Bioinformatics Advance Access published January 30, 2014

Genome Analysis

Integrative Gene Set Analysis of Multi-platform Data with Sample Heterogeneity

Jun Hu[1,2*], and Jung-Ying Tzeng[1,3*]

[1]Bioinformatics Research Center, North Carolina State University, Ricks Hall, 1 Lampe Dr., Raleigh, NC 27607

[2]Division of Bioinformatics, Omicsoft Inc., 200 Cascade Pointe Lane, Suite 101, Cary, NC 27513, USA

[3]Department of Statistics, North Carolina State University, Ricks Hall, 1 Lampe Dr., Raleigh, NC 27607

Associate Editor: Dr. Inanc Birol

ABSTRACT

Motivation: 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.

Results: We find that existing methods for 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.

Availability:
http://www4.stat.ncsu.edu/~jytzeng/Software/Multiplatform_gene_set_analysis/

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

*To whom correspondence should be addressed.

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.

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.

2 METHODS

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. 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.

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.

(A) Simulation with homogenous samples. 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).

(B) Simulation with heterogeneous samples. 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, 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.

Multi-platform methods for gene set analysis without sample heterogeneity

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.

Multi-platform methods for gene set analysis 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.,

$$T'_{m\ell} = (T_{m\ell} - \overline{T_{.\ell}})/S_{.\ell} + c_l. \quad (1)$$

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:

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

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 $\text{MAD} = 1.4286 \times median(z_{im\ell})$, where

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

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

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

  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.

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

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

- 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 $U_{k,m\ell}$.

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

Real data analysis

We performed real data analysis using the methods that have the best performances in the simulation studies (i.e., MPMWS and INT). The 1452 pathways in MSigDB were tested using both the breast cancer dataset and the KIRC datasets, which comprised genomic data from methylation, CNV, and RNA-Seq platforms.

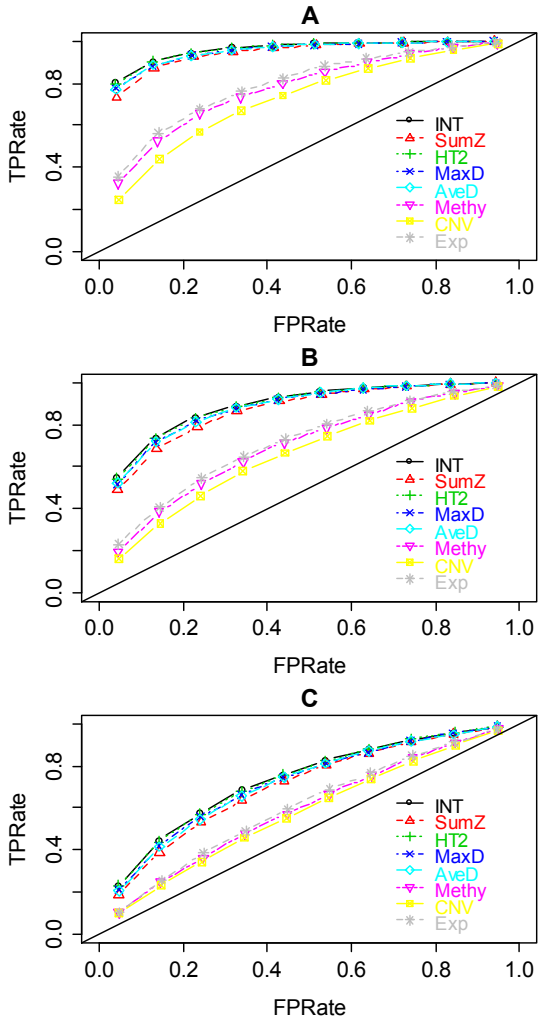

Fig. 1. ROC plots for gene set methods at different α levels (A: α = 75%; B: α = 50%; and C: α=25%). The simulated data were generated with β=0.8 and η=0.91

3 RESULTS

3.1 Multi-platform methods for gene set analysis with homogenous samples (Scenario A)

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. 1 shows the ROC plots under different proportions of causal genes in a causal set, i.e., $\alpha = 75\%, 50\%,$ and 25% for Figs. 1A, 1B, and 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 Supplementary Table 1A. From Fig. 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. 1B and 1C, the relative performance among different methods stayed the same as in Fig. 1A, except that the TPRs decreased when α decreased. The same patterns were observed for β = 0.6 (Supplementary Fig. 1; Supplementary Table 1A).

Fig. 2 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 Supplementary Table 1A). The patterns for the relative performances of different methods were observed to be similar to those of Fig. 1. As expected, all methods performed better when the difference between case and control became larger (i.e., larger β).

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 (Supplementary Fig. 2) and 0.1 (Supplementary Fig. 3); similar results were observed under these scenarios.

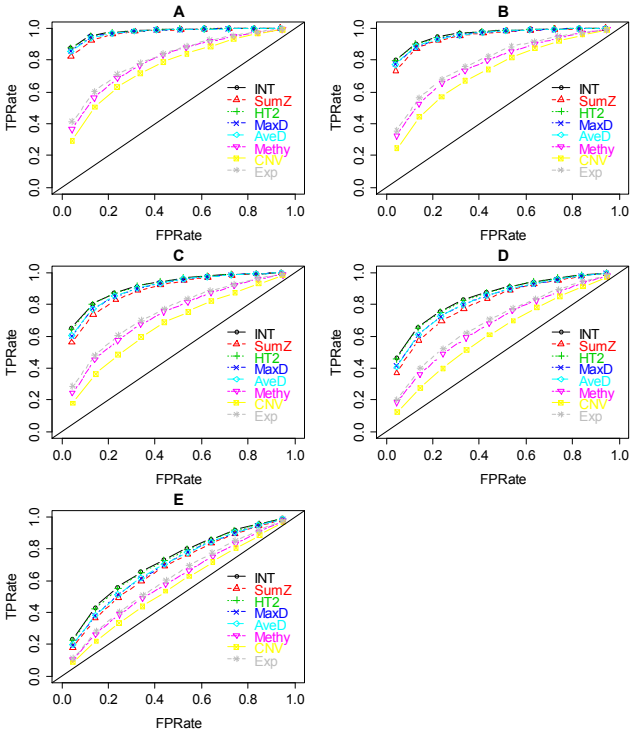

Fig. 2. ROC plots for gene set methods at different β levels (A: β = 0.9; B: β = 0.8; C: β=0.6; D: β=0.4; and E: β=0.2). The simulated data were generated with α = 75% and η = 0.91

3.2 Multi-platform methods for gene set analysis with heterogeneous samples (Scenarios B1 and B2)

To evaluate the performance under sample heterogeneity, we simulated datasets by randomly selecting γ% of case samples to be "true" cases. We focused our comparisons on the two represented approaches from Section 3.1 (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, where α = 75%, β = 0.8 and η = 0.91, and the percentage of "true" cases among all cases varies, i.e., $\gamma = (100\%, 90\%, 80\%, 60\%, 40\%,$ and $20\%)$. The corresponding AUC values are presented in Supplementary Table 1B. 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., ≤40%), MPLRS performed the best; when γ was 60%, MPLRS and MPMWS had similar power. However, when γ was high (e.g., ≥ 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 (Supplementary Fig. 4) and 0.1 (Supplementary Fig. 5)

By design, INT is good at identifying pathways with systematic changes, whereas MPMWS has robust power to detect pathways involving sample heterogeneity. In Table 1, 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., $\gamma = 20\%$).

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 (Supplementary Fig. 6).

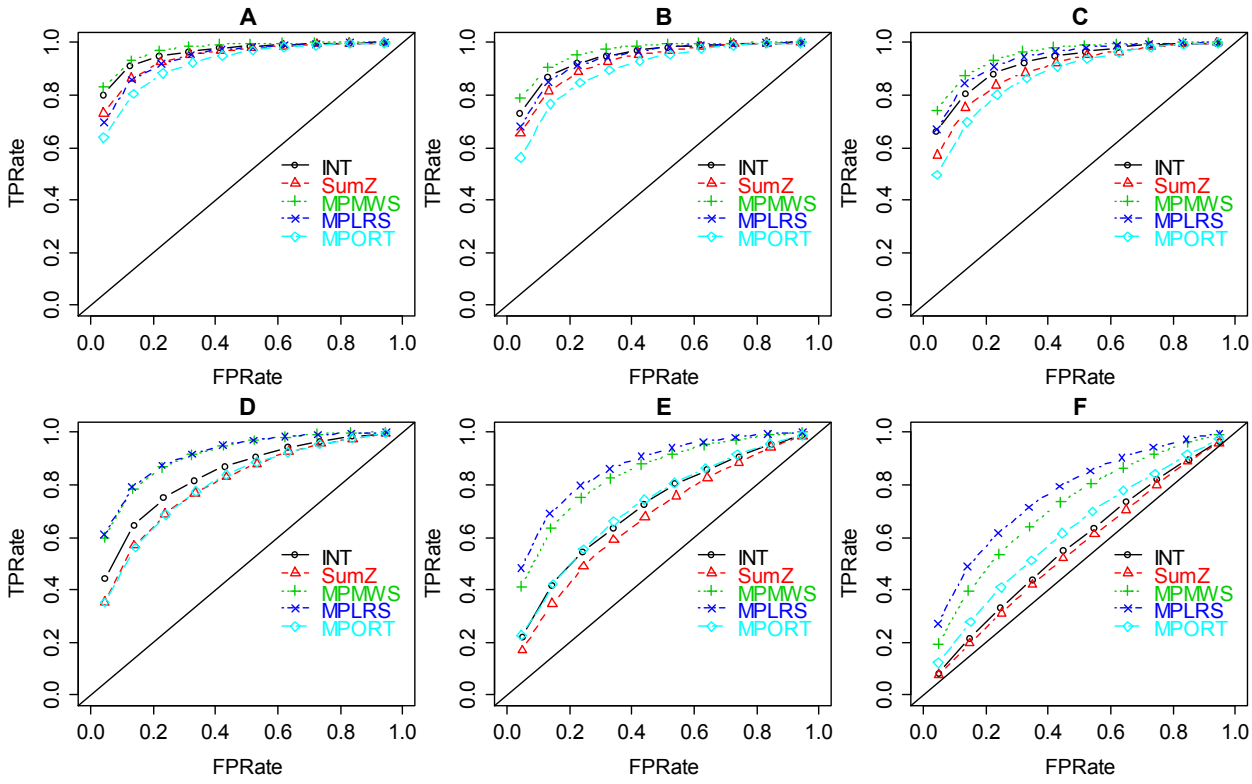

Fig. 3. ROC plots for gene set methods at different sample heterogeneity levels. (A: γ = 100%; B: γ = 90%; C: γ =80%; D: γ =60%; E: γ =40%; and F: γ=20%). The simulated data were generated with α =75%, β = 0.8, and η =0.91.

Table 1. 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)

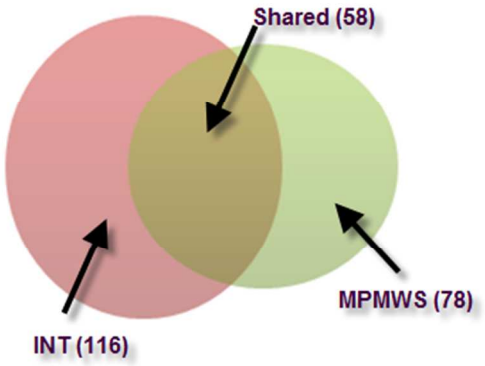

Fig. 4. Significant pathways identified by MPMWS and INT. The numbers of significant pathways are listed in parentheses.

3.3 Real data application

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. 4 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. 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 2). 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 3A). 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 3B). 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 Table 4.

4 DISCUSSSION

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. Second, 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.

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.

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/

ACKNOWLEDGEMENTS

We thank Dr. Kejun Liu at Omicsoft Inc. and Dr. Shannon Holloway at North Carolina State University for helpful discussion and invaluable suggestions.

Funding: This work is partially supported by NIH grants R01 MH074027 and P01 CA142538.

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.

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.

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.

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

Evans, D.M. and Purcell, S. (2012) Power Calculations in Genetic Studies, Cold Spring Harbor Protocols, 2012, pdb.top069559.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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, 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.