Published in final edited form as:

Methods Mol Biol. 2017 ; 1613: 125–159. doi:10.1007/978-1-4939-7027-8_7.

Extracting the Strongest Signals from Omics Data: Differentially Expressed Pathways and Beyond

Galina Glazko, Yasir Rahmatallah, Boris Zybailov, and Frank Emmert-Streib

Abstract

The analysis of gene sets (in a form of functionally related genes or pathways) has become the method of choice for extracting the strongest signals from omics data. The motivation behind using gene sets instead of individual genes is two-fold. First, this approach incorporates pre-existing biological knowledge into the analysis and facilitates the interpretation of experimental results. Second, it employs a statistical hypotheses testing framework. Here, we briefly review main Gene Set Analysis (GSA) approaches for testing differential expression of gene sets and several GSA approaches for testing statistical hypotheses beyond differential expression that allow extracting additional biological information from the data. We distinguish three major types of GSA approaches testing: (1) differential expression (DE), (2) differential variability (DV), and (3) differential co-expression (DC) of gene sets between two phenotypes. We also present comparative power analysis and Type I error rates for different approaches in each major type of GSA on simulated data. Our evaluation presents a concise guideline for selecting GSA approaches best performing under particular experimental settings. The value of the three major types of GSA approaches is illustrated with real data example. While being applied to the same data set, major types of GSA approaches result in complementary biological information.

Keywords

Omics data; Gene set analysis approaches; Hypotheses testing; Self-contained; Competitive; Differential expression; Differential co-expression; Differential variability

1 Introduction

Biological systems are living proofs of Aristotle's idea that the whole is greater than the sum of its parts. For example, cell is a product of synergistic actions of its constituents (genes, proteins, metabolites, just to name a few). Together with cellular environment this synergy defines what we call the cell type (e.g., a stem cell or dendritic cell). At the level of cell's key molecules (nucleic acids and proteins) the idea of synergy also holds true the following: genes work together in biological pathways, proteins form protein complexes, that is genes and proteins are organized in functional units acting overall differently than a single gene or a single protein would. Thus, when an investigator studies omics data, the idea to consider functional units instead of individual components comes naturally to mind. In fact, this idea was first employed for the analysis of gene expression data more than a decade ago [1]. Analyzing microarray data from diabetics vs. healthy controls Mootha and colleagues [1] did not find a single gene to be differentially expressed. However, when genes were analyzed at the pathway level using Gene Set Enrichment Analysis (GSEA) approach, it was

found that genes involved in oxidative phosphorylation showed reduced expression in diabetics although the average decrease per gene was only 20% [1]. There were two reasons behind the success of the pathway analysis approach in this case. First, the number of hypotheses to test by arranging genes into pathways is dramatically reduced, which leads to the increase in power. Second, in metabolic diseases such as diabetes changes in gene expression are moderate and therefore can be overlooked by using methods focusing on each gene individually. These two reasons explain why pathway analysis has become the method of choice in analyzing omics data in general and expression data in particular. Nowadays, we also recognize yet another important reason to employ pathway (gene set) analysis for omics data. Gene Set Analysis (GSA) approaches provide flexibility to test different statistical hypotheses, thus increasing the biological interpretability of experimental results. Here, we briefly review main Gene Set Analysis (GSA) approaches for testing differential expression of gene sets and several GSA approaches for testing statistical hypotheses beyond differential expression, which allow extracting additional biological information from the data.

We distinguish the three major types of GSA approaches that test statistically and biologically different hypotheses: (1) differential expression (DE), (2) differential variability (DV), and (3) differential co-expression (DC) of gene sets between two phenotypes. All major types of GSA approaches can be univariate (gene-level) or multivariate (accounting for intergene correlations). The chapter is organized as follows: In the first part of Subheading 2, we discuss GSA approaches developed for identification of differentially expressed pathways applicable for the analysis of microarrays and RNA-seq data (GSA-DE). The traditional GSA-DE framework aims to identify pathways with significant changes in mean gene expressions and it is well understood. In the second part of Subheading 2, DV analysis in application to gene sets (GSA-DV) is considered. The analysis of differential variability (DV) is somewhat appreciated with regards to individual genes, when the aim is to find genes with significant changes in expression variance between two phenotypes [2–6]. It was shown that many statistically significant DV genes are relevant to disease development and that DV is an indication of changes in gene regulation [2, 3]. Moreover, it was found that there are genes showing consistently higher across-sample variability in tumors of different origin as compared to normal samples [7]. These DV genes can serve as a robust molecular signature for multiple cancer types [7, 8]. Given the evidence that DV genes may play an important role in observed phenotypes, and given the popularity of GSA approaches one would expect there are many approaches implementing GSA-DV test. Our group was the first to suggest extending the DV analysis to a multivariate GSA-DV case using multivariate statistical test [9, 10]. In the same publication we further demonstrated that for three different cancer types GSA-DV approach was able to identify cancer-specific pathways, while pathways identified using conventional GSA-DE approaches were shared between the three cancer types. Thus, GSA-DV approach provides additional biological information beyond GSA-DE. It should be noted that there are other approaches claiming to perform GSA-DV test, e.g., DIRAC and EVA [11], but because they compare variability in gene ranks within a pathway between two phenotypes rather than variance estimates, these approaches are out of the scope of this chapter. We discuss two principally different GSA-DV approaches: (1) nonparametric multivariate GSA-DV approach, “radial” Kolmogorov-

Smirnov (RKS) [9] and (2) new gene-level GSA-DV test we suggest here for the first time. This gene-level GSA-DV approach applies Fisher Method (FM) [12] for combining P -values from gene-level F -test for differential variability [3]. It should be noted that currently GSA-DV approaches are applicable only to micro-array data, because RNA-seq read counts are most frequently modeled with Negative Binomial distribution that has complex dependence between mean and variance. In the third part of Subheading 2, GSA approaches estimating differential co-expression of gene sets between two phenotypes (GSA-DC) are considered. In a pathway, genes are working together, i.e., they form a co-expression network. For finding DC pathways GSA-DC approaches with or without network inference step can be employed. The most general GSA-DC approach with a network inference step is based on a Gaussian Graphical Model (GGM) [13]. In this approach, the network structure of a pathway for each phenotype is estimated and the null hypothesis to test is that the network structure across phenotypes is the same [13]. The network inference step per se is challenging because there are too many ways of estimating network structure. For example, the implementation of network inference in Bioconductor package *nethet* (that provides two-sample testing in GGMs) includes several options, such as the Graphical Lasso (GL) [14], the Meinshausen-Buhlmann approach [15], and the approach proposed by Schafer and Strimmer based on shrinkage estimation of the covariance matrix [16]. Needless to say, the *nethet* results for networks comparison will vary significantly depending on the algorithm selected for the network inference step. In addition, many approaches for network inference (e.g., GGM) require the assumption of normality that may or may not be met in the real data. This is why we present in this review only GSA-DC approaches that do not require a network inference step. The simplest GSA-DC approach, the gene sets co-expression analysis (GSCA) [17] is purely univariate. GSCA calculates the Euclidian distance between two correlation vectors (constructed from diagonal matrices of pairwise correlations for different conditions) and the significance of the difference is estimated using permutation test. The gene sets net correlations analysis (GSNCA) [18] assesses multivariate changes in the gene co-expression network between two conditions but does not require network inference step. Net correlation changes are estimated by introducing for each gene a weight factor that characterizes its cross-correlations in the co-expression networks. Weight vectors in both conditions are found as eigenvectors of correlation matrices with zero diagonal elements. Gene sets net correlations analysis (GSNCA) tests the hypothesis that for a gene set there is no difference in the gene weight vectors between two conditions [18]. The Co-expression Graph Analysis (CoGA) identifies co-expressed gene sets by statistically testing the equality in the spectral distributions [19]. For each phenotype CoGA constructs a full network from pairwise correlations between gene expressions. Then the structural properties of the two networks are compared by applying Jensen-Shannon divergence as a distance measure between the graph spectrum distributions [19, 20]. All methods are supplied with the implementation reference if available.

In Subheading 3, we first present a comparative power analysis and Type I error rates for different approaches in each major type of GSA on simulated data. Second, the value of applying the three major types of GSA approaches is illustrated with real data example, where these approaches provide different biological information obtained on the same data set.

2 Methods

2.1 Gene Set Analysis Approaches for Testing Differential Expression (GSA-DE)

There are many GSA-DE approaches readily distinguished based on the null hypothesis they test. According to Goeman and Buhlmann [21] the formulation can be either *self-contained* or *competitive*. *Self-contained* approaches compare whether a gene set is differentially expressed between different conditions, while *competitive* (e.g., GSEA) approaches compare a gene set against its complement that contains all genes except genes in the set [21, 22]. *Self-contained* approaches can be (1) univariate, in a sense that they use gene-level tests for GSA and combine univariate statistics for individual genes into a single test score [10, 23, 24]; and (2) multivariate, when a multivariate statistic is used to address the null hypothesis. In a real biological setting, moderate [25] and extensive [26] correlations between genes in gene sets are well documented [27] and that may result in a decrease of the power for gene-level tests compared to multivariate tests [24, 27–29]. In turn, *competitive* GSA approaches can be (1) “supervised,” when the class labels are known; or (2) “unsupervised,” when the enrichment score is computed for each gene set and individual sample [30]. For GSA-DE the “supervised” term indicates that the samples classification is known, while the “unsupervised” term indicates that the samples classification is unknown [30]. A number of review articles concerning the different aspects of GSA-DE approaches developed for microarrays data analysis have been published [21, 23, 31–36].

To summarize, GSA-DE approaches that test intrinsically statistically different null hypotheses developed thus far are: *self-contained* (univariate, multivariate) and *competitive* (supervised, unsupervised). Figure 1 illustrates different null hypotheses tested by various GSA-DE approaches together with R packages implementing each test. For the sake of generality, all power and Type I error rate estimates for GSA-DE approaches are presented for simulated RNA-seq counts.

Null Hypotheses—Consider two different biological phenotypes, with n_1 samples of measurements for the first and n_2 samples of the same measurements for the second. Let the two random vectors of $X = (X_1, \dots, X_{n_1})$ and $Y = (Y_1, \dots, Y_{n_2})$ represent the measurements of p gene expressions (constituting a pathway) in two phenotypes where X_i is the i th p -dimensional sample in one phenotype and Y_i is the i th p -dimensional sample in the other phenotype. Let X, Y be independent and identically distributed with the distribution functions F_x, F_y , mean vectors $\bar{\mu}_x$ and $\bar{\mu}_y$, and $p \times p$ positive-definite and symmetric covariance matrices Σ_x and Σ_y .

H_0 for self-contained tests: For multivariate self-contained tests we consider the problem of testing the general hypothesis $H_0: F_x = F_y$ against an alternative $F_x \neq F_y$, or a restricted hypothesis $H_0: \bar{\mu}_x = \bar{\mu}_y$ against an alternative $\bar{\mu}_x \neq \bar{\mu}_y$ depending on a test statistic.

Gene-level GSA approaches test a null hypothesis that the gene set-associated score does not differ between phenotypes. The score can be calculated, for example, as an L_2 -norm of the moderated t -statistics [37] or as combined P -values [24]. In all cases statistical significance is evaluated by comparing the observed score with the null distribution, obtained by permuting sample labels.

H₀ for competitive tests: The Gene Set Enrichment Analysis (GSEA) method [1, 38] is one of the most widely used competitive approaches. As a local test statistic, it uses a signal-to-noise ratio and a weighted Kolmogorov-Smirnov as a global test statistic (enrichment score, normalized to factor out the gene set size dependence) [34, 38]. Assuming a null distribution F_0^{perm} induced by permuting sample labels, GSEA evaluates significance of the global test statistic ζ_k^{GSEA} by estimating nominal P -value from F_0^{perm} [34, 38]. Thus, GSEA tests the null hypothesis that the genes in a gene set are randomly associated with the phenotype.

Most competitive GSA approaches are supervised, in a sense that sample labels are known (that is, there are at least two different phenotypes). Recently, the concept of unsupervised GSEA where an enrichment score is computed for each gene set and individual sample was introduced [30]. Essentially, unsupervised GSEA transforms a matrix of gene expressions across samples into a matrix of gene sets enrichment scores across the same samples. It makes the choice of null hypothesis flexible and context dependent. For example, Barbie et al. [39] use unsupervised competitive GSEA to test the null hypothesis that the Spearman correlation between gene sets enrichment scores is zero, while Hazeldmann et al. [30] test the hypothesis that gene set enrichment score does not differ between two phenotypes.

SELF-CONTAINED GENE-LEVEL TESTS FOR GSA—Gene-level tests for GSA can be easily designed in three steps: (1) select a gene-level score based on a univariate test statistic (e.g., a value of t -test), (2) transform a score (e.g., take an absolute value of t -statistic, or consider its P -value), and (3) summarize gene-level scores into a gene set statistic (e.g., take an average of transformed scores or use combining P -values approach) [10, 23, 24].

Gene-level GSA-DE tests that combine genes P -values: Gene-level tests for GSA that combine P -values from individual tests for microarray data were studied in [40]. As a gene-level test, the authors used an F -statistic for the correlation between the gene expression and phenotype ($F = (N - 2)[r^2 / (1 - r^2)]$ (not to confuse with F -test) and compared several approaches for combining P -values: Fisher's method (FM) [12], Stouffer's method (SM) [41], tail strength (TS) [42], and a modified tail strength statistic (MTS) [40]. It was found that FM outperformed all the other methods for combining P -values [40].

Gene-level tests for GSA that combine P -values from individual tests for RNA-seq data were studied in [24]. In what follows, we briefly reiterate the conclusions from comparative power and Type I error rate analyses of different gene-level GSA tests [24]. There are two popular univariate tests specifically developed for RNA-seq data that rely on Negative Binomial model for read counts: edgeR [43] and DESeq [44]. Empirical Bayes method eBayes [45] correctly identifies hypervariable genes and can be adapted for RNA-seq data through VROOM normalization [46]. When applied correctly the gene-level test does not per se influence the performance of a gene-level GSA approach as much as the procedure used to combine univariate statistics into a single test score does [24]. Among many approaches available for combining P -values from gene-level tests, we have shown that, similar to the results for microarray data, the safest option is to use FM [24, 47]. Here, for comparative power and Type I error rate estimates eBayes in combination with FM is selected.

Gene-level GSA-DE test that combines statistics: In the analysis of microarrays, shrinking the standard error of a test statistic (e.g., a t -test) in testing DE of individual genes improves the power of the test. Several shrinkage approaches at the level of individual genes were suggested, including the Significance Analysis of Microarrays (SAM) test [48], the regularized t -test [49], and the moderated t -test [50]. In particular, an extension of SAM test to gene set analysis (SAM-GS) has been demonstrated to outperform several conventional self-contained tests and even the original competitive GSEA approach for microarray data [10, 37, 51, 52].

SAM-GS can be applied to RNA-seq count data by using the VOOM normalization [46] prior to the test to find the log-scale counts per million (CPM) of the raw counts normalized for library sizes. The test statistic is the L_2 -norm of the moderated t -statistics for the gene expressions:

$$T_{\text{SAM-GS}} = \sum_{i=1}^p \left(\frac{X_i - Y_i}{s_i + s_0} \right)^2$$

where $\bar{X}_i$ and $\bar{Y}_i$ are respectively the mean expression levels for gene i under phenotypes X and Y , s_i is a pooled standard deviation over the samples in the two phenotype, s_0 is a small positive constant to adjust for small variability, and p is the number of genes in the gene set.

SELF-CONTAINED MULTIVARIATE TESTS FOR GSA—Based on their high power and popularity we consider two multivariate test statistics.

N-statistic: N-statistic [53, 54] tests the most general hypothesis $H: F_X = F_Y$ against a two-sided alternative $F_X \neq F_Y$:

$$N_{n_1 n_2} = \frac{n_1 n_2}{n_1 + n_2} \left[\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} L(X_i, Y_j) - \frac{1}{2n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} L(X_i, X_j) - \frac{1}{2n_2^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} L(Y_i, Y_j) \right]^{1/2}$$

Here, we consider only $L(X, Y) = \|X - Y\|$, the Euclidian distance in R^p . N-statistic was applied to microarray data and was shown to outperform other univariate and multivariate GSA-DE tests under different parameter settings [10, 28]. After VOOM normalization [46] N-statistic can also be applied to RNA-seq data and also was shown to outperform other GSA-DE tests [24, 47].

ROAST: In the context of microarray data, a parametric multivariate rotation gene set test (ROAST) has become popular for the self-contained GSA approaches [55]. ROAST uses the framework of linear models and tests whether for all genes in a set, a particular contrast of the coefficients is nonzero [55]. It can account for correlations between genes and has the flexibility of using different alternative hypotheses, testing whether the direction of changes in mean is *up*, *down*, or *mixed* (up or down) [55]. For microarrays it was shown that when correlations are low ROAST performance is similar to N-statistic [10]. Using ROAST with RNA-seq count data requires proper normalization. The VOOM normalization [46] was

proposed specifically for this purpose where log counts per million, normalized for library size are used. In addition to counts normalization, VOOM calculates associated precision weights that can be incorporated into the linear modeling process within ROAST to eliminate the mean-variance trend in the normalized counts [46].

SUPERVISED COMPETITIVE TESTS FOR GSA

ROMER: The first competitive GSA test for microarray data analysis GSEA [1] was developed a decade ago. The original GSEA was sensitive to the gene set size and the influence of other gene sets [56], so it was subsequently upgraded into GSEA-P that used a correlation-weighted KS statistic, an improved enrichment normalization, and an FDR-based estimate of significance [34, 38]. For the sake of simplicity, we will only consider the GSEA version implemented in Bioconductor package `limma` function ROMER (the rotation testing using mean ranks) [57]. ROMER is a parametric method developed originally for microarray data and uses the framework of linear models [46] and rotations instead of permutations (*see ref. 55 for more detail*). In contrast to ROAST, the `limma` implementation of ROMER does not incorporate the weights, estimated by VOOM into the linear modeling process to account for the mean-variance trend in the data.

UNSUPERVISED COMPETITIVE TESTS FOR GSA—The goal of unsupervised competitive approaches is to characterize the degree of expression enrichment of a gene set in each sample within a given data set [39]. The term “competitive” is reminiscent of the way the enrichment score is calculated: as a function of gene expression inside and outside the gene set.

Gene set variation analysis (GSVA): GSVA can be applied to microarray expression values or RNA-seq counts. Depending on the data type, expression values (counts) are first transformed using a Gaussian (or discrete Poisson) kernel into expression-level statistics [30]. The sample-wise enrichment score for a gene set is calculated using KS-like random walk statistic. An enrichment statistic (GSVA score) can be calculated as its maximum deviation from zero over all genes (similar to the original GSEA) or as the difference between the largest positive and negative deviations from zero (*see ref. 30 for more detail*).

Single sample extension of GSEA (ssGSEA): The difference between GSVA and ssGSEA stems from the way an enrichment score is calculated. In ssGSEA the enrichment score for a gene set under one sample is calculated as a sum of the differences between two weighted empirical cumulative distribution functions of gene expressions inside and outside the set [39]. The approach, together with GSVA, is implemented in the Bioconductor GSVA package [30].

2.2 Gene Set Analysis Approaches for Testing Differential Variability (GSA-DV)

It is well recognized that multivariate statistics have more power than univariate in the case of GSA-DE when intergene correlations are high [24, 27–29]; however, in the case of GSA-DV, this question was not studied at all. Here, we address this shortcoming by providing comparative power analysis for RKS, N-statistic, and gene-level approach for GSA-DV (*see below*).

Null Hypotheses

H₀ for GSA-DV: While H_0 for RKS is the same general hypothesis tested, e.g., by N-statistic, namely $H_0: F_X = F_Y$, an alternative in this case is not $F_X \neq F_Y$ or $\bar{\mu}_X \neq \bar{\mu}_Y$ but $\bar{\sigma}_X \neq \bar{\sigma}_Y$, i.e., differences in scale. N-statistic tests an alternative $F_X \neq F_Y$. Because this general alternative implicitly includes inequality of variances for distribution functions F_X and F_Y , N-statistic can also capture differences in scale, so if H_0 is rejected by N-statistic the true alternative is unknown. N-statistic is included in comparative power analysis for GSA-DV.

Gene-level GSA-DV approach we suggest here tests a null hypothesis that the gene-set-associated score does not differ between phenotypes. The score here is calculated by applying FM to combine P -values from gene-level F -test of the equality of two variances.

Gene-level GSA-DV test that combines genes P-values: To find genes with significant variability we suggest using F -test, similar to what was described for individual genes by Ho and colleagues [3]. Gene-level GSA-DV test is designed by combining P -values of individual F -tests for genes in a pathway. Because for gene level GSA-DE FM was found to be the best performing approach for combining P -values among many others [24, 40] FM is also applied here to combine P -values of F -tests. This method tests the alternative hypothesis that there are genes DV between two phenotypes.

Radial Kolmogorov Smirnov (RKS): The basic operational procedure employed in the univariate Kolmogorov-Smirnov test is to sort pooled observations in ascending order. The difficulty in extending this procedure to multivariate observations is that the notion of a sorted list cannot be immediately generalized [9]. Friedman and Rafsky suggested overcoming this difficulty using the Minimum Spanning Trees (MSTs) [9]. The multivariate generalization of KS ranks multivariate observations based on their MST. The purpose of MST ranking is to obtain the strong relation between observations differences in ranks and their distances in R^p . The ranking algorithm can be designed specifically to confine a particular alternative hypothesis more power. The general scheme is to root MST tree at a node with the largest geodesic distance and then rank the nodes in the “height directed preorder” traversal of the tree. If one is interested in a test with high power toward changes in the variance structure of the distribution, the ranking is implemented differently, aiming to give higher ranks to more distant points in R^p . That is, MST tree is rooted at the node with the smallest geodesic distance (centroid) and nodes with the largest depths are assigned higher ranks [9]. This “radial” Kolmogorov-Smirnov (RKS) test is sensitive to alternatives having similar mean vectors but differences in scale. The test statistic considering N samples under two phenotypes X and Y is the maximum absolute difference

$$D = \left| \frac{s_X^{(i)}}{N_X} - \frac{s_Y^{(i)}}{N_Y} \right|$$

where $s_X^{(i)}$ and $s_Y^{(i)}$ are respectively the number of observations in X and Y ranked lower than i , $1 \leq i \leq N$, N_X and N_Y are respectively the number of samples under phenotypes X and Y .

The null distribution of the test statistic is estimated by a permutation procedure and P -value is defined as

$$P_{\text{value}} = \frac{\sum_{k=1}^{N_{\text{perm}}} I[D_{\text{perm}}(k) \geq D_{\text{obs}}] + 1}{N_{\text{perm}} + 1}$$

where $D_{\text{perm}}(k)$ is the test statistic of permutation k , D_{obs} is the observed test statistic from the original data, N_{perm} is the number of permutations, and I is the indicator function. RKS is implemented in Bioconductor package Gene Set Analysis in R (GSAR) [10, 18].

2.3 Gene Set Analysis Approaches for Testing Differential Co-Expression (GSA-DC)

Null Hypotheses—Each individual GSA-DC approach we consider has its own null hypothesis (see below).

Gene Sets Co-Expression Analysis (GSCA): Briefly, GSCA works as follows [17]. For all $p(p-1)/2$ gene pairs, GSCA calculates intergene correlations under the two biological conditions. The test statistic is the Euclidean distance, adjusted for the size of a gene set,

$$D_{\text{GSCA}} = \sqrt{\frac{1}{p(p-1)/2} \sum_{k=1}^{p(p-1)/2} (\rho_k^{(1)} - \rho_k^{(2)})^2}$$

where k is the index of the gene pair within the gene set and $\rho_k^{(i)}$ denotes the correlation of gene pair k in condition i . GSCA tests the hypothesis $H_0: D_{\text{GSCA}} = 0$ against the alternative $H_1: D_{\text{GSCA}} > 0$.

Gene Sets Net Correlations Analysis (GSNCA): In order to quantitatively characterize the importance of gene i in a correlation network, we introduce a weight (w_i) and set w_i to be proportional to a gene's cross-correlation with all the other genes in the gene set [24]. Then, the objective is to find a weight vector w , which achieves equality between a gene weight and the sum of its weighted cross-correlations for all genes simultaneously. Thus, genes with high cross-correlations will have high weights that may indicate their regulatory importance. This problem can be formulated as a system of linear equations

$$w_i = \sum_{j \neq i} w_j r_{ij}, \quad 1 \leq i \leq p$$

where r_{ij} is the absolute correlation coefficient between genes i and j , and p is the gene set size. Equivalently, this system of linear equations can be represented in the matrix form

$$(R - I)w = w$$

where R is the correlation matrix. This is an eigenvector problem that has a unique solution when the eigenvalue $\lambda_{(R - I)} = 1$, $w > 0$. Because the matrix $(R - I)$ is not guaranteed to have

eigenvalue $\lambda_{(R-I)} = 1$, we introduce a multiplicative factor, γ , which ensures a proper scaling for eigenvalues and solves the following problem:

$$\gamma(R - I)w = w$$

The unique solution w is an eigenvector of matrix $\gamma(R - I)$ corresponding to $\lambda_{\gamma(R-I)} = 1$ [24]. As a test statistic, w_{GSNCA} , we use the L_1 norm between the scaled weight vectors $w^{(1)}$ and $w^{(2)}$ (each vector is multiplied by its norm to scale the weight factor values around one) between two conditions,

$$w_{\text{GSNCA}} = \sum_{i=1}^p |w_i^{(1)} - w_i^{(2)}|$$

This statistic tests the hypothesis $H_0: w_{\text{GSNCA}} = 0$ against the alternative $H_1: w_{\text{GSNCA}} > 0$. P -values for the test statistic are obtained by comparing the observed value of the test statistic to its null distribution, which is estimated using a permutation approach. GSNCA is implemented in Bioconductor package `GSAR` [10, 18].

Co-expression Graph Analyzer (CoGA): Let $G = (V, E)$ be an undirected graph with the adjacency matrix A . The *spectrum* of G is a set of eigenvalues of its adjacency matrix A [20]. The spectrum of a graph describes several of its structural properties, such as diameter, number of walks, and cliques. Takahashi and colleagues [20] suggested that the graph spectrum distribution is a better characterization of graph's properties than conventionally used measures such as number of edges, average path length, and clustering coefficient. Co-expression Graph Analyzer (CoGA) constructs co-expression graphs and identifies differentially co-expressed gene sets by testing the equality of the spectral distributions for two graphs by calculating Jensen-Shannon divergence between spectral densities of two adjacency matrices [19]. Let Θ measure the distance between structural properties of two graphs. CoGa tests $H_0: \Theta = 0$ against $H_1: \Theta > 0$ [19]. CoGA is implemented in Bioconductor package `CoGA` [20].

3 Data Analysis

3.1 Comparative Power Analysis and Type I Error Rate: Simulation Setup

Simulation Setup for GSA-DE—Due to the increasing popularity of RNA-seq data as compared to microarrays the simulation setup here is presented in the context of RNA-seq data. It is conventionally assumed that RNA-seq count data follow Poisson or Negative Binomial (NB) distribution. Here, the count for gene i in sample j is modeled by a random variable C_{ij} with NB distribution

$$C_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{var} = \mu_{ij}(1 + \mu_{ij}\varphi_{ij})) = \text{NB}(\mu_{ij}, \varphi_{ij})$$

where μ_{ij} and φ_{ij} are respectively the mean and dispersion parameters of gene i in sample j . For each gene, a vector of realistic values of mean count, dispersion, and gene length

information (μ_i, φ_i, L_i) is randomly picked from a pool of vectors derived from a real RNA-seq dataset. As a real dataset, we selected a subset of the Pickrell et al. [58] dataset of sequenced cDNA libraries generated from 69 lymphoblastoid cell lines that were derived from Yoruban Nigerian individuals. Samples from 58 unrelated individuals were considered (29 males and 29 females). Dispersion parameters for individual genes were estimated using the Bioconductor edgeR package [43].

Type I error rate: To simulate the null hypothesis $H_0: F_x = F_y$, we generated a dataset consisting of N samples (equally separated between two different phenotypes) and $S = 1000$ nonoverlapping gene sets, each of size p . The randomly selected parameter vector (μ_i, φ_i, L_i) is used to generate counts from the Negative Binomial distribution for gene i under all the samples in the dataset. Gene length information is used for expression normalization if necessary. To examine the effects of different sample and gene set sizes, we estimated Type I error rate under different parameter settings. We chose $p \in \{16, 60, 100\}$ and $N \in \{10, 20, 40, 60\}$. Type I error rate for a statistical test is calculated as the proportion of gene sets detected by the test. The results were averaged over ten independent datasets to obtain more stable estimates.

Detection Power: A differentially expressed (DE) gene set in real data may include up-regulated, down-regulated, and equally regulated genes between two phenotypes. To mimic real data we introduce three simulation parameters: β , the proportion of gene sets in the dataset that have truly DE genes; γ , the proportion of genes, truly DE in each gene set; and FC , the fold change in gene counts between two phenotypes. We consider $\beta \in \{0.05, 0.25\}$, $\gamma \in \{0.125, 0.25, 0.5\}$, and $FC \in [1.2, 3]$. Two different biological conditions are represented by two groups of samples with equal size $N/2$ where $N = 40$. Under each condition, $S = 1000$ nonoverlapping gene sets were formed, each consists of $p = 16$ random realizations from the Negative Binomial distribution. The power for all statistical methods was estimated by testing the hypothesis $H_0: \mu_x = \mu_y$ (or $H_0: FC = 1$) against the alternative $H_1: \mu_x \neq \mu_y$ (or $H_1: FC \neq 1$) for all gene sets. For each of the $(1 - \beta)S$ non-DE gene sets p random realizations of $NB(\mu_i, \varphi_i)$ were sampled, $1 \leq i \leq p$, under both phenotypes. For each of the βS gene sets that have truly DE genes, half of the γp DE genes in each gene set were up-regulated and half were down-regulated between the two phenotypes. Specifically, $\gamma p/2$ random realizations from $NB(\mu_i, \varphi_i)$ and $NB(FC\mu_i, \varphi_i)$ were sampled respectively under phenotype 1 and phenotype 2 for $1 \leq i \leq \gamma p/2$ and another $\gamma p/2$ random realizations from $NB(FC\mu_i, \varphi_i)$ and $NB(\mu_i, \varphi_i)$ were sampled respectively under phenotype 1 and phenotype 2 for $(\gamma p/2) + 1 \leq i \leq \gamma p$.

Simulation Setup for GSA-DV—Typically, RNA-seq counts are modeled using Poisson or Negative Binomial (NB) distribution. Since in the case of Poisson distribution variance is equal to the mean and in the case of NB distribution variance depends on the mean, there is no GSA-DV test for RNA-seq data. Therefore, we present simulation setup assuming multivariate normal distribution of gene expressions that is a standard assumption for microarray data.

Type I error rate: We generated two samples of equal size, $N/2$ from the p -dimensional normal distribution $\mathcal{N}(\mathbf{0}, I_{p \times p})$ where $I_{p \times p}$ is a $p \times p$ identity matrix and p represent the gene set size. 1000 non-overlapping gene sets were generated and Type I error rate for a statistical test is calculated as the proportion of gene sets detected by the test. We consider $p \in \{20, 60, 100\}$ and $N \in \{20, 40, 60\}$.

Detection Power: In a real gene set, the proportion of DV genes, the amount of difference in variance, and the intergene correlation vary. Therefore, three parameters: γ , the proportion of genes truly DV in a gene set, σ , the fold change in variance, and r , the strength of intergene correlation were introduced. We examine how these parameters influence the power of different tests. Two groups of samples of equal size, $N/2$ from p -dimensional normal distributions $\mathcal{N}(\mathbf{0}, \Sigma_x)$ and $\mathcal{N}(\mathbf{0}, \Sigma_y)$ to represent two biological phenotypes were generated. We consider the relationship between the covariance and correlation matrices where the correlation matrix $R = D^{-1} \Sigma D^{-1}$, $D = \sqrt{\text{diag}(\Sigma)}$ and Σ is the covariance matrix.

Let Σ_x and Σ_y be $p \times p$ positive definite and symmetric covariance matrices. The diagonal elements of Σ_x are equal to 1 and off-diagonal elements are equal to r . Matrix Σ_y is defined as

$$\Sigma_y = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A is a $\gamma p \times \gamma p$ matrix with $A_{ij} = \sigma$ for $i = j$ and $A_{ij} = r\sigma$ for $i \neq j$, B and C are respectively $\gamma p \times (1-\gamma)p$ and $(1-\gamma)p \times \gamma p$ matrices where $B_{ij} = C_{ij} = \sqrt{\sigma} r$ for all i and j , and D is a $(1-\gamma)p \times (1-\gamma)p$ matrix with $D_{ij} = 1$ for $i = j$ and $D_{ij} = r$ for $i \neq j$. We consider the parameters $\gamma \in \{0.25, 0.5, 0.75, 1\}$, $r \in \{0.1, 0.5, 0.9\}$, $\sigma \in [1, 5]$, $p = 20$, and $N = 40$.

Simulation Setup for GSA-DC—Since GSA-DC approaches are not yet frequently applied to RNA-seq data here again the simulation setup is presented for microarray data, assuming multivariate normal distribution of gene expressions. Let X and Y be independent p -dimensional vectors with distribution functions $F_x = \mathcal{N}(\mathbf{0}, \Sigma_x)$ and $F_y = \mathcal{N}(\mathbf{0}, \Sigma_y)$.

Type I error rate: To simulate the null hypothesis $H_0: \Sigma_x = \Sigma_y$, we generated two samples of equal size, $N/2$ from the p -dimensional normal distribution $\mathcal{N}(\mathbf{0}, I_{p \times p})$ where $I_{p \times p}$ is a $p \times p$ identity matrix. We generated 1000 gene sets and Type I error rate for a statistical test is the proportion of gene sets detected by the test. We consider $p \in \{20, 100, 200\}$ and $N \in \{20, 40, 60\}$.

Detection Power: In a real biological setting, the proportion of co-expressed genes in a gene set varies and intergene correlations vary in strength. Therefore, two parameters: γ , the proportion of genes truly co-expressed in a gene set, and r , the strength of intergene correlation were introduced. We examine how these parameters influence the power of different tests. We simulated two groups of samples of equal size, $N/2$ ($N = 40$) from p -dimensional normal distributions $\mathcal{N}(\mathbf{0}, \Sigma_x)$ and $\mathcal{N}(\mathbf{0}, \Sigma_y)$ to represent two biological

phenotypes where $p \in \{20, 100, 200\}$. We test the null hypothesis $H_0: \Sigma_X = \Sigma_Y$, against the alternative $\Sigma_X \neq \Sigma_Y$. To ensure that Σ_X and Σ_Y are positive-definite and symmetric, two different scenarios for the alternative hypothesis were studied.

First, Σ_X was set to $I_{p \times p}$ and Σ_Y was set such that its elements are

$$\sigma_{ij} = \begin{cases} r & i \neq j, \forall i, j \leq \gamma p \\ 0 & i \neq j, \forall i, j > \gamma p \\ 1 & i = j. \end{cases}$$

We consider $\gamma \in \{0.25, 0.5, 0.75, 1\}$ and $r \in \{0.1, 0.2, \dots, 0.9\}$. Figure 2 (parts A and B) depicts the covariance matrices Σ_X and Σ_Y under this scenario for $p = 20$ and $\gamma = 0.25$. Dark and light colors represent high and low correlations, respectively. This design presents a gene set with low intergene correlations under one phenotype (Fig. 2A) and one group of highly co-expressed genes under the second phenotype (Fig. 2B).

Second, both Σ_X and Σ_Y are set such that they have diagonal blocks of equal size βp , where β is the ratio of block size to gene set size. For each of the diagonal blocks, the first scenario is reproduced. Therefore, each diagonal block has $\gamma \beta p$ genes with intergene correlation specified by r while all the other genes in the block have zero correlations. The locations of the $\gamma \beta p$ co-expressed genes inside each block are assigned differently for Σ_X and Σ_Y under the alternative hypothesis. While for Σ_X these genes occupy the upper-left corner of the block, for Σ_Y they occupy the lower-right corner. Figure 2 (C, D) depicts this scenario for $p = 20$, $\beta = 0.25$, and $\gamma = 0.6$. Dark and light colors represent high and low correlations, respectively. Depending on γ , the diagonal blocks in Σ_X and Σ_Y may have a few common genes (when $\beta > 0.5$) or may be exclusive (when $\beta \leq 0.5$). We consider the case $\beta = 0.25$ and let $\gamma = 0.6, 0.4$, and 0.5 respectively when $p = 20, 100$, and 200 to allow $\gamma \beta p$ to be an integer number. These settings yield diagonal blocks of 3, 10, and 25 genes respectively when $p = 20, 100$, and 200 . All intergene correlations outside the diagonal blocks are set to zero. This setup presents a gene set with low intergene correlations except for a selected group of highly co-expressed genes and the membership of the genes in this group is changing between the two phenotypes.

3.2 Comparative Power Analysis and Type I Error Rate: Results

Results for GSA-DE

Type I error rate: Table 1 presents the estimates of the attained significant levels for all GSA tests considered ($\alpha = 0.05$). Overall, self-contained and competitive tests control Type I error rate near nominal $\alpha = 0.05$. For more detailed discussion of Type I error rates for self-contained and competitive GSA-DE tests, see [47].

Power: Figure 3 presents the power estimates when $H_1: \mu_X \neq \mu_Y$ is true ($N = 20, p = 16$) (see ref. 47 for more detail). Self-contained methods have higher power than competitive methods and because they test a hypothesis about a single gene set by considering only its gene expressions and ignoring the rest of the dataset, they are not affected by the proportion of gene sets in the dataset that have truly differentially expressed genes (β parameter).

Overall, all self-contained GSA-DE tests (ROAST, N-statistic, SAM-GS, eBayes_FM) have virtually the same power. It should be noted that the simulation setup here does not include intergene correlations. This is why there is no difference in power of multivariate and univariate self-contained approaches. For simulation setup that includes intergene correlations, we refer the reader to [10, 28]. The power of ROMER demonstrates dependence on the proportion of truly DE genes in a gene set (parameter γ). While the power is relatively low at $\gamma = 0.125$, it increases drastically at higher γ values. Competitive methods have slightly lower power for higher values of β especially ROMER. This observation can be explained by the fact that competitive methods are influenced by adding more genes to the dataset where adding non-DE genes enhances their power [36], while adding DE genes may decrease it.

The lack of power under all settings demonstrated by unsupervised competitive methods (especially GSVA) can be explained by the sample-wise ranking they perform to calculate the enrichment scores for gene sets [47]. While half of the genes in the gene set are up-regulated under one phenotype, the other half are up-regulated under the other phenotype. This setup maintains a stable enrichment score for the gene set under all samples and hence the gene set is found non-DE between the two phenotypes. When all DE genes in the gene set are up-regulated under one phenotype only, samples under that phenotype would have had higher gene set enrichment scores compared to the samples under the second phenotype. To substantiate this explanation with simulation results, we consider two hypothetical cases of expression patterns in a gene set consisting of 16 genes. In the first case, all DE genes in a gene set are up-regulated in phenotype 1 compared to phenotype 2. These genes normally have higher ranks in samples under phenotype 1 compared to samples under phenotype 2, and hence the gene set has higher enrichment score under phenotype 1 as compared to phenotype 2. This case is expected to demonstrate high power as shown in Fig. 4. Consider the second case where DE genes in a gene set are equally divided into up-regulated genes between phenotype 1 and phenotype 2, similar to the simulation setup that produced Fig. 3. While the up-regulated genes under phenotype 1 have higher ranks under phenotype 1 as compared to phenotype 2, the up-regulated genes under phenotype 2 are exactly the opposite. This case yields high (however lower than the first case) enrichment score for the gene set under all samples. Due to the expected small difference (if any) in average enrichment score between the two phenotypes, low power is expected (see Fig. 4). Since it is more likely to have both up-regulated and down-regulated genes between two phenotypes in a real gene set than having all up-regulated or down-regulated genes, the power of supervised competitive methods is likely to be consistently lower than other methods for real expression data. It should be noted that the authors of the ssGSEA method expected their enrichment score to be slightly more robust and more sensitive to differences in the tails of the distributions compared to the Kolmogorov–Smirnov-like statistic [39]. The simulation results in Fig. 4 confirm this expectation.

Results for GSA-DV

Type I error rate: Table 2 presents the estimates of the attained significant levels for all GSA-DV approaches considered ($\alpha = 0.05$). Overall, RKS test is slightly more conservative than N-statistic and gene-level GSA-DV test that combines F -tests P -values with FM.

Power: Figure 5 presents the power estimates for the three GSA-DV approaches considered against the alternative hypothesis $\bar{\sigma}_x \neq \bar{\sigma}_y$. It appears that in contrast with GSA-DE approaches, where multivariate tests always outperform univariate tests when correlation increases, multivariate N-statistic and RKS have lower power than gene-level GSA-DV test that combines F -test P -values with FM in all settings. Gene-level GSA-DV test has the highest power, RKS test has an intermediate power, and N-statistic has the lowest power in all settings (Fig. 5). This pattern can be explained by the fact that it is much easier to satisfy the alternative hypothesis tested by the gene-level GSA-DV under our simulation setup than the alternatives tested by both N-statistic and RKS. N-statistic and RKS both test $H_0: F_x = F_y$, with different alternatives $F_x \neq F_y$ and $\bar{\sigma}_x \neq \bar{\sigma}_y$, respectively. Thus, the rejection of H_0 in the case of N-statistic can happen when $\bar{\mu}_x \neq \bar{\mu}_y$, $\bar{\sigma}_x \neq \bar{\sigma}_y$ or other higher order moments of F_x, F_y are not equal. The rejection of H_0 in the case of RKS test is supposedly happened when $\bar{\sigma}_x \neq \bar{\sigma}_y$, but not necessary so because the RKS test is just “more sensitive” to “differences in scale” as compared to “shift differences” [9]. It means that both tests are sensitive to not strictly one alternative, while gene-level GSA-DV test that combines F -test P -values with FM is sensitive to only the case when genes in a gene set are DV genes between two conditions. Figure 6 illustrates this point by showing the estimated power when the alternative hypothesis $\bar{\mu}_x \neq \bar{\mu}_y$ is true. The power trend is just the opposite of the trend presented in Fig. 5. Here, N-statistic has the highest power, RKS test has an intermediate power, and gene-level GSA-DV test has the lowest power in all settings (Fig. 6).

Results for GSA-DC

Type I error rate: Table 3 presents the estimates of the attained significant levels for the three GSA-DC tests considered ($\alpha = 0.05$). Overall, all tests control Type I error rate near nominal $\alpha = 0.05$.

Power: Figure 7 presents power estimates under the first simulation setup (see the simulation setup for GSA-DC) for different parameter settings. For each parameter setting, the results are obtained from 1000 independent gene sets. First, consider the case when only 25% of genes in a gene set are co-expressed ($\gamma = 0.25$). This case is highly plausible in real expression data since only a few genes in a gene set are expected to be highly co-expressed [25, 27]. GSNCA has the highest power followed respectively by GSCA and CoGA for all settings ($p = \{20, 100, 200\}$). Second, consider the case when 50% of genes in a gene set are co-expressed ($\gamma = 0.5$). While all tests show similar power when the size of gene set is relatively small ($p = 20$), GSNCA outperforms both GSCA and CoGA when the size of gene set is relatively large ($p = 100$ and $p = 200$). Third, consider the case when 75% of genes in a gene set are co-expressed ($\gamma = 0.75$). While GSCA and CoGA outperform GSNCA when the size of gene set is relatively small ($p = 20$), all tests have virtually the same power when the number of genes is relatively large ($p = 100$ and $p = 200$). Fourth, consider the case when 100% of genes in a gene set are co-expressed ($\gamma = 1$). GSCA and CoGA have similar power and GSNCA has virtually no power.

The statistic used in GSCA depends on the average pairwise correlation difference between the two phenotypes. Hence, power increases when γ becomes higher as shown in Fig. 7. Similar argument can be applied to CoGA where larger γ causes larger changes in the

spectral distribution of the correlation matrix in one phenotype as compared to the other. When intergene correlation (r) is uniformly low in one phenotype and uniformly high in another phenotype ($\gamma = 1$ and r is high), eigenvectors corresponding to the largest eigenvalues for both correlation matrices remain unchanged while the eigenvalues (spectral distribution) change. Therefore, GSNCA does not detect changes regardless of the value of r when $\gamma = 1$, while CoGA shows high power. This case illustrates the fundamental difference between GSNCA and both GSCA and CoGA. Both GSCA and CoGA detect any differences in pairwise correlations, while GSNCA detects differences in the co-expression structure, i.e., when some pairwise correlations change relative to others in the same phenotypes. The greatest change in the co-expression structure between two phenotypes in the first simulation setup occurs when $\gamma = 0.5$ and hence GSNCA is expected to show highest power as shown in Fig. 7.

Figure 8 presents power estimates under the second simulation setup (see simulation setup for GSA-DC) for different parameter settings. When $p = 20$ and $\gamma = 0.6$ (diagonal block size $\gamma\beta p = 3$), GSCA outperforms GSNCA. When $p = 100$ and $\gamma = 0.4$ (diagonal block size = 10), both GSCA and GSNCA show similar power. When $p = 200$ and $\gamma = 0.5$ (diagonal block size = 25), GSNCA outperforms GSCA. The increment in the size of the diagonal block of differential correlations results in increased detection power when the gene set size increases. When $p = 200$ and $\gamma = 0.5$, power follows similar pattern to what has been shown in Fig. 7 when $\gamma = 0.5$, i.e., GSNCA outperforms GSCA. The difference in power between GSNCA and GSCA when $p = 20$ and $\gamma = 0.6$ follows a similar pattern to what has been observed in Fig. 7 when $\gamma = 0.75$ and could be attributed to the correlation matrix in one phenotype moving closer to a uniformly high correlation pattern. CoGa has almost no power for all settings. This is explained by the fact that unlike eigenvectors the eigenvalues remain unchanged when the number of pairwise intergene correlations with value r remains unchanged but the set of pairwise correlations having value r differs between phenotypes.

3.3 Application to Expression Data

We illustrate the use of GSA-DE, GSA-DV, and GSA-DC tests applied to the NCI-60 cell lines (p53) dataset. The p53 dataset comprises 50 samples of NCI-60 cell lines differentiated based on the status of the TP53 gene: 17 cell lines carrying normal (wild type, WT) TP53 gene and 33 cell lines carrying mutated TP53 (MUT) [38, 59]. For this data set, probe-level intensities were quantile normalized and transformed to the log scale. Gene sets were taken from the C2 pathways set of the molecular signaturedatabase (MSigDB version 5.1) [38, 60, 61]. Pathways with less than 10 or more than 500 genes were discarded and the resulted dataset comprised 4256 gene sets.

Results for GSA-DE—To find pathways, differentially expressed between cancer cell lines with and without p53 mutation we applied SAM-GS. We choose SAM-GS because it tests a fairly simple null hypothesis, namely whether the difference in moderated t -statistics averaged over all pathway genes, is zero between two phenotypes. SAM-GS detected 44 gene sets at the given significance level ($P < 0.001$) (Table 4). All but one detected pathways were significantly enriched with p53 target genes (Table 4). This is not a surprise because if

the expression level of a regulator changes, so do the levels of the regulated genes, leading to significant differences in the average expression of pathways, enriched with p53 targets.

Results for GSA-DV—To find pathways with differential variability between cancer cell lines with and without p53 mutation, the gene-level test combining F -test P -values was applied. It detected only three pathways, between WT and MUT phenotypes at a significance level $P < 0.001$. These pathways are “BANDRES RESPONSE TO CARMUSTIN WITHOUT MGMT 24HR UP,” “BANDRES RESPONSE TO CARMUSTIN MGMT 48HR UP,” and “ONGUSAHA BRCA1 TARGETS DN.” These pathways are not significantly enriched in p53 targets. The first two pathways represent cellular response to carmustine treatment that involves regulation of complex pathways responsible for cell death [62]. All of them employ directly or indirectly expression of p53 gene and expectedly mutation in this gene results in different variability in these pathways. The “ONGUSAHA BRCA1 TARGETS DN” pathway consists of BRCA1 target genes [63] and because the p53 protein regulates BRCA1 transcription, mutation in p53 interferes with gene’ functions, in particular regulation of BRCA1. This may cause indirect mixed effects on regulation of BRCA1 targets.

Results for GSA-DC—To find pathways, differentially co-expressed between cancer cell lines with and without p53 mutation GSNCA test was applied. GSNCA detected only four pathways differentially co-expressed between two phenotypes at a significance level $P < 0.001$. Two of them (“KEGG PEROXISOME,” “REACTOME NOREPINEPHRINE NEUROTRANSMITTER RELEASE CYCLE”) are related to crucial metabolic processes such as fatty acid oxidation, biosynthesis of ether lipids, and free radical detoxification and release of noradrenalin synaptic vesicle. One is related to changes in DNA methylation and histone acetylation (“ZHONG RESPONSE TO AZACITIDINE AND TSA UP”) and one with changes in gene expressions related to intercellular matrix (“PEDERSEN METASTASIS BY ERBB2 ISOFORM 4”). These pathways are also not significantly enriched in p53 target genes.

In addition to detecting DC pathways GSNCA identifies hub genes—genes with the largest weights in each pathway. Hub genes provide useful biological information beyond the test result that a pathway is differentially co-expressed between two conditions. For example, pathway KEGG PEROXISOME (Fig. 9) presents genes that play key roles in redox signaling and lipid homeostasis. For p53 wild-type data, hub gene is MVK (mevalonate kinase Fig. 9A), which encodes the peroxisomal enzyme mevalonate kinase, a key early enzyme in isoprenoid and sterol synthesis. When p53 is mutated (Fig. 9B), hub gene becomes ACOX1 (acyl-coenzyme A oxidase 1, palmitoyl) that is the first enzyme of the fatty acid beta-oxidation pathway which catalyzes the desaturation of acyl-CoAs to 2-transenoyl-CoAs. That is in p53 MUT phenotype a shift from isoprenoid and sterol synthesis to fatty acid beta-oxidation pathway may happen. For more discussion of hub genes the reader is referred to [18].

4 Conclusion

In this chapter, we provided an in-depth review of univariate and multivariate Gene Set Analysis approaches (GSA-DE, GSA-DV, GSA-DC) for testing different statistical hypotheses. A comparative power analysis and Type I error rate estimates for different approaches in each major type of GSA provide concise guidelines for selecting GSA approaches that are best performing under particular experimental settings. An example was presented applying the methods GSA-DE, GSA-DV, GSA-DC on a p53 data set. This analysis demonstrated that different GSA types are allowing to obtain new and complementary biological information for the same underlying data set.

Acknowledgments

We would like to thank **Bárbara Macías Solís** for proof reading of the manuscript. Support has been provided in part by the Arkansas INBRE program, with grants from the National Center for Research Resources (P20RR016460) and the National Institute of General Medical Sciences (P20 GM103429) from the National Institutes of Health. Large-scale computer simulations were implemented using the High Performance Computing (HPC) resources at the UALR Computational Research Center supported by the following grants: National Science Foundation grants CRI CNS-0855248, EPS-0701890, MRI CNS-0619069 and OISE-0729792.

References

1. Mootha VK, et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 2003; 34(3):267–273. [PubMed: 12808457]
2. Bar HY, Booth JG, Wells MT. A mixture-model approach for parallel testing for unequal variances. *Stat Appl Genet Mol Biol.* 2012; 11(1) p. Article 8.
3. Ho JW, et al. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics.* 2008; 24(13):i390–i398. [PubMed: 18586739]
4. Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. *Genetics.* 2013; 193(1):95–108. [PubMed: 23150607]
5. Mar JC, et al. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet.* 2011; 7(8):e1002207. [PubMed: 21852951]
6. Xu Z, et al. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol.* 2011; 7:468. [PubMed: 21326235]
7. Bravo HC, et al. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinform.* 2012; 13:272.
8. Dinalankara W, Bravo HC. Gene expression signatures based on variability can robustly predict tumor progression and prognosis. *Cancer Informat.* 2015; 14:71–81.
9. Friedman JH, Rafsky LC. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann Stat.* 1979; 7(4):697–717.
10. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis for self-contained tests: complex null and specific alternative hypotheses. *Bioinformatics.* 2012; 28(23):3073–3080. [PubMed: 23044539]
11. Afsari B, Geman D, Fertig EJ. Learning dysregulated pathways in cancers from differential variability analysis. *Cancer Informat.* 2014; 13(Suppl 5):61–67.
12. Fisher, R. Statistical methods for research workers. Oliver and Boyd; Edinburg: 1932.
13. Stadler N, Mukherjee S. Multivariate gene-set testing based on graphical models. *Biostatistics.* 2015; 16(1):47–59. [PubMed: 24974316]
14. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics.* 2008; 9(3):432–441. [PubMed: 18079126]

15. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat.* 2006; 34(3):1436–1462.
16. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol.* 2005; 4(1) Article 32.
17. Choi Y, Kendziora C. Statistical methods for gene set co-expression analysis. *Bioinformatics.* 2009; 25(21):2780–2786. [PubMed: 19689953]
18. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics.* 2014; 30(3):360–368. [PubMed: 24292935]
19. Santos Sde S, et al. CoGA: an R package to identify differentially co-expressed gene sets by analyzing the graph spectra. *PLoS One.* 2015; 10(8):e0135831. [PubMed: 26313749]
20. Takahashi DY, et al. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS One.* 2012; 7(12):e49949. [PubMed: 23284629]
21. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics.* 2007; 23(8):980–987. [PubMed: 17303618]
22. Tian L, et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A.* 2005; 102(38):13544–13549. [PubMed: 16174746]
23. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinform.* 2009; 10(1):47.
24. Rahmatallah Y, Emmert-Streib F, Glazko G. Comparative evaluation of gene set analysis approaches for RNA-Seq data. *BMC Bioinform.* 2014; 15(1):397.
25. Montaner D, et al. Gene set internal coherence in the context of functional profiling. *BMC Genomics.* 2009; 10:197. [PubMed: 19397819]
26. Gatti DM, et al. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics.* 2010; 11:574. [PubMed: 20955544]
27. Tripathi S, Emmert-Streib F. Assessment method for a power analysis to identify differentially expressed pathways. *PLoS One.* 2012; 7(5):e37510. [PubMed: 22629411]
28. Glazko GV, Emmert-Streib F. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics.* 2009; 25(18):2348–2354. [PubMed: 19574285]
29. Wang X, et al. Linear combination test for hierarchical gene set analysis. *Stat Appl Genet Mol Biol.* 2011; 10(1) Article 13.
30. Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* 2013; 14:7.
31. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012; 8(2):e1002375. [PubMed: 22383865]
32. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Brief Bioinform.* 2014; 15(4):504–518. [PubMed: 23413432]
33. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform.* 2008; 9(3): 189–197. [PubMed: 18202032]
34. Tamayo P, et al. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res.* 2012; 25(1):472–487. [PubMed: 23070592]
35. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One.* 2013; 8(11):e79217. [PubMed: 24260172]
36. Tripathi S, Glazko GV, Emmert-Streib F. Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. *Nucleic Acids Res.* 2013; 41(7):e82. [PubMed: 23389952]
37. Dinu I, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinform.* 2007; 8:242.
38. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005; 102(43):15545–15550. [PubMed: 16199517]

39. Barbie DA, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009; 462(7269):108–112. [PubMed: 19847166]
40. Fridley BL, Jenkins GD, Biernacka JM. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*. 2010; 5(9)
41. Stouffer, S., DeVinney, L., Suchmen, E. *The American soldier: adjustment during army life*. Vol. 1. Princeton University Press; Princeton, NJ: 1949.
42. Taylor J, Tibshirani R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*. 2006; 7(2):167–181. [PubMed: 16332926]
43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26(1):139–140. [PubMed: 19910308]
44. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010; 11(10):R106. [PubMed: 20979621]
45. Smyth, G. Limma: linear models for microarray data. In: Smyth, G, Gentleman, R, Carey, V, Dudoit, S, Irizarry, R., Huber, W., editors. *Bioinformatics and computational biology solutions using R and bioconductor*. Springer; New York: 2005. p. 397–420.
46. Law CW, et al. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014; 15(2):R29. [PubMed: 24485249]
47. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief Bioinform*. 2016; 17(3):393–407. [PubMed: 26342128]
48. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001; 98(9):5116–5121. [PubMed: 11309499]
49. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*. 2001; 17(6):509–519. [PubMed: 11395427]
50. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3:3.
51. Dinu I, et al. Gene-set analysis and reduction. *Brief Bioinform*. 2009; 10(1):24–34. [PubMed: 18836208]
52. Liu Q, et al. Comparative evaluation of gene-set analysis methods. *BMC Bioinform*. 2007; 8:431.
53. Baringhaus L, Franz C. On a new multivariate two-sample test. *J Multivar Anal*. 2004; 88:190–206.
54. Klebanov L, et al. A multivariate extension of the gene set enrichment analysis. *J Bioinforma Comput Biol*. 2007; 5(5):1139–1153.
55. Wu D, et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*. 2010; 26(17):2176–2182. [PubMed: 20610611]
56. Damian D, Gorfine M. Statistical concerns about the GSEA procedure. *Nat Genet*. 2004; 36(7):663. author reply 663. [PubMed: 15226741]
57. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; 43(7):e47. [PubMed: 25605792]
58. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010; 464(7289):768–772. [PubMed: 20220758]
59. Olivier M, et al. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat*. 2002; 19(6):607–614. [PubMed: 12007217]
60. Liberzon A, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011; 27(12):1739–1740. [PubMed: 21546393]
61. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res*. 2012; 40(17):e133. [PubMed: 22638577]
62. Bandres E, et al. Gene expression profile induced by BCNU in human glioma cell lines with differential MGMT expression. *J Neuro-Oncol*. 2005; 73(3):189–198.
63. Ongusaha PP, et al. BRCA1 shifts p53-mediated cellular outcomes towards irreversible growth arrest. *Oncogene*. 2003; 22(24):3749–3758. [PubMed: 12802282]

Fig. 1. Schematic overview illustrating the breakup of the GSA-DE methods into different categories based on the null hypotheses they test

Fig. 2.

The correlation matrices of the two simulation setups with sample size $N=40$ and gene set size $p=20$. Parts (A) and (B) respectively represent the correlation matrices of two conditions when the alternative hypothesis of setup 1 is true and $\gamma=0.25$. Parts (C) and (D) respectively represent the correlation matrices of two conditions when the alternative hypothesis of setup 2 is true, $\beta=0.25$ and $\gamma=0.6$. *Dark* and *light colors* respectively represent high and low correlation coefficients

Fig. 3.

The power of different DE tests to detect differential expression between two phenotypes of samples when the alternative hypothesis $\bar{\mu}_x - \bar{\mu}_y$ is true with different settings (values of β , γ and FC). The gene set size is $p = 16$ and the sample size in each group is $N/2$ ($N = 20$). Half of the $\gamma \times p$ DE genes in a gene set are up-regulated under one phenotype and the other half are up-regulated under the other phenotype

Fig. 4.

The power of unsupervised competitive tests (GSVA and ssGSEA) to detect differences between two phenotypes when the alternative hypothesis $\bar{\mu}_x - \bar{\mu}_y$ is true with different settings (values of β , γ , and FC). The gene set size $p = 16$ and the sample size in each group is $N/2$ ($N = 20$). In case 1, all the γp DE genes in a gene set are up-regulated in phenotype 1 as compared to phenotype 2. In case 2, half of the γp DE genes in a gene set are up-regulated in phenotype 1 and the other half are up-regulated in phenotype 2. Both GSVA and ssGSEA have much higher power under case 1

Fig. 5. The power of three GSA-DV tests to detect differential expression between two phenotypes of samples when the alternative hypothesis $\sigma_x \neq \sigma_y$ is true with different settings (values of β , γ , and σ). The gene set size $p = 20$ and the sample size in each group is $N/2$ ($N = 20$)

Fig. 6.

The power of three GSA-DV tests to detect differential expression between two phenotypes of samples when the alternative hypothesis $\bar{\mu}_x \neq \bar{\mu}_y$ is true with different settings (values of β , γ , and σ). The gene set size $p = 20$ and the sample size in each group is $N/2$ ($N = 20$)

Fig. 7.

The power of three GSA-DC tests to detect differential expression between two phenotypes of samples when the alternative hypothesis of the first simulation setup is true with different settings (values of γ and r). The gene set size $p = \{20, 100, 200\}$ and the sample size in each group is $N/2$ ($N = 40$)

Fig. 8.

The power of three GSA-DC tests to detect differential expression between two phenotypes of samples when the alternative hypothesis of the second simulation setup is true with different settings (values of β , γ , and r). The gene set size $p = \{20, 100, 200\}$ and the sample size in each group is $N/2$ ($N = 40$)

Fig. 9. MST2s of pathway “KEGG PEROXISOME” co-expression network under both (A) wild-type (WT) and (B) mutated (MUT) p53 phenotypes

Table 1

Estimated Type I error rates for GSA-DE methods, $\alpha = 0.05$

Method placement	Self-contained		N-statistic		SAM-GS		ROAST			
	Competitive		GSVA		ssGSEA		ROMER			
	Combined <i>P</i> -value		eBayes_FM		-		-			
	<i>P</i> = 16									
<i>P</i> = 60										
<i>P</i> = 100										
<i>N</i> = 10	Self.	0.049	0.044	0.084	0.048	0.045	0.042	0.048	0.045	0.041
	Comp.	0.025	0.042	0.047	0.017	0.047	0.050	0.013	0.045	0.047
	Comb.	0.047	-	-	0.042	-	-	0.044	-	-
<i>N</i> = 20	Self.	0.052	0.046	0.044	0.055	0.050	0.047	0.051	0.055	0.050
	Comp.	0.040	0.047	0.051	0.038	0.041	0.054	0.037	0.050	0.053
	Comb.	0.048	-	-	0.051	-	-	0.054	-	-
<i>N</i> = 40	Self.	0.054	0.054	0.051	0.047	0.047	0.044	0.050	0.053	0.055
	Comp.	0.051	0.044	0.050	0.057	0.048	0.045	0.060	0.049	0.052
	Comb.	0.051	-	-	0.047	-	-	0.055	-	-
<i>N</i> = 60	Self.	0.051	0.051	0.052	0.046	0.047	0.048	0.049	0.054	0.054
	Comp.	0.060	0.046	0.051	0.061	0.051	0.049	0.066	0.047	0.050
	Comb.	0.052	-	-	0.046	-	-	0.055	-	-

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Estimated Type I error rates for GSA-DV methods, $\alpha = 0.05$

Method	N-statistic			F-test			RKS		
p	20	60	100	20	60	100	20	60	100
$N = 20$	0.050	0.047	0.050	0.044	0.040	0.043	0.036	0.025	0.049
$N = 40$	0.052	0.060	0.045	0.057	0.040	0.052	0.047	0.038	0.039
$N = 60$	0.053	0.035	0.049	0.053	0.054	0.056	0.041	0.038	0.034

Table 3

Estimated Type I error rates for GSA-DC methods, $\alpha = 0.05$

Method	GSCA			GSNCA			CoGA		
p	20	100	200	20	100	200	20	100	200
$N = 20$	0.057	0.050	0.044	0.052	0.042	0.045	0.053	0.048	0.056
$N = 40$	0.046	0.045	0.059	0.036	0.051	0.051	0.043	0.052	0.050
$N = 60$	0.052	0.049	0.047	0.054	0.047	0.054	0.043	0.048	0.050

Table 4

Pathways differentially expressed between p53^{WT} and p53^{MUT} cell lines^{a)}

Gene set name	Size	TP53 targets	P _{hypergeo}	Target genes
1 KEGG_P53_SIGNALING_PATHWAY	53	34	1.1E-26	SEN_TSC2_CDK1_RCHY1_IGFBP3_SERPINB5_PPM1D_BID_MDM4_BAX_MDM2_TP53B_PMAIP1_ATM_ATR_CHEK2_APAF1_CDKN2A_RRM2_CDKN1A_BBC3_GADD45A_TNFRSF10B_DDB2_SIAH1_PTEN_CDK2_CHEK1_SERPINE1_TP53_CCNE2_CCNB1_CCNG1_CCNG2
2 KEGG_AMYTROPHIC_LATERAL_SCLEROSIS_ALS	48	12	9.3E-05	MAPK11_BID_MAPK13_BAX_DAXX_MAPK14_MAPK12_APAF1_GPX1_BCL2_BCL2L1_TP53
3 BIOARTA_CHEMICAL_PATHWAY	22	12	4.7E-09	PTK2_BCL2_BID_BCL2L1_STAT1_PRRCA_APAF1_BAX_CASP6_TP53_ATM_PARP1
4 BIOARTA_ATM_PATHWAY	19	15	1.4E-14	JUN_CHEK2_MAPK8_MRE11A_BRCA1_REL_A_CHEK1_MDM2_ABL1_CDKN1A_TP53_GADD45A_ATM_RAD51_NFKBIA
5 BIOARTA_CERAMIDE_PATHWAY	22	6	3.4E-03	BCL2_MAPK8_RELA_BAX_MAPK3_MAPK1
6 BIOARTA_HIVNEF_PATHWAY	56	17	1.6E-07	CHUK_CFLAR_PRRCD_PTK2_BID_MDM2_CASP6_RBI_DAXX_PRRDC_NFKBIA_MAPK8_RELA_APAF1_TRAF1_BCL2_PARP1
7 BIOARTA_P53HYPOXIA_PATHWAY	21	17	1.0E-16	MAPK8_CSNK1D_CSNK1A1_EP300_BAX_HIF1A_HIC1_MDM2_TAF1_CDKN1A_TP53_NQO1_GADD45A_HSP90AA1_IGFBP3_ATM_RPA1
8 BIOARTA_IL22BP_PATHWAY	13	2	2.3E-01	STAT1_STAT6
9 BIOARTA_P53_PATHWAY	16	12	2.0E-11	BCL2_E2F1_APAF1_BAX_CDK2_MDM2_RBI_CDKN1A_TP53_GADD45A_ATM_PCNA
10 BIOARTA_BAD_PATHWAY	26	6	8.3E-03	IGF1R_BCL2_BCL2L1_BAX_MAPK3_MAPK1
11 SA_G1_AND_S_PHASES	13	6	1.3E-04	E2F1_CDK2_MDM2_TP53_CDKN2A_CDKN1A
12 SA_PROGRAMMED_CELL_DEATH	10	6	2.0E-05	BCL2_BAX_BCL2L1_BAK1_BID_APAF1
13 PID_P73PATHWAY	69	41	5.0E-30	TP53I3_PML_GRAMD4_WT1_EP300_GDF15_RCHY1_BAX_MDM2_PLK3_CCNB1_S100A2_AHP_MAPK11_KAT5_SERPINE1_WWOX_BRCA2_CCNA2_PIN1_MYC_CDK1_MAPK14_BBC3_ABL1_CDK2_PLK1_SFN_RELA_ITCH_SPL_RBI_RAD51_TP63_CHEK1_BAK1_UBE4B_CCNE2_BUB1_FOXO3_CDKN1A
14 PID_HDAC_CLASSIII_PATHWAY	18	11	4.1E-09	FOXO1_CREBBP_HDAC4_BAX_TP53_XRCC6_FOXO3_EP300_KAT2B_CDKN1A_TUBB2A
15 PID_REG_OR_PATHWAY	78	36	1.9E-21	BAX_NR4A1_MAPK8_HDAC1_STAT1_HDAC2_EGR1_NCOA2_GSK3B_SMARCD1_RELA_SUMO2_TP53_SMARCA4_TBP_SFN_MDM2_MAPK3_CREBBP_EP300_FOS_MAPK9_MAPK10_NR3C1_HSP90AA1_TSG101_MAPK1_MAPK14_MAPK11_NCOA1_SMARCC1_JUN_AFP_CREB1_CDKN1A_CDK5

	Gene set name	Size	TP53 targets	P_{hypergeo}	Target genes
16	PID_P53_DOWNSTREAM_PATHWAY	110	97	1.6E-99	DDIT4_FDXR_SERPINB5_IGFBP3_GPX1_ATE3_MAP4K4_BNIP3L_TSC2_BCL2_TP63_TP53_MMP2_TNFRSF10D_SPI_BBC3_PRKAB1_TYRPI_CEBPZ_NFYB_MLH1_PCNA_SMARCA4_BAK1_JUN_CARM1_VCAN_TAF9_APP_CSE1L_IRF5_PRDM1_NFYA_CCNG1_MDM2_MET_PTEN_TFDP1_CAV1_CCNB1_CX3CL1_ARID3A_PML_DDX5_BDKRB2_TP53I3_HIC1_GADD45A_TGFA_APC_NFYC_SERPINE1_PRMT1_BTG2_SH2D1A_BAX_TRIAP1_RB1_VDR_KAT2A_TRRAP_TNFRSF10B_EP300_HTT_NDRG1_MSH2_PPIR13B_DDB2_CASPI0_GDF15_LIF_CASP6_EGFR_PLK3_SNAI2_DKK1_CTSD_EPHA2_COL18A1_RCHY1_PCBP4_SF_N_BCL2L2_E2F1_BCL6_BID_ST00A2_BCL2L1_DUSP5_CDKN1A_APAF1_CREBBP_TP53BP2_HDAC2_MCL1_EDN2_PMAIP1_NR4A1_VDR_NCOA1_THRB_RARA_TGFB1_SREBF1_BCL2_MEDI
17	PID_RXR_VDR_PATHWAY	25	9	3.0E-05	NR4A1_VDR_NCOA1_THRB_RARA_TGFB1_SREBF1_BCL2_MEDI
18	PID_TAP63_PATHWAY	48	29	7.6E-22	EP300_NOO1_SERPINB5_YWHAQ_S100A2_CDKN1A_PML_BBC3_GADD45A_CHUK_TP63_SPI_MDM2_NOC2L_PMAIP1_ITCH_PRKCD_IGFBP3_GPX2_TFAP2C_TP53I3_CDKN2A_VDR_PLK1_BAX_IKKBK_FDXR_GDF15_ABL1_NEDD8_PPM1D_HIPK2_CHEK2_TP53_TRIM28_CCNG1_HUWE1_CSNK1E_SKP2_ATM_CSE1L_CSNK1D_CSNK1A1_CDK2_PIN1_CHEK1_MDM4_KAT5_CCNA2_SMYD2_EP300_KAT2B_KAT8_DAXX_DYRK2_RPL11_PRKCD_MDM2_CDKN2A_ATR_ABL1_CSNK1G2_GSK3B_CREBBP_MAPK14_MAPK9_RPL23_USP7_RCHY1_UBE2D1_YY1_MAPK8
19	PID_P53_REGULATION_PATHWAY	46	43	1.0E-46	EP300_NOO1_SERPINB5_YWHAQ_S100A2_CDKN1A_PML_BBC3_GADD45A_CHUK_TP63_SPI_MDM2_NOC2L_PMAIP1_ITCH_PRKCD_IGFBP3_GPX2_TFAP2C_TP53I3_CDKN2A_VDR_PLK1_BAX_IKKBK_FDXR_GDF15_ABL1_NEDD8_PPM1D_HIPK2_CHEK2_TP53_TRIM28_CCNG1_HUWE1_CSNK1E_SKP2_ATM_CSE1L_CSNK1D_CSNK1A1_CDK2_PIN1_CHEK1_MDM4_KAT5_CCNA2_SMYD2_EP300_KAT2B_KAT8_DAXX_DYRK2_RPL11_PRKCD_MDM2_CDKN2A_ATR_ABL1_CSNK1G2_GSK3B_CREBBP_MAPK14_MAPK9_RPL23_USP7_RCHY1_UBE2D1_YY1_MAPK8
20	PID_RB_1PATHWAY	58	28	1.3E-17	SMARCB1_CDKN1A_JUN_CREBBP_TBP_SMARCA4_MAPK14_TFDP1_DNMT1_CTBP1_PAX3_MET_MAPK11_SKP2_RBBP4_ABL1_EP300_HDAC3_CDKN2A_CCNA2_E2F1_CDK2_HDAC1_TAF1_MAPK9_MDM2_RB1_CEBPB
21	REACTOME_SIGNALING_BY_ERBB2	85	19	5.4E-06	CDKN1A_CHUK_CREB1_EGFR_FOXO1_FOXO3_MTOR_GRB2_NR4A1_HSP90AA1_MDM2_PRKCA_PRKCD_MAPK1_MAPK3_PTEN_RPS27A_TSC2_CDK1
22	REACTOME_P13K_EVENTS_IN_ERBB2_SIGNALING	36	12	3.6E-06	CDKN1A_CHUK_CREB1_EGFR_FOXO1_FOXO3_MTOR_GRB2_NR4A1_MDM2_PTEN_TSC2
23	REACTOME_P13K_AKT_ACTIVATION	31	10	3.3E-05	CDKN1A_CHUK_CREB1_FOXO1_FOXO3_MTOR_NR4A1_MDM2_PTEN_TSC2
24	REACTOME_AKT_PHOSPHORYLATES_TARGETS_IN_THE_CYTOSOL	11	4	5.5E-03	CDKN1A_CHUK_MDM2_TSC2
25	REACTOME_GAB1_SIGNALOSOME	30	12	3.7E-07	CDKN1A_CHUK_CREB1_EGFR_FOXO1_FOXO3_MTOR_GRB2_NR4A1_MDM2_PTEN_TSC2
26	REACTOME_DOWNSTREAM_SIGNAL_TRANSDUCTION	81	18	1.0E-05	CDKN1A_CHUK_CREB1_FOXO1_FOXO3_MTOR_GRB2_NR4A1_MDM2_PTEN_TSC2
27	REACTOME_PIP3_ACTIVATES_AKT_SIGNALING	22	10	8.7E-07	CDKN1A_CHUK_CREB1_FOXO1_FOXO3_MTOR_NR4A1_MDM2_PTEN_TSC2
28	REACTOME_INTRINSIC_PATHWAY_FOR_APOPTOSIS	26	13	4.3E-09	E2F1_BBC3_APAF1_NMT1_PMAIP1_MAPK8_BAK1_BAX_BCL2_BCL2L1_BID_TFDP1_TP53
29	GAZDA_DIAMOND_BLACKfan_ANEMIA_MYELOID_UP	26	4	1.0E-01	BAX_TNFRSF10B_DDB2_SRSF1
30	GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_BLACK_UP	20	3	1.6E-01	PLAGL1_CDKN1A_UBB
31	NUNODA_RESPONSE_TO_DASATINIB_IMATINIB_DN	13	6	1.3E-04	CDKN1A_IKKBK_CASP10_STAT6_STAT1_BAX
32	GALLUZZI_PERMEABILIZE_MITOCHONDRIA	38	19	1.0E-12	NR4A1_BBC3_PRKCD_SERPINB5_FDXR_BAK1_SIVA1_BCL2L1_PPID_BAX_BID_STK11_MAPK8_ABL1_TP53_MCL1_BCL2_PMAIP1_GSK3B
33	DUTTA_APOPTOSIS_VIA_NFKB	30	9	1.5E-04	BAX_TP53_MDM2_CFLAR_BCL2L1_AFP_TNFRSF10B_TRAFI1_BCL2
34	GALLUZZI_PREVENT_MITOCHONDIAL_PERMEABILIZATION	20	9	3.4E-06	BCL2_MAPK14_MCL1_BCL2L1_BAK1_MUC1_TN_X_BAX_BCL2L2
35	SCHAVOLT_TARGETS_OF_TP53_AND_TP63	14	11	6.1E-11	CDKN1A_SERPINB5_BAX_TP53I3_SF_N_PMAIP1_EPHA2_VCAN_FDXR_MDM2_PCNA
36	AMUNDSON_DNA_DAMAGE_RESPONSE_TP53	15	8	2.4E-06	LIF_DDB2_MDM2_CDKN1A_CCNG1_CTSD_BTG2_PPM1D

Gene set name	Size	TP53 targets	P_{hypergeo}	Target genes
37 FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_OK_VS_DONOR_DN	25	5	2.8E-02	BAX_FOS_XRCC6_MYC_EIF2AK2
38 MA_MYELOID_DIFFERENTIATION_UP	35	9	5.6E-04	BNIP3L_PPM1G_CDKN1A_UBB_MT1A_RB1_CCT5_MDM2_NCL
39 GENTILE_UV_LOW_DOSE_UP	24	8	1.6E-04	CDKN1A_SOX4_BTG2_FDXR_SAT1_GDF15_PMAIP1_CSNNK1G2
40 INGA_TP53_TARGETS	15	12	5.3E-12	MDM2_SEN_BAX_TNFRSF10B_FOS_PCNA_CCNG1_PMAIP1_BBC3_IGFBP3_CDKN1A_GADD45A
41 DELLA_RESPONSE_TO_TSA_AND_BUTYRATE	19	5	8.8E-03	HSPB1_CDKN1A_PRKCD_NR4A1_GADD45A
42 ONO_FOXP3_TARGETS_DN	34	5	8.8E-02	PARP1_CCNA2_CDKN1A_E2F1_BCL2
43 WARTERS_RESPONSE_TO_IR_SKIN	37	13	7.2E-07	MDM2_DDB2_FDXR_TRIAP1_PPM1D_TP53TG1_BTG2_GDF15_BBC3_CDKN1A_CCNG1_STR13_SERPINB5
44 WARTERS_IR_RESPONSE_SGY	21	12	2.3E-09	MDM2_CDKN1A_GDF15_FDXR_BBC3_DDB2_PPM1D_PLK3_ATF3_BTG2_PCNA_GADD45A

^{a)} P -values were calculated using hypergeometric test. The names of p53 target genes for a pathway are listed