier Robust T-statistics), and *MPLRS* (Multi-Platform Likelihood Ratio Statistics). We test their performances along with existing multi-platform GSA methods using both simulated and real dataset. We find that MPMWS method outperforms other multi-platform GSA methods, and it has satisfactory power and robust performance regardless the degree of heterogeneity.

In Chapter 4, using Affymetrix and Illumina RNA-Seq as representative platforms, we check the concordance and discordance between microarray and RNA-Seq. We find that gene expression level, gene length and GC content can all contribute to the platform bias between microarray and RNA-Seq. Then we compare different BER methods for their abilities to remove the platform bias between microarray and RNA-Seq.

(Note: Chapter 2 and Chapter 3 are reprinted from Hu, J. and Tzeng, J.Y. 2014 Integrative Gene Set Analysis of Multi-platform Data with Sample Heterogeneity, *Bioinformatics*.)

Integrative Analysis of Multi-Platform Genomic Data

by
Jun Hu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
Dr. Jung-Ying Tzeng Committee Chair	Dr. Alison Motsinger-Reif
_____	_____
Dr. Jeffrey A Yoder	Dr. Zhao-Bang Zeng

DEDICATION

This dissertation is dedicated to my parents, who taught me that hard work, persistence and determination are the keys to success. It is also dedicated to my sister and grandparents, for their endless encouragements and support.

BIOGRAPHY

Jun Hu was born in Shanghai, China. Jun and his family lived in Wuhan, China when he was young. In 1997, he graduated from Wuhan University in China with a B.S. degree on biochemistry.

From 1997 to 2000, Jun studied at Rutgers University where he received a M.S. degree on molecular biology in 2000. Jun got his second M.S. degree on computer sciences from New Jersey Institute of Technologies in 2001. Since then, Jun has worked in New Jersey Medical School of Rutgers University and the Cancer Institute of New Jersey as a senior bioinformatician before he joined Omicsoft Corporation, a startup bioinformatics software company located in the research triangle of North Carolina, in 2010.

Jun transferred to the Bioinformatics Research Center of North Carolina State University in 2010, to continue pursuing a Ph.D. degree on bioinformatics. Under the direction of Dr. Jung-Ying Tzeng, he conducted research to develop novel statistical methods on multi-platform high-throughput genomic data analysis.

To date, Jun has authored and co-authored more than 20 publications in peer reviewed journals. Jun's research interests include multi-platform high-throughput genomic data analysis algorithms; gene expression regulatory motifs and gene expression regulation networks.

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor, Dr. Jung-Ying Tzeng, for her patience, encouragement and excellent guidance throughout my study at North Carolina State University. Dr.Tzeng's insightful criticisms and sage advice are essential to this dissertation.

I would also like to thank my dissertation committee members, Dr. Alison Motsinger-Reif, Dr. Jeffrey A Yoder and Dr. Zhao-Bang Zeng for their direction and invaluable advice on my research projects.

I also thank Dr. Spencer Muse, Dr. Jeffrey L. Thorne, Ms. Siarra Dickey, other faculties and stuffs at Bioinformatics Research Center (BRC) for their support and kindness. Their hard works have made BRC a nice study and research environment for students.

To all my friends, I am grateful for all the encouragement and happiness that you bring to me. You make this long journey more joyful.

Last and foremost, I would like to thank my sister and dear parents for their understanding, unlimited love and support throughout my whole life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 BACKGROUND

1.1 Multi-platform genomic data

Since the 1990s, high-throughput technologies, such as microarray, next generation sequencing (NGS), mass-spectrometry based large scale proteomics and metabolomics have revolutionized the way scientists study the biological systems. With the help of these new technologies, scientists are able to study thousands of genes in a single experiment and improve their research productivity by magnitudes. The results of high-throughput studies enable the researchers to have a more profound understanding of biological processes at a fundamental level. For this presented thesis, the main focuses are on microarray and NGS (or high-throughput sequencing), two most dominant technologies of all high-throughput technologies.

DNA-microarray was developed on the same principle of southern blotting, where DNA fragments are fixed onto a supporting substance (e.g., nitrocellulose membrane) and probed with another DNA sequence. For DNA-microarray, tens of thousands of DNA oligonucleotides with known sequences are manufactured and spotted onto a glass slide (array). Then the array is used to probe a DNA mixtures (which are often cDNAs reversed-transcribed from RNAs) from testing samples. Each sequence on the slide could hybridize to its complement sequence in the DNA mixture. Based on the amount of the hybridized sequences, we can find out the relative compositions of the DNA mixture.

Although DNA microarrays are most frequently used to profile RNA gene expression levels, over the years, other types of microarray have been designed to test other aspects of

the biological system. Infinium HumanMethylation450 is one type of methylation array that can test 450,000 methylation sites across the human genome. Human SNP Array 6 from Affymetrix is designed to genotype millions of single nucleotide polymorphisms (SNPs) and check copy numbers variations (CNVs) at the same time. Promoter arrays are used during chromatin immunoprecipitation on microarray (CHIP-Chip) experiments to test the binding of transcription factors. Protein array are also available, e.g., antibody arrays are used to profile the protein concentrations in cell lysates.

NGS was evolved from the traditional DNA sequencing where nucleotide sequences are determined using random chain termination method during double strand DNA synthesis (Sanger method) (Sanger and Coulson, 1975; Sanger, et al., 1977). Inspired by the huge success of Human Genome Project (HGP), the tremendous demand for fast, accurate and low-cost sequencing methods has leaded to the development of several different NGS strategies such as 454 sequencing, Illumina, SOLID, Ion Torrent, etc. These NGS methods can sequence thousands or even millions of reads in parallel. It took HGP three billion dollars and 10 years to sequence the first human genome. Now a single lab can get a new genome sequenced within weeks with only several thousand dollars.

Similar to microarray, NGS have also been applied to various pipelines to access the RNA expression (RNA-Seq), miRNA expression (miRNA-Seq), methylation (e.g., BS-seq) or transcription factor binding (CHIP-seq). Compared to microarray technology, NGS often have higher sensitivity and broader detection ranges. Since it sequences the nucleotides at base pair level, it is ideal to detect novel genomic aberrations such as gene insertions and deletions (indels), SNPs, alternative-splicing or gene fusions.

Microarray, NGS and other high-throughput technologies have generated huge amount of data. Over the last two decades, high-throughput data analysis has been the focal research area for bioinformatics and biostatistics. Numerous algorithms and software packages have been developed to deal with data from different high-throughput platforms. However, how to compare or utilized datasets from different platforms is often overlooked. With the advancement of technology and decreasing of cost, we see more and more large-scale studies, e.g., The Cancer Genome Atlas (TCGA) and The Encyclopedia of DNA Elements (ENCODE). The researchers are able to study the same samples with multiple high-throughput platforms to study sample genome compositions, gene expressions and their gene regulations mechanisms at the same time. With so many diverse data types from different platforms, it becomes challenging to properly integrate, analyze, interpret the results and obtain biological insights with all these invaluable data. And this dissertation aims to develop some new statistical methods to perform analysis using multi-platform genomic data.

1.2 Gene set analysis

Gene set analysis (GSA) is an analytic method originally developed for high-throughput gene expression profiling data. It works on a set of related genes with biological similarities or functional associations, such as genes in one KEGG pathway (Kanehisa and Goto, 2000) or those related to the same Gene Ontology (Ashburner, et al., 2000) term. During high-throughput data analysis, many changed genes often fail to be identified by differential expression gene (DEG) analysis because they only have moderate changes and/or small sample sizes. GSA methods could be helpful for these genes since they gain additional power

by analyzing all the genes in the set together. It is especially powerful for genes with weak/moderate but coordinate changes in the sets. Moreover, since the sets used for testing often have real biological connections, it becomes much easier to interpret the analysis outcomes. It also helps to establish a link between the genes in the set with their potential biological functionalities.

1.2.1 Statistical methods

Many GSA methods have been developed over the years. They are mostly applied on the gene expression data, the dominant high-throughput data type for both microarray and NGS. These methods can be generally divided into "competitive" and "self-contained" (Maciejewski, 2013; Nam and Kim, 2008; Tian, et al., 2005) depending on their null hypothesis and the selection strategy of genes used for statistical analysis. Self-contained methods only consider the genes within the set and examine the null hypothesis that these genes have no association with the phenotype. Whereas competitive methods tests the null hypothesis that the genes within the set and those outside the sets have the same association with phenotype. Globaltest (Goeman, et al., 2004), ANCOVA (Mansmann and Meister, 2005) and SAM-GS (Dinu, et al., 2007) are examples of self-contained GSA methods, while examples for competitive GSA methods include PAGE (Kim and Volsky, 2005), GSA algorithm (Efron and Tibshirani, 2007) and the popular GSEA algorithm (Subramanian, et al., 2005).

Over-representation analysis (ORA) is a competitive GSA method. It assesses the over or under-representation of functional genes from the gene set compared to the reference

null sets. ORA generally use Fisher's exact test or hyper-geometric test for statistical inferences. Because of its simplicity, many public tools like DAVID (Huang da, et al., 2009; Huang da, et al., 2007) and GoStats (Falcon and Gentleman, 2007) have implemented ORA to mine gene annotation databases such as Gene Ontology (Ashburner, et al., 2000), KEGG (Kanehisa and Goto, 2000) or MSigDB (Subramanian, et al., 2005). However, ORA tends to suffer from low power, its scores also vary dramatically depending on the number of shared genes between testing gene set and the reference null sets.

Gene Set Enrichment Analysis (GSEA) is one of the most popular GSA methods. First, the genes are ranked based on the correlation between their expressions and the phenotype classification. Then for each gene set, the cumulative sums over these ranked genes were computed. When a gene is in the gene set, the sum is increased, the sum is decreased otherwise. Magnitude of increment depends on correlation of gene with phenotype. The enrichment score (ES) for each gene set is recorded as the maximum deviation from zero. This ES score is a weighted Kolmogorov–Smirnov (KS) statistic comparing the ranks of genes in the gene set with the uniform null distribution. (Note: An early version by Moore *et al*. uses an un-weighted version of KS-test (Mootha, et al., 2003). It does not consider the correlation of gene with phenotype.) The significance of ES is then tested by empirical phenotype-based permutation test. ES for gene sets are adjusted to account for the different sizes of the gene sets. FDR can be used to control the overall false discovery rate.

GSEA uses KS statistic, a nonparametric statistic which is known to lack sensitivity. Several alternative strategies to compute ES were proposed by various groups. Irizarry et al.

proposed to use standard one sided z-test or chi-square test (Irizarry, et al., 2009) and demonstrated that these simple parametric methods outperformed GSEA in a microarray dataset. Efron and Tibshirani tested mean, mean absolute values, maxmean against GSEA's ES values under different simulation conditions. They found that maxmean outperformed other statistic (Efron and Tibshirani, 2007). Parametric analysis of gene set enrichment (PAGE) calculates a Z score by comparing the mean fold change for all genes in the testing gene set against the mean fold change of all the genes in experiment. This Z score follows a normal distribution under the null hypothesis. And this property is used to infer the statistical significance of the Z score (Kim and Volsky, 2005). Using simulation and several real datasets, PAGE has been demonstrated to have better sensitivity than GSEA, it is also more computational efficient since no permutation is need to derive the p-values.

Globaltest is a self-contained GSA method. It performs a generalized linear modeling between expression matrix and the phenotype (Goeman, et al., 2004). Globaltest assumes that all the coefficients in the generalized linear model are zero and this null hypothesis is evaluated with a score test. The score test guarantees good power for gene sets when genes in the set have small or moderate effect to the phenotype. This is a desire feature for analysis. Similar multivariate ANCOVA-based approach has also been proposed by Mansmann and Meister (Mansmann and Meister, 2005).

Dinu *et al.* extended the individual-gene analysis method, significance analysis of microarray (SAM) (Tusher, et al., 2001), to gene set analyses (SAM-GS) (Dinu, et al., 2007). The null hypothesis of SAM-GS is that the mean vector of expression of genes in a gene set does not differ by the phenotype of interest. Instead of using multi-variant Hotelling's T^2,

which might not be valid for large gene sets with more genes than the number of samples, a permutation based test like SAM is used on the L_2-norm of the mean-vector differences. Comparative analysis of self-contained GSA methods by Liu et al. has shown that Globaltest, ANCOVA and SAM-GS show comparable performances in terms of size and power if data are properly standardized to stabilize per-gene variance. SAM-GS were found to have slightly higher power for gene sets with small p-values (Liu, et al., 2007).

Both the competitive and self-contained methods have their limitations. Competitive methods can cause the "zero-sum game" behavior (Allison, et al., 2006). For example, in an extreme case where all the genes are down-regulated, some gene sets can be considered up-regulated when comparing with the background distribution. In addition, the assumption that genes are independent is often not valid in real life data sets. This invalid assumption of independent gene sampling hurts most of the competitive methods. Self-contained methods use only the genes contained in the given gene set. However, a single extreme DEG can make the whole gene set significant whereas the whole gene set is not really 'enriched' with DEGs.

1.2.2 None-expression platforms

GSA methods are also commonly used for genotyping, CNV, methylation and other none-expression high-throughput dataset. For some non-expression high-throughput platforms, e.g. protein arrays, most of the GSA methods originally developed for gene expression dataset can be applied directly. However, for many other platforms, transformation of original features is often needed before applying GSA methods.

Gene wide association studies (GWAS) has been a hot research area since the late last decade. Many methods have been developed to perform GSA on GWAS data. These GSA methods can not only gain power by examining groups of variants sharing a common gene set (e.g., pathway), but also help the scientist better understand the biological themes relevant to complex diseases (Wang, et al., 2011). Although some of these GSA methods use raw genotyping data of individual subjects directly, the SNP p-values are more commonly used to assess enrichment score for each get set. In general, the most significant SNPs of the gene (Min-P) or the average of p-values of selective SNPs (Avg-P) are used as the representative p-value of the gene, and then GSEA or other GSA methods developed for expression data can be used. For example, Wang *et al.* modified the GSEA algorithm and used it for SNP analysis (Wang, et al., 2007). The min-P strategy is first used to select the representative SNP among all SNPs within 500kb of a gene, and then weighted KS statistic is used to compute ES of a gene set. Similar strategy was used by Jia *et al.* to analysis schizophrenia GWAS data (Jia, et al., 2010).

There are other strategies to select representative SNPs for a gene. Aligator algorithm (Holmans, et al., 2009) first selects the significant SNPs with a pre-specified p-value cut-off. And the genes containing these significant SNPs are deeming significant. For each gene set, the over-representation of significant genes was tested against simulated null distributions. Similarity, SNP ratio test calculates the ratio of the number of significant SNPs in a gene set to the total number of SNPs in the set and derive p-values using simulated null datasets (O'Dushlaine, et al., 2009). Weng *et al.* extended the idea of Aligator by adopting an adaptive truncated product statistic to select the representative SNPs of each gene (Weng, et al., 2011).

Gene set ridge regression in association studies (GRASS) is a self-contained GSA method that takes raw genotype data as input. It uses regularized regression to select representative eigenSNPs for each gene, then tests the joint association of genes in gene set with disease risk (Chen, et al., 2010).

Many studies have applied GSA on high-throughput methylation data (Booth, et al., 2012; Irizarry, et al., 2009). Methylation data has some similarities to genotyping data in terms of GSA strategy, because one single gene can also have many methylation sites, just like SNPs. In majorities of these studies, differentially methylated genes (DMGs) were first identified from high-throughput methylation data. Then ORA and other GSA methods were applied to identify gene sets that were enriched with DMGs.

1.3 Sample heterogeneity

Sample heterogeneity refers to the molecular and cellular differences among biological samples. Different from monogenic diseases where a single dysfunctional gene is the major causal factor, complex diseases like cancer are heterogeneous and multifactorial. They are often the outcome of the alteration of multiple regulatory pathways and/or due to the interplay between different genes and the environment (Di Camillo, et al., 2012; Sole, et al., 2009).

Cancer samples from different sources or individuals often have very different genomic differences even if they shared the same phenotype (diseases). Clinically, cases with different genotypes, genomic copy numbers or expression patterns often lead to different disease prognosis and treatment strategies (Fisher, et al., 2013; Russnes, et al., 2011). During

data analysis, the heterogeneity of a dataset creates some major challenges. Traditional analytical methods that search for common changes of genes across a class of samples, such as t-test or linear modeling, often fail to work or have little power when only a subset of samples have changed (e.g., CNV data set from cancer samples) (Tomlins, et al., 2005).

Several statistical methods have been proposed to address sample heterogeneity. We illustrate these different methods using an example case below. Assume that we try to detect up-regulated genes in a two-class (case vs. control) microarray cancer dataset at the existence of heterogeneity. Assume that there are M genes measured from n_0 control samples and n_1 case samples (i.e., in total, $n = n_0 + n_1$ samples). Let x_{im} be the observed value of the expression variable for gene m of sample i from control samples, and y_{jm} be the observed value of the expression variable for gene m of sample j from cases samples. Then the t-statistic can be expressed as $T_m = (\overline{y_m} - \overline{x_m})/s_m$ where for gene m, $\overline{x_m}$ is the mean of expression values of control samples, $\overline{y_m}$ is the mean of expression values of case samples and s_m is the pooled standard deviation from both case and case samples.

Tomlins *et al.* proposed to use cancer outlier profile analysis (COPA) (MacDonald and Ghosh, 2006; Tomlins, et al., 2005). The COPA statistic is defined as

$$C_m = \frac{q_{r\left(\{y_{km}\}_{1\le k\le n_1} - median_m\right)}}{mad_m}$$

$$mad_m = 1.4826 \times median\left(\{y_{km} - median_m\}_{1\le k\le n_1}, \{x_{km} - median_m\}_{1\le k\le n_0}\right)$$

where $q_{r(.)}$ Is the rth percentile of the data, $median_m$ is the median of the pooled samples for gene m, and mad_m is the median absolute deviation of the pooled samples from gene m. The selection of r depends on the dataset, 75%, 90% or 95% percentiles are commonly used

for different studies. A cutoff (default is 5) is then applied to filter for 'outlier' genes. The idea of COPA is quite straight forward. It assumes that normal(control) samples are not likely to have extreme values, and COPA picks the extreme percentiles (outliers) that will not be "diluted" by heterogeneities in the cancer samples. The use of median and mad rather than means and standard deviations provides additional robustness to the method.

COPA only uses a single percentile during analysis, Tibshirani and Hastie extended COPA by using all the outlier detected (Tibshirani and Hastie, 2007). Their outlier sum (OS) statistic is defined as $OS_m = \frac{\sum_{k=1}^{n_1}[(y_{km} - median_m) \times I(y_{km} > q_{75_m} + IQR_m)]}{mad_m}$ where $I(A)$ is an indicator function of event A, q_{75_m} is the 75th percentile of gene m for the pooled samples, and IQR_m is the inter-quartile range of the pooled samples. This OS method is further improved by Wu. He claims that using median and mad from pooled samples might overestimate the normal group mean owing to the contamination by disease samples. He proposed to use outlier robust t-statistics (ORT) where the outliers are defined relative to the normal samples instead of the pooled samples (Wu, 2007).

$$ORT_m = \frac{\sum_{k=1}^{n_1}\left[\left(y_{km} - median_{m\_control}\right) \times I\left(y_{km} > q_{75_{m\_control}} + IQR_{m\_control}\right)\right]}{1.4826 \times median\left(\left\{y_{km} - median_{m\_case}\right\}_{1 \le k \le n_1}, \left\{x_{km} - median_{m\_control}\right\}_{1 \le k \le n_0}\right)}$$

where $median_{m\_case}$ and $median_{m\_control}$ are the median of case and control sample for gene m respectively; $q_{75_{m\_control}}$ is the 75th percentile of gene m for the control samples, and $IQR_{m\_control}$ is the inter-quartile range of the control samples.

In COPA, OS and ORT, the outliers are defined arbitrarily using a user specified cutoff. This could lead to the loss of analysis power or even mix some normal into the cases

if the cutoff is not chosen correctly. To solve this issue, change point detection algorithms were introduced to pick the ideal cutoff. For example, cancer likelihood ratio statistics (LRS) assumes that the case samples are a mixture of true cases and some normal samples. LRS first sorts the case data from the smallest to the largest ($y_{(k)}$ is the kth order statistic, assume we search for up-regulated genes). Then the point in case samples where true cases and all normal samples have the best separation is located by formula below (Hu, 2008).

$$LRS_m = \max_k \left(\frac{(n_0+k)*mean_m - \sum_1^{n_0} x_m - \sum_1^k y_{(k),m}}{\sqrt{\frac{(n_0+k)*(n_1-k)}{n}}} \right) where\ 1 \le k \le n_1$$

From both simulation and real datasets, LRS has been shown to have better power than COPA, ORT or OS. Similar strategy have also been used by maximum ordered subset t-statistics (MOST) (Lian, 2008). Other non-parametric methods to detect the cutoff have also been developed. E.g., Non-parametric change-point statistics (NPCPS) is based on a modified Kolmogorov statistic, it compares the data distribution of control and case samples to detect the existence of possible change-point in the case group (Wang, et al., 2011). Compared with other methods look for a significant change of mean or median, like t statistics, OS or LPS, NPCPS focuses on change in data distribution, its results often are orthogonal to those from parametric methods.

1.4 Batch effect and platform specific bias

Batch effects are commonly observed during experiments using high-throughput technology (Leek, et al., 2010; Luo, et al., 2010). In a broad term, they are the systematic non-biological differences among different batches of samples in the experiments.

1.4.1 Common sources

Batch effect can be caused by varieties of factors, e.g. the differences in sample preparation and processing protocols (source bias), different platform, chip type and lot (e.g. array quality may vary from lot to lot), different people/instrument or laboratory (Irizarry, et al., 2005) that performs the experiment or even different dates that the samples are processed (Luo, et al., 2010). Although some of the batch effects can be minimized through careful experimental planning, many others cannot be avoided. For example, in large scale studies, the samples often come from different hospitals, and then processed by different lab scientists and on different dates, these issues are often out of the control of researchers.

Platform bias can be seen as a special type of batch effect. It is a systematic bias between two or more groups of samples when the data measurements are acquired from different high-throughput platforms. For example, microarrays from different vendors have different probe designs. Even for the same vendor, different generation of the same type array can have substantial different probes. The differences on chemistry and kinetics of experiment assay of different platforms can lead to large platform specific biases. In the early days, microarrays designed by different companies showed considerable divergence even with the same samples (Tan, et al., 2003). Newer generations of arrays have demonstrated

better concordance, but the platform bias due to probe sequences still exists, especially for low expression genes (Canales, et al., 2006; Eklund and Szallasi, 2008).

To adjust signal intensities for experimental artifacts between all the samples, various normalization methods have been developed. These normalization procedures often re-calibrate and/or homogenize the data based on their intensity value distributions. They can compensate the measures for the effects of the differences in procedures among the samples, thus increase the precision of measurements. For Affymetrix expression microarray, commonly used normalization methods include dChip (Li and Wong, 2001), MAS5 (Pepper, et al., 2007), RMA (Irizarry, et al., 2003) and gcRMA (Gharaibeh, et al., 2008; Wu, et al., 2004). LOWESS-based methods are often used for cDNA two-color microarrays (Yang, et al., 2002). Upper quartile (UQ), DESeq (Anders and Huber, 2010), trimmed mean of M values (TMM) (Robinson and Oshlack, 2010) and reads per kilobase per million mapped reads (RPKM) normalization (Mortazavi, et al., 2008) are often used for RNA-Seq dataset (Dillies, et al., 2013). Some of the batch effect could be removed during normalization. However, because normalization methods are not specifically designed for adjusting data to remove the systematic differences between multiple groups of samples, significant batch effects still remain after normalization for most data sets. MAQC-II project and other studies have demonstrated that normalization is not sufficient for removing batch effects and other statistical procedures must be applied (Johnson, et al., 2007; Luo, et al., 2010; Sun, et al., 2011).

1.4.2 Batch effect removal methods

Multiple batch effect removal (BER) methods have been presented in the literatures. Batch mean centering (BMC) transforms the data by subtracting the sample mean for each batch. It is also named as zero-mean, or one-way analysis of variance adjustment, etc. Gene normalization (Gnorm) goes a step further. It performs Z-transformation standardization of the data using the standard deviation from each batch. Both BMC and Gnorm are special cases of the location and scale (L/S) adjustments methods.

Ratio-based methods (e.g. Ratio-G and Ratio-A) scale the sample expression intensities by reference expression intensities. If there are several reference control samples within each batch, the pooled reference intensities are calculated using either the arithmetic mean (Ratio-A) or geometric mean (Ratio-G) of the control samples.

Singular value decompositions (SVD) was used to correct for systematic biases in different batches of microarray expression data (Alter, et al., 2000; Nielsen, et al., 2002). SVD modifies the data by filtering out eigengenes and eigenarrays (eigensamples) that represent noise or experimental artifacts. Benito *et. al.* developed DWD (Distance weighted Discrimination). DWD is an adaptive support vector machine (SVM) method which finds an ideal separation hyper-plane and removes biases by projecting different batches of data on this hyper-plane (Monica Benito, *et. al* , 2004). SVD and DWD are more power methods compared with BMC and Gnorm, but they are also more complex, and require median to large sample size per batch to work properly. In addition, DWD can only be applied to two batches at a time. For more than two batches, a step-wise method can be applied, but this further increases the inconvenience using this method.

COMBAT (Combining Batches of Gene Expression Data) uses empirical Bayes approach to estimate parameters of the model and correct for both additive and multiplicative batch effects (W.E. Johnson, et. al., 2007). COMBAT assumes that factors contribute to batch effects often affect many genes in similar ways (i.e. higher/lower expression level, larger variability, etc.). It estimates the L/S model parameters that represent the batch effects by "pooling/borrowing information" across genes in each batch. This strategy enables COMBAT to provide more robust estimation for each gene compared with SVD and DWD, especially for batches with small sample size.

In many cases, the true sources of batch effects are either unknown or cannot be adequately explained or modeled with known factors. Surrogate variable analysis (SVA) can be used to estimate the sources of batch effects directly from the high-throughput data. Then these SVA variables can be added back into the linear model for batch effect removal (Leek, et al., 2012; Leek and Storey, 2007).

Most of the presented methods were developed for expression microarrays, however, many of them can be used for other microarray and NGS platforms as well. Successful application of different BER method on RNA-Seq (Taub, et al., 2010) and miRNA-Seq (Guo, et al., 2014) have been reported. Platform specific methods have also been proposed. For example, Scharpf *et al.* has developed a multilevel model specifically to address batch effects in copy number estimation with SNP arrays (Scharpf, et al., 2011).

References

Allison, D.B., *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus, *Nature reviews. Genetics*, **7**, 55-65.

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

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

Ashburner, M., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature genetics*, **25**, 25-29.

Booth, M.J., *et al.* (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution, *Science*, **336**, 934-937.

Canales, R.D., *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms, *Nature biotechnology*, **24**, 1115-1122.

Chen, L.S., *et al.* (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data, *The American Journal of Human Genetics*, **86**, 860-871.

Di Camillo, B., *et al.* (2012) Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment, *PloS one*, **7**, e32200.

Dillies, M.A., *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, *Briefings in bioinformatics*, **14**, 671-683.

Dinu, I., *et al.* (2007) Improving gene set analysis of microarray data by SAM-GS, *BMC bioinformatics*, **8**, 242.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes, *Annals of Applied Statistics*, **1**, 18.

Eklund, A.C. and Szallasi, Z. (2008) Correction of technical bias in clinical microarray data improves concordance with known biological information, *Genome biology*, **9**, R26.

Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association, *Bioinformatics*, **23**, 257-258.

Fisher, R., Pusztai, L. and Swanton, C. (2013) Cancer heterogeneity: implications for targeted therapeutics, *British journal of cancer*, **108**, 479-485.

Gharaibeh, R.Z., Fodor, A.A. and Gibas, C.J. (2008) Background correction using dinucleotide affinities improves the performance of GCRMA, *BMC bioinformatics*, **9**, 452.

Goeman, J.J., *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome, *Bioinformatics*, **20**, 93-99.

Guo, Y., *et al.* (2014) Statistical strategies for microRNAseq batch effect reduction, *Translational Cancer Research*, **3**, 260-265.

Holmans, P., *et al.* (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder, *American journal of human genetics*, **85**, 13-24.

Hu, J. (2008) Cancer outlier detection based on likelihood ratio test, *Bioinformatics*, **24**, 2193-2199.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, **4**, 44-57.

Huang da, W., *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, *Nucleic acids research*, **35**, W169-175.

Irizarry, R.A., *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249-264.

Irizarry, R.A., *et al.* (2009) The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores, *Nature genetics*, **41**, 178-186.

Irizarry, R.A., *et al.* (2009) Gene set enrichment analysis made simple, *Statistical methods in medical research*, **18**, 565-575.

Irizarry, R.A., *et al.* (2005) Multiple-laboratory comparison of microarray platforms, *Nature methods*, **2**, 345-350.

Jia, P., *et al.* (2010) Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data, *Schizophrenia research*, **122**, 38-42.

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

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research*, **28**, 27-30.

Kim, S.Y. and Volsky, D.J. (2005) PAGE: parametric analysis of gene set enrichment, *BMC bioinformatics*, **6**, 144.

Leek, J.T., *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments, *Bioinformatics*, **28**, 882-883.

Leek, J.T., *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data, *Nature reviews. Genetics*, **11**, 733-739.

Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genet*, **3**, 1724-1735.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 31-36.

Lian, H. (2008) MOST: detecting cancer differential gene expression, *Biostatistics*, **9**, 411-418.

Liu, Q., *et al.* (2007) Comparative evaluation of gene-set analysis methods, *BMC bioinformatics*, **8**, 431.

Luo, J., *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data, *Pharmacogenomics J*, **10**, 278-291.

MacDonald, J.W. and Ghosh, D. (2006) COPA--cancer outlier profile analysis, *Bioinformatics*, **22**, 2950-2951.

Maciejewski, H. (2013) Gene set analysis methods: statistical models and methodological differences, *Briefings in bioinformatics*.

Mansmann, U. and Meister, R. (2005) Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach, *Methods Inf Med*, **44**, 449-453.

Mootha, V.K., *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature genetics*, **34**, 267-273.

Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature methods*, **5**, 621-628.

Nam, D. and Kim, S.Y. (2008) Gene-set approach for expression pattern analysis, *Briefings in bioinformatics*, **9**, 189-197.

Nielsen, T.O., *et al.* (2002) Molecular characterisation of soft tissue tumours: a gene expression study, *Lancet*, **359**, 1301-1307.

O'Dushlaine, C., *et al.* (2009) The SNP ratio test: pathway analysis of genome-wide association datasets, *Bioinformatics*, **25**, 2762-2763.

Pepper, S.D., *et al.* (2007) The utility of MAS5 expression summary and detection call algorithms, *BMC bioinformatics*, **8**, 273.

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

Russnes, H.G., *et al.* (2011) Insight into the heterogeneity of breast cancer through next-generation sequencing, *The Journal of Clinical Investigation*, **121**, 3810-3818.

Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, *J Mol Biol*, **94**, 441-448.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463-5467.

Scharpf, R.B., *et al.* (2011) A multilevel model to address batch effects in copy number estimation using SNP arrays, *Biostatistics*, **12**, 33-50.

Sole, X., *et al.* (2009) Biological convergence of cancer signatures, *PloS one*, **4**, e4544.

Subramanian, A., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

Sun, Z., *et al.* (2011) Batch effect correction for genome-wide methylation data with Illumina Infinium platform, *BMC Med Genomics*, **4**, 84.

Tan, P.K., *et al.* (2003) Evaluation of gene expression measurements from commercial microarray platforms, *Nucleic acids research*, **31**, 5676-5684.

Taub, M., Bravo, H. and Irizarry, R. (2010) Overcoming bias and systematic errors in next generation sequencing data, *Genome Med*, **2**, 1-5.

Tian, L., *et al.* (2005) Discovering statistically significant pathways in expression profiling studies, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13544-13549.

Tibshirani, R. and Hastie, T. (2007) Outlier sums for differential gene expression analysis, *Biostatistics*, **8**, 2-8.

Tomlins, S.A., *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer, *Science*, **310**, 644-648.

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

Wang, K., Li, M. and Bucan, M. (2007) Pathway-based approaches for analysis of genomewide association studies, *American journal of human genetics*, **81**, 1278-1283.

Wang, L., *et al.* (2011) Gene set analysis of genome-wide association studies: Methodological issues and perspectives, *Genomics*, **98**, 1-8.

Wang, Y., *et al.* (2011) Non-parametric change-point method for differential gene expression detection, *PloS one*, **6**, e20060.

Weng, L., *et al.* (2011) SNP-based pathway enrichment analysis for genome-wide association studies, *BMC bioinformatics*, **12**, 99.

Wu, B. (2007) Cancer outlier differential gene expression detection, *Biostatistics*, **8**, 566-575.

Wu, Z., *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays, *Journal of the American statistical Association*, **99**, 909-917.

Yang, Y.H., *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic acids research*, **30**, e15.

CHAPTER 2 MULTI-PLATFROM GENE SET ANALYSIS

Abstract

Gene set analysis is a popular method for large-scale genomic studies. Because genes that have common biological features are analyzed jointly, gene set analysis often achieves better power and generates more biologically informative results. With the advancement of technologies, genomic studies with multi-platform data have become increasingly common. Several strategies have been proposed that integrate genomic data from multiple platforms to perform gene set analysis. To evaluate the performances of existing integrative gene set methods under various scenarios, we conduct a comparative simulation analysis based on the TCGA breast cancer data set. Our analysis results demonstrate that multi-platform methods outperform single platform methods.

2.1 Introduction

High-throughput, genome-wide assays, such as microarray and next-generation sequencing, have become more reliable and affordable. With the ever-increasing throughput and the scale of omics studies, more and more projects choose to measure multiple genomic features (e.g., gene expression, methylation, gene mutation, copy number, promoter binding, and protein expression) on the same samples. Evaluating multiple genome features can lead to a better examination of functional responses and provide a comprehensive understanding of the underlying biological mechanisms. In recent years, well-known, large-scale projects, such as the Cancer Genome Atlas (TCGA) (2012), the Cancer Cell Line Encyclopedia(CCLE), and the Encyclopedia of DNA Elements (ENCODE), have generated genomic profiles across

multiple platforms. In addition, more and more recent projects in the Gene Expression Omnibus (GEO) contain multi-platform data. With diverse data types from different platforms, it becomes challenging to properly integrate, analyze, and interpret the results to obtain biological insights. Gene set analysis is a powerful strategy developed to analyze large-scale profiling data. Instead of studying one gene at a time, gene set analysis focuses on a set of related genes, such as genes in one KEGG pathway (Kanehisa and Goto, 2000) or those related to the same Gene Ontology (Ashburner, et al., 2000) term. Joint analysis of genes in a set often improves power, especially when the signals of individual genes are moderate. Because the set itself often has biological meanings, gene set analysis also facilitates the interpretation of experiment results and helps to identify important biological findings (Ramanan, et al., 2012). Many methods have been developed to perform gene set analysis in a single platform, for example, GSEA (Subramanian, et al., 2005), GSA (Efron and Tibshirani, 2007), and Globaltest (Goeman, et al., 2004). Several review articles have been published that discuss the performances of different gene set methods (Ackermann and Strimmer, 2009; Goeman and Buhlmann, 2007; Hung, et al., 2012; Maciejewski, 2013).

Gene set analysis on multi-platform genomic data is gaining momentum. Approaches can be roughly classified into three different categories, characterized by how the multi-platform information is integrated. The first type performs a gene set analysis on each platform and then combines the single platform information, such as p-values, (e.g., (Jia, et al., 2012)). Such a strategy is commonly used when the multi-platform data are from similar but not identical samples. The second strategy, such as that employed in the SumZ approach of Xiong, et al., (2012), first sums the gene-specific association score of each platform to

compute a multi-platform score for each gene and then uses the gene scores to perform gene set analysis. The third strategy is similar to the second except that it directly derives the multi-platform gene scores using data from all platforms simultaneously. One representative approach is the integrative approach (INT) proposed by Tyekucheva et al., which uses a logistic regression with all multi-platform values of a gene as predictors and takes the model deviances as the gene scores for downstream gene set analysis (Tyekucheva, et al., 2011). Bayesian methods have also been developed to analyze multi-platform genomic data, e.g., iBAG (Wang, et al., 2013) and PARADIGM (Vaske, et al., 2010). Compared with traditional gene set methods, Bayesian methods often use extensive knowledge of the biological relationships among different data platforms and/or the interactions between studied genes.

2.2 Methods

2.2.1 TCGA data sets

We downloaded the TCGA breast cancer data from the National Cancer Institute (NCI) ftp site in January 2013. We focused on the level 3 gene summary data from RNA-Seq (RNA Sequencing), methylation, and copy number variation (CNV) and extracted 530 common samples (480 case samples and 50 control samples) and 10371 common genes shared among the three platforms. For RNA-Seq data, the log2(RPKM) (i.e., reads per kilo base per million) were used as gene expression values. Before the log2 transformation, a minimal value (0.0001) was added to prevent infinite values. For methylation, the mean beta value of all of the probes mapped to a gene were first computed and then converted into an M value for each gene (Du, et al., 2010). The CNV values were provided in log2 format. Within each

platform, the data were standardized to have mean 0 and standard deviation 1. The TCGA breast cancer data were used to perform simulations and a real data analysis.

2.2.2 Simulations design

We generated simulated data based on the TCGA breast cancer dataset, which contains 480 cancer samples and 50 control samples (i.e., the case proportion η =91%). First, we created 207 non-overlapping gene sets by randomly drawing genes from the 10371 genes without replacement. The sizes of the 207 gene sets were randomly determined based on the size distribution of the MSigDB canonical pathways (Subramanian, et al., 2005). The genomic data for cases and controls were simulated using the scheme described in the Tyekucheva study (Tyekucheva, et al., 2011). In short, we first shuffled the case-control labels to remove any association that may exist in the original data. Then, we randomly selected 10 gene sets as causal gene sets and "spiked in" signals into the causal gene sets as detailed below. We performed 300 replicates for each simulation scenario.

Given a causal gene set, we randomly selected α% (25%, 50%, or 75%) of the genes as causal genes. For each causal gene, one platform was randomly selected as causal and Δk was added to the genomic values of the causal platform for cases. The value of Δk was derived such that the two-sample t-test between cases and controls had power β (0.2, 0.4, 0.6, 0.8, or 0.9).

2.2.3 Multi-platform methods for gene set analysis

The general steps of integrative gene set analysis start with computing gene-specific association scores (gene scores in short) of multi-platform data and then using these scores to perform gene set analysis. For the gene set analysis, we conducted the gene set tests using R function "geneSetTest" from the R/Bioconductor package "limma" (Smyth, 2005) and obtained p-values for each gene set. The ranks of the gene scores were used instead of the actual scores (Michaud, et al., 2008). We selected different thresholds of p-value cutoff and computed the true positive rate (TPR), i.e., the percentage of the causal gene sets truly identified, and the false positive rate (FPR), i.e., the percentage of non-causal gene sets falsely identified as causal gene sets. We plotted the receiver operating characteristic (ROC) curves to compare the performances of the different methods using R. Below; we describe how different methods obtain the multi-platform gene scores considered in the simulation study.

- Integrative (INT) analysis (Tyekucheva, et al., 2011):

  For each gene, regress the disease status on the genomic variables from all platforms using a logistic regression model. The multi-platform gene scores are computed by taking the differences of the deviances between the null models (excluding genomic predictors) and the full models (including all genomic predictors).

- Hotelling's T2 (HT2):

  For each gene, perform the Hotelling's T2 test to conduct a case-control comparison using the genomic variables from all platforms (Xiong, et al., 2002). The multi-platform gene scores are the Hotelling's T2 statistics.

- SumZ (Xiong, et al., 2012):

  For each gene at each platform, calculate the association score (t-statistics). Next, use permutations to obtain the null distribution of the t-statistics within each platform. Then, standardize the t-statistics of each gene based on the null distributions. Finally, for each gene, obtain the gene scores by taking the sum of the standardized values across different platforms.

- Deviance summarization:

  For each gene at each platform, fit the logistic regression under the null model (i.e., excluding the genomic variable) and under the full model (i.e., including the genomic variable). Next, obtain the deviance difference between the two models. Finally, for each gene, take the average of the deviance difference across platforms as the multi-platform gene scores (referred to as AveD). The method of MaxD is obtained in the same fashion except that the maximum is used rather than the average.

- Single platform method (benchmark):

  For each gene at each platform, perform the same analysis as described in "deviance summarization". Then, obtain the single-platform gene scores by taking the deviance difference between the null model and the full model. We applied this strategy on methylation, CNV, and RNA-Seq expression platforms and referred to the corresponding methods as Methy, CNV, and Exp, respectively.

2.3 Results

We evaluated the abilities of AveD, MaxD, INT, SumZ, and Hotelling's T2 (HT2) to correctly identify causal gene sets under various parameter settings. Single platform methods Methy, CNV, and Exp were used to benchmark the performances of the multi-platform methods. Fig. 2.1 shows the ROC plots under different proportions of causal genes in a causal set, i.e., $\alpha = 75\%, 50\%,$ and 25% for Figs. 2.1A, 2.1B, and 2.1C, respectively, while fixing β (power) at 0.8 and η (percentage of case samples) at 0.91. The corresponding AUCs (Area under the curve) are summarized in Table 2.1. From Fig. 2.1A, it is clear that multi-platform methods outperformed single platform methods. Among the multi-platform methods, Hotelling's T2 and INT had similar performances, and these methods had the best performances among all methods. AveD and MaxD had slightly lower TPRs than INT, and SumZ followed closely. In Figs. 2.1B and 2.1C, the relative performance among different methods stayed the same as in Fig. 2.1A, except that the TPRs decreased when α decreased. The same patterns were observed for β = 0.6 (Fig. 2.2; Table 2.1).

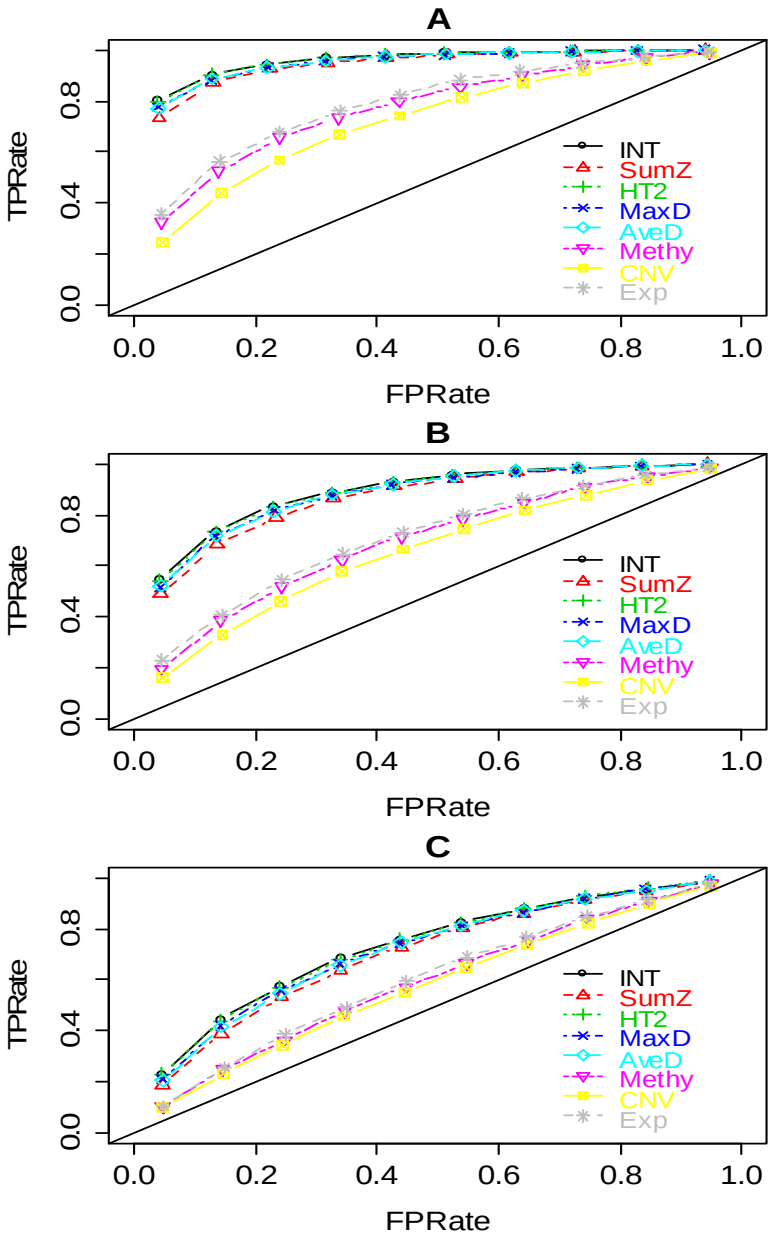


Fig. 2.1. ROC plots for gene set methods at different α levels (A: α = 75%; B: α = 50%; and C: α=25%). The simulated data were generated with β=0.8 and η=0.91

Table 2.1. AUC for the performances of different methods at various α (percentage of genes that are causal) and β (power) settings.

α values	β values	INT	SumZ	HT2	MaxD	AvgD	Methy	CNV	Exp
25%	0.6	0.623	0.593	0.621	0.603	0.603	0.519	0.500	0.527
50%	0.6	0.759	0.732	0.757	0.745	0.745	0.606	0.561	0.613
75%	0.6	0.833	0.809	0.832	0.821	0.821	0.670	0.613	0.683
25%	0.8	0.670	0.644	0.669	0.657	0.655	0.539	0.525	0.548
50%	0.8	0.813	0.792	0.812	0.803	0.804	0.637	0.602	0.650
75%	0.8	0.870	0.857	0.869	0.862	0.862	0.710	0.666	0.725
75%	0.9	0.885	0.875	0.884	0.881	0.880	0.732	0.696	0.740
75%	0.4	0.771	0.735	0.769	0.750	0.751	0.616	0.562	0.630
75%	0.2	0.657	0.619	0.653	0.630	0.631	0.544	0.510	0.557

Summary AUC results for Fig. 1, Fig. 2 and Supplementary Fig. 1.

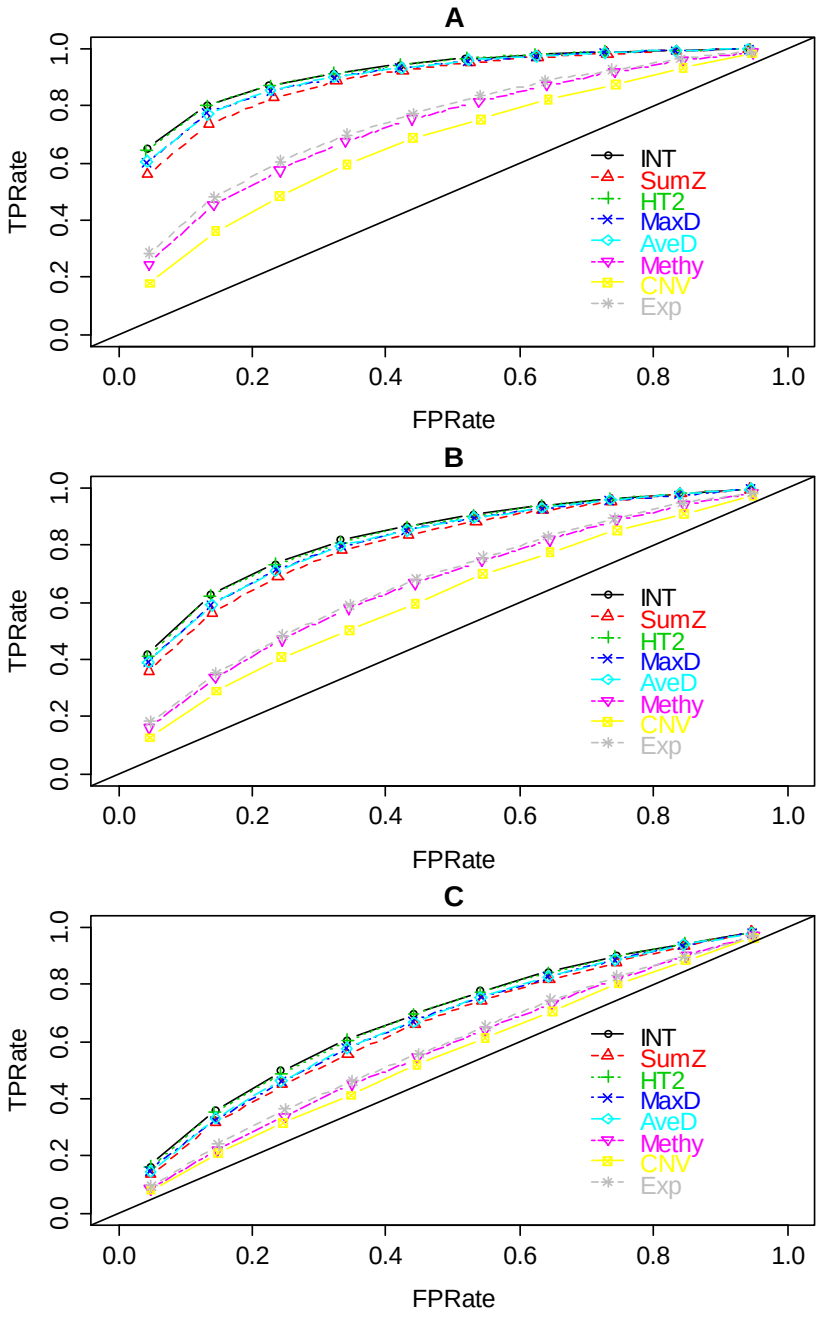


Fig. 2.2. ROC plots for gene set methods at different α levels (A: α = 75%; B: α = 50%; and C: α=25%). The simulated data were generated with β=0.6 and η= 0.91.

Fig. 2.3 shows the ROC plots under different β levels, i.e., $\beta = (0.9, 0.8, 0.6, 0.4,$ and $0.2)$ when α = 75% and η = 0.91. (The AUC values are shown in Table 2.1). The patterns for the relative performances of different methods were observed to be similar to those of Fig. 2.1. As expected, all methods performed better when the difference between case and control became larger (i.e., larger β).

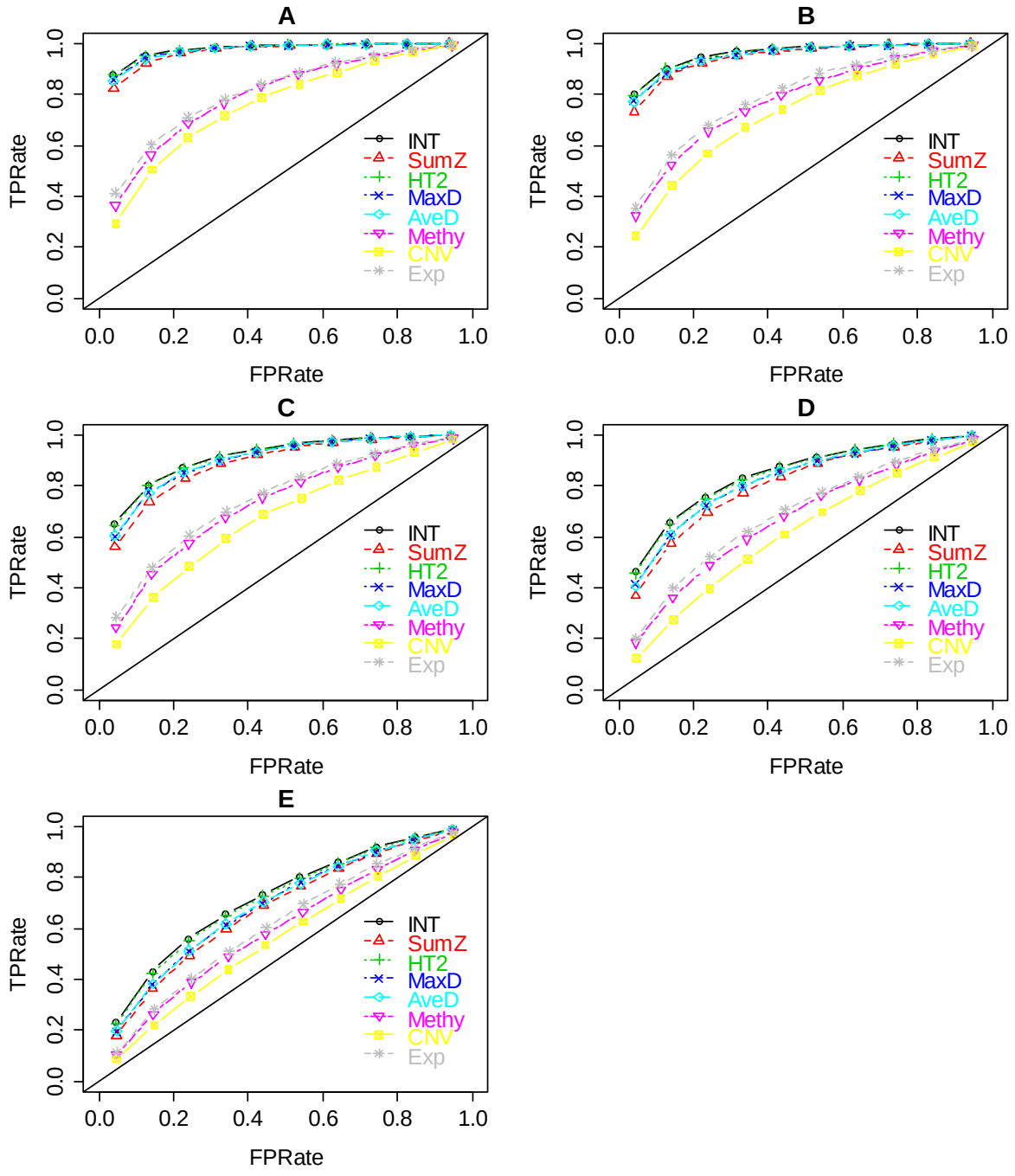


Fig. 2.3. ROC plots for gene set methods at different β levels (A: β = 0.9; B: β = 0.8; C: β=0.6; D: β=0.4; and E: β=0.2). The simulated data were generated with α = 75% and η = 0.91

The case proportion, η, is known to affect the power of statistical methods (Evans and Purcell, 2012). We repeated the studies for η = 0.5 (Fig. 2.4) and 0.1 (Fig. 2.5); similar results were observed under these scenarios

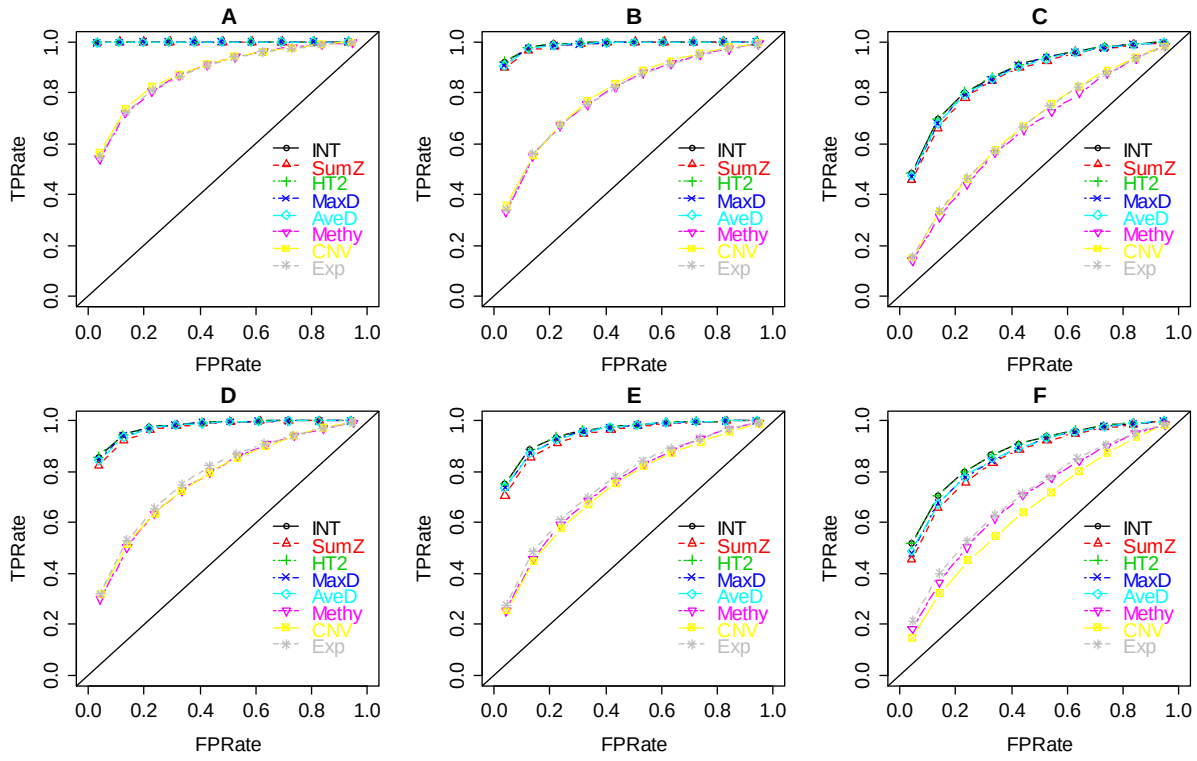


Fig. 2.4. ROC plots for gene set methods when prevalence η=0.50.

A: α=75%, β = 0.8; B: α=50%, β = 0.8; C: α=25%, β = 0.8;D: α=50%, β = 0.6; E: α=50%, β = 0.4; and F: α=50%, β = 0.2.

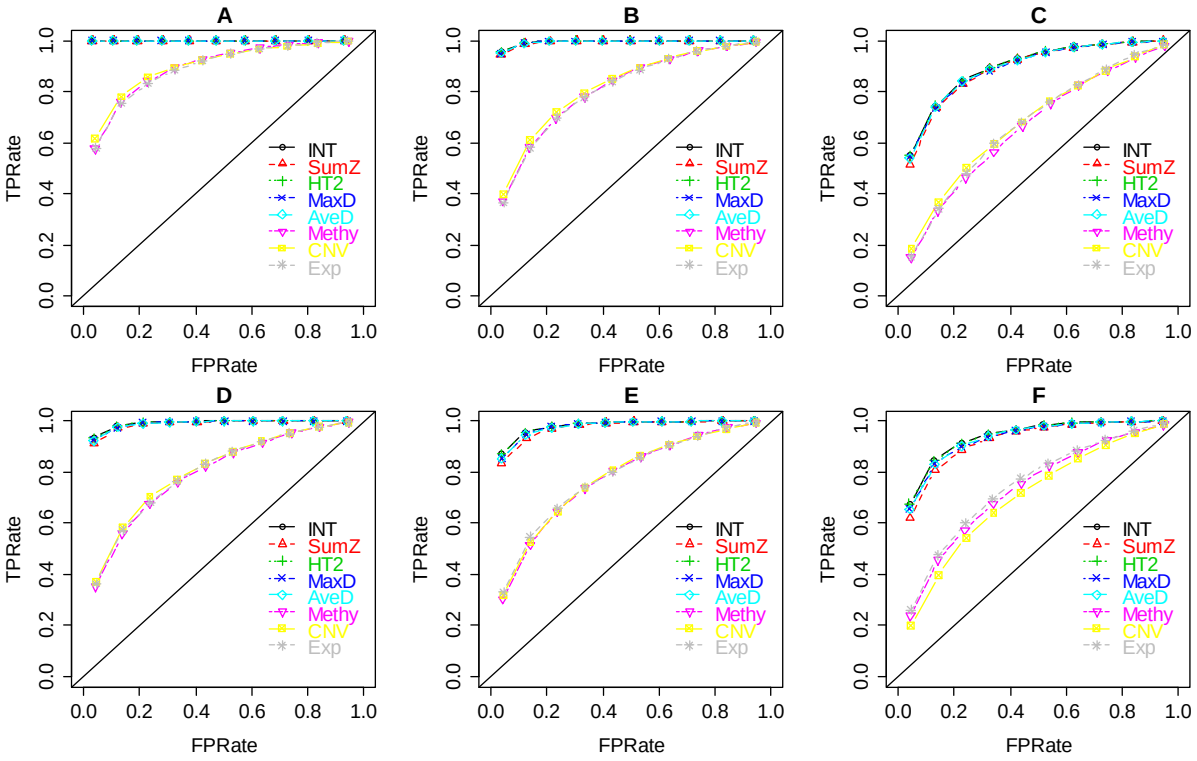


Fig. 2.5 ROC plots for gene set methods when prevalence η=0.10.

A: α=75%, β = 0.8; B: α=50%, β = 0.8; C: α=25%, β = 0.8; D: α=50%, β = 0.6; E: α=50%, β = 0.4; and F: α=50%, β = 0.2.

2.4 Discussion

In the presented work, we compared different multi-platform methods for gene set analysis using extensive simulated studies. First, when there is no sample heterogeneity, we found that INT and Hotelling's T2 method had the best performances compared to other methods. INT might have wider applicability compared to Hotelling's T2 because it can accommodate covariates.

In our analysis, we ignored the issues of missing values by focusing on genes with complete observations in all platforms. In reality, missing data are commonly observed in large-scale studies because of the experimental conditions, individual sample differences or platform constraints. When a considerable amount of data are missing, removing all the samples or genes with missing data could lead to substantial loss of information. To address this issue, imputing can be used to fill in the missing values. Performing self-contained gene set analysis tests is another strategy (Tyekucheva, et al., 2011). Further research is needed to characterize the patterns of missing data on different platforms, understand their impact on the gene set analysis, and develop the proper statistical methods for missing data.

References

(2012) Comprehensive molecular portraits of human breast tumours, Nature, 490, 61-70.

Ackermann, M. and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis, BMC bioinformatics, 10, 47.

Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature genetics, 25, 25-29.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289-300.

Du, P., et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC bioinformatics, 11, 587.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes, Annals of Applied Statistics, 1, 18.

Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics, 23, 980-987.

Goeman, J.J., et al. (2004) A global test for groups of genes: testing association with a clinical outcome, Bioinformatics, 20, 93-99.

Hung, J.H., et al. (2012) Gene set enrichment analysis: performance evaluation and usage guidelines, Briefings in bioinformatics, 13, 281-291.

Jia, P., Liu, Y. and Zhao, Z. (2012) Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer, BMC systems biology, 6 Suppl 3, S13.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, Nucleic acids research, 28, 27-30.

Maciejewski, H. (2013) Gene set analysis methods: statistical models and methodological differences, Briefings in bioinformatics.

Michaud, J., et al. (2008) Integrative analysis of RUNX1 downstream pathways and target genes, BMC genomics, 9, 363.

Ramanan, V.K., et al. (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development, Trends in genetics : TIG, 28, 323-332.

Smyth, G.K. (2005) Limma: linear models for microarray data, Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor, 397--420.

Subramanian, A., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

Tyekucheva, S., et al. (2011) Integrating diverse genomic data using gene sets, Genome biology, 12, R105.

Vaske, C.J., et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, Bioinformatics, 26, i237-245.

Wang, W., et al. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data, Bioinformatics, 29, 149-159.

Xiong, M., Zhao, J. and Boerwinkle, E. (2002) Generalized T2 test for genome association studies, American journal of human genetics, 70, 1257-1268.

Xiong, Q., et al. (2012) Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets, Genome research, 22, 386-397.

CHAPTER 3 MULTI-PLATFROM GENE SET ANALYSIS WITH HETEROGENEOUS SAMPLES

Abstract

We find that existing methods for multi-platform gene set analysis are less effective when sample heterogeneity exists. To address this issue, we develop three methods for multi-platform genomic data with heterogeneity: two non-parametric methods, *MPMWS* (Multi-Platform Mann-Whitney Statistics) and *MPORT* (Multi-Platform Outlier Robust T-statistics), and a parametric method, *MPLRS* (Multi-Platform Likelihood Ratio Statistics). Using simulations, we show that the proposed MPMWS method has higher power for heterogeneous samples and comparable performance for homogeneous samples when compared to existing methods. Our real data applications to two TCGA datasets also suggest that the proposed methods are able to identify novel pathways that are missed by other strategies.

3.1 Introduction

Sample heterogeneity refers to molecular and cellular differences among biological samples. Such differences are commonly encountered in complex diseases like cancer, where cases with different genotypes, genomic copy numbers, or expression patterns often lead to different disease progressions and treatment strategies (Fisher, et al., 2013; Russnes, et al., 2011). Several methods have been developed to address sample heterogeneity, e.g., cancer outlier profile analysis (COPR) (MacDonald and Ghosh, 2006), outlier sum (OS) (Tibshirani and Hastie, 2007), outlier robust t-statistics (ORT) (Wu, 2007), cancer likelihood ratio

statistics (LRS) (Hu, 2008), and non-parametric change-point statistics (NPCPS) (Wang, et al., 2011). While the superiority of these methods over ordinary analysis has been demonstrated with heterogeneous data in a single platform, to the best of our knowledge, there are no corresponding gene set approaches for multi-platform heterogeneous data. The impact of sample heterogeneity on multi-platform analyses can be more substantial than on single platform analyses. First, the level of heterogeneity can be different from platform to platform, e.g., platforms such as somatic mutations and DNA methylation have much higher diversity (heterogeneity) among individuals and samples than DNA copy number (Aryee, et al., 2013; Chin, et al., 2011). In addition, the heterogeneous subsets can be different from one platform to another, e.g., some samples might have changes on platform A but no changes on platform B, while different subsets of samples have changes on platform B but not on platform A. Such a scenario may lead to power loss due to the attenuation of signals when the association is evaluated across platforms. In contrast, a multi-platform method that can tackle platform-specific heterogeneous data would be able to identify the signals when integrating information across platforms.

In this study, we perform simulation studies to systematically evaluate different integrative methods under a range of scenarios. We observe that the true positive rates and the true negative rates of existing multi-platform gene set methods decrease dramatically when heterogeneity exists. These results motivated us to construct three methods to account for sample heterogeneity in multi-platform gene set analysis: *MPMWS* (Multi-Platform Mann-Whitney Statistics), *MPORT* (Multi-Platform Outlier Robust T-statistics), and *MPLRS*

(Multi-Platform Likelihood Ratio Statistics). We use simulations and real data analyses to demonstrate the utility of these methods under various conditions.

3.2 Methods

3.2.1 TCGA data sets

We use the same TCGA datasets as Chapter 2.2.1.We downloaded the TCGA breast cancer data from the National Cancer Institute (NCI) ftp site in January 2013. We focused on the level 3 gene summary data from RNA-Seq (RNA Sequencing), methylation, and copy number variation (CNV) and extracted 530 common samples (480 case samples and 50 control samples) and 10371 common genes shared among the three platforms. For RNA-Seq data, the log2(RPKM) (i.e., reads per kilo base per million) were used as gene expression values. Before the log2 transformation, a minimal value (0.0001) was added to prevent infinite values. For methylation, the mean beta value of all of the probes mapped to a gene were first computed and then converted into an M value for each gene (Du, et al., 2010). The CNV values were provided in log2 format. Within each platform, the data were standardized to have mean 0 and standard deviation 1. The TCGA breast cancer data were used to perform simulations and a real data analysis. We also performed a data analysis on the TCGA KIRC (Kidney Renal Clear Cell Carcinoma) data set, for which we applied the same procedures of data processing and obtained 486 common samples (463 case samples and 23 control samples) and 11182 common genes shared among the three platforms of methylation, CNV, and RNA-Seq data.

3.2.2 Simulations design

We generated simulated data based on the TCGA breast cancer dataset, which contains 480 cancer samples and 50 control samples (i.e., the case proportion η =91%). First, we created 207 non-overlapping gene sets by randomly drawing genes from the 10371 genes without replacement. The sizes of the 207 gene sets were randomly determined based on the size distribution of the MSigDB canonical pathways (Subramanian, et al., 2005). The genomic data for cases and controls were simulated using the scheme described in the Tyekucheva study (Tyekucheva, et al., 2011). In short, we first shuffled the case-control labels to remove any association that may exist in the original data. Then, we randomly selected 10 gene sets as causal gene sets and "spiked in" signals into the causal gene sets as detailed below. We performed 300 replicates for each simulation scenario.

We considered two scenarios (referred to as Scenarios B1 and B2) to simulate datasets with sample heterogeneity. In Scenario B1, we followed the simulation scheme for Scenario A in chapter 2.2.2, except we randomly selected γ% (20%, 40%, 60%, 80%, 90%, or 100%) of the case samples as "true" cases for each causal gene. In other words, we only "spiked in" Δk signals into the (randomly selected) causal platform of the causal gene for the "true" cases. Because the causal platform of a causal gene was randomly selected, the causal genes in a platform are different from each other (though there may be some overlaps).

In Scenario B1, there is only a single causal platform for each causal gene for the "true" cases. In real biological situations, we often see genes that have changes in multiple platforms. To account for these scenarios, we considered Scenario B2, in which each causal gene is allowed to have changes in more than one platform. Specifically, let w be the number

of causal platforms of a casual gene; then, the probability of $w = (1, 2, 3)$ is $(4/8, 3/8, 1/8)$, respectively. That is, we first determined the number of causal platforms from Binomial (3, ½) and then converted $w = 0$ to $w = 1$. We then added Δk values to the genomic data of the causal platform(s) of a causal gene for the "true" cases.

3.2.3 Multi-platform gene set analysis methods accounting for sample heterogeneity

We constructed three multi-platform methods to address sample heterogeneity. Specifically, we extended two current methods designed for single platform analysis to the multi-platform setting, i.e., MPORT (based on ORT of Wu (2007)) and MPLRS (based on the LRS of Hu (2008)). We also developed a non-parametric method, MPMWS, which obtains the gene scores based on the Mann-Whitney statistics and does not assume symmetric distributions for the genomic variables.

The general procedure of multi-platform heterogeneous methods is as follows. Assume that there are M genes and L platforms measured from n_0 control samples and n_1 case samples (i.e., in total, $n = n_0 + n_1$ samples). Let $x_{im\ell}$ be the observed value of the genomic variable for gene m and platform ℓ of sample i. For each gene, use the single platform method to compute association statistic $T_{m\ell}$ for platform ℓ. Next, similar to the SumZ method, use permutations to obtain a null distribution of the statistics for platform ℓ. Finally, calculate the standardized gene statistics within platform ℓ, denoted by $T'_{m\ell}$, using the mean and standard deviation (denoted by $\overline{T_{.\ell}}$ and $S_{.\ell}$, respectively) obtained from the permuted null distribution, i.e.,

$$T'_{m\ell} = (T_{m\ell} - \overline{T_{\cdot\ell}})/S_{\cdot\ell\cdot} + c_{l\cdot} \qquad (1)$$

As is done in the SumZ implementation, these scores are made positive by adding a constant, c_l, that is the absolute value of the most negative score across the platform. This translation makes all of the $T'_{m\ell}$ values positive but does not change the shape of their distribution. Then, the sum of the standardized gene statistics from each platform defines the multi-platform gene scores:

$$G_m = \sum_{\ell=1}^{L} T_{m\ell}{}'. \qquad (2)$$

The MPORT, MPLRS, and MPMWS methods differ only in how $T_{m\ell}$ is obtained. We show the formula for computing $T_{m\ell}$ when detecting "up-regulated" genes. (Here, the term "up-regulated" indicates the increase of numerical values rather than the biological "turning on" of the gene.) The approaches can be extended to detecting down-regulated genes by reversing the signs of the observed values.

- MPORT: $T_{m\ell}$ is computed using the outlier robust t-statistics (ORT) method (Wu, 2007). For each gene at each platform, calculate the mean absolute deviance (MAD) by $\text{MAD} = 1.4826 \times median(z_{im\ell})$, where

$$z_{im\ell} = \begin{cases} |\, x_{im\ell} - median_{control}| \;\; (if\ i\ is\ a\ control\ sample) \\ |\, x_{im\ell} - median_{case}| \;\; (if\ i\ is\ a\ case\ sample) \end{cases}. \qquad (3)$$

  For up-regulated genes, $T_{m\ell}$ is computed from the case samples using the ORT method:

$$T_{m\ell} = T_{ORT} = \frac{\sum[(x_{im\ell} - median_{case}) \times I(x_{im\ell} > q_{75_{case}} + IQR_{case})]}{MAD}, \qquad (4)$$

where $I(A)$ is an indicator function of event A, $q_{75_{case}}$ is the 75th percentile of $x_{im\ell}$ for the case samples, and IQR_{case} is the inter-quartile range of the case samples.

- MPLRS: $T_{m\ell}$ is computed using the LRS method (Hu, 2008). For up-regulated genes, the genomic data are sorted from the smallest to the largest under the constraint that all controls are ranked lower than cases.

$$S_{k,m\ell} = \sum_{i=1}^{k} x_{im\ell} \text{ where } (n_0 + 1 \leq k < n), \text{ and } \quad (5)$$

$$T_{m\ell} = T_{LRS} = \max_k \left(\frac{\frac{kS_{n,m\ell}}{n} - S_{k,m\ell}}{\sqrt{k\left(1-\frac{k}{n}\right)}} \right) \quad (6)$$

- MPMWS: $T_{m\ell}$ is computed using the nonparametric Mann-Whitney change point detection method implemented in R package CPM (Ross, 2013; Ross, et al., 2011). The genomic data are sorted from the smallest to the largest under the constraint that all controls are ranked lower than cases; the Mann-Whitney U statistic, $U_{k,m\ell}$, for each case sample is computed; and $T_{m\ell}$ is selected as the largest $U_{k,m\ell}$.

$$T_{m\ell} = T_{MWS} = \max_k \left(U_{k,m\ell} \right) \text{ where } (n_0 + 1 \leq k < n) \quad (7)$$

3.3 Results

To evaluate the performance under sample heterogeneity, we simulated datasets by randomly selecting $\gamma\%$ of case samples to be "true" cases. We focused our comparisons on the two represented approaches from Section 2.2.3 (i.e., INT and SumZ) and the three proposed methods for sample heterogeneity, i.e., MPLRS, MPORT, and MPMWS. The results of

Scenario B1 are shown in Fig. 3.1, where $\alpha = 75\%$, $\beta = 0.8$ and $\eta = 0.91$, and the percentage of "true" cases among all cases varies, i.e., $\gamma = (100\%, 90\%, 80\%, 60\%, 40\%, \text{ and } 20\%)$. The corresponding AUC values are presented in Table 3.1.

Table 3.1 AUC for the performances of different methods at various heterogeneity levels (γ values).

γ values	INT	SumZ	MPMWS	MPLRS	MPORT
100%	0.870737	0.855134	0.881578	0.858055	0.835458
90%	0.857839	0.838386	0.873798	0.855071	0.817425
80%	0.836551	0.811316	0.865067	0.853439	0.793585
60%	0.765627	0.730307	0.830807	0.835133	0.731692
40%	0.649072	0.611792	0.768069	0.794834	0.656148
20%	0.515046	0.496794	0.650281	0.699035	0.561845

$\alpha = 75\%$ and $\beta = 0.8$.

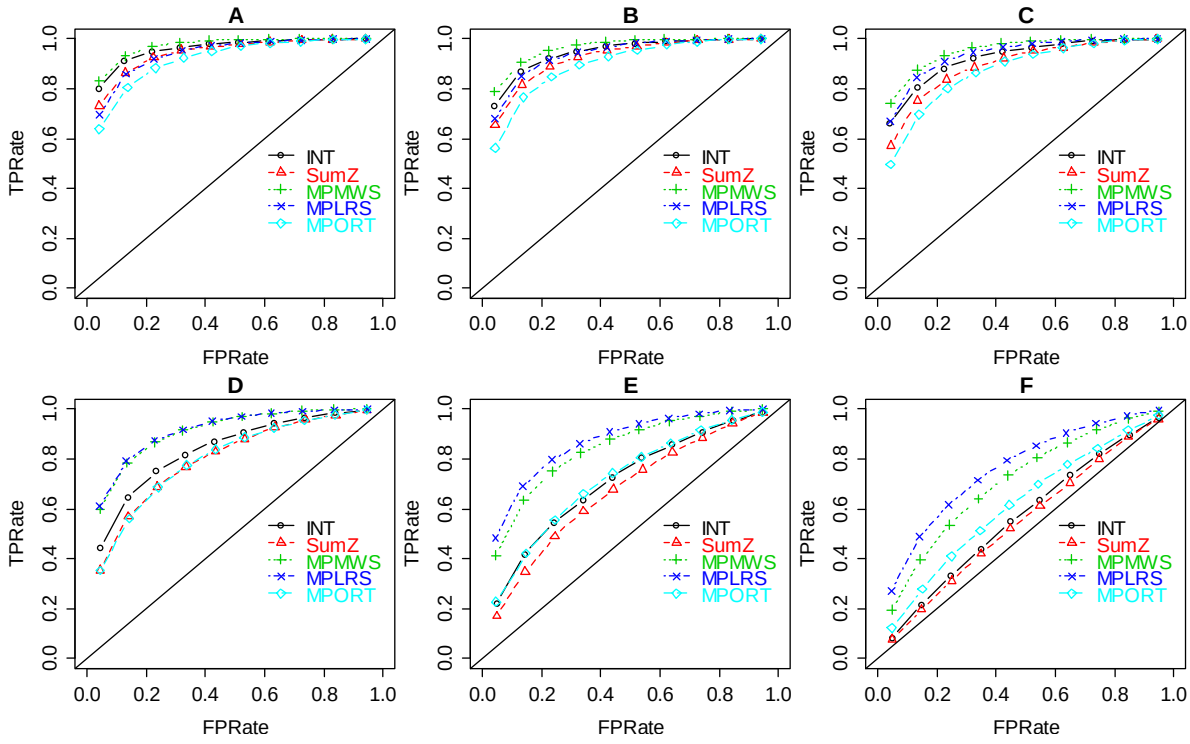


Fig. 3.1. ROC plots for gene set methods at different sample heterogeneity levels. (A: γ = 100%; B: γ = 90%; C: γ =80%; D: γ =60%; E: γ =40%; and F: γ=20%). The simulated data were generated with α =75%, β = 0.8, and η =0.91.

We see that INT and SumZ, which are designed for multi-platform homogeneous data, quickly lost power as γ decreased. In contrast, MPLRS and MPMWS retained good power when γ decreased. However, the relative performance between MPLRS and MPMWS depended on γ. When γ was low (e.g., $\leq$40%), MPLRS performed the best; when γ was 60%, MPLRS and MPMWS had similar power. However, when γ was high (e.g., $\geq$ 80%), MPLRS had less TPRs than MPMWS, sometimes even less than INT. MPORT performed inferior to MPLRS and MPMWS, and its power advantages over INT and SumZ did not show until γ became small, i.e., 20%~40%. Because γ is unknown in practice, MPMWS

appears to be the most robust choice; it yielded the highest or the second highest TPRs regardless of the γ values. Although the method is designed to account for sample heterogeneity, it had similar power to INT when samples were homogeneous (γ = 100%). This behavior is likely attributable to the fact that the genomic variables of certain platforms tended to deviate away from normal distributions, e.g., methylation values, and the non-parametric MPMWS is robust against non-normality. Finally, the improved TPR obtained using MPLRS and MPMWS with heterogeneous samples was observed when we repeated the analysis for α = 50% and η = 0.5 (Fig. 3.2) and 0.1 (Fig. 3.3)

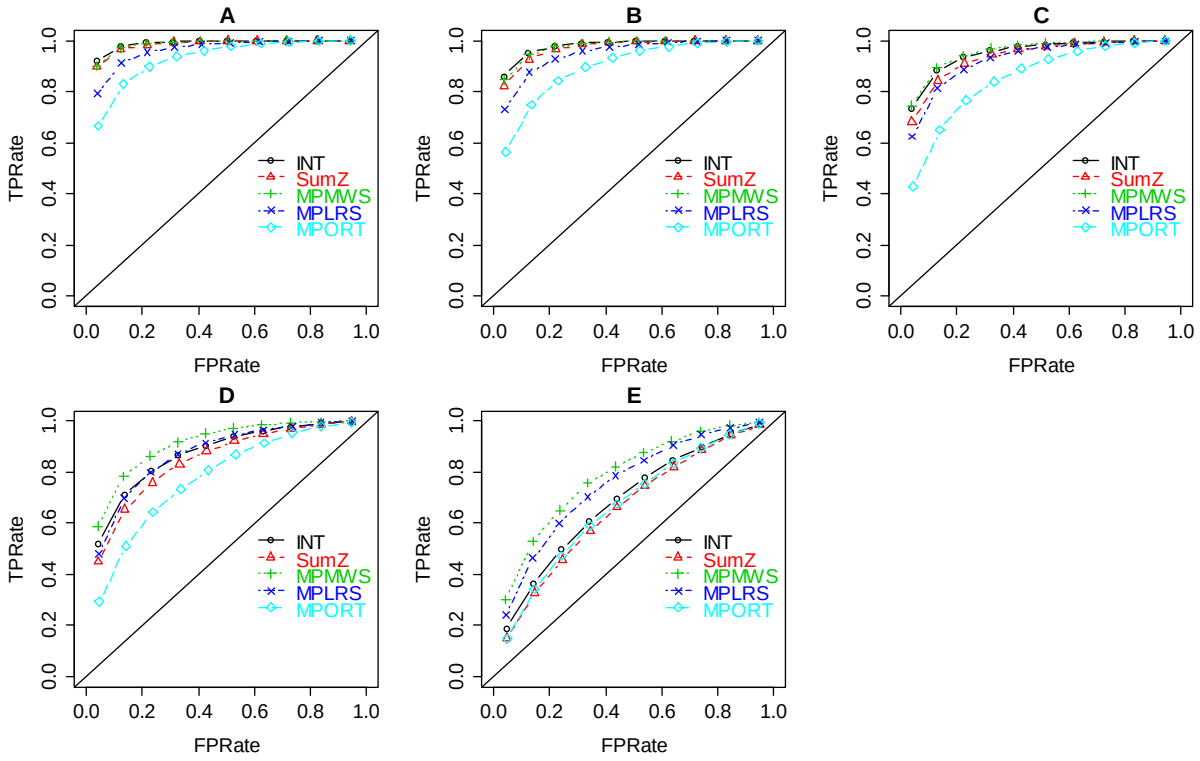


Fig. 3.2. ROC plots for gene set methods at different sample heterogeneity levels. (A: γ = 100%; B: γ =80%; C: γ =60%; D: γ =40%; and E: γ=20%). The simulated data were generated with α=50%, β = 0.8, and η=0.5.

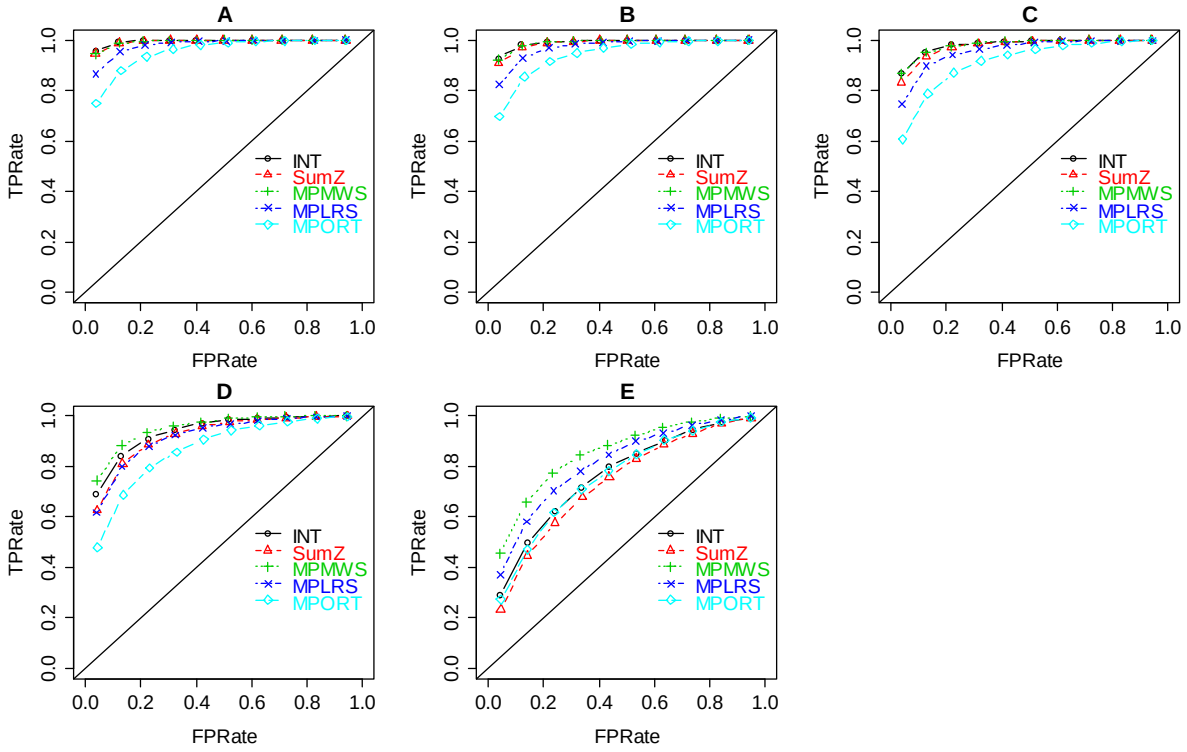


Fig. 3.3. ROC plots for gene set methods at different sample heterogeneity levels. (A: γ = 100%; B: γ =80%; C: γ =60%; D: γ =40%; and E: γ=20%). The simulated data were generated with α=50%, β = 0.8, and η=0.1.

By design, INT is good at identifying pathways with systematic changes, whereas MPMWS has robust power to detect pathways involving sample heterogeneity. In Table 3.2, we show the number of significant pathways and the number of true-positive (TP) pathways identified by INT and MPMWS. We observe that both methods identified many common significant/TP pathways. In addition, there was a high percentage of TPs among the common significant pathways, especially when the heterogeneity level was not extremely high. The results also show that each method identified some unique significant/TP pathways that were missed by the other method. For INT, the proportion of TPs among the unique pathways

became smaller as the heterogeneity increased. For MPMWS, the corresponding TP proportion stayed roughly constant until very severe heterogeneity (e.g., $\gamma = 20\%$).

Table 3.2. The average number of significant gene sets identified by INT and MPMWS at different heterogeneity levels

Significant Gene Set	Common Pathways			INT only			MPMWS only		
γ values	Positive	TP	TP (%)	Positive	TP	TP (%)	Positive	TP	TP (%)
100%	7.33	6.92	94.41%	8.52	1.08	12.68%	9.04	1.37	15.15%
90%	6.59	6.17	93.63%	8.75	1.13	12.91%	9.51	1.72	18.09%
80%	5.79	5.36	92.57%	9.09	1.24	13.64%	9.94	2.05	20.62%
60%	3.61	3.14	86.98%	9.54	1.28	13.42%	11.06	2.86	25.86%
40%	1.7	1.17	68.82%	9.78	1.04	10.63%	11.5	2.96	25.74%
20%	0.76	0.19	25.00%	9.72	0.63	6.48%	10.66	1.77	16.60%

Ten gene sets out of 207 gene sets were selected as causal, and the results were averaged over 300 repeats. (TP: True Positive)

The analyses above were performed under Scenario B1, where each causal gene only had one causal platform. We repeated the same analyses under Scenario B2, where each causal gene had at least 1 causal platform. We obtained very similar results as observed in Scenario B1 (Fig. 3.4)

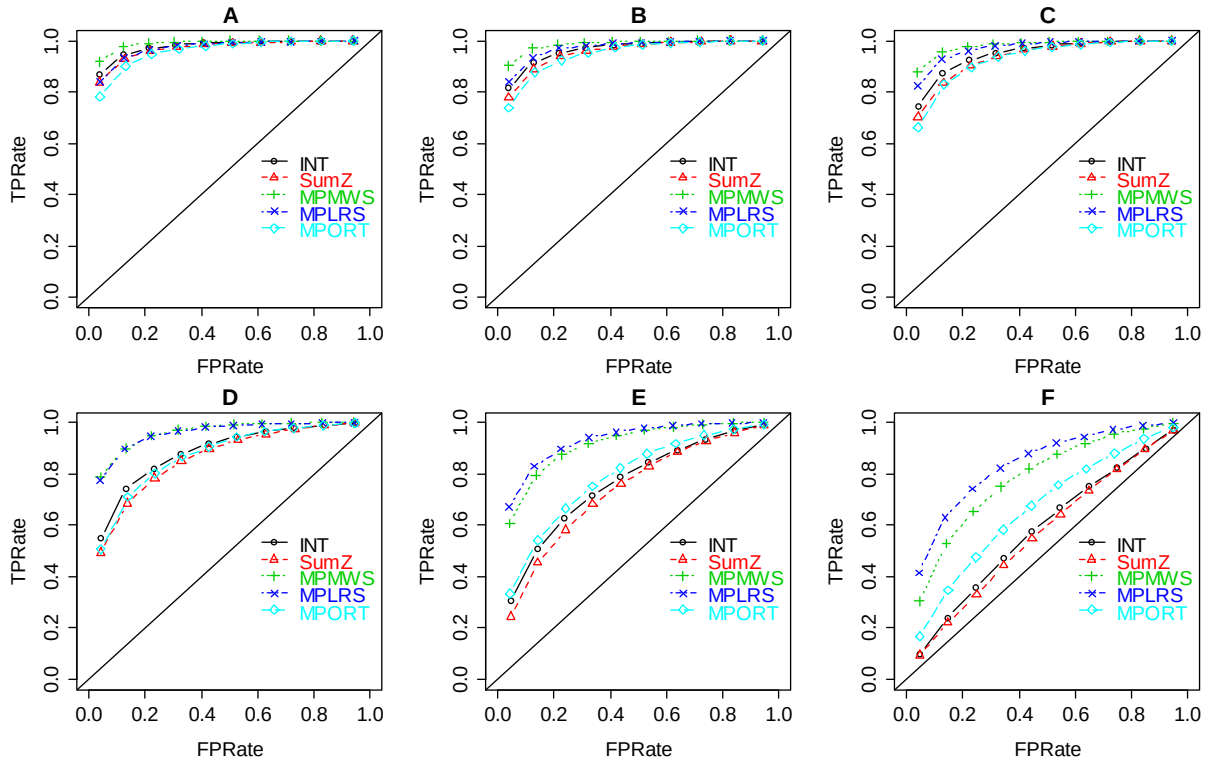


Fig. 3.4. ROC plots for gene set methods at different sample heterogeneity levels when multiple platform changes are allowed. (A: γ = 100%; B: γ =80%; C: γ =60%; D: γ =40%; and E: γ=20%). The simulated data were generated with α=75%, β = 0.8, and η=0.91.

We first considered the TCGA breast cancer data set containing methylation, CNV, and RNA-Seq measurements. We performed multi-platform gene set analyses on the 1452 MSigDB pathways using MPMWS and INT (i.e., the top two methods from Scenarios B1 and B2). Unlike the simulated gene sets, pathways in MSigDB often share common genes and can have significant overlaps. Fig. 3.5 shows the number of pathways identified by each method and their overlaps at FDR (False Discovery Rate) 0.05 using the Benjamini and Hochberg's FDR procedure (Benjamini and Hochberg, 1995). The numbers of significant pathways identified by INT and MPMWS were 116 and 78, respectively.

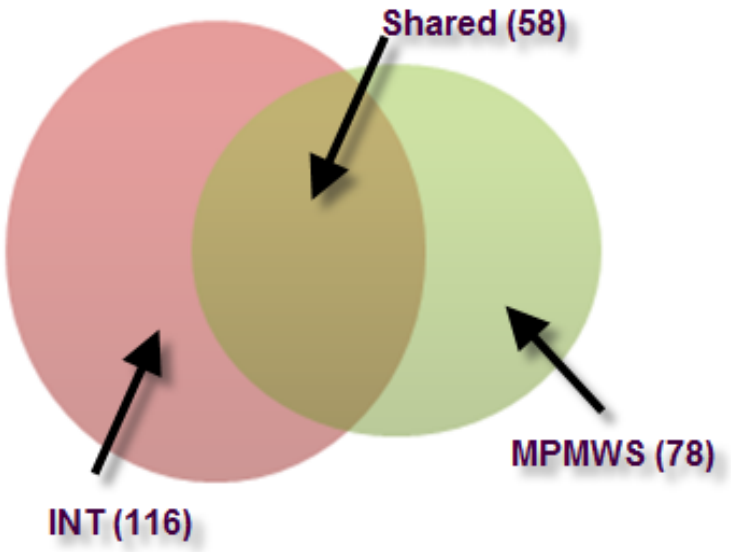


Fig. 3.5. Significant pathways identified by MPMWS and INT. The numbers of significant pathways are listed in parentheses.

Comparing the significant findings from MPMWS and INT, we found that a majority (58 and 74%) of the pathways were shared between the two methods (Supplementary Table 3.1). This includes many well-known pathways related to breast cancer, e.g., PKL1 (King, et al., 2012; Wierer, et al., 2013) and the cell cycle pathway (Caldon, et al., 2006). As was observed in the simulation study, there were quite a few overlaps between MPMWS and INT. However, some significant pathways that were identified by one method had large p-values in the other method. For example, the pathways of DNA replication and DNA strand elongation are important for breast cancer (Lomonosov, et al., 2003; Thomassen, et al., 2009); they were identified by INT but missed by MPMWS (Supplementary Table 3.2A). In contrast, the BAF complex (Hargreaves and Crabtree, 2011; Kadoch, et al., 2013), the well-

known tumor suppressors, and the G1 pathway (Thomassen, et al., 2008), a known breast cancer related pathway, were found to be significant by MPMWS but not by INT (Supplementary Table 3.2B). These results agree with the observations in the simulation study: INT and MPMWS appear to identify different types of signals and can be used together in real practice.

We applied multi-platform gene set analyses on a second TCGA dataset, i.e., the KIRC dataset. We observed similar results as for the breast cancer data and reported the detailed results in Supplementary Tables 3.3A, 3.3B and 3.3C.

3.4 Discussion

To account for sample heterogeneity, we proposed and tested three different strategies, MPMWS, MPORT, and MPLRS, for multi-platform gene set analysis. We found that the non-parametric MPMWS method had satisfactory TPRs and robust performance regardless of the degree of heterogeneity. Finally, based on the results of the simulations and the real data applications, we recommend using both MPMWS and INT: The significant gene sets identified by both methods are more likely to be true positives, while each approach is able to identify orthogonal yet relevant biological gene sets. It might worth following up with these orthogonal findings combining with additional biological information so to minimize the false positives.

We performed the tests assuming that genes are uncorrelated within and across platforms. This assumption may not be valid in real practice, especially for genes within the same gene sets. Inter-gene correlation is known to inflate the false discovery rate of single-

platform gene set analysis, and several methods have been proposed to address this issue (Gatti, et al., 2010; Wu and Smyth, 2012). In addition, the genomic variables of a gene from different platforms can also be highly correlated with each other. For example, copy number change can lead to a change of transcript level; and a high methylation level of the gene promoter region often leads to down regulation of transcription. It is worth future studies to evaluate how inter-gene and inter-platform correlations will affect multi-platform gene set analysis.

The R code for all of the methods and test data sets are available on the website: http://www4.stat.ncsu.edu/~jytzeng/Software/Multiplatform_gene_set_analysis/

References

(2012) Comprehensive molecular portraits of human breast tumours, Nature, 490, 61-70.

Ackermann, M. and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis, BMC bioinformatics, 10, 47.

Aryee, M.J., et al. (2013) DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases, Science translational medicine, 5, 169ra110.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289-300.

Caldon, C.E., et al. (2006) Cell cycle control in breast cancer cells, Journal of cellular biochemistry, 97, 261-274.

Chin, L., et al. (2011) Making sense of cancer genomic data, Genes & development, 25, 534-555.

Du, P., et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC bioinformatics, 11, 587.

Fisher, R., Pusztai, L. and Swanton, C. (2013) Cancer heterogeneity: implications for targeted therapeutics, British journal of cancer, 108, 479-485.

Gatti, D.M., et al. (2010) Heading down the wrong pathway: on the influence of correlation within gene sets, BMC genomics, 11, 574.

Hargreaves, D.C. and Crabtree, G.R. (2011) ATP-dependent chromatin remodeling: genetics, genomics and mechanisms, Cell research, 21, 396-420.

Hu, J. (2008) Cancer outlier detection based on likelihood ratio test, Bioinformatics, 24, 2193-2199.

Kadoch, C., et al. (2013) Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy, Nature genetics, 45, 592-601.

King, S.I., et al. (2012) Immunohistochemical detection of Polo-like kinase-1 (PLK1) in primary breast cancer is associated with TP53 mutation and poor clinical outcom, Breast cancer research : BCR, 14, R40.

Lomonosov, M., et al. (2003) Stabilization of stalled DNA replication forks by the BRCA2 breast cancer susceptibility protein, Genes & development, 17, 3017-3022.

MacDonald, J.W. and Ghosh, D. (2006) COPA--cancer outlier profile analysis, Bioinformatics, 22, 2950-2951.

Ross, G.J. (2013) cpm: Sequential Parametric and Nonparametric Change Detection.

Ross, G.J., Tasoulis, D.K. and Adams, N.M. (2011) Nonparametric Monitoring of Data Streams for Changes in Location and Scale, Technometrics, 53, 379-389.

Russnes, H.G., et al. (2011) Insight into the heterogeneity of breast cancer through next-generation sequencing, The Journal of Clinical Investigation, 121, 3810-3818.

Smyth, G.K. (2005) Limma: linear models for microarray data, Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor, 397--420.

Subramanian, A., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

Thomassen, M., Tan, Q. and Kruse, T.A. (2008) Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer, BMC cancer, 8, 394.

Thomassen, M., Tan, Q. and Kruse, T.A. (2009) Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis, Breast cancer research and treatment, 113, 239-249.

Tibshirani, R. and Hastie, T. (2007) Outlier sums for differential gene expression analysis, Biostatistics, 8, 2-8.

Tyekucheva, S., et al. (2011) Integrating diverse genomic data using gene sets, Genome biology, 12, R105.

Wang, Y., et al. (2011) Non-parametric change-point method for differential gene expression detection, PloS one, 6, e20060.

Wierer, M., et al. (2013) PLK1 Signaling in Breast Cancer Cells Cooperates with Estrogen Receptor-Dependent Gene Transcription, Cell reports, 3, 2021-2032.

Wu, B. (2007) Cancer outlier differential gene expression detection, Biostatistics, 8, 566-575.

Wu, D. and Smyth, G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation, Nucleic acids research, 40, e133.

Xiong, Q., et al. (2012) Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets, Genome research, 22, 386-397.

CHAPTER 4 PLATFORM BIAS BETWEEN MICROARRAY and RNA-SEQ

Abstract

Both microarray and RNA-Seq are popular high-throughput platforms for quantifying gene expressions. Results from Microarray and RNA-Seq have been compared by many studies. However, majorities of these studies focus on the concordance of microarray and RNA-Seq, few of them address the discordance between the two platforms. Using TCGA and GTEx datasets, we perform a systematical analysis of platform bias between microarray and RNA-Seq platforms. We find that gene expression level, gene length and GC content can all contribute to the platform bias. We then test different batch effect removal methods for their abilities to remove the platform bias.

4.1 Introduction

Since the late 90th, microarray technology has been the dominate method to large scale gene profiles studies. In general, oligo-nucleotides (probes) are designed to match selected regions on the genes. They are synthesized and then deposited onto the arrays. The cDNA transcripts will hybridize to its complement sequences on the microarray chip during the assay, the microarray signals from the probe hybridization can then be used to quantify thousands of gene expression at the same time (Lockhart, et al., 1996; Schena, et al., 1995).

Using next generation sequencing (NGS), RNA sequencing (RNA-Seq) is a relative new technology to study the transcriptome. Compared with the traditional microarray

platforms, RNA-Seq often has broad signal detection ranges, especially for low expression genes (Marioni, et al., 2008; Nagalakshmi, et al., 2008; Wang, et al., 2009). Because it detects the RNA molecule at the base level, it offers more flexibility to detect novel transcripts isoform, splice junction, transcript insertion-deletions (indels) or gene fusions (Marioni, et al., 2008; Sultan, et al., 2008; Wang, et al., 2009). Since the latter stage of the last decade, with the steady improvement of RNA-Seq technology and the decreasing cost, it has gain tremendous momentum to become a leading strategy for gene expression profiling.

We are in the transition period between these two technologies right now. Although microarray is no longer the cutting edge technology, it remains a popular choice for many gene expression profiling projects because of its low cost and mature pipeline (Fu, et al., 2009). More importantly, over the last decade, hundreds of thousands of samples have been profiled using microarray. Thus, how to compare the results of RNA-Seq with the existing microarray results becomes an important research area. It will not only help to better understand the pros and cons of each technology, but also help to filter problematic data and yield more reliable findings

Multiple studies comparing RNA-Seq and microarray have shown good concordance between these two platforms (Marioni, et al., 2008; Sirbu, et al., 2012; Zhang, et al., 2012). For example, Marioni et al. found that transcript measurements by Illumina RNA-Seq and Affymetrix microarray have high correlations (>70%), and there are more than 80% overlap of differential expressed genes (DEG) identified by these two platforms (Marioni, et al., 2008). Similar results have also been observed by using TCGA data, where they found above 0.8 Spearman correlations between Affymetrix microarray and RNA-Seq signals and about

60% overlap of DEGs between Agilent and RNA-Seq platforms (Guo, et al., 2013). Majorities of these studies were focused on the consistency of microarray and RNA-Seq; however, there are significant amount of genes shown poor concordance between microarray and RNA-Seq. This is likely due to the platform differences between microarray and RNA-Seq. Each platform has its platform bias. For microarray, the signals are detected when labeled DNA molecules are hybridized with the probes on the array. Different gene and probe sequences will affect its binding affinity and the detection signal. Thus, it is not surprising to see that the gene specific bias from different GC content (Wu, et al., 2004). For RNA-Seq, GC content, the lengths of the transcripts and the fragmentation pattern can lead to gene specific bias (Dohm, et al., 2008; Finotello, et al., 2014; Risso, et al., 2011; Roberts, et al., 2011; Ross, et al., 2013). Depend on the experimental protocol, 3' to 5'-end bias could exist for both microarray(Spiess, et al., 2003) and RNA-Seq (Mortazavi, et al., 2008). Some of the biases are specific to one platform, e.g., bad designed probes of microarrays targeted to inter-genic region (Miller, et al., 2014; Zhao, et al., 2014), and cross-hybridization of the probes are issues specific for microarray, whereas the gene length bias tends to affect RNA-Seq more often (Oshlack and Wakefield, 2009).

Taking advantage of the recent large scale studies such as TCGA (2008) and Genotype-Tissue Expression (GTEx) (2013), where hundreds of samples are run in parallel with both microarray and RNA-Seq, we systematically identified and examined those genes with bad concordance between these two platforms. Affymetrix gene express arrays and Illumina RNA-Seq data are used as the representative platform for each technology since they are the most dominant platforms for microarray and RNA-Seq. We find that gene

expression levels, gene length and GC content, are major contributors of platform specific bias. To correct for these platform specific bias, we compared several popular batch effect removal (BER) methods known for microarray studies and evaluated their abilities to remove platform bias between microarray and RNA-Seq.

4.2 Methods

4.2.1 Data sets

TCGA GBM (Glioblastoma multiforme) and OV (Ovarian serous cystadenocarcinoma) level3 Affymetrix expression data (HT_HG-U133A) and Illumina RNA-Seq read counts were downloaded from TCGA website (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/). There are 11312 common genes from 156 common samples shared by RNA-Seq and Affymetrix microarray from GBM datasets. OV dataset has 11312 common genes from 267 common samples. The RPKM (reads per kilobase of transcript per million reads mapped) values were computed as the gene read counts divided by corresponding gene lengths.

GTEx Affymetrix raw signal (cel) files and raw RNA-Seq reads (fastq) files were downloaded from NCBI SRA (Sequence Read Archive). The Affymetrix microarray probe-set signals were extracted using Omicsoft internal script which was similar to Affymetrix MAS5 algorithm [Affymetrix® Microarray Suite version 5] and then summarized as gene expression intensities. For RNA-Seq, fastq files were aligned to the human genome (version NCBI B37.3) using Omicsoft Sequence Aligner (Hu, et al., 2012) with UCSC human gene models. For each gene, its read counts and RPKM values were computed using RSEM

algorithm (Li and Dewey, 2011). There are 25126 common genes with 807 common samples shared between RNA-Seq and Affymetrix expression platform (HuGene-1_1-st). To test the potential bias from low expression genes, we generate a "good signal" GTEx data set (referred as GS-GTEx dataset) with 11275 common genes by filtering out genes with weak expression (median RNA-Seq log RPKM values less than 1). For all datasets, raw reads counts and RPKM values are log2 transformed before usage.

4.2.2 Batch effect removal methods

BER methods were used to remove the platform specific bias between microarray and RNA-Seq datasets. A detailed review has been provided in Chapter 1. We select the following five BER methods because they represent different statistical categories of BER, they are also widely used during microarray expression data analysis. We used R package *inSilicoMerging* (v1.6.0) which implements all these five BER methods.

- BMC (Batch Mean Centering) transforms the data by subtract the sample mean for each batch.
- Gnorm (Gene Normalization) performs Z-transformation of data from each batch.
- DWD (Distance weighted Discrimination) is an adaptive SVM (Support Vector Machine) method which find an ideal separation hyper-plane and removes biases by projecting different batches of data on this hyper-plane (Benito, et al., 2004).
- COMBAT (Combining Batches of Gene Expression Data) uses empirical Bayes approach to estimate parameter of the model and correct for both additive and multiplicative batch effects (Johnson, et al., 2007).

- XPN (Cross-Platform Normalization) identifies homogeneous blocks gene and samples and use them to perform cross platform normalization (Shabalin, et al., 2008).

4.2.3 Clustering analysis

We formed a data matrix by combining gene expression data from both microarray and RNASeq using matched samples in GTEx dataset. We performed Principal Component Analysis (PCA) using Omicsoft Array Studio (v7.0.2) on the data matrix either before or after BER. Four tissues (adipose tissue, heart, lung and muscle) were selected for analysis. We generated 3D PCA plots to show the analysis outcomes, where each dot represent a sample. We use different shape to indicate samples of different platforms, and different colors to indicate different tissues. This analysis was repeated on GS-GTEx dataset and also ranks of gene expression from GS-GTEx.

4.2.4 Variance analysis

The GS-GTEx dataset were first quantile normalized. For each gene in the dataset, we regressed log expression intensities on covariates of platform, subject and tissue using R linear regression function ‘lm’. The percentages of sum of square contributed by platform were computed from the datasets before and after BER. R (v3.0.1) was used to carry out computation and plotting.

4.2.5 Differential expressed gene analysis

To test the potential effects of BER methods on DEG, we compared the DEGs between lung and adipose tissue identified from GS-GTEx dataset before and after BER.

4.3 Results

4.3.1 Gene expression concordance analysis

We compared the gene expressions correlations between Affymetrix microarray and RNA-Seq using TCGA GBM dataset. For each gene, we computed its correlation between microarray and RNA-Seq (in terms of gene counts or RPKM) across matched samples (gene level correlations) (Fig. 4.1A). For each pair of matched sample, we computed the correlation between microarray and RNA-Seq across all genes (sample level correlations). Similar to previous findings (Guo, et al., 2013), we see high correlations at sample level with either Spearman or Pearson correlations coefficients. Gene level correlations are also strong overall, but the spread is much more diverse, many genes show weak or even negative correlations. We repeated the concordance analysis for TCGA OV dataset (Fig 4.2). It shows similar patterns. We also found that the results from Spearman or Pearson correlation were comparable and we used Pearson correlation exclusively in later analysis. Using different measurement metric for expression level, e.g., RKPM (Fig. 4.1B and Fig. 4.2B) or gene counts (Fig. 4.1A and Fig. 4.2A) in RNA-Seq, does not affect the analysis outcomes much either.

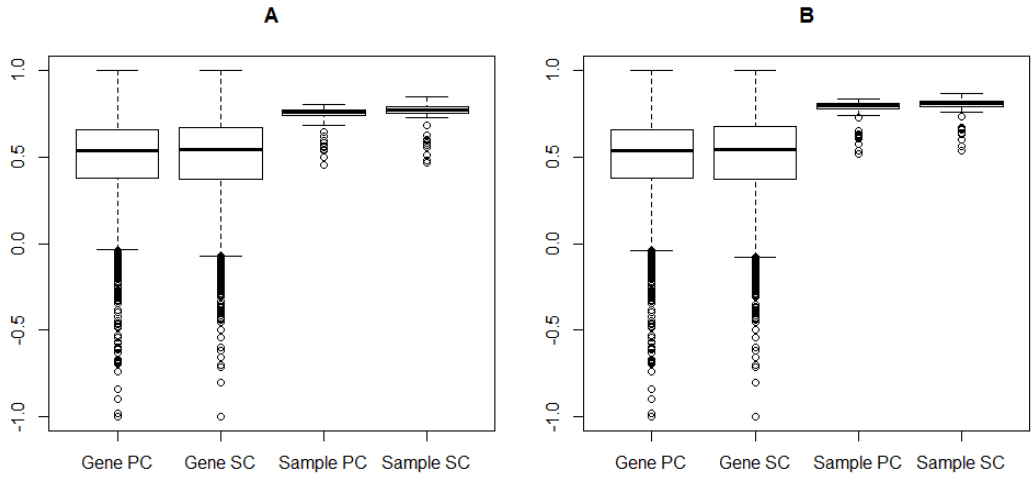


Fig. 4.1. Correlation between RNA-Seq and microarray for TCGA GBM dataset. A: RNA-Seq gene counts versus Affymetrix expression B: RNA-Seq gene RPKM versus Affymetrix expression , PC: pearson correlation, SC: spearman correlation (all values are in log2 scale)

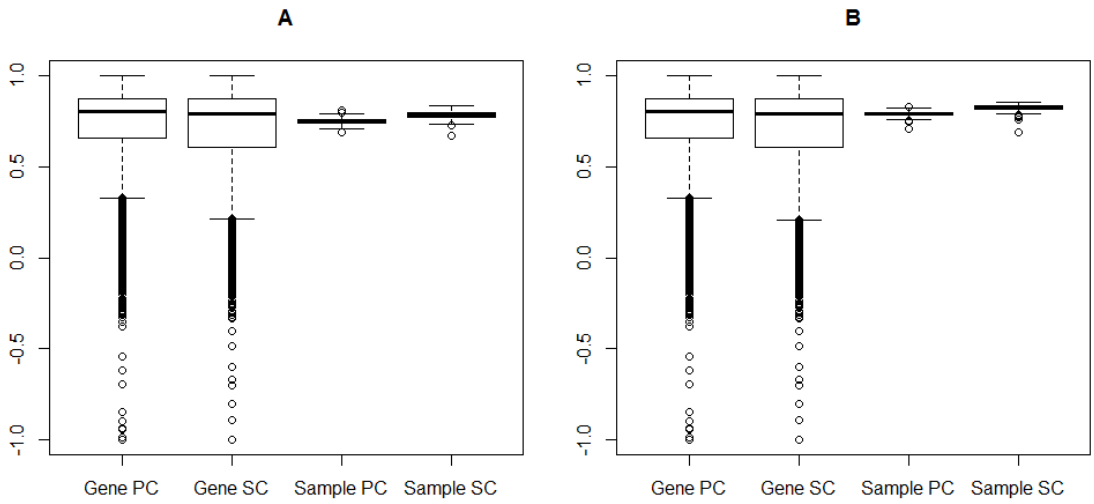


Fig 4.2. Correlation between RNA-Seq and microarray for TCGA OV dataset. A: RNA-Seq gene counts versus Affymetrix expression B: RNA-Seq gene RPKM versus Affymetrix expression , PC: pearson correlation, SC: spearman correlation (all values are in log2 scale)

Many factors can contribute to the high variations of gene level correlations. Microarray tends to have smaller detection range compared to RNA-Seq, especially for those low expression genes. One example is shown in Fig 4.3, where we randomly selected a single sample (or sample) from GBM dataset and plot the matched intensity values from microarray and RNA-Seq for demonstration purpose. We can see that microarray have little detection power for genes with intensity values around 2.0 whereas RNA-Seq can still detect signals properly near this signal region. This likely contributes to the poor correlations between microarray and RNA-Seq results at low expression range. We tested this hypothesis by comparing the mean signal distributions of the bottom 10% of genes with the worst correlations against that of all the genes using TCGA GBM dataset. From the density plots,

we can see that most low correlation genes have much lower expression values signals for both RNA-Seq (Fig. 4.4B vs. 4.4A) and microarray platforms (Fig. 4.4D vs. 4.4C). For example, in RNA-Seq, the peak is around 10 for majorities of genes whereas low-correlated genes have mean expression around 0. We get a p-value < 2.2e-16 from Wilcoxon test comparing the correlations distributions of low expression genes with all the genes. Similar results were also observed in GTEx dataset (Figure 4.5A,B,C,D).

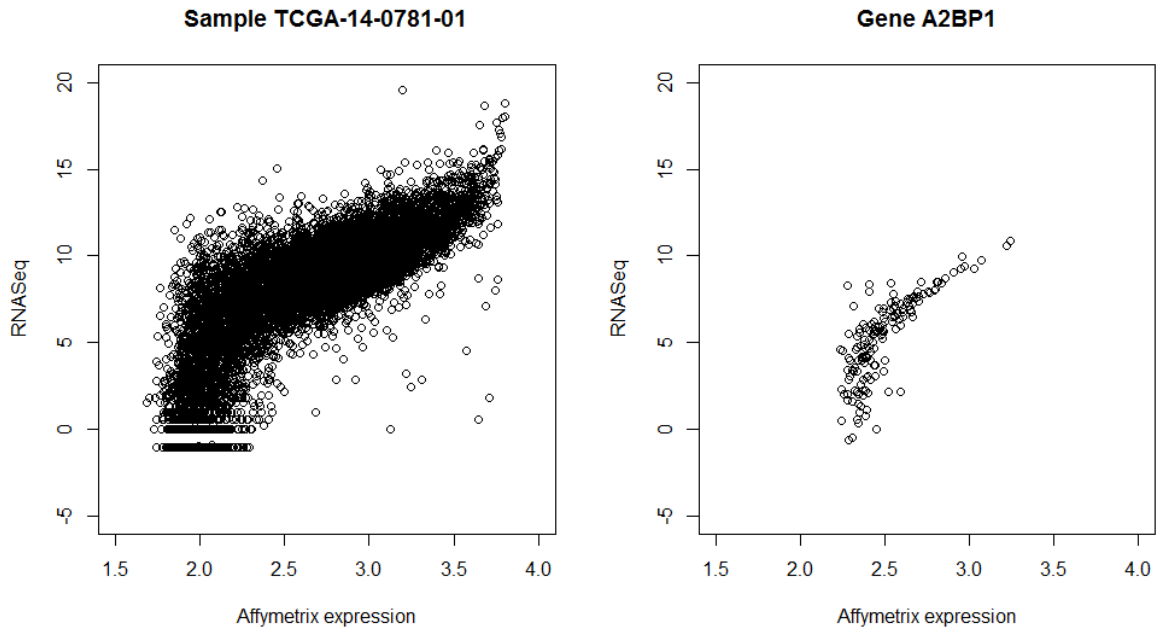


Fig. 4.3. Gene expression pattern of microarray and RNA-Seq

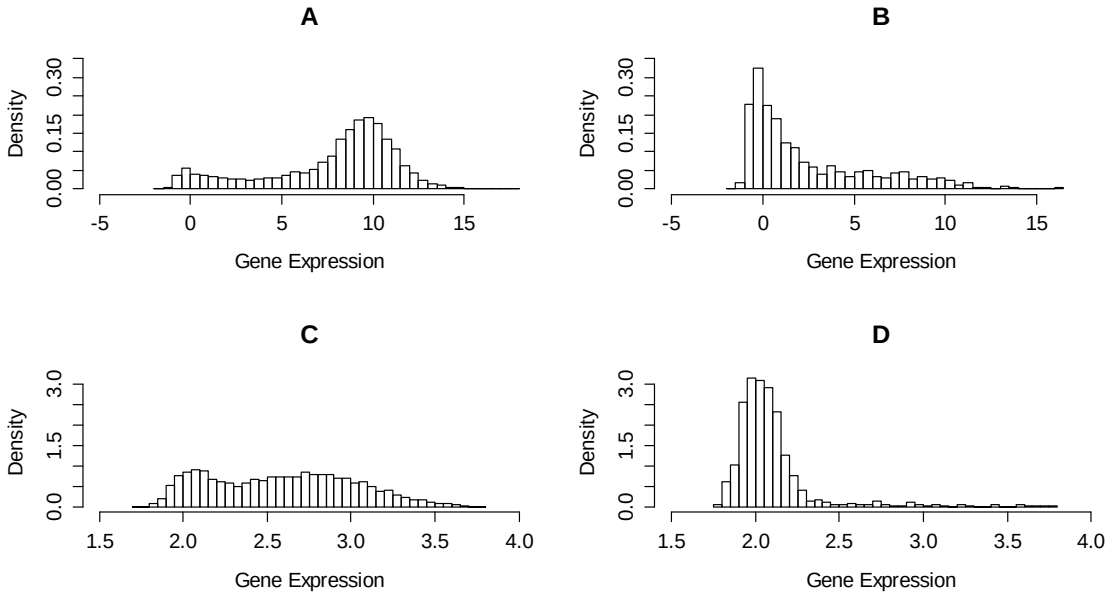


Fig. 4.4. Gene expression levels of low correlation genes in GBM dataset. A: RNA-Seq gene expression level of all genes. B: RNA-Seq gene expression levels of low-expression genes. C: Microarray gene expression levels of all genes D: Microarray gene expression levels of low expression genes. (log count intensity values were used for RNA-Seq)

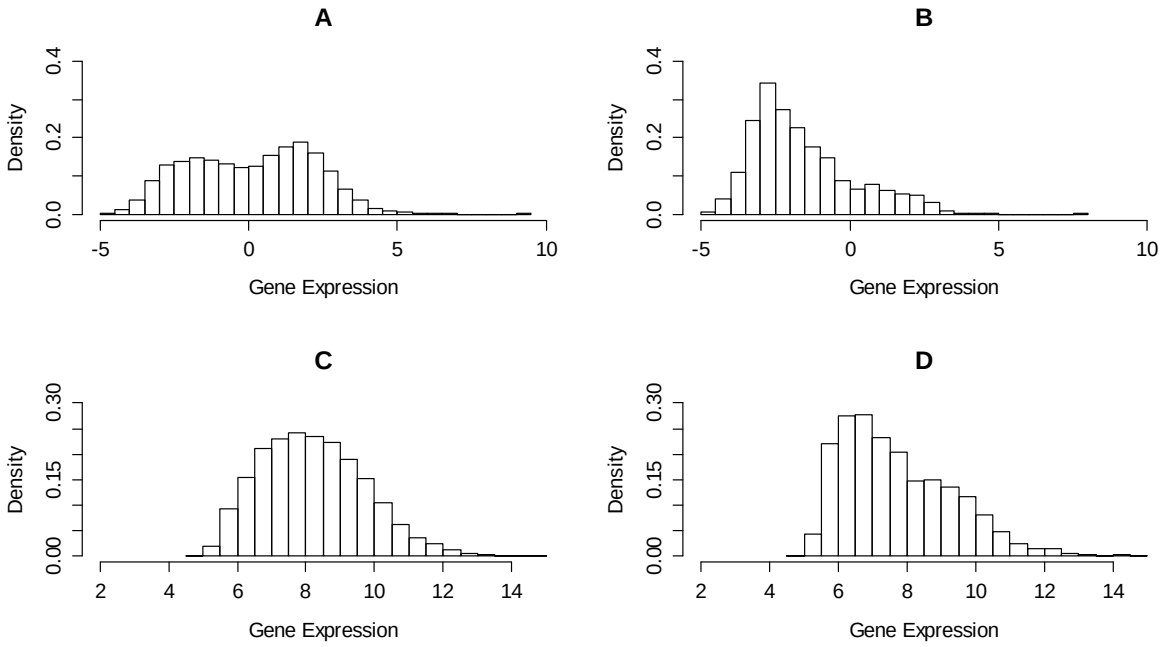


Fig. 4.5. Gene expression levels of low correlation genes in GTEx dataset. A: RNA-Seq gene expression level of all genes. B: RNA-Seq gene expression levels of low-expression genes. C: Microarray gene expression levels of all genes D: Microarray gene expression levels of low expression genes. (log RPKM values were used for RNA-Seq)

Gene lengths and GC content have also been known to affect microarray and NGS signal detection. We examined the distribution of gene lengths (and GC content) between low correlation genes and the distribution of all genes to gain more insight on the relationship between these factors and platform-specific bias. Overlay the distribution of low correlation genes with the distributions of all the genes, we found that peak of low correlation genes shifts left for gene lengths which indicates they have shorter lengths (Fig. 4.6B vs. 4.6A). And for GC content, the low-correlation genes shifts slightly right, which indicates slightly higher GC contents (Fig. 4.6D vs. 4.6C).

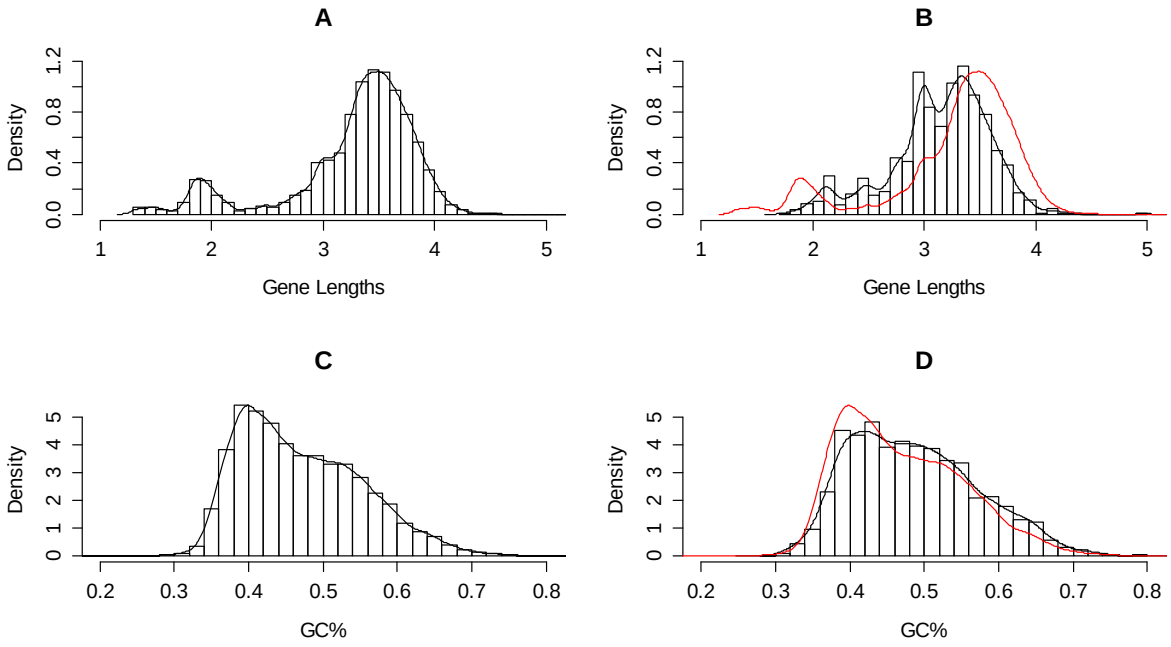


Fig. 4.6. Gene length and GC content affect platform bias. A: Gene length distribution of all genes. B: Gene length distribution of low expression genes. (Red lines are the distribution of A). C: GC content distribution of all genes. D. GC content distribution of low expression genes. (Red lines in B and D are density lines from A and C correspondingly.)

4.3.2 Platform bias and clustering analysis

We examined the effect of platform bias when performing cross platform study using data from both Affymetrix microarray and RNA-Seq platforms. PCA were performed using the combined data matrix of RNA-Seq and microarray data from GTEx dataset. From Fig. 4.7, we can see that the first principal component (explained about 38% of the variance) is mainly caused by different platforms. And the biological differences (e.g., tissue types) mainly contribute to the second and third principal components. The samples from the same tissue with different platforms are clearly separated in the PCA plots. To avoid the scenario that such separation is caused by the low expression genes that microarrays fail to detect, we repeated the same analysis on the 11275 good expression genes using GS-GTEx dataset (Supplementary Fig. 4.1). The variations explained by platform (i.e., the first PC) decreased from 38% to 31%, which is still high. Clearly, the platform bias is not limited to low expression genes only. To rule out the possibility that such outcomes are caused by the scale difference or the different range of the expression values between the two platforms, for GS-GTEx dataset, we performed PCA using the ranks of the gene expression in each sample, similar results were observed (Supplementary Fig. 4.2).

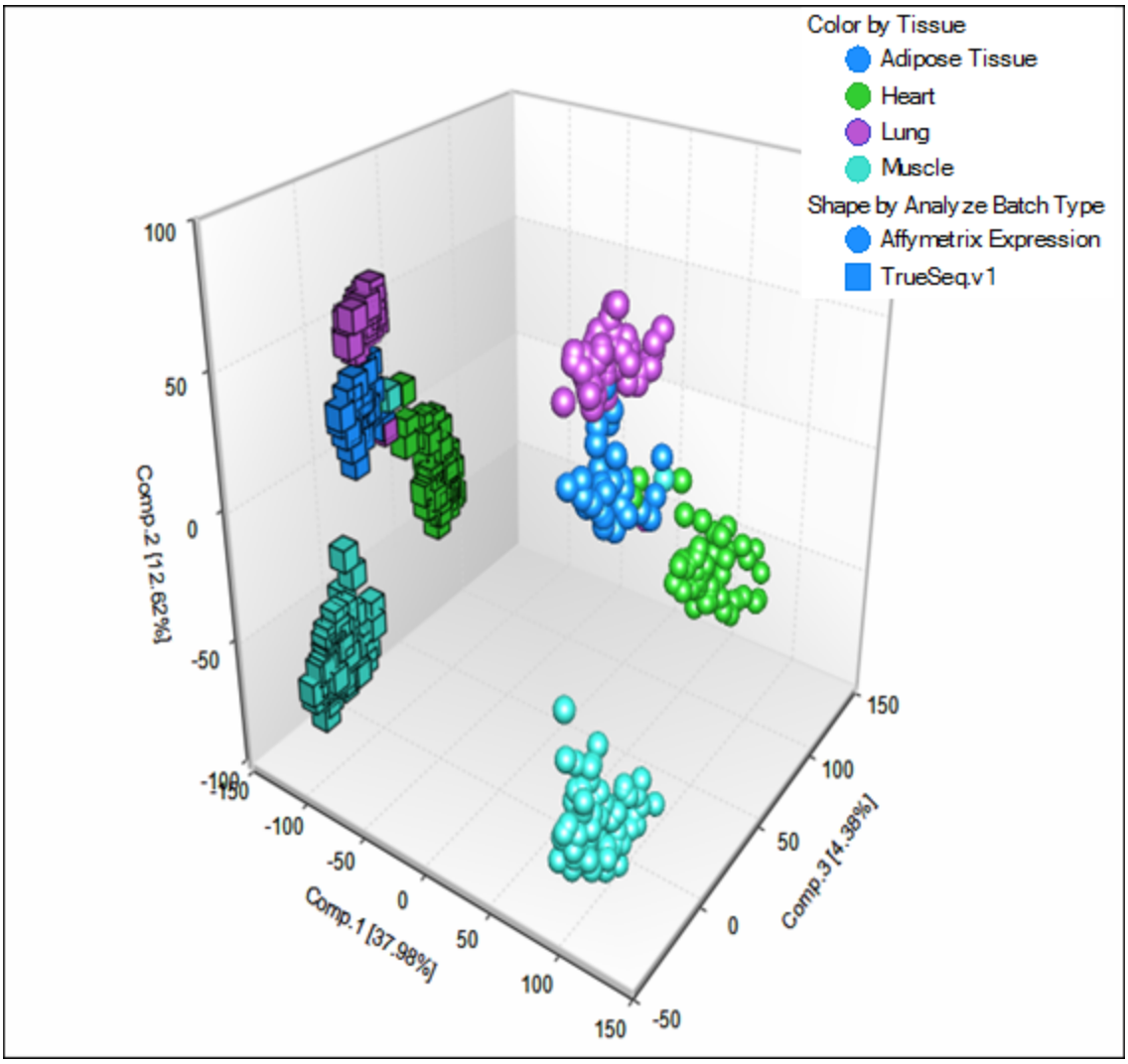


Fig 4.7. 3D PCA plot of GTEx dataset

4.3.3 BER method performances to remove platform bias

Many BER methods have been developed for microarray data sets. They have also been applied successfully to combine datasets from different microarray platforms during cross platform analysis. We benchmarked five popular BER methods (BMC, Gnorm, DWD, COMBAT and XPN) of their abilities to remove platform bias during cross platform analysis of data from both microarray and RNA-Seq. From the PCA outcomes before and after BER using the GS-GTEx dataset, we can see all these methods can successfully remove the platform bias, and the sample from both Affymetrix microarray (sphere shape) and RNA-Seq (cubic shape) can be clustered together (Supplementary Fig. 4.3, 4.4, 4.5, 4.6, 4.7) after BER procedure. The samples from different tissue are still well separated into different clusters. This indicates that the meaningful biological variations are mostly kept by BER methods.

Next, we performed the variance analysis (detailed in the method section) to compare the percentage of variations attribute to platform differences before or after applying BER methods. The results showed that variations due to platform differences reduced from 18% to near zero after BER (Fig. 4.8). The gene level correlations between matched samples also significantly improved after applying BER methods (Fig. 4.9). Compared with the BMC and Gnorm, which improve the median correlation from 65% to 80%; complex methods, DWD, COMBAT, and XPN perform even better, improved the correlation from 65% to around 90%.

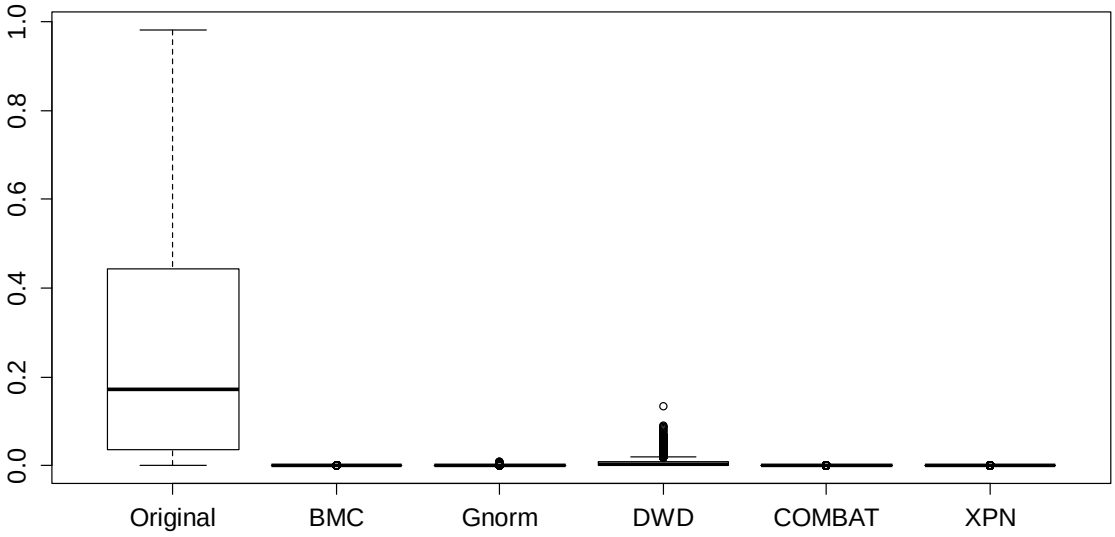


Fig. 4.8. Percent of variances from platform before/after BER step

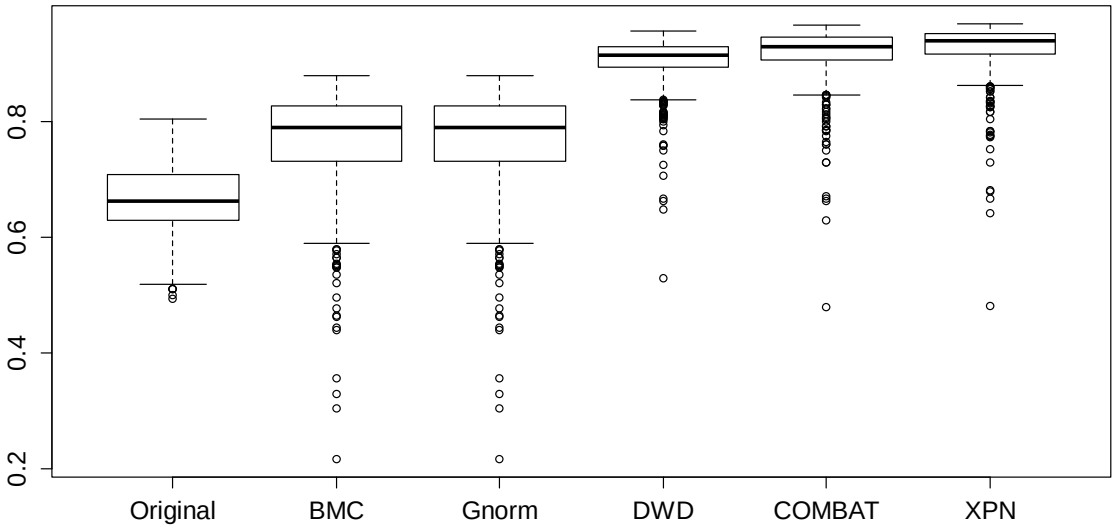


Fig. 4.9. Correlations of matched GTEx samples before or after BER.

DEG analysis is one major application of gene expression profiling. We compared lung tissue vs. adipose tissue in GS-GTEx dataset and identified 6328 significantly changed genes from microarray and 7253 from RNA-Seq before BER. After applying BER methods, Gnorm identified 6393 DEGs from microarray and 7410 from RNA-Seq. XPN identified 6344 DEGs from microarray and 7295 from RNA-Seq. And mostly of these DEGs are shared with DEGs from the data before BER (overlap larger than 98%). BMC, Combat and DWD find the same DEGs as those without BER.

4.4 Discussion

In the presented study, using Affymetrix microarray and Illumina RNA-Seq dataset from some of the recent large scale studies, such as TCGA and GTEx, we systematically studied the platform bias between microarray and RNA-Seq. As previous studies have shown, most genes show highly consistent expressions between microarray and RNA-Seq. However, we have also found that a fair amount of genes have different expression patterns between these two platforms, which indicates the existence of platform specific bias. The platform bias can mix with real biological signals, decrease the power of experiments and even lead to false positives or wrong interpretation of the results. For example, our study shows that platform specific bias distorts the clustering analysis. Therefore, it is essential to discover the sources of the platform bias, identify genes that are affected by platform bias and use the statistical methods to correct for these biases when performing cross-platform studies.

Focusing on those genes with inconsistent expressions, we have discovered that they significantly enrich low expression genes. It also tends to have shorter lengths and slightly

higher GC contents. The relationship between weak signal and platform bias is very likely related to the different detection ranges of microarray and/or RNA-Seq platforms. Especially for microarray platform, it is known to have limited detection power for low expression genes. Shorter gene length usually also could lend to less RNA-Seq reads detected (Oshlack and Wakefield, 2009). Although we used RPKM values which corrected some length bias. There could be some length effect still remains which affect the signal (Bullard, et al., 2010). The relationship of GC contents to mammalian gene expression is still under debate. Many studies have shown the positive correlation of gene expression and its GC content whereas other studies have found weak or opposite correlation (Arhondakis, et al., 2008; Kudla, et al., 2006; Semon, et al., 2005). Although our finding is statistically significant, compared with gene expression level itself and gene lengths, it is a much weaker factor.

The platform bias is not limited to weakly expressed genes. Even after the removal of most low expression genes, the platform bias remain the first principle component of our PCA analysis. The low-correlations high expression genes vary from microarray platform to platform. This is likely caused by bad probe designs in microarray or different experimental factors that lead to different hybridization patterns. However, more datasets and further studies are needed to elucidate the exact reason for each gene at each platform.

Platform bias can be thought as a special case of batch effects. We thus tested multiple BER methods for their ability to remove platform bias between microarray and RNA-Seq. Comparing the PCA clustering results before and after BER, we can see that the five tested BER methods successfully reduce the platform bias. This is further validated by our variance analysis. The variances due to platform are almost reduced to zero by most

methods. And the correlations between matched samples (i.e., sample-level correlation) also significantly improved after BER. DWD, COMBAT and XPN improve the correlations more than Gnorm and BMC; it is likely due to these complex methods that remove platform bias beyond using standardization can not only deal with additive bias, but also bias in more complex patterns. By design, some BER methods, e.g., BMC, will not affect the relative gene expression values in each platform. Thus, they will not affect DEG analysis. For those methods that do, our analysis indicates that the tested BER procedures have minimal affects to DEG analysis on both microarray and RNA-Seq.

In general, BER methods should be only applied to biologically similar batches. Otherwise, meaningful biological differences could be treated as batch difference and also got removed. This also applies to the application of BER methods on platform bias.

Our studies focused on several Affymetrix microarray platforms and Illumina RNA-Seq data, both of which are representative platforms from microarray and RNA-Seq. Other platforms for microarray and RNA-Seq were also commonly available, e.g., Illumina microarrays, Ion Torrent or SOLID platform for RNA-Seq. The strategies and methods that we have used could be easily extended to these platforms as well. However, for microarray based on two color hybridizations technology, like Agilent two color arrays, usually the data generated are the ratios from the intensities of two color channels, and their values and distributions are quite different from the intensity values measured directly by one channel microarray or RNA-Seq. For these platforms, it is probably more proper to compare the results at ratio level or DEGs rather than expression level directly.

References

(2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature*, **455**, 1061-1068.

(2013) The Genotype-Tissue Expression (GTEx) project, *Nature genetics*, **45**, 580-585.

Arhondakis, S., Clay, O. and Bernardi, G. (2008) GC level and expression of human coding sequences, *Biochemical and biophysical research communications*, **367**, 542-545.

Benito, M.*, et al.* (2004) Adjustment of systematic microarray data biases, *Bioinformatics*, **20**, 105-114.

Bullard, J.H.*, et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC bioinformatics*, **11**, 94.

Dohm, J.C.*, et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic acids research*, **36**, e105.

Finotello, F.*, et al.* (2014) Reducing bias in RNA sequencing data: a novel approach to compute counts, *BMC bioinformatics*, **15 Suppl 1**, S7.

Fu, X.*, et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics, *BMC genomics*, **10**, 161.

Guo, Y.*, et al.* (2013) Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data, *PloS one*, **8**, e71462.

Hu, J.*, et al.* (2012) OSA: a fast and accurate alignment tool for RNA-Seq, *Bioinformatics*, **28**, 1933-1934.

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

Kudla, G., *et al.* (2006) High guanine and cytosine content increases mRNA levels in mammalian cells, *PLoS biology*, **4**, e180.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC bioinformatics*, **12**, 323.

Lockhart, D.J., *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature biotechnology*, **14**, 1675-1680.

Marioni, J.C., *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome research*, **18**, 1509-1517.

Miller, J.A., *et al.* (2014) Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq, *BMC genomics*, **15**, 154.

Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature methods*, **5**, 621-628.

Nagalakshmi, U., *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*, **320**, 1344-1349.

Oshlack, A. and Wakefield, M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology, *Biology direct*, **4**, 14.

Risso, D., *et al.* (2011) GC-content normalization for RNA-Seq data, *BMC bioinformatics*, **12**, 480.

Roberts, A., *et al.* (2011) Improving RNA-Seq expression estimates by correcting for fragment bias, *Genome biology*, **12**, R22.

Ross, M.G., *et al.* (2013) Characterizing and measuring bias in sequence data, *Genome biology*, **14**, R51.

Schena, M., *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

Semon, M., Mouchiroud, D. and Duret, L. (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance, *Human molecular genetics*, **14**, 421-427.

Shabalin, A.A., *et al.* (2008) Merging two gene-expression studies via cross-platform normalization, *Bioinformatics*, **24**, 1154-1160.

Sirbu, A., *et al.* (2012) RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering, *PloS one*, **7**, e50986.

Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data, *BMC bioinformatics*, **14**, 91.

Spiess, A.N., Mueller, N. and Ivell, R. (2003) Amplified RNA degradation in T7-amplification methods results in biased microarray hybridizations, *BMC genomics*, **4**, 44.

Sultan, M., *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, *Science*, **321**, 956-960.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews. Genetics*, **10**, 57-63.

Wu, Z., *et al.* (2004) A model-based background adjustment for oligonucleotide expression arrays, *Journal of the American statistical Association*, **99**, 909-917.

Zhang, W., *et al.* (2012) Effector CD4+ T cell expression signatures and immune-mediated disease associated genes, *PloS one*, **7**, e38510.

Zhao, S., *et al.* (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells, *PloS one*, **9**, e78644.

CHAPTER 5 SUMMARY AND FUTURE DIRECTIONS

With the advancement of technology, unit costs of high-throughput technologies decrease steadily while their throughputs keep increasing. This leads to the growing number of projects that biomedical samples are probed by multiple high-throughput platforms. Multi-platform genomic data analysis is gaining momentum in recent years because of the availability of these multi-platform data. Our research efforts presented in this dissertation address some of the important statistically issues/demands from multi-platform genomic data analysis. We have systematically studied different strategies of multi-platform gene set analysis and developed novel methods to perform gene set analysis when sample heterogeneity exists. We have also studied the sources that can contribute to the platform biases between expression microarray and RNA-Seq, two dominate platforms for gene expression profiling.

5.1 Multi-platform gene set analysis

During multi-platform gene set analysis, we summarize the features of each platform by gene. If a gene has multiple features, (e.g., a gene has multiple probe sets from microarray), usually the mean or median value is used. For certain platforms, like SNP or methylation array, this over-simplified strategy could decrease the power of analysis or lead to bias during real data analysis. One gene can have tens or even hundreds of SNPs or methylation sites. The biological effects of these sites could vary dramatically. For example, SNPs or methylation level at the transcription factor binding site often have a much bigger effect on gene expression than those downstream. To prevent other non-causal sites from diluting the

signals of causal SNPs or methylation site, more advanced feature selection strategy could be adopted to select representative sites for the gene. We feel some methods from single platform gene set analysis could be usful for multi-platform. We could use existing biological knowledge to guide the selection, or select significant features pass a pre-specified cutoff, or use eigen-feature (e.g., eigenSNPs) for each gene (Chen, et al., 2010). More studies are needed to test how well these methods work on multi-platform data.

In our multi-platform gene set analysis, we assume that there is no inter-gene and inter-platform correlation. This assumption is often not valid in real biological scenarios. Inter-gene correlation is known to inflate the false discovery rate of single-platform gene set analysis (Gatti, et al., 2010). Several methods have been proposed. Wu *et al.* developed Camera to estimate a variance inflation factor and used it to adjust for inter-gene correlation during gene set analysis (Wu and Smyth, 2012). Yaari *et al.* further extended this idea in QuSAGE to quantify gene-set activity with a complete probability density function (Yaari, et al., 2013). Inter-gene correlation could be more complex in multi-platform genomic data since the correlation structure often varies across different platforms. More efforts are needed to either extend the existing strategies to multi-platform or develop novel methods. Moreover, for multi-platform genomic data, the genomic variables from different platforms can also be highly correlated. E.g., high methylation level of gene promoter region often leads to down-regulation of transcription; the amount of one miRNA can directly affect the transcription levels of its target genes. Future research on how inter-gene and inter-platform correlations affect multi-platform gene set analysis will be crucial to the development of stable and powerful methods.

Inter-platform correlation also affects gene set analysis on multi-platform heterogeneous data. For all three methods that we developed in Chapter 3, platform specific genomic scores are first computed and then summarized as multi-platform gene level statistic. This approach offers the flexibility to apply platform specific methods on each platform to increase power. However, when inter-platform correlation exists, it could inflate the false positive rates. Further improvement of the algorithms to detect and correct inter-platform correlation should be useful to increase the specificity of the method.

5.2 Platform specific bias

Different microarray expression platforms have quite different probe designs, so the low correlation genes identified from one microarray platform often cannot be generalized to other microarray platforms. The good news is that majorities of expression profiling experiments by microarray are based on limited types of the arrays from several large vendors, like Affymetrix, Agilent or Illumina. Works are underway to identify all the low correlation genes for each of these platforms. To move one step further, it will be interesting to see which probe(s) on the microarray contribute most to the inconsistency between expression array and RNA-Seq. And we expect that the signals reported from microarray could get improved after removing those bad designed probes.

Allele specific expression (ASE) (Buckland, 2004; Yan, et al., 2002) is another factor that might contribute to the discordance between RNA-Seq and expression microarray platform. Probes from expression array are usually designed to detect one type of allele of the human genome. Depend on the position of the allele on the probe; the transcripts that have a

different allele might fail to bind to the probe. And this will lead to bias between microarray and RNA-Seq. Further studies are needed to find out how much of the platform bias is caused by ASE. These findings can not only help us to identify bad probes on the microarray. More importantly, it can also be used to identify genes and loci that are affected by ASE.

5.3 Other multi-platform genomic data analyses

Certainly, the application of multi-platform genomic data is not limited to the areas that we discussed in this dissertation. Novel statistical methods have been developed to utilize multi-platform genomic data on biomarker identification (Wu, et al., 2012; Zhang, et al., 2012), clinical prediction (Srivastava, et al., 2013; Wang, et al., 2013) and disease classification (2013; Kandoth, et al., 2013). Successful methods that can properly integrate and analyze multi-platform data will be essential to find the regulation mechanism of the biological system, to help diagnosis of complex diseases like cancer, or even provide key guidance for personalized medicine.

References

(2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia, *The New England journal of medicine*, **368**, 2059-2074.

Buckland, P.R. (2004) Allele-specific gene expression differences in humans, *Human molecular genetics*, **13 Spec No 2**, R255-260.

Chen, L.S., *et al.* (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data, *The American Journal of Human Genetics*, **86**, 860-871.

Gatti, D.M., *et al.* (2010) Heading down the wrong pathway: on the influence of correlation within gene sets, *BMC genomics*, **11**, 574.

Kandoth, C., *et al.* (2013) Integrated genomic characterization of endometrial carcinoma, *Nature*, **497**, 67-73.

Srivastava, S., *et al.* (2013) Integrating multi-platform genomic data using hierarchical Bayesian relevance vector machines, *EURASIP journal on bioinformatics & systems biology*, **2013**, 9.

Wang, W., *et al.* (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data, *Bioinformatics*, **29**, 149-159.

Wu, D. and Smyth, G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation, *Nucleic acids research*, **40**, e133.

Wu, S., *et al.* (2012) Multiple-platform data integration method with application to combined analysis of microarray and proteomic data, *BMC bioinformatics*, **13**, 320.

Yaari, G., *et al.* (2013) Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations, *Nucleic acids research*, **41**, e170.

Yan, H., *et al.* (2002) Allelic variation in human gene expression, *Science*, **297**, 1143.

Zhang, S., *et al.* (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data, *Nucleic acids research*, **40**, 9379-9391.

APPENDICES

Appendix A – Supplementary tables from chapter 3

Supplementary Table 3.1. Significant pathways identified by both INT and MWMPS.

Pathway Name	INT	MPMWS
KEGG_PROTEASOME	1.27E-08	9.65E-04
KEGG_CELL_CYCLE	8.36E-06	1.15E-07
BIOCARTA_PROTEASOME_PATHWAY	9.92E-08	1.52E-02
PID_FANCONI_PATHWAY	6.51E-03	1.14E-02
PID_AURORA_B_PATHWAY	1.19E-03	1.52E-02
PID_ATR_PATHWAY	5.78E-07	4.05E-04
PID_PLK1_PATHWAY	8.76E-07	2.92E-03
PID_FOXM1PATHWAY	4.49E-03	5.23E-03
PID_A6B1_A6B4_INTEGRIN_PATHWAY	2.51E-02	1.52E-02
MIPS_20S_PROTEASOME	2.24E-04	3.05E-02
MIPS_PA28_20S_PROTEASOME	2.05E-04	2.04E-02
MIPS_PA700_20S_PA28_COMPLEX	3.41E-09	3.95E-03
MIPS_SIN3_ING1B_COMPLEX_II	1.78E-02	2.73E-02
MIPS_LARC_COMPLEX	3.17E-02	1.52E-02
MIPS_CEN_COMPLEX	1.67E-02	2.37E-02
MIPS_ALL_1_SUPERCOMPLEX	1.20E-02	3.12E-02
REACTOME_COPI_MEDIATED_TRANSPORT	6.99E-03	1.05E-02
REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	4.65E-08	1.05E-02
REACTOME_G0_AND_EARLY_G1	2.99E-03	9.65E-04
REACTOME_CELL_CYCLE	2.25E-19	4.01E-13
REACTOME_ORC1_REMOVAL_FROM_CHROMATIN	9.10E-08	5.23E-03
REACTOME_P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT	1.35E-08	4.67E-03
REACTOME_CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OF_CDC6	2.76E-08	6.86E-03
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	1.58E-04	1.67E-03
REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC	3.41E-07	5.23E-03
REACTOME_GLYCOLYSIS	4.07E-02	4.83E-02
REACTOME_CELL_CYCLE_MITOTIC	8.13E-19	4.01E-13
REACTOME_CELL_CYCLE_CHECKPOINTS	7.10E-11	1.45E-05
REACTOME_CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION_	2.62E-06	3.80E-03
REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE	2.22E-06	1.14E-02
REACTOME_MRNA_PROCESSING	4.77E-04	3.95E-03
REACTOME_M_G1_TRANSITION	7.10E-11	3.62E-03
REACTOME_MRNA_SPLICING	1.38E-05	9.29E-03
REACTOME_G1_S_TRANSITION	2.05E-11	4.34E-05
REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX	2.04E-07	1.52E-02
REACTOME_SYNTHESIS_OF_DNA	7.40E-11	3.62E-03
REACTOME_AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_COP1	7.67E-08	1.11E-02
REACTOME_MITOTIC_G1_G1_S_PHASES	5.15E-13	1.09E-08
REACTOME_REGULATION_OF_MITOTIC_CELL_CYCLE	1.35E-08	3.21E-04
REACTOME_MITOTIC_M_M_G1_PHASES	4.37E-20	1.19E-10
REACTOME_ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX	7.18E-10	2.16E-03
REACTOME_REGULATION_OF_MRNA_STABILITY_BY_PROTEINS_THAT_BIND_AU_RICH_ELEMENTS	3.18E-08	6.09E-04
REACTOME_DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0	9.10E-08	4.67E-03
REACTOME_DNA_REPLICATION	1.60E-20	5.51E-11
REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION	8.68E-04	2.86E-02
REACTOME_APOPTOSIS	3.01E-05	2.24E-02
REACTOME_HIV_INFECTION	4.58E-05	1.36E-02
REACTOME_HOST_INTERACTIONS_OF_HIV_FACTORS	7.05E-04	5.23E-03
REACTOME_APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20	1.17E-07	5.25E-03
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	9.67E-05	1.81E-02
REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS	1.96E-08	2.46E-03

REACTOME_AUTODEGRADATION_OF_CDH1_BY_CDH1_APC_C	4.70E-06	1.44E-02
REACTOME_MITOTIC_PROMETAPHASE	7.01E-09	5.38E-07
REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1	1.73E-07	5.25E-03
REACTOME_G2_M_CHECKPOINTS	2.20E-05	7.94E-03
REACTOME_S_PHASE	1.35E-10	4.05E-04
REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	4.72E-06	1.36E-02
REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	1.46E-08	2.29E-03

Pathway names and FDR adjusted p-values from INT and MPMWS are shown in the table. For both methods, pathways with p-values less than 0.05 were selected.

Supplementary Table 3.2A: Significant Pathways from INT.

Pathway Name	INT	MPMWS
KEGG_OXIDATIVE_PHOSPHORYLATION	1.83E-02	7.68E-01
KEGG_PHENYLALANINE_METABOLISM	1.89E-02	2.17E-01
KEGG_DNA_REPLICATION	4.97E-03	5.74E-01
BIOCARTA_SALMONELLA_PATHWAY	3.17E-02	3.22E-01
SIG_REGULATION_OF_THE_ACTIN_CYTOSKELETON_BY_RHO_GTPASES	1.51E-03	3.90E-01
PID_E2F_PATHWAY	3.72E-02	2.02E-01
MIPS_55S_RIBOSOME_MITOCHONDRIAL	1.09E-02	3.86E-01
MIPS_39S_RIBOSOMAL_SUBUNIT_MITOCHONDRIAL	3.73E-02	6.86E-01
MIPS_RC_COMPLEX_DURING_S_PHASE_OF_CELL_CYCLE	4.39E-02	6.97E-01
MIPS_RC_COMPLEX_DURING_G2_M_PHASE_OF_CELL_CYCLE	3.33E-02	4.66E-01
MIPS_C_COMPLEX_SPLICEOSOME	4.89E-02	2.02E-01
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	1.24E-03	3.16E-01
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION	7.70E-05	2.60E-01
REACTOME_DOWNSTREAM_SIGNALING_EVENTS_OF_B_CELL_RECEPTOR_BCR	4.68E-03	2.23E-01
REACTOME_MRNA_SPLICING_MINOR_PATHWAY	9.98E-04	2.46E-01
REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	1.89E-02	2.55E-01
REACTOME_POL_SWITCHING	1.41E-02	3.76E-01
REACTOME_REPAIR_SYNTHESIS_FOR_GAP_FILLING_BY_DNA_POL_IN_TC_NER	4.67E-02	3.57E-01
REACTOME_LAGGING_STRAND_SYNTHESIS	5.81E-03	5.60E-01
REACTOME_INHIBITION_OF_REPLICATION_INITIATION_OF_DAMAGED_DNA_BY_RB1_E2F1	4.13E-02	3.38E-01
REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	3.33E-02	2.57E-01
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION	1.11E-02	4.50E-01
REACTOME_ANTIGEN_PROCESSING_UBIQUITINATION_PROTEASOME_DEGRADATION	1.60E-02	3.22E-01
REACTOME_TELOMERE_MAINTENANCE	3.13E-03	4.68E-01
REACTOME_EXTENSION_OF_TELOMERES	1.56E-03	4.00E-01
REACTOME_DNA_STRAND_ELONGATION	2.34E-05	2.02E-01

Pathway names and FDR adjusted p-values from INT and MPMWS are shown in the table. Selected pathways have p-values less than 0.05 in INT but larger than 0.2 in MPMWS.

Supplementary Table 3.2B: Significant Pathways from MPMWS.

Pathway Name	INT	MPMWS
BIOCARTA_G1_PATHWAY	7.62E-01	3.72E-02
BIOCARTA_CELLCYCLE_PATHWAY	2.99E-01	2.37E-02
PID_TELOMERASEPATHWAY	7.07E-01	4.43E-02
PID_P38ALPHABETADOWNSTREAMPATHWAY	2.39E-01	2.33E-02
MIPS_BAF_COMPLEX	2.44E-01	1.80E-02
MIPS_BRM_SIN3A_COMPLEX	2.18E-01	1.80E-02
MIPS_BRM_SIN3A_HDAC_COMPLEX	2.46E-01	4.83E-02
MIPS_NK_3_GROUCHO_HIPK2_SIN3A_RBPA48_HDAC1_COMPLEX	4.41E-01	2.55E-02
REACTOME_G1_PHASE	2.95E-01	3.99E-02
REACTOME_POST_TRANSLATIONAL_MODIFICATION_SYNTHESIS_OF_GPI_ANCHORED_PROTEINS	3.61E-01	4.65E-02

Pathway names and FDR adjusted p-values from INT and MPMWS are shown in the table. Selected pathways have p-values less than 0.05 in MPMWS but larger than 0.2 in INT.

Supplementary Table 3.3A. Significant pathways identified by both INT and MWMPS from TCGA KIRC dataset.

Pathway Name	INT	MPMWS
KEGG_NON_HOMOLOGOUS_END_JOINING	3.02E-02	2.10E-02
PID_TCPTP_PATHWAY	4.52E-02	1.86E-02
PID_HIF1APATHWAY	1.01E-02	6.82E-03
MIPS_POLYCOMB_REPRESSIVE_COMPLEX_1	5.17E-03	5.43E-03
MIPS_RNASE_MRP_COMPLEX	1.36E-02	4.15E-02
MIPS_ALL_1_SUPERCOMPLEX	1.73E-02	1.53E-02
REACTOME_TERMINATION_OF_O_GLYCAN_BIOSYNTHESIS	3.64E-02	3.92E-02
REACTOME_SIGNALING_BY_NOTCH1	4.22E-02	3.90E-03
REACTOME_ACTIVATED_POINT_MUTANTS_OF_FGFR2	3.35E-02	7.18E-03
REACTOME_OXYGEN_DEPENDENT_PROLINE_HYDROXYLATION_OF_HYPOXI A_INDUCIBLE_FACTOR_ALPHA	1.40E-02	2.11E-03
REACTOME_SIGNALING_BY_ACTIVATED_POINT_MUTANTS_OF_FGFR1	9.64E-03	2.14E-03
REACTOME_SIGNALING_BY_FGFR3_MUTANTS	1.34E-02	1.49E-02
REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S	3.41E-02	3.96E-02
REACTOME_FGFR2C_LIGAND_BINDING_AND_ACTIVATION	3.65E-02	6.27E-03
REACTOME_FGFR4_LIGAND_BINDING_AND_ACTIVATION	3.69E-02	2.92E-02
REACTOME_FGFR1_LIGAND_BINDING_AND_ACTIVATION	1.04E-02	2.94E-03

The p-values presented are not adjusted.

Supplementary Table 3.3B. Significant pathways identified by INT from TCGA KIRC dataset.

Pathway Name	INT	MPMWS
KEGG_TYROSINE_METABOLISM	4.71E-02	5.13E-01
KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_KERATAN_SULFATE	1.50E-02	2.05E-01
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES	3.90E-02	2.92E-01
KEGG_ECM_RECEPTOR_INTERACTION	2.14E-03	2.76E-01
BIOCARTA_EPHA4_PATHWAY	4.03E-02	4.93E-01
PID_INTEGRIN1_PATHWAY	7.37E-05	2.15E-01
PID_INTEGRIN_CS_PATHWAY	1.91E-02	4.29E-01
REACTOME_METABOLISM_OF_STEROID_HORMONES_AND_VITAMINS_A_AND_D	4.36E-02	5.59E-01
REACTOME_VITAMIN_B5_PANTOTHENATE_METABOLISM	3.89E-02	2.10E-01
REACTOME_SIGNALING_BY_NOTCH4	4.95E-02	2.33E-01
REACTOME_PLATELET_ADHESION_TO_EXPOSED_COLLAGEN	4.42E-02	5.57E-01
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	1.00E-03	2.02E-01
REACTOME_NCAM1_INTERACTIONS	4.64E-02	8.29E-01
REACTOME_ADHERENS_JUNCTIONS_INTERACTIONS	3.37E-02	6.15E-01
REACTOME_RNA_POL_III_CHAIN_ELONGATION	4.90E-02	5.24E-01
REACTOME_TANDEM_PORE_DOMAIN_POTASSIUM_CHANNELS	5.33E-03	6.70E-01
REACTOME_METABOLISM_OF_PORPHYRINS	4.23E-02	2.01E-01

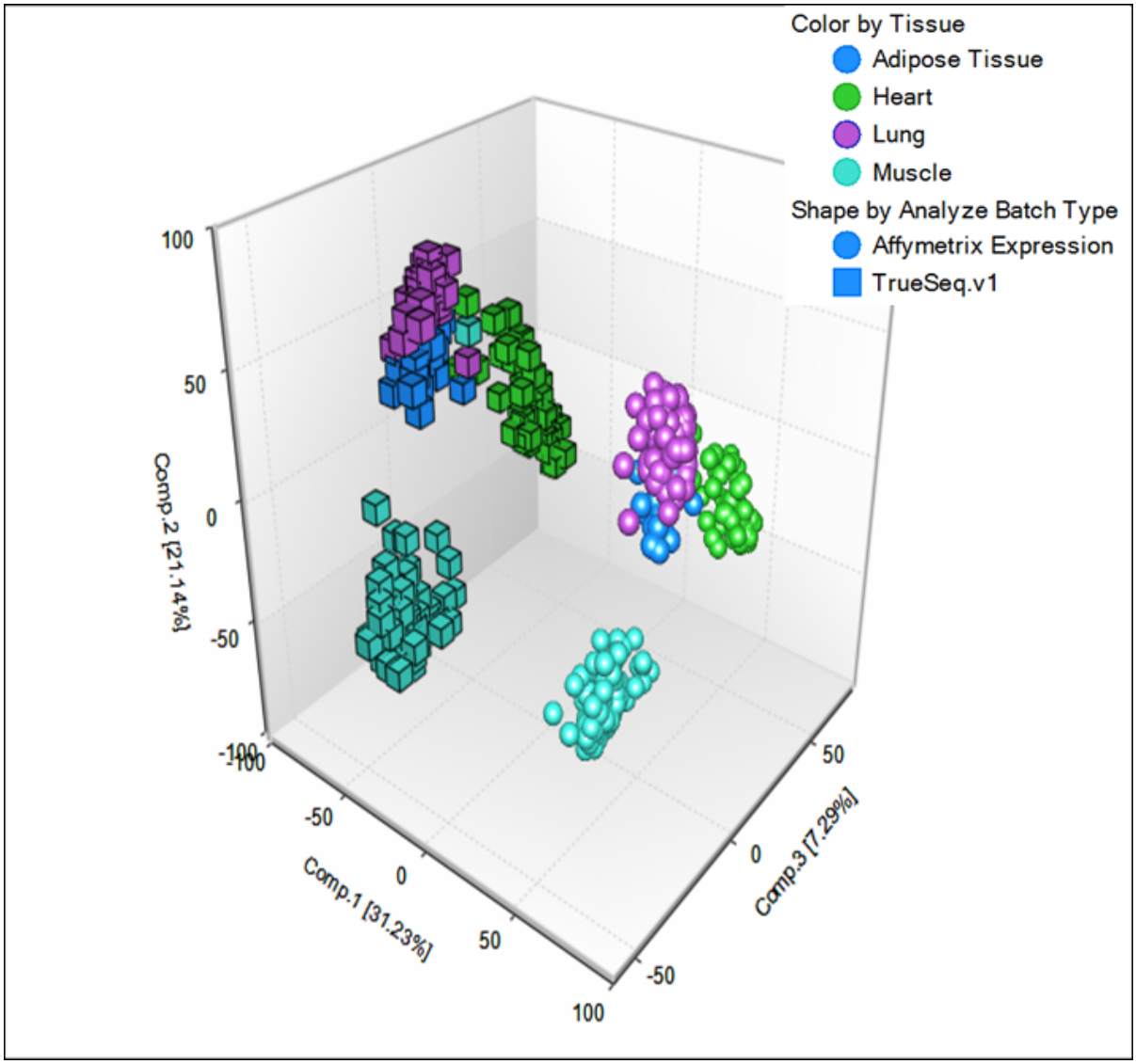
The p-values presented are not adjusted.

Supplementary Table 3.3C. Significant pathways identified by MWMPS from TCGA KIRC dataset.

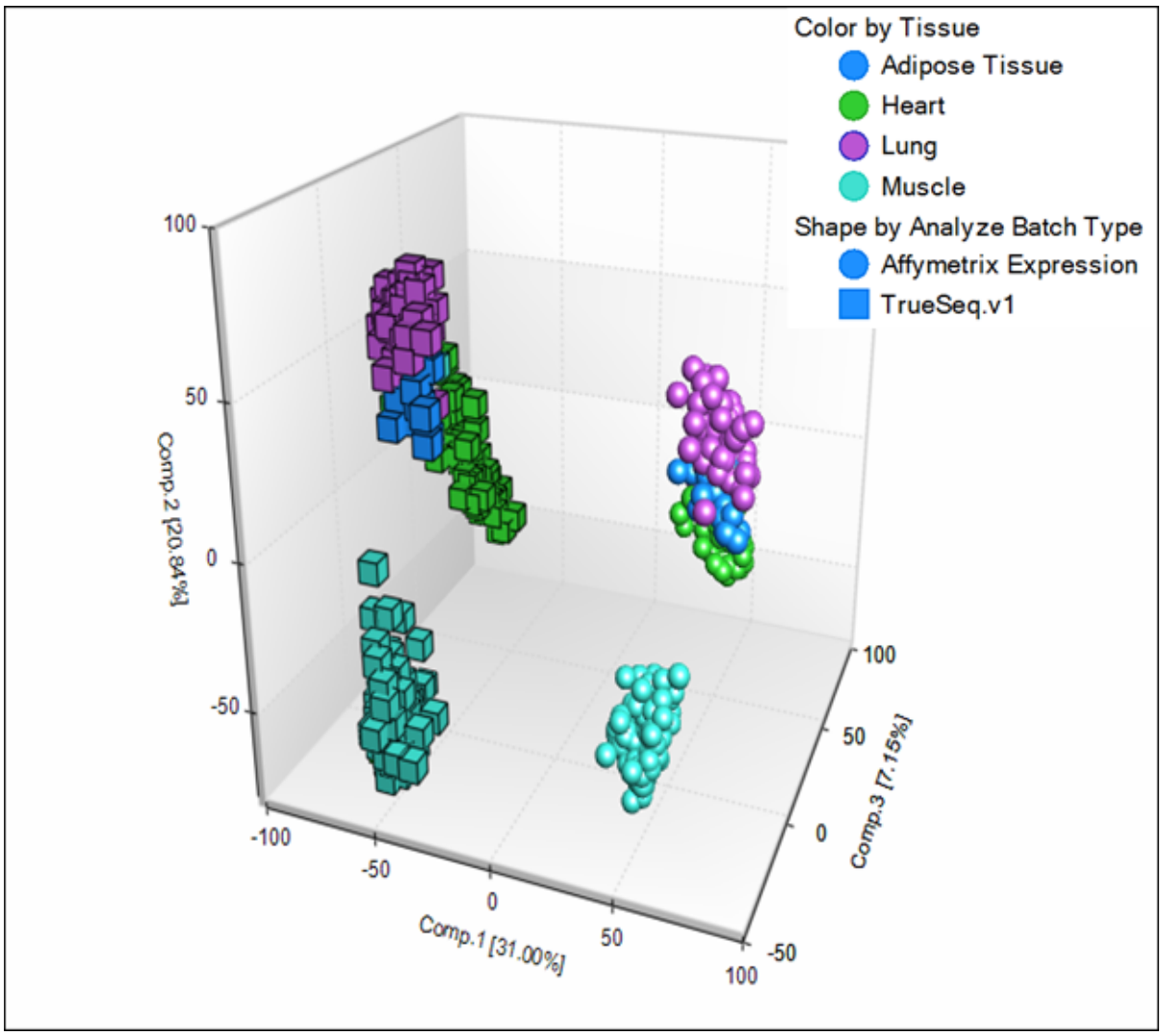
Pathway Name	INT	MPMWS
BIOCARTA_TID_PATHWAY	2.70E-01	4.08E-02
BIOCARTA_RNA_PATHWAY	3.00E-01	1.61E-02
BIOCARTA_TALL1_PATHWAY	4.34E-01	4.79E-02
BIOCARTA_TNFR2_PATHWAY	4.15E-01	1.40E-02
BIOCARTA_TOLL_PATHWAY	2.77E-01	3.91E-02
BIOCARTA_VEGF_PATHWAY	2.74E-01	1.10E-02
ST_TUMOR_NECROSIS_FACTOR_PATHWAY	2.44E-01	8.08E-03
SIG_CD40PATHWAYMAP	8.41E-01	3.59E-02
ST_GRANULE_CELL_SURVIVAL_PATHWAY	6.69E-01	6.91E-03
PID_HDAC_CLASSII_PATHWAY	2.51E-01	3.77E-02
PID_CD40_PATHWAY	8.93E-01	6.49E-03
PID_AVB3_OPN_PATHWAY	5.81E-01	1.14E-02
PID_HDAC_CLASSI_PATHWAY	2.52E-01	4.14E-02
PID_ERA_GENOMIC_PATHWAY	3.08E-01	3.28E-02
MIPS_FA_COMPLEX	4.73E-01	2.84E-02
REACTOME_O_LINKED_GLYCOSYLATION_OF_MUCINS	2.40E-01	1.41E-02
REACTOME_REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN	2.12E-01	1.82E-02
REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY	3.70E-01	2.90E-02
REACTOME_METABOLISM_OF_PROTEINS	2.90E-01	1.14E-03
REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	2.91E-01	2.80E-02
REACTOME_JNK_C_JUN_KINASES_PHOSPHORYLATION_AND_ACTIVATION_MEDIATED_BY_ACTIVATED_HUMAN_TAK1	4.43E-01	4.95E-02
REACTOME_TRAF6_MEDIATED_INDUCTION_OF_TAK1_COMPLEX	2.67E-01	1.30E-02
REACTOME_REGULATION_OF_GLUCOKINASE_BY_GLUCOKINASE_REGULATORY_PROTEIN	7.10E-01	3.94E-02

The p-values presented are not adjusted.

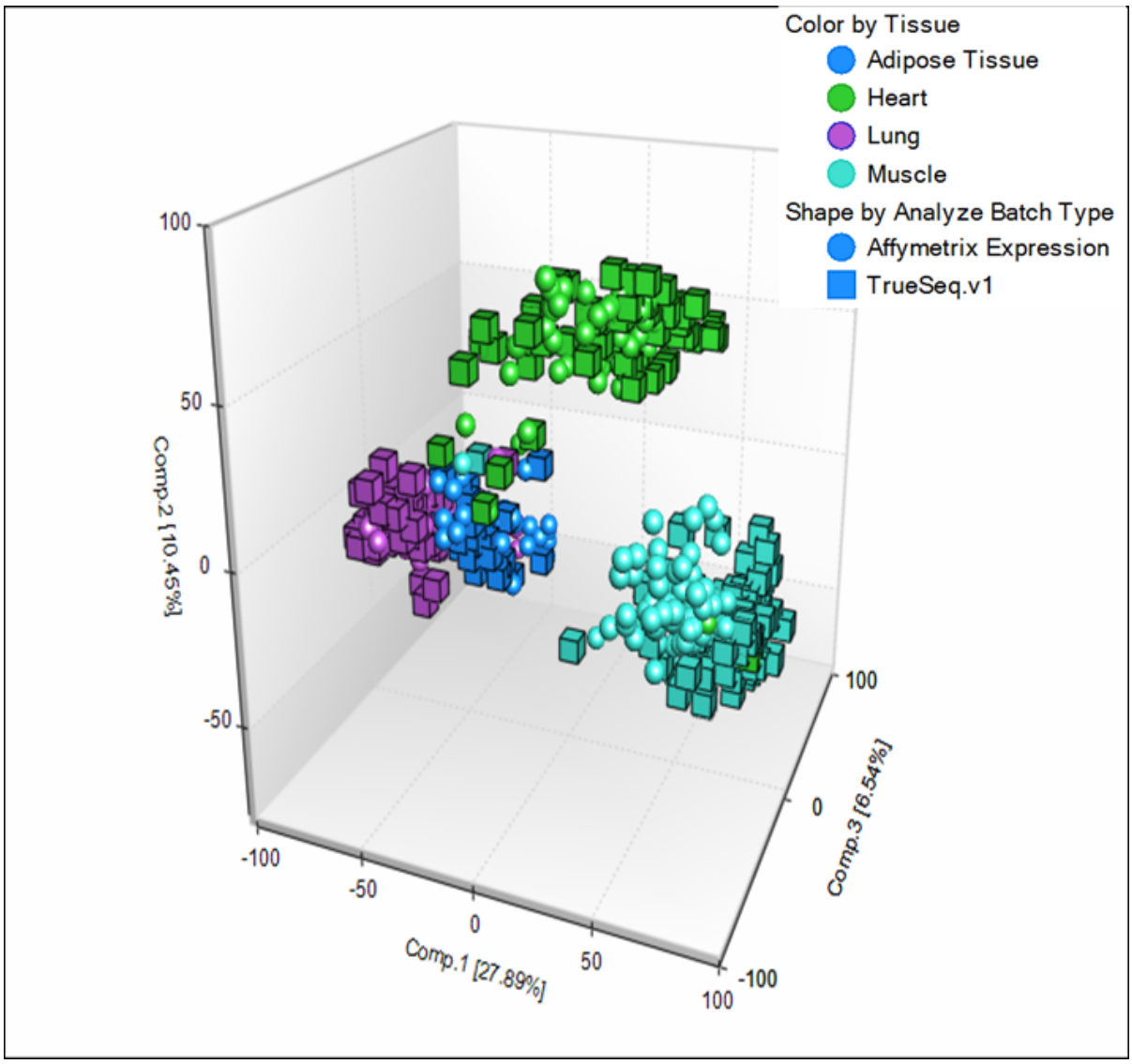
Appendix B – Supplementary figures from chapter 4



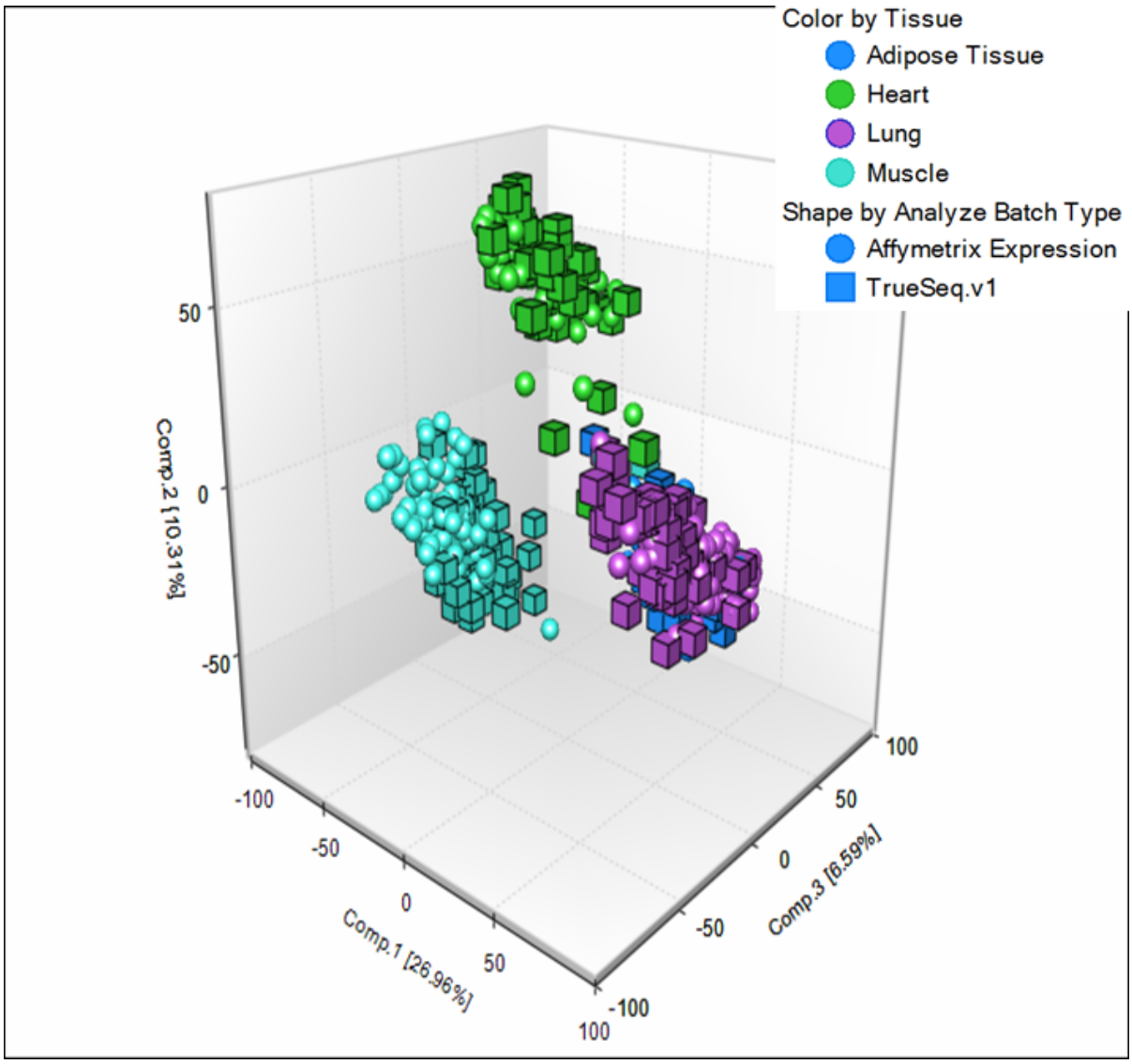
Supplementary Fig. 4.1. 3D PCA plot of GS-GTEx dataset



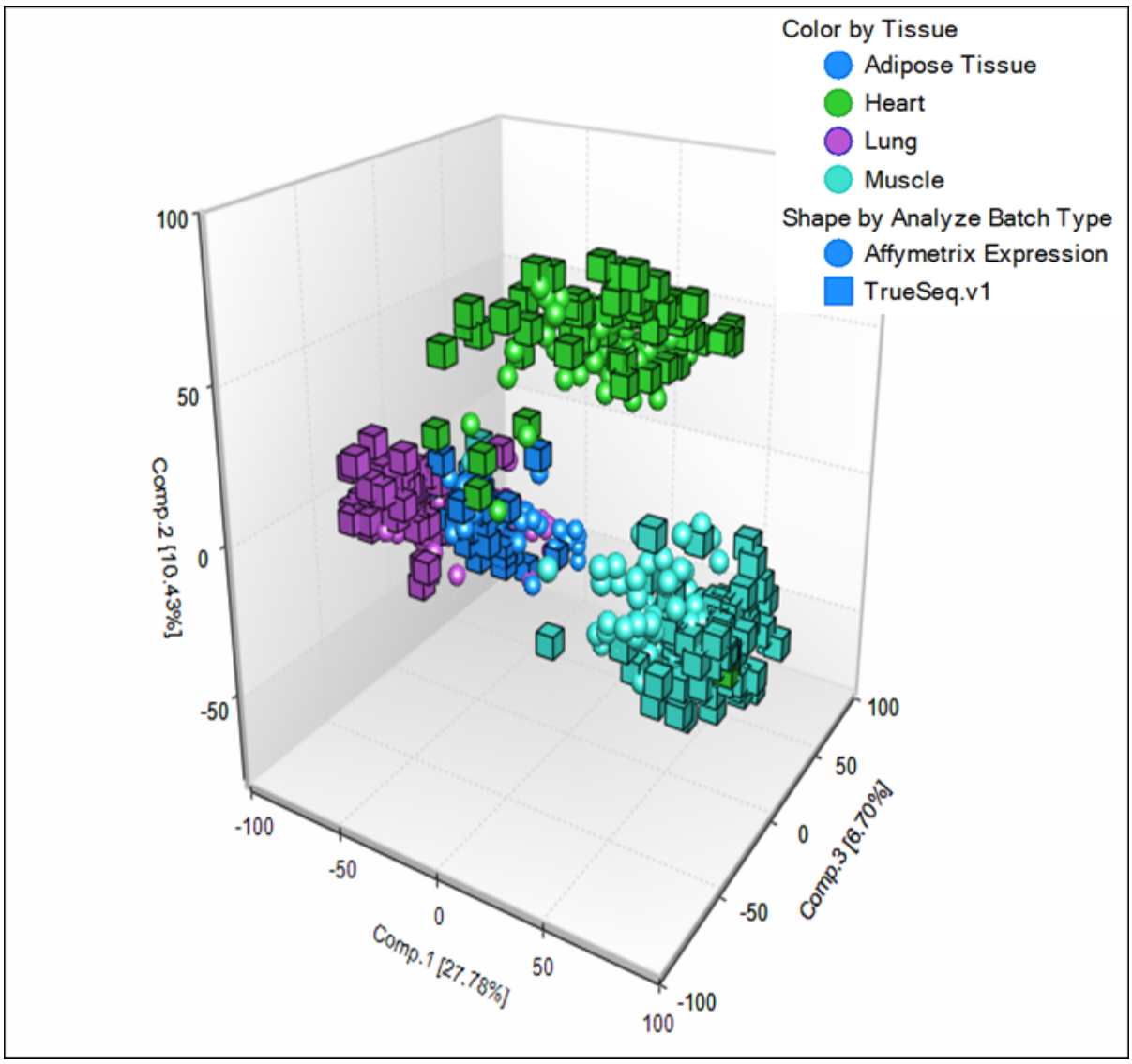
Supplementary Fig. 4.2. 3D PCA plot of GS-GTEx dataset using ranks



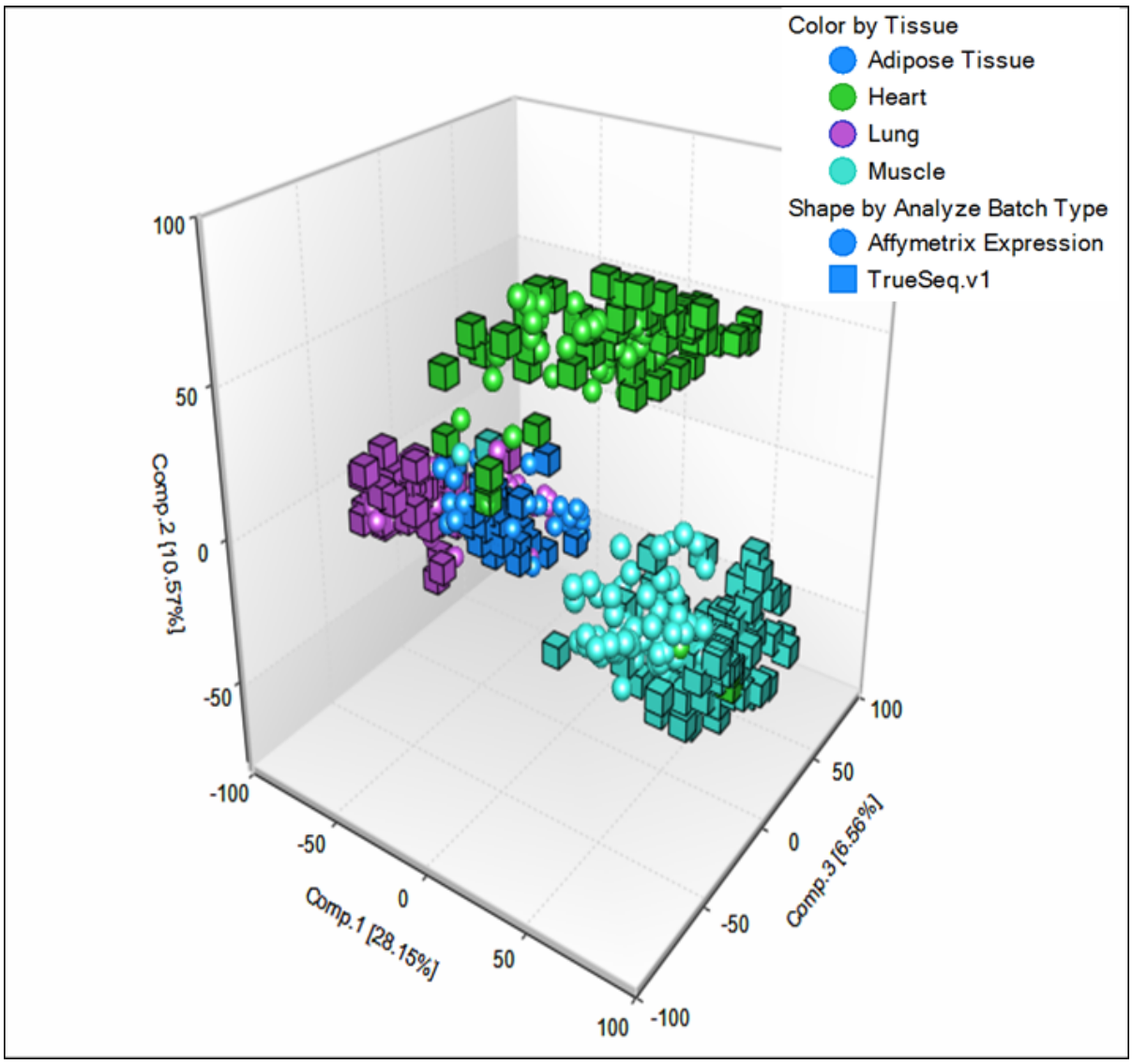
Supplementary Fig. 4.3. 3D PCA plot of GS-GTEx dataset after BMC



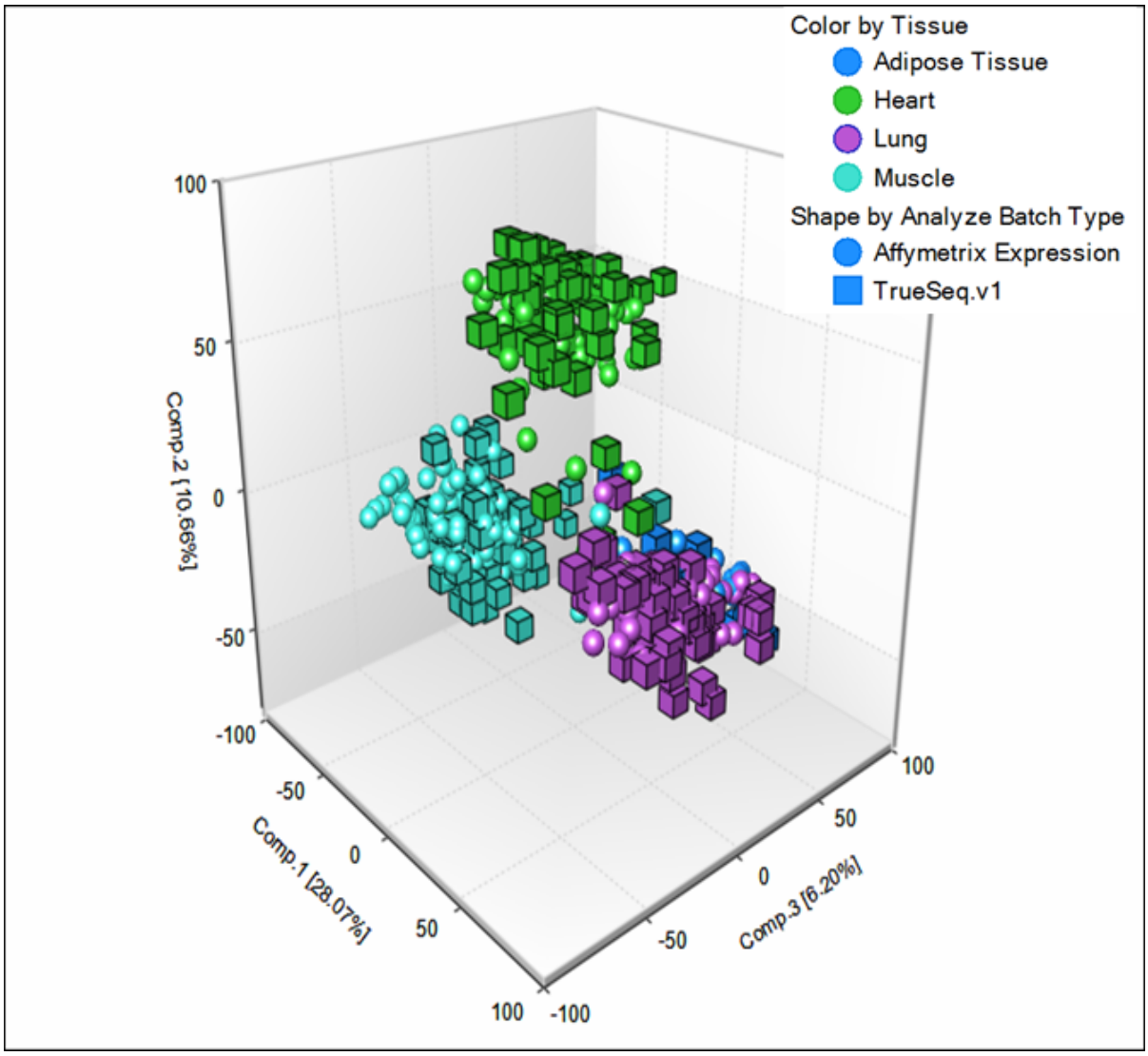
Supplementary Fig. 4.4. 3D PCA plot of GS-GTEx dataset after Gnorm



Supplementary Fig. 4.5. 3D PCA plot of GS-GTEx dataset after DWD



Supplementary Fig. 4.6. 3D PCA plot of GS-GTEx dataset after COMBAT



Supplementary Fig. 4.7. 3D PCA plot of GS-GTEx dataset after XPN