University of Cincinnati

Date: 7/1/2020

I, Jacek Biesiada, hereby submit this original work as part of the requirements for the degree of Master of Science in Biostatistics (Environmental Health).

It is entitled:

Shiny Application for Enrichment and Topological Pathway Analysis

Student's name: **Jacek Biesiada**

This work and its defense approved by:

Committee chair: Mario Medvedovic, Ph.D.

Committee member: Jaroslaw Meller, Ph.D.

UNIVERSITY OF Cincinnati

36808

Shiny application for enrichment and topological pathway analysis

A thesis submitted to the
Graduate School
of the University of Cincinnati
in partial fulfillment of the
requirements for the degree of

Master of Science

in the Department of Environmental
and Public Health Sciences
of the College of Medicine
by

Jacek Biesiada

Ph.D. University of Silesia
October 2000

Committee Chair: M. Medvedovic, Ph.D.

EApp (Enrichment Application) Shiny application was developed to facilitate gene enrichment and topological pathway analysis of omics data by scientists without coding skills to support basic biological questions about biological processes and at the same time give people without coding skills the ability to perform those analyses. The application accommodates various external sources of data such as eset with expression and genes signature data, iLINCS project data, GRAIN project data, and raw data (separate data frames with rpkms, count data, and phenotypic data necessary for creating gene signatures). Shiny application provides the user with a graphical user interface that selects appropriate methods based on the source of data and calculated benchmarks. The case study included in thesis were based on TCGA data, and the benchmarks performed in the thesis provide guidelines for using set-based and network-based enrichment methods implemented in the web server.

Contents

Table of Figures

Acknowledgment

I would like to express my warmest thanks to my advisor, Dr. Mario Medvedovic, for his mentorship, encouragement, and continuous support of my study.

Besides my advisor, I would like to thank you to my thesis committee, Dr. Mario Medvedovic and Jarek Meller, for their insightful comments, and fruitful discussion.

Also I thank my friend Robert Tamer for fruitful discussion and support during writing of this thesis.

Introduction

Identification of enriched/active pathways from gene expression data is an essential problem, which, when solved, allows one to gain underlying biological function of cellular and other systems, based on the observed patterns of differential gene expression [1]. This thesis describes a shiny web-based application that combines and integrates several existing methods for enrichment analysis. A case study and several benchmarks are used to compare the results of these different methods and provide guidance for users. This application will allow a broader group of users to perform different types of gene set enrichment analyses and create publication-ready visualizations of the results. Clustering of enriched pathways and differentially expressed genes is one such visualization that addresses basic questions such as: i) what biological processes/pathways a differentially expressed gene is involved in? ii) what evidence supports the enriched pathways? iii) how known gene-gene interactions and transcriptional regulatory modules can explain the observed differential expression patterns ? [1]. The user will be able to input data from various external sources: local ExpressionSets with expression and genes signature data, data from iLINCS project [2], data from GREIN project [3], and finally raw data (separate data frames with rpkms, count data, and phenotypic data necessary for creating gene signatures). This chapter introduces some general aspects of gene set analysis and describes established classes of methods used for enrichment analysis.

After almost two decades of research, numerous methods for pathway analysis have been developed, which could be divided into three groups [1, 2] referred to by Kathri, first-, second-, and third-generation (Figure 1)[1]. Huang used similar classification, but categories were named: singular enrichment analysis (SEA); gene set enrichment analysis (GSEA); and modular enrichment analysis (MEA)[3]. Another simpler classification comes from Geistlinger [4], where methods are divided into two categories: *set-based* and *network-based* methods. *Set-based* methods are the

[1] Figure is taken from:

1. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS computational biology, 2012. **8**: p. e1002375-e1002375.

methods, which ignore interaction between genes and *network-based* methods incorporate known interactions between genes. Set-based category is equivalent to first- (over-representation analysis method), second- (Function Class Scoring methods) generation method from Kathri classification. *Network-based* methods are equivalent to third-generation called Pathway Topology-Based, and this name will be used in this thesis interchangeably. In this thesis we will follow Geistlinger categorization for description of methods, because it naturally represents implementation of Enrichment Browser [4], which was used as a back-end for the shiny front–end implementation.

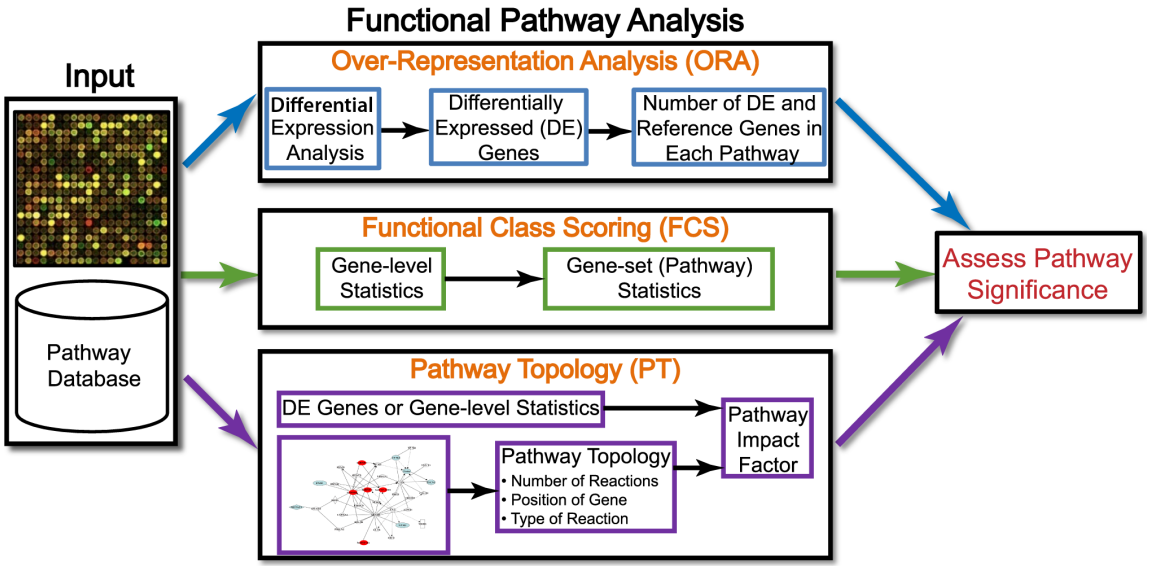

Figure 1. Overview of existing pathway analysis methods using gene expression data as an example.

The first generation methods, also called over-representation analysis (ORA) methods, statistically evaluate the set of differentially expressed genes for their over-representation among pathway genes. This method usually consists of three steps. The first step creates an input list of genes based on specific significance cutoff from expression experiment. Then the number of overlapping genes between the input gene list and the set of the pathway genes is counted for every pathway in the database is used to evaluate pathways for under- or over-representation by a statistical test. The ORA [5] methods disregard the effect size and significance of observed (p-value or fold-changes in expressions level), which is recognized as the main

disadvantage. A second limitation is the arbitrary nature of the p-value or fold-change value cut-offs that define the input gene list. A third limitation of ORA methods comes from the assumption that expression changes in genes within the same pathway are independent which can lead to false positive results. The fourth limitation is related to the assumption of independence between pathways, for example GO defines biological process, as assemblies of molecular functions where information shared between pathways could be redundant, which are not independent either.

The second generation, refered to as Function Class Scoring (FCS) methods, address the three limitations of ORA algorithms. The Function Class Scoring (FCS) methods, similarly to ORA methods, use molecular measurements such as gene expression to identify significant pathways, but at the same time try to use changes in individual genes to evaluate effects on pathways. The FCS method calculates gene-level statistics, followed by calculation of pathway level statistics, and in the final step assesses pathway significance. The gene-level statistics express correlation of molecular measurements with phenotype by ANOVA, Q-statistics, signal-to-noise ratio, t-test, and Z-score. The pathway level statistics aggregate gene-level statistics to single pathway score using Kolmogorov-Smirnov statistic, sum, mean, or median of gene-level statistic, the Wilcoxon rank sum, and the max-mean statistic. The FCS methods do not require an arbitrary threshold for dividing expression data into significant and non-significant subsets of genes. The main limitation of those methods similar to the ORA algorithm is an assumption about the independence of genes and pathways in a database. The second limitation comes from the fact that the most algorithms look at relative changes and do not emphasize the genes with the most significant changes.

The last group of methods called Pathway Topology-Based (PTB) methods incorporates information on how genes interact (e.g., activation, inhibition, acetylation, regulation, etc.) within the pathway during pathway level statistics calculation. Knowledge about gene interaction is taken from resources such as KEGG [6-9], MetaCyc [10-12], Reactome [13], RegulonDB [14], PantherDB [15], STRING [16, 17], WikiPathways [18], PID [19], PathCards [16], and MSigDB [6, 17]. The Pathway Topology Based (PTB) method analysis usually consists of three steps [18]: initially, expression data and pathway topology data are used to calculate the gene-level statistics; the next step aggregates gene level statistics into pathway-level statistics; finally, the pathway-level statistics is compared with the

reference distribution under the null hypotheses of no enrichment. Similarly to the previous method the PTB analysis methods can be divided based on three main criteria [19]: the null hypothesis they test; the method for identification of differentially expressed genes (DEGs) before pathway analysis; and the number of variables in the model.

Goeman [20, 21] makes a distinction between two null hypotheses and defines them based on a subset of differentially expressed genes. Let G be the set of genes in the pathway and G^C its complement. With these two complementary sets of genes, the competitive null hypothesis is:

$$H_0^{comp}: \text{The genes in G are at most as often differentially expresssed as the genes } G^C$$

and the self-contained null hypothesis is:

$$H_0^{self}: \text{No genes in G are differentially expressed}$$

The self-contained null hypothesis expects none of the genes from the pathway to be differentially expressed, and uses sample randomization in the assessment of the statistical significance. This method requires a sufficient sample size for sample randomization, but at the same time gives the advantage of p-value, which relates to true biological replications of the experiment to new subjects. Similarly to self-contained method the competitive null hypothesis performs permutation, but in this case algorithm permutes gene labels for each pathway, and compares the set of genes in the pathway with a set of genes that are not in the pathway. In contrast, a self-contained null hypothesis permutes class labels (i.e., phenotypes) for each sample and compares the set of genes in a given pathway with itself, while ignoring the genes that are not in the pathway [1].

Several limitations of the PTB method could be listed [22]: 1) Some PTB methods do not recognize the direction and type of the connections between the pathway components; 2) Most of the PTB methods do not take into account how the pathways are interconnected, the consequences of those interactions and gene set overlaps in analyzed pathways; 3) Methods also do not take into consideration dynamic changes underlying biological processes such as spatial, temporal, and multiple states and variants that a pathway component can have.

The lack of reproducibility of different methods could be overcome by consensus methods [23, 24]. Ensemble methods can find a common list of genes across various studies, which are further used for enrichment analysis. Consensus methods on the level of pathway analysis can combine and rank results based on different criteria such as sum, mean, median, min [4], and majority voting score [25, 26] on the levels of ranking. On the level of p-values, methods such as Fisher's method, Logit method, Summation of Z method, average method, Summation method, Wilkinson's method [25, 26]can be used to re-rank individual rankings. The consensus methods have shown to improve individual rankings by increasing confidence in specific target pathways and removing irrelevant pathways from the top of the ranking [4].

The enrichment methods implemented in the web application – Enrichment Application (EApp), set-based enrichment analysis: RS [27], ORA (FisherExact) [4, 27], SAFE [28], GSEA [29], PADOG [30], ROAST [31], CAMERA [32], GSA [33], GSVA [34], GLOBALTEST [35], EBM [36], MGSA [37, 38], SAMGS [39] , and net-based enrichment algorithm: SPIA [40], PathNet [37], CePa [41], GGEA [42], DEGraph [43], GANPA [44], TOPOLOGYGSA [45], NETGSA [46]. Extended information about classification, version, type of pathway representation, bioconductor, type of tested hypothesis can be found in table included in Appendix A.

This thesis continues with the following chapters: Chapter 2 includes a description of algorithms used in shiny application. Chapter 3 presents a case study for set-based, network-based, and consensus enrichment analysis. Chapter 4 presents a benchmark analysis in the sense of runtime, fraction of significant gene sets, and phenotype relevance. Finally, Chapter 5 describes the achieved results and further improvement.

Set-based enrichment analysis

ORA – Over Representation Analysis.

The simplest and most popular methods testing for overrepresentation start from creating a contingency table for differentially expressed genes based on a specific cut-off and test whether the differentially expressed gene set is overrepresented. In most cases a 2x2 contingency table [21] is used, where m_{GD} represents counts of differentially expressed genes and genes in pathways (gene sets), m_{GD}^{C} represents counts of no-differentially expressed genes and genes in pathways, and similarly for genes not in the pathway (gene sets). C stands for complements of the set. Gene set represents a pathway or group of genes related to the disease.

Table 2x2 to assessing overrepresentation

	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{GD}	m_{GD^c}	m_G
Not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
Total	M_D	m_{D^c}	M

The p-value for gene set statistic is calculated using a test for independence in the 2x2 table represented above. The most commonly used tests are the χ^2 test, the hypergeometric test (Fisher's exact test), and the binomial z-test for proportions.

SAFE - Significance Analysis of Function and Expression.

SAFE [28, 47] is a resampling method for enrichment analysis in gene expression experiments. Estimation of local and pathway level p-values is accomplished in two steps. The first step includes calculation of association

between genes expression and group response defined by user. The user can choose between several tests such as Weltch test, different version of t-test , for 2-sample designs, one-way ANOVAs, and simple linear regressions. In the second step for pathway level statistics, the user can choose among the Wilcoxon rank sum, a Fisher's Exact test statistic, Pearson's chi-squared type statistic, or t-statistic. For adjusting for multiple comparisons SAFE uses several strategies based on permutation and bootstrap. The package allows reporting false discovery rate based on Benjamini-Hochberg procedure [48] or Yekutieli-Benjamini [49], family-wise error rate (FWER) estimated by Bonferroni correction, Holm's step-down procedure [50], or Westfall-Young [51] method. In the article written by Barry [47] it was shown that SAFE with bootstrap is more powerful for controlling Type I error. The bootstrap methods extends [52] implemented in safe methods for controlling multiple comparison problems.

The authors included two methods in the package. The first method, "bootstrap.t" invokes pivot tests to look for the exclusion of a null value from Gaussian confidence intervals computed from the resampled mean and variance of the global statistic. The second method, "bootstrap.q" invokes tests based on the exclusion of a null value from the alpha-quantile interval of the resampled global statistic.

GSEA - Gene Set Enrichment Analysis

GSEA [29] is one of the first frequently used method from the category of Function Class Scoring which uses a weighted Kolmogorov-Smirnov statistic to test whether the ranks of the p-values of genes in a pathway resembles a uniform distribution.

GSEA can be applied to expression profiles from samples belonging to two distinct classes. Genes are ranked based on the association between their expression and the phenotype by using fold change of expression or another suitable measure such as z-score, p-value (Figure 2A).

GSEA method is evaluating whether genes from the pathway of interest are randomly distributed through expression profile of the experiment. We expect that for enriched pathways, that genes will be concentrated at the top or bottom ranked genes from the experiment.

GSEA method calculates enrichment statistics in three steps:

Step 1: Enrichment Score calculation. Enrichment score (ES) reflects the degree to which a set S is overrepresented in the set of all ranked genes set D in total of N elements. The bigger the peaks on the edges (top or bottom) of the entire ranked list L, the bigger the ES will be. The score is calculated as the cumulative sum over ranked list L based on r_j which represents strength of association, and could be expressed by signal-to-noise ratio, absolute value of signal-to-noise ratio, difference of expression means between classes, ratio of expression means of two classes, log2 of ratio, and t-test statistic. The value increases when a gene is in the set of genes representing a pathway (S), and decreases when the gene is not in S. The value of the increment depends on the correlation of the gene with the phenotype, and is positive for genes in the list, and negative for genes not in the list. The enrichment score is defined as weighted Kolmogorov–Smirnov-like statistic and expresses the maximum deviation from zero encountered in the random walk (Fig. 2B), and is defined as follows:

$$ES(S,D) = \max_{1 \le j \le N} \left\{ \sum_{\substack{g_j \in S \\ j \le D}} \frac{|r_j|}{N_R} - \sum_{\substack{g_j \notin S \\ j \le D}} \frac{1}{N - |S|} \right\}$$

where $|S|$ is the number of genes in S, N_R is a sum of absolute values of ranking metrics for all genes in the gene set S.

Step 2: Estimation of Significance Level of ES. Statistical significance is estimated by permutation test, method allows use gene and sample permutation. The best method uses permutation based on phenotype, which preserves the correlation structure between genes, and for small sample size algorithm allows permutations for genes labels. The empirical, nominal p-value of the observed ES is then calculated relative to this null distribution.

Step 3: Adjustment for Multiple Hypothesis Testing. The procedure described above is repeated for all gene sets in the database, and thus requires adjustment for multiple hypothesis testing. The false discovery rate [53, 54] is calculated for every normalized enrichment score (NES). The

FDR estimates the probability that a set with a given *NES* represents a false positive finding. It is computed by comparing the tails of the observed and null distributions for the *NES*.

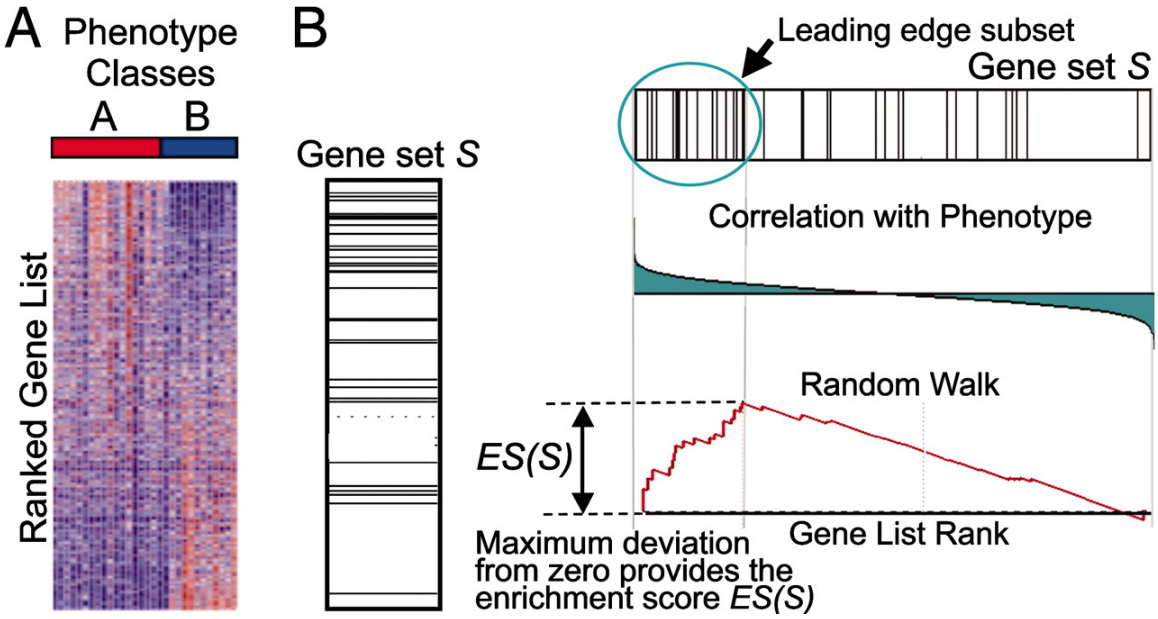

Figure 2. A GSEA overview illustrating the method[2] (*A*) An expression dataset sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the dataset, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

SAMGS - Significance Analysis of Microarray for Gene Sets

SAMGS [39]- Significance Analysis of Microarrays on Gene Sets, extends the SAM method for single genes to gene set analysis. Gene set statistics used in the SAM-GS algorithm is defined as L2-norm of the t-like-statistic vector $d = (d_1, d_2, \ldots, d_{|s|})$, the length of the line segment joining the two phenotypes' mean gene-expression vectors of set S. The null hypothesis tests that there is no difference in mean of genes in the analyzed pathway. We cannot apply Hotelling's T^2 for a two-sample mean test because $|S| > n_1 + n_2 - 2$, where n_1 and n_2 are the sample sizes in the two groups defining the phenotype D.

[2] A. Subramanian et al. PNAS 2005;102:43:15545-15550

SAM-GS Steps:

1) For each of the N genes, calculate the statistic d as in SAM for an individual-gene analysis:

$$d_i = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

2) Compute the SAMGS test statistic corresponding to set S:

$$SAMGS = \sum_{i=1}^{|S|} d_i^2$$

3) Permute the labels of the phenotype D and repeat 1) and 2). Repeat until all (or a large number of) permutations are considered.
4) Statistical significance for the association of S and D is obtained by comparing the observed value of the SAMGS statistic from 2) and its permutation distribution from 3).

s(i) is a pooled standard deviation over the two groups of the phenotype, and s_0 is a small positive constant that adjusts for the small variability encountered in microarray data.

PADOG - P*athway* A*nalysis with* D*own-weighting of* O*verlapping* G*enes*

PADOG [30] is method that computes a gene set score as the mean of absolute values of weighted moderated gene t-scores [55]. The difference between standard t-score and moderated T-score comes from variance calculation. For moderated t-test, variance is based on all selected genes, whereas for standard t-test, it is calculated separately for each gene. The weighting procedure is designed to emphasize genes uniquely (less commonly) represented in pathways, versus genes that appear in many gene sets. The gene set score is calculated by the formula:

$$S_O(GS_i) = \frac{1}{N(GS_i)} \sum_{g \in GS_i} |T(g)| * w(g)$$

,where $w(g) = 1 + \sqrt{\frac{\max(f) - f(g)}{\max(f) - \min(f)}}$, $f(g)$ is the frequency of gene g across all gene sets to be analyzed. Finally, p-value is calculated by permutation procedure and comparing with original value:

$$P_{PADOG}(GS_i) = \frac{\sum_{ite} I\left(S^*_{ite}(GS_i) \geq S^*_O(GS_i)\right)}{N_{ite}}$$

,where I is a function that returns 1 when the argument is true and 0 otherwise, and represents the standardized score obtained with the ite-th permutation of the samples for gene set GS_i.

ROAST – ROtAtion gene Set Tests

ROAST *[31]* represents one of the methods for assessing differentially expressed genes without permutation strategy, which unrealistically assume independence of genes. Null hypothesis for gene set is $H_O: \beta_g = 0$ for all genes; alternative hypothesis tests if at least one gene is down ($H_O: \beta_g < 0$), up ($H_O: \beta_g > 0$), or any direction (mixed $H_O: \beta_g \neq 0$) regulated. The authors define several gene set statistics inspired by previous works of: Ackermann, Strimmer [56]; Jiang, Gentleman [57]; and Efron and Tibshirani [33]. A linear model representing the problem could be written as $E(y_g) = X\alpha_g$, where X represents design matrix, α_g unknown coefficient vector. By contrast, a statistical problem could be represented by $\beta_g = c^t\alpha_g$ and t-statistics $t_g = \frac{\widehat{\beta_g}}{s_g\sqrt{v}}$, where $\hat{\beta}_g = c^t\alpha_g$ is the least squares estimator of β_g, s_g is the residual of SD for genes g, and $v = c^T(X^TWX)^{-1}c$ the unscaled standard deviation of $\hat{\beta}_g$. Finally, rotation test described by Langsrud [58] could be applied, which is replacing a commonly used permutation procedure. Rotation test performs B (typically around 10000) random rotations of the vector of residuals of $\boldsymbol{u}_g$ to a random point $\boldsymbol{u}^*_g$ on the sphere d+1 with radius ρ_g. Sphere radius is defined as: $\rho_g^2 = \boldsymbol{u}_g^T * \boldsymbol{u}_g$, d=n-p, n is the number of samples, and p is the number of groups. Vector of residuals $\boldsymbol{u}_g$ is estimated during QR decomposition, and elements of that vector are independent, normally distributed with mean zero, and variance equal to variance for given gene g. For every random rotation, the moderated t-statistics and gene set statistic T are computed, and after a large number of rotations, the Monte Carlo p-value is computed. The authors implemented several different pathway level statistics T, for example floor-mean T statistics. Details of the numerical procedure could be found in the original

article *[31]*. The rotation test could be applied for small datasets and allows for any experimental design that can be expressed as a linear model, and can also incorporate array weights and correlated samples.

CAMERA – Correlation Adjusted MEan RAnk gene set test

CAMERA[32] gene set test with adjustment for inter-gene correlation. Recent studies show that competitive tests are sensitive to inter-gene correlation and small correlations can significantly overestimate the false discovery rate (FDR) [33, 59-61]. The authors are using an extended version of two-side t-test and Wilcoxon–Mann–Whitney (WMW) rank sum test allowing for correlation. In the first test we can see additional factor in the denominator of T-statistics called variance inflation factor (VIF), which is accounting for correlation $VIF = 1 + (m_1 - 1)\hat{\rho}$, where m_1,m_2 – sample size of group 1 and 2, respectively. $\hat{\rho}$-is average of all pairwise correlation between genes in the first group. T-statistics take the form:

$$T = \frac{\bar{z_1} - \bar{z_2}}{s_p \sqrt{\frac{VIF}{m_1} + \frac{1}{m_2}}}$$

$\bar{z}_1, \bar{z}_2$-means in the groups, sp is the pooled residual standard deviation. In the case of WMW test standard square root of variance $\mathrm{var}(\mathrm{RankSum}) = m_1 m_2 (m_1 + m_2 + 1)/12$ is replaced by:

$$var(Rank; sum) = \frac{m_1 m_2}{2\pi} \left\{ \begin{array}{c} sin^{-1} 1 + (m_2 - 1) \sin^{-1} \frac{1}{2} \\ +(m_1 - 1)(m_2 - 1) \sin^{-1} \frac{\rho}{2} + (m_1 - 1) \sin^{-1} \frac{\rho + 1}{2} \end{array} \right\}$$

CAMERA is a competitive gene set test with improved estimation of type I error regardless of inter-gene correlation and still retains good statistical power.

GSA – gene-set analysis

GSA[33] – gene set analysis approach is similar to the GSEA approach, but instead Kolmogorov-Smirnov statistics the authors use "maxmean statistic"

calculated as follows: compute the average of positive parts of each z_i in *S,* and also negative parts, and choose the one that is larger in absolute value. It can be shown by analytic calculations and simulations that maxmean statistics is more powerful than statistic used in GSEA. To estimate the p-value for a given dataset, the algorithm combines randomization and permutation ideas into a method called "Restandarization". Randomization refers to the effect related to sampling a subset of genes, and permutation refers to the reshuffling effect related to columns (samples). This process can be thought of as a method of adjusting permutation values S^* to account for the overall distribution of the individual scores s_i. Similarly to other re-sampling methods, a restandardized p-value for testing the enrichment of gene set S is obtained by comparing its actual enrichment score S to a large number of iteration N_{ite} "of S**" simulations:

$$p_S = \#(S^+ values\ exceeding\ S)/N_{ite}$$

,where $S^+ = mean_S + \frac{stdev_S}{stdev^*}(S^* - mean^*)$ with ($mean_S$, $stdev_S$) and ($mean^*$,$stdev^*$) the overall means and standard deviation from randomization and permutation versus permutation process. S represents the average of "maxmean statistics".

GSVA - gene set variation analysis

The Gene Set Variation Analysis (GSVA) [34] method tests a null hypothesis that there is no difference between genes inside and outside the gene set. To achieve this goal, the GSVA algorithm calculates sample-wise gene set enrichment scores as a function of genes inside and outside the gene set.

In the first step, GSVA evaluates whether a gene *i* is highly or lowly expressed in sample *j* in the context of the sample population distribution. To do so we need to bring distinct expression profiles to a common scale by applying a discrete Poisson kernel:

$$z_{ij} = \hat{F}_r(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\sum_{y=0}^{x_{ij}} \frac{e^{-(x_{ij}+r)}(x_{ij}+r)^y}{y!}$$

, where r=0.5 is scaling parameter in order to set the mode of the Poisson kernel at each $x_{i\,k}$. To avoid effects of outliers z_{ij} is converted to ranks and

centered on the median $r_{ij} = \left|\frac{p}{2} - z_{(i)j}\right|$. Now, we can calculate the contribution to the enrichment statistics (ES)(also called GSVA score):

$$v_{ij}(\ell) = \frac{\sum_{i=1}^{\ell}\left|r_{ij}\right|^{\tau} I(g_{(i)} \in \gamma_k)}{\sum_{i=1}^{p}\left|r_{ij}\right|^{\tau} I(g_{(i)} \in \gamma_k)} - \frac{\sum_{i=1}^{\ell} I(g_{(i)} \notin \gamma_k)}{p - |\gamma_k|}$$

The last step involves calculation of enrichment statistics (ES). The algorithm assumes two alternative scores maximum deviation from zero: $ES_{jk}^{max} = v_{jk}\left[\arg\max_{\ell=1,..,p}\left|v_{jk}(\ell)\right|\right]$ and second difference between maximum and minimum deviation:

$$ES_{jk}^{diff} = \left|ES_{jk}^{+}\right| - \left|ES_{jk}^{-}\right| = \max_{\ell=1,....=,p}\left(0, v_{jk}(\ell)\right) - \min_{\ell=1,...,p}\left(0, v_{jk}(\ell)\right)$$

This statistic has a biological interpretation; it emphasizes genes in pathways that are concordantly activated in one direction only. This statistic has approximately normal shape. In the results for pathways containing genes strongly acting in both directions, the deviations will cancel each other out and show little or no enrichment. If the relevant gene sets are not separated into "up" and "down" behavior, the max deviation score should be used.

Globaltest – test based on regression model

GLOBALTEST [35] – For the global test of a group of genes to be used to predict the clinical outcome, the gene expression patterns must differ for different clinical outcomes and this problem could be addressed by building a generalized linear model.

Testing whether there is a predictive effect of the gene expressions on the clinical outcome is equivalent to testing the hypothesis:

$$\mathrm{H_0}: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

that all regression coefficients are zero. However, it is possible to reduce this problem to check the null hypothesis $\mathrm{H_0}: \tau^2 = 0$, where $\tau^2$is the variance of distribution of regression coefficients $\beta_1, \beta_2, \ldots, \beta_m$ with expectation zero.

To resolve the discussed hypothesis the authors suggest using score test discussed in Le Cessie and Van Houwelingen [62] and Houwing-Duistermaa [63]. A score test for $\tau^2 = 0$ can be calculated by taking the derivative of the loglikelihood concerning τ^2 at $\tau^2 = 0$ divided by the standard deviation of this derivative under H_0. This yields to T-test statistic, which is asymptotically normally, distributed if H_0 is true:

$$T = \frac{(Y-\mu)'R(Y-\mu) - \mu_2 trace(R)}{(2\mu_2^2 trace(R^2) + (\mu_4 - 3\mu_2^2)\sum_i R_{ii}^2)^{1/2}}$$

where $R = \left(\frac{1}{m}\right) XX'$ is n x n matrix proportional to the covariance matrix of the random effects r, $\mu = h^{-1}(\alpha)$ is expectation of Y under H_O and μ_2, μ_4 are the second and fourth central moment of Y under H_O.

EBM – Empirical Brown Method

EBM[36] – Combined correlation measurements for genes in the pathways give information on how strongly the expression pattern could explain phenotype/pathway. Combining p-values from multiple statistical tests is a standard solution in statistics, but this procedure is non-trivial for dependent p-values. In the developed package, the authors discuss several methods such as Fisher's, Kost's [64], and an empirical adaptation of Brown's [65], [36] method which is appropriate for the large and correlated datasets found in high-throughput biology. During analysis, the technique is deriving combined p-values by associating the expression levels of all genes in the dataset and pathway. If genes are members of the pathway, the correlation between genes themselves is excluded. Then, we combined these p-values for each of the gene sets using both Fisher's and the Empirical Brown's Method. To assess the association and statistical significance on the gene set, the authors were using the following permutation test:

$$P_{PERM} = \frac{\sum_{m=1}^{M} I(\Psi_m^* \geq \Psi)}{M}$$

, where M-number of iteration, Ψ_m^*, Ψ combined p-values, I – indicator function.

MGSA - model-based gene set analysis

MGSA[38] - Model-based gene set analysis (MGSA) represent biological categories in the terms of the activation or deactivation as response to gene expression. Bayesian network allows us embeds all categories together at one and automatically takes into account overlaps between categories. Such a Bayesian network could be represented in figure 3[3].

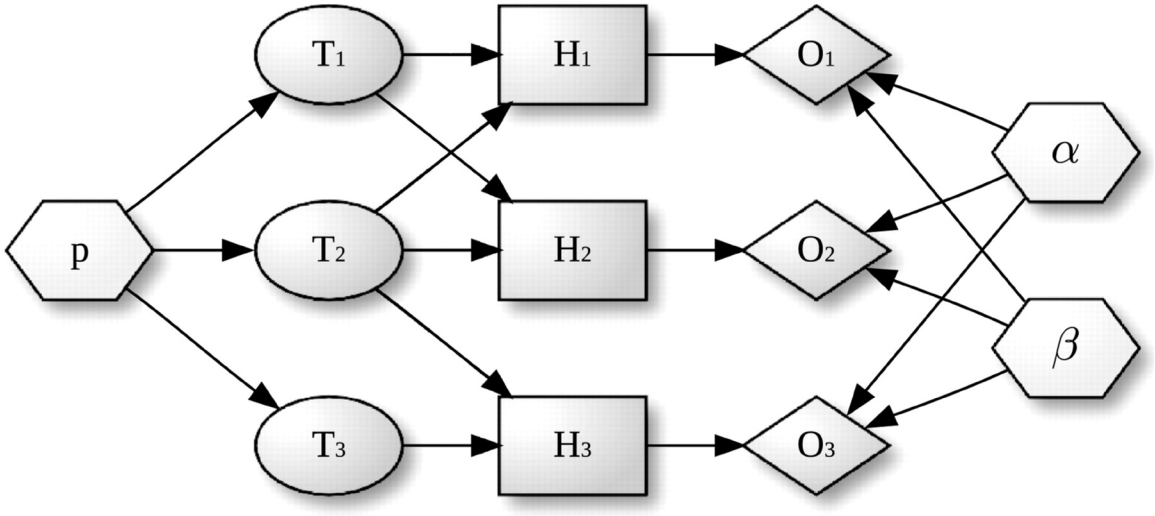

Figure 3. A Bayesian network to model gene response with gene categories.

Gene categories, or terms (T_i, ellipses) can be either on (1) or off (0). Terms that are on (1) activate the hidden state (H_J, rectangles) of all genes annotated to them, and the other genes remain off (0). The observed states (O_j, diamonds) of the genes are noisy observations of their true hidden state. The parameters of the model (light gray nodes) are the prior probability (P) of each term to be active, the false positive rate α, and the false negative rate, β.

---

[3] From: GOing Bayesian: model-based gene set analysis of genome-scale data
Nucleic Acids Res. 2010;38(11):3523-3532. doi:10.1093/nar/gkq045
Nucleic Acids Res | © The Author(s) 2010. Published by Oxford University Press..

The joint probability distribution for this Bayesian network could be written as:

$$P(T,H,O) = P(T)P(H|T)P(O|H) = P(T)\prod_{i=1}^{n} P(H_i|T)P(O_i|H_i)$$

The authors suggest modeling P(T) as Bernoulli distribution with $P(T_j = 1) = p$ and denote $m_{x|T} = |\{j|T_{j=x}\}|$ number of terms that have state x for a given T, then:

$$P(T) = p^{m_{1|t}}(1-p)^{m_{0|T}}$$

The second part could be modeled by:

$$P(O|T) = \prod_{i=1}^{n} P(H_i\,|T)P(O_i|H_i) = \alpha^{n_{10|T}}(1-\alpha)^{n_{00|T}}\beta^{n_{11|T}}(1-\beta)^{n_{01|T}}$$

The notation $n_{xy|T} = |(i|O_i = x \wedge H_i = y)|$ means the number of genes having observed activation x and true activation y according to the state of T. For example, $n_{01|T}$ corresponds to the number of genes observed to be not differentially expressed, but whose true activation state is on. The authors chose for the transition $H \to O$ two Bernoulli distributions: $P(O_i = 1|H_i = 0) = \alpha$ and $P(O_i = 0|H_i = 1) = \beta$. All parameters are estimated by Markov chain Monte Carlo method (MCMC) [66], because marginal posteriors for network cannot be derived analytically.

Novelty of the method comes from the Bayesian approach, that model the data with all categories simultaneously, rather than using hypothesis testing on each category. This type of approach avoids an issue of multiple testing. The pathway score from the Bayesian approach is simply the probability of a category to be active, and it reverses the value given by hypothesis-based procedure, where a low p-value indicates high confidence.

GRS – Generalized Random Set

This method initially was described by Sengupta [67], Newton [68] and also implemented in the CLEAN package [27]. Generalized Random Set improves the random set (RS) methodology and does not require

specification of a significance cutoff for either the query signature or the reference datasets.

This method tests that average gene-level evidence across the category C is not at random. Enrichment score is calculated as the average value of gene-level statistics s_g, such as fold change, p-value or other measures of association such as Pearson correlation, Spearman correlation, ranks associated with gene-level scores. The enrichment score is represented by average value:

$$\bar{X} = \frac{1}{m}\sum_{g \in C} s_g$$

where *m* is the number of genes in category C. It is useful to treat the random set C as drawn uniformly at random without replacement from the $\binom{G}{m}$ subsets of *m* distinct genes from the population of G genes. This could be seen as the permutation scheme in which gene-level scores are randomly shuffled among the gene labels. The enrichment score has approximately a Gaussian distribution with mean and variance:

$$\mu = E(\bar{X}) = \frac{1}{G}\sum_{g \in G} s_g$$

and

$$\sigma^2 = var(\bar{X}) = \frac{1}{m}\left(\frac{G-m}{G-1}\right)\left\{\left(\frac{\sum_{g=1}^{G} s_g^2}{G}\right) - \left(\frac{\sum_{g=1}^{G} s_g}{G}\right)^2\right\}$$

The authors propose to use the standardized category enrichment score $Z = (\bar{X} - \mu)/\sigma$, which is equal to zero for the category C not enriched for differentially expressed genes. This approach significantly simplifies analysis, especially in the case of multiple categories, because Z is computable without using permutation. Large values of Z favor the enrichment hypothesis.

The authors using random set address two problems related to multiple-category inference; namely, that equally enriched categories are not detected with equal probability if they are of different sizes, and also that there is dependence among category statistics owing to shared genes.

Random-set enrichment calculations do not require Monte Carlo for implementation.

The null hypothesis states there is no enrichment of differentially expressed genes among the genes in genes set (pathway). The authors introduce probabilities of differential expression for each gene in two datasets (one representing functional category, and the other representing genes of interest) to avoid discretization to binary categories: “differentially” or “non-differentially” expressed.

Network-based enrichment analysis

GGEA - Gene Graph Enrichment Analysis,

Gene Graph Enrichment Analysis (GGEA) Gene Graph Enrichment Analysis (GGEA) [42] method is designed to identify enriched gene sets, based on knowledge built-in to gene regulatory networks. The GGEA method consists of three main steps presented in the figure 4 below. First, the gene set is mapped onto the underlying regulatory network. In the second step, consistency scores are computed for each sub-network with mapped expression levels and significance level. Third, the significance of the scores is estimated via re-sampling, and evaluated to rank the gene sets. Differential expression of genes is combined into pairs of the fold change and p-value of significance from t-test in order to simultaneously summarize and express whether the transcriptional activity of a particular gene is reduced or enhanced. Regulatory interactions of the gene regulatory network are represented by transition with an input place for the regulator and an output place for its target, as well as an associated effect (activation, inhibition) and the direction of the interaction. To evaluate agreement between pathway (regulatory network) and expression of their element, the algorithm calculates a normalized consistency score:

$$\bar{S} := \frac{S}{|T_u|} = \frac{\sum_{t \in T_u} C(t)}{|T_u|}$$

,where $C(t) := cons(de_0, f_t(de_i))$ represents similarity between the predicted and measured taken on the output place of transition t. Significance level is calculated by the permutation procedure.

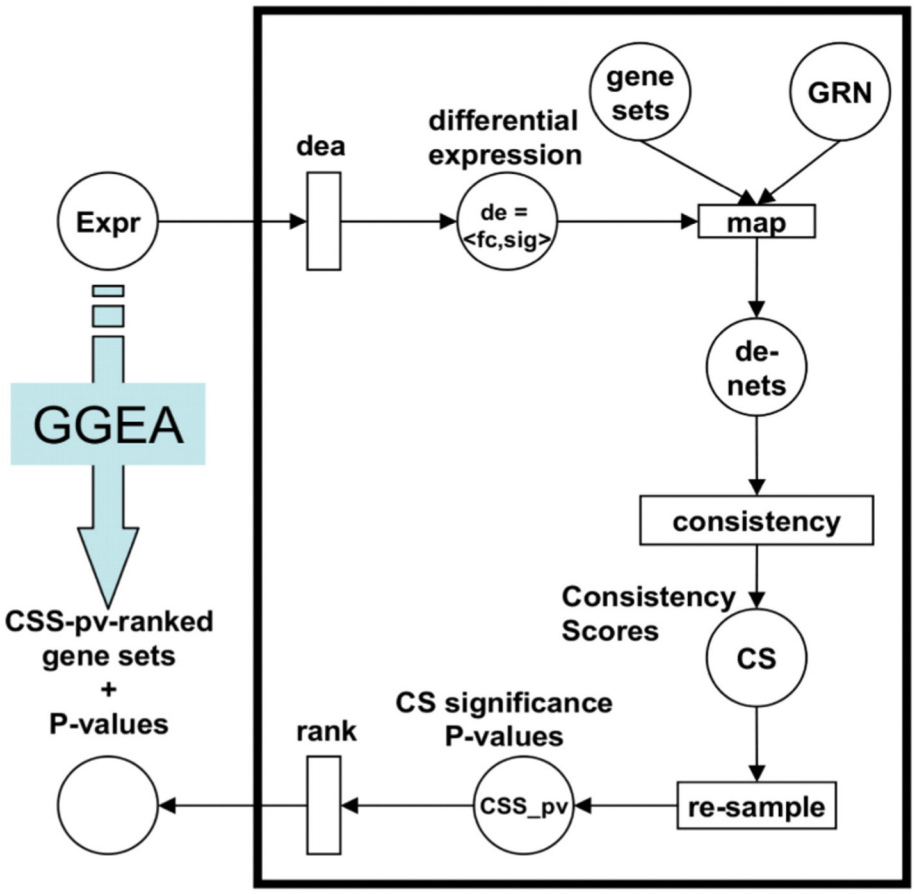

Figure 4 . Key Steps of GGEA algorithm[4]

SPIA: Signaling Pathway Impact Analysis

The first method which was exploiting knowledge from pathways, is called signaling pathway impact analysis (SPIA) [40]. The algorithm combines evidence obtained from ORA method with measures coming from perturbation evidence which comes from the importance of the position of

[4] From: From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems
Bioinformatics. 2011;27(13):i366-i373. doi:10.1093/bioinformatics/btr228

the DE genes in the given pathway as well as their fold-changes. Similar to previous methods, the p-value is estimated by resampling methodology. Final P_G p-value consists of two parts P_{NDE} – the over-representation of DE genes in the given pathway, and P_{PERT}-which measures based on the amount of permutation measured in each-pathway and finally combined by Fisher formula:

$$P_G = c_i - c_i * \ln c_i$$

where $c_i = P_{NDE}(i) \cdot P_{PERT}(i)$. P_{PERT} is measured based on the gene perturbation factor:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \frac{PF(g_j)}{N_{ds}(g_j)}$$

the $\Delta E(g_i)$ represents the normalized estimated of expression change of the gene g_i (log fold-change if two conditions are compared). The second part represents normalized by the number of downstream genes of each such gene $N_{ds}(g_j)$ sum of perturbation factors of the genes g_j directly upstream of the target gene g_i,. The β reflects the type of interaction: +1 for induction (activation), −1 for repression and inhibition, as described by each pathway. Also, β will have non-zero value only for the genes that directly interact with the gene g_i according to the pathway description.

Perturbation p-value is calculated by bootstrap procedure:

1. Initialize with k=1.
2. Calculate total accumulation across each pathway: $T_A(k) = \sum_i Acc(g_{ik})$ where net accumulation of the perturbation is equal $Acc = B(I - B)^{-1}\Delta E$, with random sample drawing each time.
3. Repeat procedure above a large number of time (N_{ite}=2000)
4. Center value by subtracting median of T_A, the resulting corrected values are denoted as $T_{A,c}(k)$. Net accumulation is also shifted – $t_{A,c}$.
5. If $t_{A,c}>0$ we conclude that pathway is activated and in the opposite situation pathway is inhibited.
6. Finally P_{PERT} is computed as:

$$P_{PERT} = \begin{cases} 2\dfrac{\sum_k I\left(T_{A,c}(k) \geq t_{A,c}\right)}{N_{ite}} & if\ t_{A,c} > 0 \\ 2\dfrac{\sum_k I\left(T_{A,c}(k) \leq t_{A,c}\right)}{N_{ite}} & otherwise \end{cases}$$

PathNet: Pathway analysis using Network information

The PathNet [37] method applies ORA on combined evidence of the observed signal and the signal implied by connected neighbors in the pooled network created based on all KEGG pathways. Based on the pooled network, the authors created an adjacency matrix $A_{ij,}$ when the element is equal to 1 when interaction exists, and equal to 0 when interaction does not exist. Also, diagonal elements are set to 0 to avoid self-interaction. This method combines direct (difference from treatment) and indirect evidence. The indirect evidence considers the differential expression of the neighbors of gene *i* in the *pooled pathway*. This is characterized by Indirect Evidence Score (SI_i), which consolidate the topological information of the pathways:

$$SI_i = \sum_{j \in G, i \neq j} A_{ij} * \left(-\log\left(p_j^D\right)\right)$$

,where G denotes the set of all genes present in the *pooled pathway,* and A_{ij} is adjacent matrix defined above. Indirect Evidence Score is used to estimate the indirect evidence (p_i^I) p-value values by counting the number of random scores larger than the actual scores, as follows:

$$P_i^I \approx \frac{1}{N} \sum_{n=1}^{N} \begin{cases} 1 & if\ SI_i^R > SI_i \\ 0 & otherwise \end{cases}$$

Finally, direct and indirect p-values are combined by Fisher method [65], using the following equation:

$$p_i^C = \int_{-\ln\left(p_i^D * p_j^I\right)}^{\infty} P(\chi_4^2)$$

where $P(\chi 4)$ denotes the probability density function of the $\chi 2$ distribution with 4 degrees of freedom. Note that even if the pD- and pI-values were correlated, they could still be combined using a modified version of Fisher's method. Finally, we selected genes with $p_i^C < 0.05$ as differentially expressed and used the hypergeometric test to calculate pathway enrichment.

DEGraph: Differential expression testing for gene graphs

DEGraph [43] implements recent hypothesis testing methods which directly assess whether a particular gene network is differentially expressed between two conditions. This is to be contrasted with the more classical two-step approaches, which first test individual genes, then test gene sets for enrichment in differentially expressed genes. These recent methods take into account the topology of the network to yield more powerful detection procedures. DEGraph provides methods to easily test all KEGG pathways for differential expression on any gene expression dataset, as well as tools to visualize the results. The algorithm consists of two main steps. In the first, graph-based dimensionality reduction allows for reducing dimension, and the second step allows for applying multivariate test of means. Another way to state the procedure as described by the authors is: (1) project the vectors of covariates in a new space of lower dimension that preserves the distribution shift, that is, the distance between the expression measures of the two groups, and (2) apply the multivariate statistic in this new space.

Let's introduce main concepts needed to realize dimensionality reduction with graphs. The gene network is represented by graph $\mathcal{g} = (\mathcal{E}, \upsilon)$, where $|\upsilon| = p$ is the number of nodes and $\mathcal{E}$ is the set of edges. Let $\delta \in \mathbb{R}^P$ denote the mean shift, that is, the vector of differences in mean expression measures for these p genes between the two populations of interest. We expect the shift δ to minimalize $E_G(\delta)$, where E_G is coherent with the graph G and is a function of the shift of expression measures, and the structure of the network

We can use a recursive procedure to find the subset of eigen vectors that minimalizes energy defined by the network structure and expression shift. This step will act as the dimensionality reduction procedure, where the final subset with dimension k << p, and will minimalize energies that:

$$u_i = \begin{cases} \underset{f \in \mathbb{R}^p}{\operatorname{argmin}} E_G(f) \\ \text{such that } u_i \perp u_j, j < i \end{cases}$$

E_G represents a semi-definitive matrix in the form $E_G = \delta^T \mathcal{L} \delta = \delta^T U \Lambda U^T \delta$, where U is an orthogonal matrix, and Λ is a diagonal matrix with

elements $E_G(u_i) = \lambda_i$, i=1,…,p. $\mathcal{L}$ is graph Laplacian, defined as $\mathcal{L} = D - A$ or $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$. An undirected graph, which represents the network structure, is defined by the adjacency matrix A, where $a_{ij} = 1$ if connection exists, and and $a_{ij} = 0$ otherwise. Degree matrix D=Diag(A1), where 1 is a unit column-vector, and $D_{ii} = d_i$. Finally, energy can be represented as:

$E(\delta) = \sum_{i,j\in v}(\delta_i - \delta_j)^2$ or $E(\delta) = \sum_{i,j}\left(\frac{\delta_i}{d_i} - \frac{\delta_j}{d_j}\right)^2$, depends of the form a Laplacian.

The multivariate test of means is realized by applying Hotteling's T^2 – test. The Hotteling's T^2 test statistic is based on squared Mahalanobis norm of the sample mean shift and is given by $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^T\hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$ and could be represented in reduced graph space by:

$$\bar{T}_k^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T U_{[k]}\left(U_{[k]}^T \hat{\Sigma} U_{[k]}\right)^{-1} U_{[k]}^T(\bar{x}_1 - \bar{x}_2)$$
$$= \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T U 1_k U^T\left(U 1_k U^T \hat{\Sigma} U 1_k U^T\right)^{+} U 1_k U^T(\bar{x}_1 - \bar{x}_2)$$

where A^+ denotes the generalized inverse of a matrix A, p x k matrix $U_{[k]}$ denotes the restriction of U to its k columns, and 1k is a p x p diagonal matrix, with i-th diagonal element equal to one if I < k, and zero otherwise.

TopologyGSA: Topology-based Gene Set Analysis

The authors of TopologyGSA [45] use Gaussian graphical models to explicitly incorporate the dependence structure among genes by the topology of pathways and information about topology is embedded into covariance matrix. The analysis is designed to be used to detect changes in a pathway for different experimental conditions. The algorithm is able to detect strength of the relations among genes and also changes in expression for a different experimental condition, which is done by multivariate hypothesis testing.

The first step in the algorithm consists of conversion of the pathway into a graphical model represented by directed acyclic graph (DAG), followed by conversion into a moral graph.

Given the graph structure and two conditions, we can formulate the Gaussian graphical model as:

$$M_1(G) = \{Y \sim N_P(\mu_1, \Sigma_1), \Sigma_1^{-1} \in S^+(G)\}$$
$$M_2(G) = \{Y \sim N_P(\mu_2, \Sigma_2), \Sigma_2^{-1} \in S^+(G)\}$$

,where μ_1, μ_2 represents mean expression for case, control respectively. Σ_1, Σ_2 are the corresponding covariances, both constrained to have the same structure as specified a priori reflecting gene network interaction. p is the number of vertices (genes) on the graph, and S+(G) is the set of symmetric positive definite matrices with null elements corresponding to the missing edges of G.

We are interested in testing for differential expression in a given pathway. This is accomplished in two step. First, include testing equality of the strength of the relations among genes, which is equivalent to testing equality of variance (in this case, the equality of two concentration matrices – inverse of the covariance matrices). Testing could be performed with likelihood ratio test for the equality of two covariance matrices, or replaced by the more robust Wald test [69]. Second step for differential expression testing depends on the results of the previous test. If $\Sigma_1 = \Sigma_2$ we can perform multivariate analysis of variance (MANOVA) [70], otherwise we have to solve Behrens-Fisher problem, which could be achieved using Yao, Johansen methods [71] and other [72].

GANPA: Gene Association Network-based Pathway Analysis

Gene Association Network-based Pathway Analysis [44] is a weighted version of GSEA algorithm. Both GASE and weighted-GSEA statistics are presented below:

$$S_{GSEA} = \min_{j \in L}\left(\sum_{\substack{G_i \in S \\ i \le j}} \frac{|r_i|}{N_R} - \sum_{\substack{G_i \notin S \\ i \le j}} \frac{1}{N_{miss}}\right),$$
$$S_{W-GSEA} = \min_{j \in L}\left(\sum_{\substack{G_i \in S \\ i \le j}} \frac{|r_i| * W_i}{N_{RW}} - \sum_{\substack{G_i \notin S \\ i \le j}} \frac{1}{N_{miss}}\right),$$

where $N_R = \sum_{G_i \in S}|r_i|$, $N_{RW} = \sum_{G_i \in S}|r_i| * W_i$, and maxdev(x) is a maximum deviation function. The weight w_i is estimated as the difference between the number of connections and expected number of connections based on the assumption that expected values follow a hypergeometric

distribution. The number of associations between G_i and the K genes in S, is designated as X_i, and the number of associations between G_i and the N genes in the genome, is designated as M_i. Then

$$w_i = X_i - E(X_i) = X_i - \frac{M_i K}{N}$$

$$W_i = \log_2(w_i I_A(w_i) + 2)$$

the weight W_i defined in this way has a minimum value of 1, which is a basic-level weight for genes within a pathway, and I_A is an indicator function. Weights X_i, M_i were estimated based on: BioGrid[73], HPRD[74], DIP [75] ,MINT[76] , IntAct [77]and Reactome[13].

CePa: Centrality-based Pathway enrichment

The Centrality-based Pathway enrichment method [41] extends the original pathway enrichment method (ORA) by introducing network centralities as the weight of nodes which have been mapped from differentially expressed genes in pathways. The main limitation of many enrichment methods comes from the fact they treat genes identically in pathways. Based on pathway structure we can identify how member genes interact with each other. Obviously, changes in expression level of key genes will have more effect in the pathway than insignificant genes. In CePa the importance of genes in pathways is assessed by network centralities. Graph theory provides different measurements to quantify the importance of nodes in these situation genes, such as degree of centrality, which quantifies the number of interacting neighbors; or betweenness, which defines the numbers of information streams passing through a given gene. In computer science language, betweenness expresses the number of shortest paths that pass through vertex – gene. Nodes with high centralities represent metabolites, proteins, or genes which are essential for specific biological processes represented by biological networks in a steady state.

The significance of a pathway is evaluated by a pathway-level statistic, defined as:

$$s = \sum_{i=1}^{n} w_i d_i$$

,where d_i is indicator function, and identifies whether the i^{th} node is affected ($d_i = 1$) or not affected ($d_i = 0$). w_i is the weight of the i^{th} node (reflecting the importance of the node), n is the number of nodes in the pathway. To calculate betweenness [78] W of node i, one first counts the number of shortest paths between two nodes going through node i. Let w_i be the ratio of this number to the total number of shortest paths existing between those two nodes. The sum of w_i over all pairs of nodes in the network represents the betweenness W_i' of the node i. The authors of the CePa algorithm were using the quantity W_i, the scaled W_i' with respect to the maximum possible W(=(n-1)(n-2)/2) in a network having n nodes. The equation for betweenness is given by:

$$W_i = \frac{2W_i'}{(n-1)(n-2)}$$

W_i is positive and always less than or equal to 1 for any network. To calculate the betweenness of the whole graph we need to find an average of the differences of all W_i from the largest value among the n nodes of the graph. To assess the significance of the pathways, the algorithm performs 1000 simulations for the p-value calculation, and the false discovery rate (FDR) is calculated by Benjamini-Hochberg (BH) [48] process.

NetGSA: Network-based Gene Set Analysis

The Network-based Gene Set Analysis (NetGSA) [46, 79, 80] algorithm tests changes in the network structure under different experimental or disease conditions. The authors propose a latent variable model to directly incorporate the underlying gene network, and present test statistics for assessing the significance of arbitrary sub-networks based on the theory of mixed linear models.

The NetGSA method could be represented by the equation:

$$X = \Psi\beta + \Pi\gamma + \varepsilon$$

where X represents data, β is the vector of fixed effects, and γ and ε are random effects and random errors, respectively. The influence matrices

further determine the design matrices Ψ and Π in the mixed linear model. Formally, the influence matrix under each condition represents the effect of each gene on all the other genes in the network and is calculated from the adjacency matrix.

A variety of hypotheses about fixed effect parameters of mixed linear models can be tested by considering tests of the form:

$$H_0: l\beta = 0 \quad vs. \quad H_1: l\beta \neq 0$$

Here l is in general any linear combination of β_S which meets the estimability requirement of [81]. Proposed latent variable model is able to incorporate the change in network structures, correlations among genes, and consider variability in the expression levels of the genes in different conditions.

Description of Enrichment App

EApp (Enrichment Analysis application) is a shiny application deployed via shiny server. The application workflow consist of three possible steps: i). Data upload and preprocessing; ii). Pathway analysis using set-based and network-based methods iii). Visualization and exporting the enrichment analysis results. Data can be uploaded from different resources such as: local ExpressionSets with expression and genes signature data, data from iLINCS project [2], data from GREIN project [3], and finally raw data (separate data frames with rpkms, count data, and phenotypic data necessary for creating gene signatures), example of the menu is presented in figure 5.

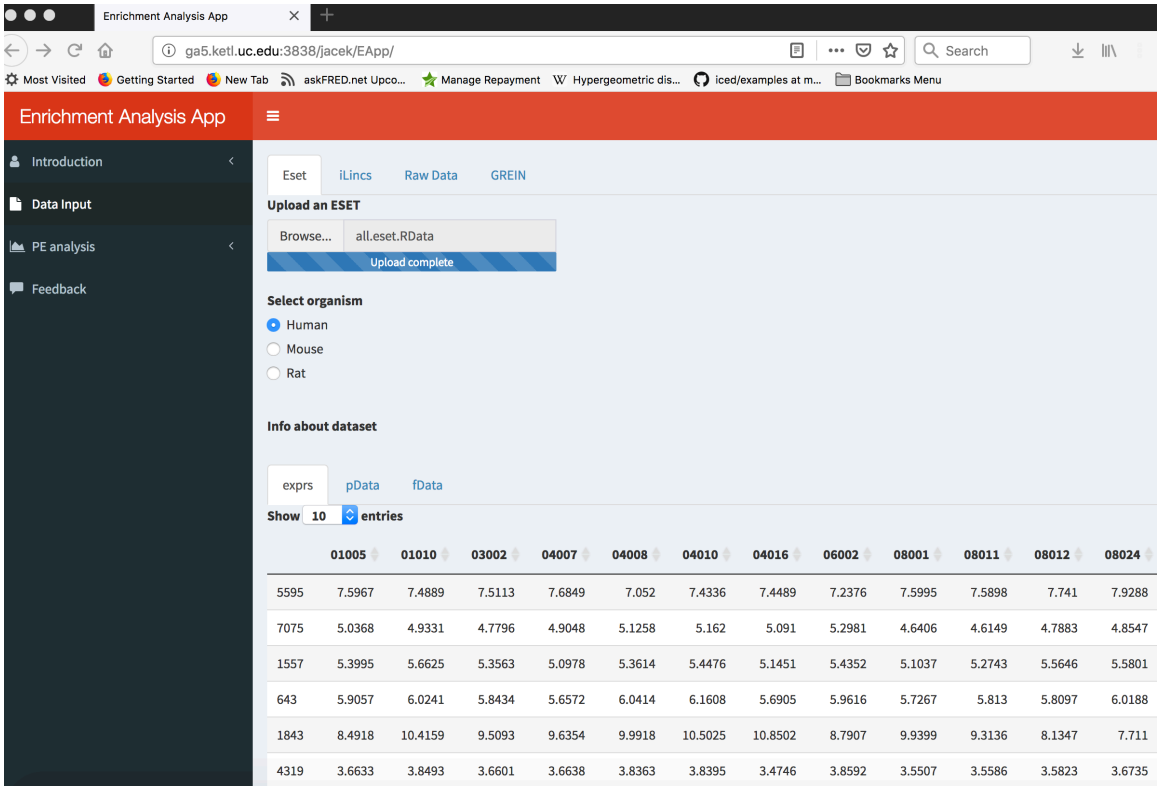

Figure 5. Enrichment Analysis Application (EApp). Data input menu, with example of available sources of data.

Availability of enrichment methods is limited by information included in data. For example, if data include only information about fold changes and related p-values, we can not perform analysis with sample permutation. Data such as signature deposited in iLINCS portal meet this criterion, and consequently could be only analyzed with Fisher's Exact, Random Set, and network-based methods such as SPIA, PathNet, and CePa . Other methods described elsewhere in this thesis are available for data structures in local eset, constructed from raw data, and GREIN. After validation of input parameters such as type of genome, number of permutations in enrichment analysis, value of FDR used for filtering genes used in analysis, selection of database with pathways, the user is able to perform analysis with a preselected algorithm. The results are presented as a table, which could be downloaded as an Excel file. The visualization section allows for examining the interaction matrix, representing information about existence of the gene in particular pathway (1 – in gene is in pathway, 0 if not). The same information is represented by a heatmap created and displayed in the application with the java script plugin Morpheus, which allows interactively modifying and visualizing results. All results could be downloaded in Excel or picture format for inclusion in publication.

Case study with set-, network-based methods available in EApp.

In the following, we demonstrate the application of the EApp (Enrichment Analysis Application shiny server) to RNA-seq data from the Cancer Genome Atlas (TCGA) [82]. In our example we will use data from Uterine Corpus Endometrial Carcinoma (UCEC) [83], and Kidney Renal Clear Cell Carcinoma (KIRC) [84]. Data on uterine corpus endometrial carcinoma (UCEC) [83], which is one of the most common cancers of the female reproductive system, were analyzed by Geistlinger [4] using all data including a 554 UCEC tumor and 35 adjacent normal samples. Our analysis will also include a subset of data stored in iLINCS portal based on a signature derived from 338 UCEC samples, where 5 were solid tissue normal and the rest were tumors. For the long calculation we will also use UCEC data stored in GSEABenchmarkeR R package [85] by Geistlinger [4]. These data were stored as eset with counts, normalized rpkms, meta data, and signature calculated with edgeR algorithm.

Kidney Renal Clear Cell Carcinoma (KIRC)[84] is a second dataset from TCGA project that is considered in evaluation. Renal Clear Cell Carcionma accounts for 75% of all renal cancers.

Our results for UCEC data with set-based methods based on standard signature (100 top genes) using KEGG database confirms the findings from Geistlinger's study. Similarly, we were able identify the pathways involved in cancer development and cell cycle, such as: *p53 signaling pathway (FisherExact(4), RS(4)) , MicroRNAs in cancer(RS(5))*[5]. *Wnt signaling pathway* [86] (hsa04310) was also identified by Fisher Exact method using the full UCEC set (554 samples) using data stored in eset with all 589 samples. This analysis also identified *p53 signaling pathway at position nineteen,* and *MicroRNAs in cancer* on position twenty-five. Algorithms such as CAMERA, CePa, DEGRAPH, EBM, GANPA, GLOBALTEST, MGSA, PADOG, ROAST, and SAMGS failed to find any important pathways due to granularity problems [87] where many sets had the same p-value. The PathNet method used with signatures derived from iLINCS portal, because of a similar problem wasn't able to identify any significant pathway, and in this case an interaction matrix was created based on 978 iLINCS gene signature. Using the PathNet method with global interaction matrix and bigger UCEC set (eset with 589 samples), we were able to identify: *Wnt signaling pathway*[6] *[hsa04310 (20)], MicroRNAs in cancer [*hsa04115(42)*], p53 signaling pathway [*hsa05200(46)*]*[5]. Similar results were achieved with SAFE, where additionally method was able on the top of the list identify hsa05226 - *Gastric cancer pathway* [hsa05226(25)][5].

Several interesting findings were identified by GSVA, GSEA such as: *Ras signaling pathway* (25)[5], *p53 signaling pathway (19), MicroRNAs in cancer (25)*[5], respectively.

Most of the implicated pathways were previously independently identified [88] including: *Notch Signaling pathway, cell cycle, Wnt signaling pathway, MicroRNAs in cancer, HIF-1 signaling pathway, Proteoglycans in cancer.*

Results were generated using EApp, with parts being generated directly with R because of time constraints. Run time for most of the methods take between several minutes to several hours depending on the number of permutations performed and size of dataset.

[5]In the parenthesis is a position on the enrichments results list.

Several important processes are associated with KIRC [89, 90] such as: *HIF-1 signaling pathway, VEGF signaling pathway, MAPK signaling pathway (hsa0401), TGF-β signaling pathway, Citrate cycle, mTOR–PI3K pathway, programmed death-1 receptor pathway (PD-1 is an "anti-immune" protein*, the stimulation of which suppresses the immune system and decreases the number of cytotoxic T cells attacking foreign antigens and cancer cells) , and the glutaminase pathway.

Mentioned processes were identified by several methods. For example out of 318 pathways represented in KEGG (number in parenthesis represents position on the list.):

1. *HIF-1 signaling pathway* was identified by RS (36), GSEA(36), PADOG(152), ROAST(46), SAFE(15), CAMERA(84), EBM(97), GGEA(68), PathNet (2), GSVA(33).

2. *Natural killer cell mediated cytotoxicity* was identified by RS(9), SAFE(2), CAMERA(18), GGEA(24), PATHNET(7), GSVA(9), SPIA(10,6 with 1000 steps in permutation) , GGEA (24), GSVA (9).

3. *PI3K-Akt signaling pathway* was identified by ROAST (3), SAFE(38), EBM(2) – problem with p-value – many methods has the same p-value), GGEA(15,11 with 10000 steps), PATHNET(17).

4. *mTOR signaling pathway* was determined [hsa04150] by EBM(42), MGSA(42) or *MAPK signaling pathway* (hsa0401) – MGSA(2).

5. *Citrate cycle process* was picked out GSEA(15), PADOG(25), SAFE(50), SAMGS(15), CAMERA (7), GSVA(15).

Globaltest, SAMGS ,MGSA, EBM, GGEA, ROAST, PADOG suffered during analysis from granularity problem. For example: Globaltest assigned 0.001 to all results for the first 304 pathways. Similarly, ROAST ranked top value of p-value 0.001 to 228 pathways (algorithm used 1000 permutations).

When using the dataset deposited in iLINCS database, the algorithm based on Fisher Exact test was not able to find any interesting pathways. The Random Set algorithm was able to find on position 7 Natural killer cell mediated cytotoxicity pathway followed by citrate cycle at position 15, HIF-1 signaling pathway at position 35.

An important factor with large datasets is runtime of the algorithm. For example, the UCEC dataset with 1000 steps for network based methods calculation takes from 27 seconds (globaltest) up 41 minutes for GANPA. Similarly, for the KIRC dataset with 1000 steps in permutation results, can be generated in 27 seconds for GGEA, but takes 42 minutes for CePa. When we increase steps to 10000 permutations most algorithms behave linearly with time. For example, topologygsa with 1000 steps takes 21 minutes, where for 10000 steps takes 3.27 hours. One way to resolve the granularity problem is to increase the number of steps in the permutation procedure. For instance, increasing to 10000 steps for PADOG results in 4.25 days of calculations.

In summary, using a fixed length of the top 100 genes from iLINCS signature for both renal and uterine cancer datasets, we were able to find meaningful pathways related to etiology of the disease and findings were comparable to results achieved with the complete dataset with KEGG database.

Evaluation of Pathway Analysis Methods

The several authors have evaluated Enrichments Methods [28, 91, 92] and commented on the problems with the lack of a gold standard in cases of predefined experimental datasets used in evaluation. The experimental data cannot be replaced by approximated or simulated data because real life problems are very complex and often do not follow a normal distribution. Tarca et al. [30, 93] compiled 42 microarray datasets from GEO. For every dataset in metadata they also assigned a specific KEGG pathway related to disease etiology. These datasets are available in the *Bioconductor* packages KEGGdzPathwaysGEO and KEGGandMetacoreDzPathwaysGEO. An additional resource for evaluating Enrichment Methods based on RNA-seq platform is deposited in GSEABenchmarkeR package and includes 24

cancer types consistently preprocessed by Rathman et al. [94] and annotated with MalaCards [95] database of human diseases.

To evaluate Enrichment Methods we can use criteria such as runtime, fraction of statistically significant gene sets, and phenotype relevance. Runtime is an important criterion in the applicability of the method. Runtime depends on implementation, the computational intensiveness of the calculation of gene and pathway level statistics, and finally, the number of permutations used in the estimation of FDR.

The criterion of fraction of statistically significant gene sets evaluates the ability to find relevant pathways, which could be impaired by an unrealistic assumption of independence between genes [21], as well as biases and inaccuracies incorporated in the permutation procedure [33, 96]. This criterion is defined as a number of significant sets – pathways with given threshold of FDR and allows evaluation of resulting fractions of significant gene sets in comparison to other methods. Too many significant pathways/gene sets will suggest a lack of specificity of the method, suboptimal system setup, or other technical problems mentioned above.

The phenotype relevance investigates if there is any association between top-ranked gene sets and the investigated phenotype. Tarca [30, 93] assigned target pathways to each dataset in the GEO2KEGG compendium. A more comprehensive and systematic approach is used in the case of MalaCards, where for every phenotype we have assigned a ranked list of important pathways based on experimental evidence and literature. To evaluate the ability to recover gene sets with high relevance for disease, we can calculate the relevance score of a gene set ranking and compare with a random distribution. The relevance score is calculated based on weights of relevance from enrichment analysis, where $w = 1-r/N$, N – number of pathways used in enrichment analysis and r position in the ranking. Those weights are multiplied by the corresponding relevance score, summed, and divided by the optimal relevance score. A higher value shows better recovery of information.

We tested the runtime of 20 implemented methods (12 set-based and 8 network-based methods). Estimated runtime is presented in Figure 6, with the set-based method on the left panel, and the network-based enrichment methods on the right. Runtime of the methods depends mainly on whether permutation testing is used to estimate gene set significance. A second important factor influencing runtime is quality of implementation. On average, set-based methods are significantly faster than network-based

methods, with the exception of GGEA, DEAGRAPH, and GANPA, where time is comparable to PADOG and slightly faster then GSEA. Time for set-based methods vary on average from 0.5 second up to 5 minutes (300 seconds), whereas for network-based methods runtime ranges from 3 seconds to 8.5 hours (31622 seconds).

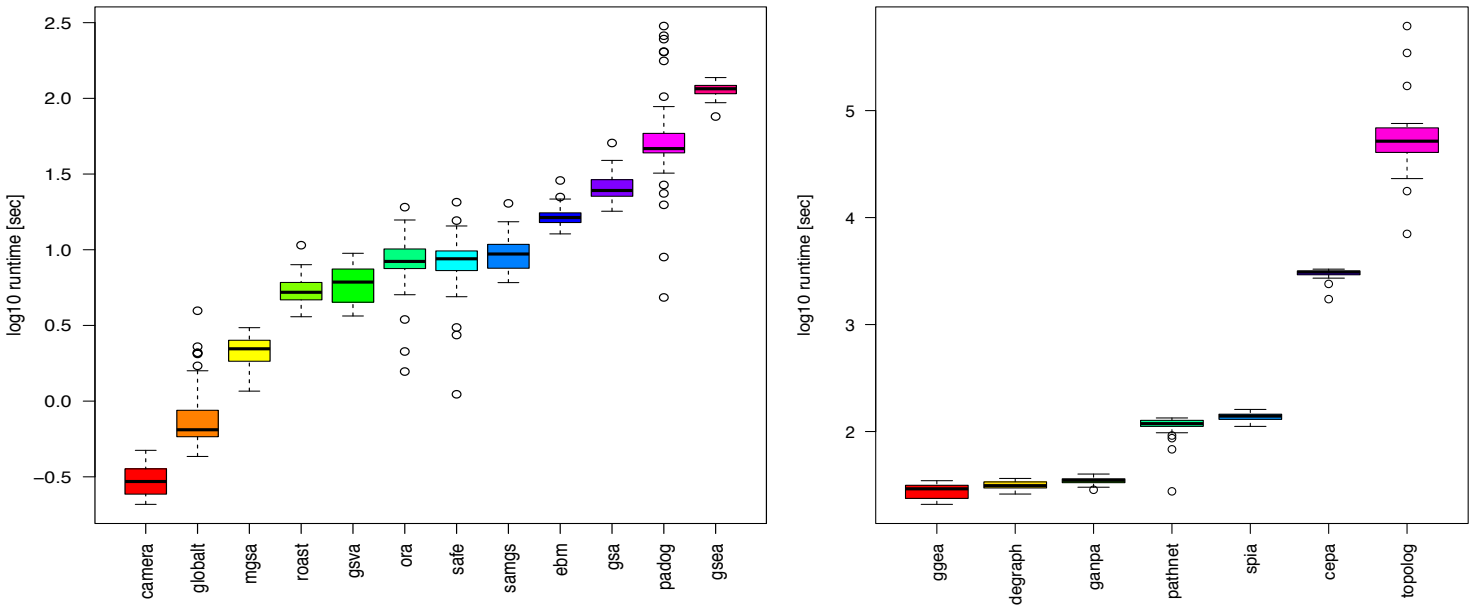

Figure 6. Elapsed processing times (*y*-axis, log-scale) when applying the enrichment methods indicated on the *x*-axis to the 42 datasets of the GEO2KEGG microarray compendium. Gene sets were defined according to KEGG (323 gene sets). Left panel represents runtime for set-based enrichment method and right panel represents runtime for network-based enrichment method.

Assessing similarity between the Enrichment Analysis rankings and the precompiled relevance rankings gives additional information about certain EA methods if those tend to produce rankings of higher phenotype relevance. Results for datasets combined in GEO2KEGG by Tarca are represented in figure 7. When a box-plot is located in the upper part of the y-axis, that method tends to recover more phenotype-relevant gene sets than a method closer to the bottom of y-axis.

Competitive methods (ORA, GSEA, SAFE, CAMERA, PADOG) tend to rank phenotype-relevant pathways/gene sets systematically higher

than self-contained methods (GLOBALTEST, SAMGS). Surprisingly, network-based methods were not significantly better than set-based. The DEGRAPH method was the worst out of all methods, and only GGEA, PATHNET methods were comparable to set-based methods.

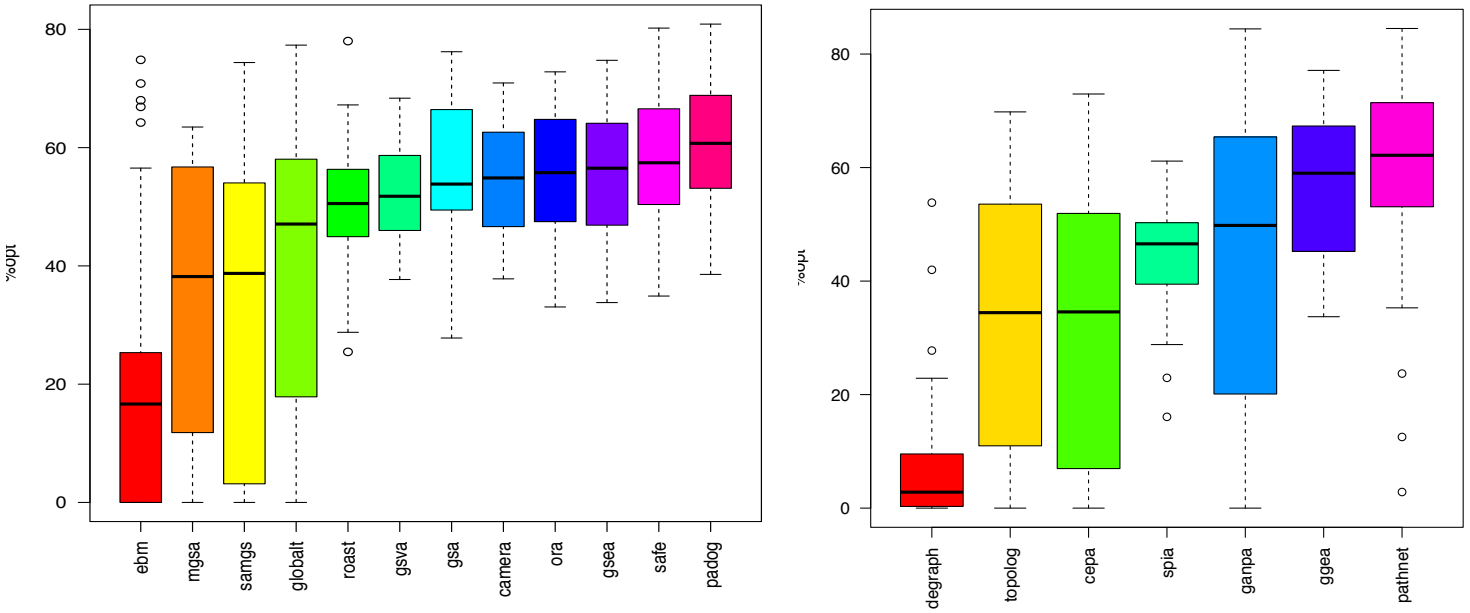

Figure 7. Phenotype relevance. Percentage of the optimal phenotype relevance score (*y*-axis) when applying methods to the GEO2KEGG microarray compendium (top, 42 datasets) Gene sets were defined according to KEGG (323 gene sets) The phenotype relevance score of a method *m* applied to a dataset *d* is the sum of the gene set relevance scores, weighted by the relative position of each gene set in the ranking of method *m*. Left panel represents phenotype relevance for set-based enrichment method and right panel represents phenotype relevance for network-based enrichment method.

This criterion of fraction of statistically significant gene sets, is the last criterion to be evaluated in the Genes Set Enrichment method performance. This criterion finds different subsets of significant genes sets and these results could be correlated with the type of null hypothesis tested by the method, and the performance of the given method at particular parameters. A fraction that is too high could suggest granularity problems or lack of specificity. Results are presented in figure 8. Competitive methods are reporting much smaller fractions of significant gene sets. ORA, SAFE, and EBM reported not a single significant gene set where two methods from set-based (GLOBAL, SAMGS) and three network-based methods (CePa, GANPA, TOPOLOGYGSA) reported all sets to be significant at FDR=0.05.

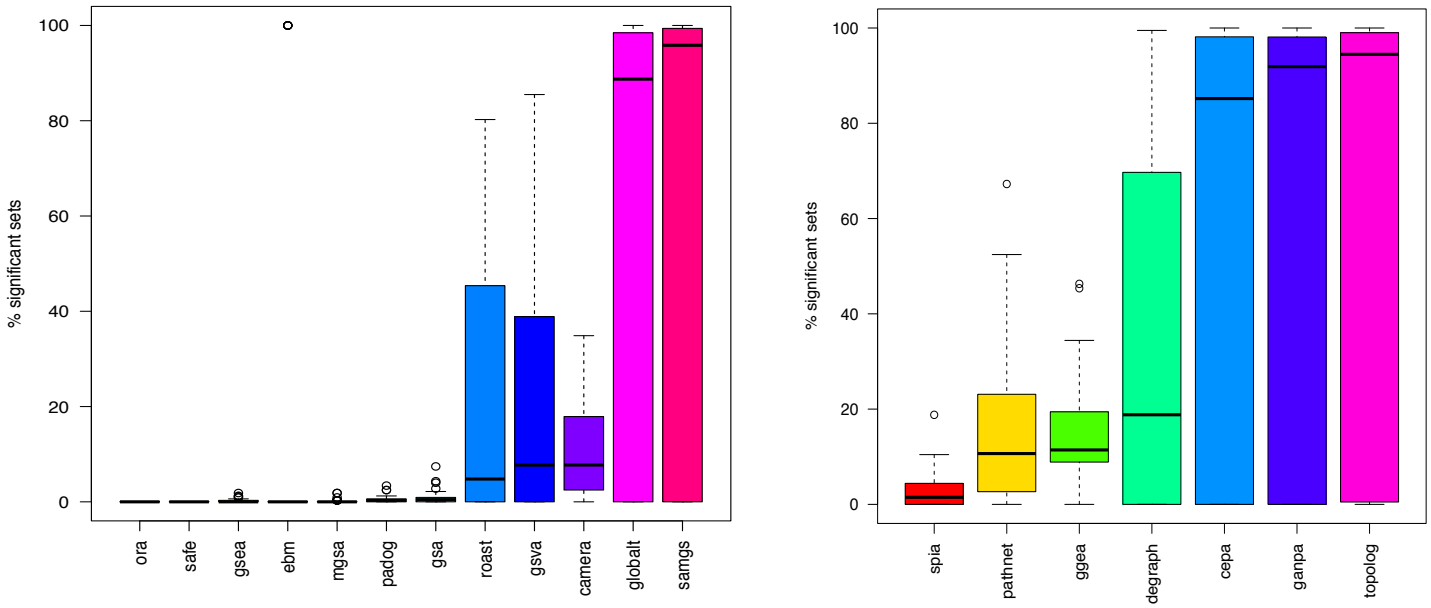

Figure 8. Statistical significance. Percentage of significant gene sets (FDR <0.05, *y*-axis) when applying methods to the GEO2KEGG microarray compendium (top, 42 datasets. Gene sets were defined according to KEGG (left, 323 gene sets). Left panel represents percent of significant sets for set-based enrichment method and right panel represents statistical significance for network-based enrichment method.

Conclusion

EApp (Enrichment Analysis application) shiny application is accessible by iLINCS project and also as an independent server under URL: htpp://gimm12.ketl.uc.edu:3090/EApp. Shiny application allows using an extended number of enrichment analysis methods to users with limited coding skills.

Based on our and previous studies [85], we recommend methods based on Fisher Exact test / ORA methods for explanatory analysis, which are characterized by short runtime and simple interpretation, and give very similar results to more complex and advanced methods.

Alternatives for ORA methods will be ROAST and CAMERA, for which results are comparable, and in some cases significantly faster while not requiring re-sampling.

For identifying relevant pathways PADOG (**P***athway* **A***nalysis with* **D***own-weighting of* **O***verlapping* **G***enes)* method tends to rank relevant gene sets systematically higher than other methods that account for gene set overlaps, A similar situation was observed for the PathNet method, but PADOG was slightly better in runtime and fraction of significant gene sets. Unfortunately, there are few methods, which account for genes sets overlap, such as GSEA with integration of gene Appearance Frequency (GSEA-AF). Evaluation was not possible because GSEA-AF is lacking in R implementation

Unexpectedly, network-based methods were not significantly better than set-based, which could be related to incompleteness and quality of pathways stored in databases and complexity of biological processes represented by the datasets used in analysis.

Further work will include several improvements and implementation of approximated fastest methods suitable for large datasets and replacing the permutation method:

1. Perform more independent benchmarking following GSEABenchmarkeR using TCGA data [85]
2. Add to server other methods based on diffusion kernels [97]
3. Implement approximated method for methods, which require full data not only gene score statistics, such as: GSEA, SPIA.
4. Add for iLINCS signatures methods which allow gene permutation such as: FGSEA - GSEA with adaptive multi-level split Monte-Carlo scheme [98] or npGSEA (non-permutation GSEA) [87]
5. Implement consensus mode for enrichment analysis and comparison results with single method approach.
6. Add methods with correlation network analysis such as: WGNCA [99] to automatically visualize interaction genes in pathways and between pathways.

Appendix A.

Comparison of methods used in EApp.

Name	Availability	Pathway representation	Cathegory	Package	Year	Database support	Null hypothesis	Reference
ORA/FisherExact	Bioconductor	GS	ORA	CLEAN	2007	KEGG, GO	competitive	[21,27]
safe	Bioconductor	GS	FCS	safe	2005	KEGG, GO	competitive/self-contained	[28,47]
RS	CRAN	GS	FCS	CLEAN	2007	KEGG, GO	competitive	[27,68]
padog	Bioconductor	GS	FCS	PADOG	2012	KEGG, GO	competitive	[30]
roast	Bioconductor	GS	FCS	limma	2010	KEGG, GO	self-contained	[31]
camera	Bioconductor	GS	ORA	limma	2012	KEGG, GO	competitive	[32]
gsa	CRAN	GS	FCS	GSA	2007	KEGG, GO	competitive	[33]
gsva	Bioconductor	GS	FCS	GSVA	2013	KEGG, GO	self-contained	[34]
globaltest	Bioconductor	GS	FCS	globaltest	2006	KEGG, GO	self-contained	[35]
samgs	Bioconductor	GS	FCS	limma	2007	KEGG, GO	self-contained	[39]
ebm	Bioconductor	GS	FCS	EmpiricalBrownsMethod	2016	KEGG, GO	competitive/self-contained	[36]
mgsa	Bioconductor	GS	ORA	mgsa	2010	KEGG, GO		[37,38]
SPIA	Bioconductor	PT-based	PTB	spia	2009	KEGG	competitive	[40]
GSEA	Bioconductor	GS	FCS	Enrichment Browser, fgsea, npgsea	2005	KEGG	competitive/self-contained	[29,87,98]
PathNet	Bioconductor	PT-based	PTB	PathNet	2012	KEGG	competitive	[37]
DEGraph	Bioconductor	PT-based	PTB	DEGraph	2010	KEGG	competitive/self-contained	[43]
TopologyGSA	CRAN	PT-based	PTB	TopologyGSA	2010	KEGG	competitive/self-contained	[45]
GANPA	CRAN	PT-based	PTB	GANPA	2011	KEGG	self-contanied	[44]
CePa	CRAN	PT-based	PTB	CePa	2012	KEGG	self-contained/competitive	[41]
NetGSA	CRAN	PT-based	PTB	netgsa	2016	KEGG	competitive	[46]
GGEA	Bioconductor	PT-based	PTB	ggea	2011	KEGG	competitive/self-contained	[42]

References:

1. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS computational biology, 2012. **8**: p. e1002375-e1002375.
2. Nguyen, T., C. Mitrea, and S. Draghici, *Network-Based Approaches for Pathway Level Analysis.* Current Protocols in Bioinformatics, 2018. **61**(1): p. 8.25.1-8.25.24.
3. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**: p. 1-13.
4. Geistlinger, L., G. Csaba, and R. Zimmer, *Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis.* BMC Bioinformatics, 2016. **17**(1): p. 45.
5. Liberzon, A., et al., *The Molecular Signatures Database Hallmark Gene Set Collection.* Cell Systems, 2015. **1**(6): p. 417-425.
6. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic acids research, 2015. **44**(D1): p. D457-D462.
7. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG.* Nucleic acids research, 2013. **42**(D1): p. D199-D205.
8. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic acids research, 2011. **40**(D1): p. D109-D114.
9. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic acids research, 2004. **32**(suppl1): p. D277-D280.
10. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes.* Nucleic acids research, 2017. **46**(D1): p. D633-D639.
11. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic acids research, 2011. **40**(D1): p. D742-D753.
12. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic acids research, 2009. **38**(suppl1): p. D473-D479.
13. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Research, 2018. **46**(D1): p. D649-D655.
14. Santos-Zavaleta, A., et al., *RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12.* Nucleic Acids Research, 2018. **47**(D1): p. D212-D220.
15. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome research, 2003. **13**(9): p. 2129-2141.

16. Snel, B., et al., *STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.* Nucleic acids research, 2000. **28**(18): p. 3442-3444.
17. Szklarczyk, D., et al., *STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Research, 2018. **47**(D1): p. D607-D613.
18. Slenter, D.N., et al., *WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research.* Nucleic acids research, 2018. **46**(D1): p. D661-D667.
19. Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic acids research, 2009. **37**(Database issue): p. D674-D679.
20. Langsrud, V., *Rotation tests.* Statistics and Computing, 2005. **15**(1): p. 53-60.
21. Goeman, J.J. and P. Bühlmann, *Analyzing gene expression data in terms of gene sets: methodological issues.* Bioinformatics (Oxford, England), 2007. **23**: p. 980-987.
22. García-Campos, M.A., J. Espinal-Enríquez, and E. Hernández-Lemus, *Pathway Analysis: State of the Art.* Frontiers in physiology, 2015. **6**: p. 383-383.
23. Lascorz, J., et al., *Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies.* PLoS One, 2011. **6**(4): p. e18867-e18867.
24. Lascorz, J.s., K. Hemminki, and A. Fˆrsti, *Systematic enrichment analysis of gene expression profiling studies identifies consensus pathways implicated in colorectal cancer development.* Journal of Carcinogenesis, 2011. **10**(1): p. 7-7.
25. Alhamdoosh, M., et al., *Combining multiple tools outperforms individual methods in gene set enrichment analyses.* Bioinformatics, 2017. **33**(3): p. 414-424.
26. Alhamdoosh, M., et al., *Easy and efficient ensemble gene set testing with EGSEA.* F1000Res, 2017. **6**: p. 2010.
27. Freudenberg, J.M., et al., *CLEAN: CLustering Enrichment ANalysis.* BMC Bioinformatics, 2009. **10**(1): p. 234.
28. Barry, W.T., A.B. Nobel, and F.A. Wright, *Significance analysis of functional categories in gene expression studies: a structured permutation approach.* Bioinformatics, 2005. **21**(9): p. 1943-9.
29. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545.
30. Tarca, A.L., et al., *Down-weighting overlapping genes improves gene set analysis.* BMC Bioinformatics, 2012. **13**(1): p. 136.
31. Wu, D., et al., *ROAST: rotation gene set tests for complex microarray experiments.* Bioinformatics, 2010. **26**(17): p. 2176-82.
32. Wu, D. and G.K. Smyth, *Camera: a competitive gene set test accounting for inter-gene correlation.* Nucleic Acids Research, 2012. **40**(17): p. e133-e133.
33. Efron, B. and R. Tibshirani, *On testing the significance of sets of genes.* Ann. Appl. Stat., 2007. **1**(1): p. 107-129.
34. Hänzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-Seq data.* BMC Bioinformatics, 2013. **14**(1): p. 7.

35. Goeman, J.J., et al., *A global test for groups of genes: testing association with a clinical outcome.* Bioinformatics, 2004. **20**(1): p. 93-9.
36. Poole, W., et al., *Combining dependent P-values with an empirical adaptation of Brown's method.* Bioinformatics (Oxford, England), 2016. **32**(17): p. i430-i436.
37. Dutta, B., A. Wallqvist, and J. Reifman, *PathNet: a tool for pathway analysis using topological information.* Source Code for Biology and Medicine, 2012. **7**(1): p. 10.
38. Bauer, S., J. Gagneur, and P.N. Robinson, *GOing Bayesian: model-based gene set analysis of genome-scale data.* Nucleic acids research, 2010. **38**(11): p. 3523-3532.
39. Dinu, I., et al., *Improving gene set analysis of microarray data by SAM-GS.* BMC Bioinformatics, 2007. **8**(1): p. 242.
40. Tarca, A.L., et al., *A novel signaling pathway impact analysis.* Bioinformatics, 2009. **25**(1): p. 75-82.
41. Gu, Z., et al., *Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes.* BMC systems biology, 2012. **6**: p. 56-56.
42. Geistlinger, L., et al., *From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems.* Bioinformatics, 2011. **27**(13): p. i366-i373.
43. Jacob, L., P. Neuvial, and S. Dudoit, *More power via graph-structured tests for differential expression of gene networks.* Ann. Appl. Stat., 2012. **6**(2): p. 561-600.
44. Fang, Z., W. Tian, and H. Ji, *A network-based gene-weighting approach for pathway analysis.* Cell research, 2012. **22**(3): p. 565-580.
45. Massa, M.S., M. Chiogna, and C. Romualdi, *Gene set analysis exploiting the topology of a pathway.* BMC systems biology, 2010. **4**: p. 121-121.
46. Ma, J., A. Shojaie, and G. Michailidis, *Network-based pathway enrichment analysis with incomplete network information.* Bioinformatics, 2016. **32**(20): p. 3165-3174.
47. Barry, W.T., A.B. Nobel, and F.A. Wright, *A statistical framework for testing functional categories in microarray data.* Ann. Appl. Stat., 2008. **2**(1): p. 286-315.
48. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
49. Yekutieli, D. and Y. Benjamini, *Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics.* Journal of Statistical Planning and Inference, 1999. **82**(1): p. 171-196.
50. Holm, S., *A Simple Sequentially Rejective Multiple Test Procedure.* Scandinavian Journal of Statistics, 1979. **6**(2): p. 65-70.
51. Westfall, P.H. and S.S. Young, *p Value Adjustments for Multiple Tests in Multivariate Binomial Models.* Journal of the American Statistical Association, 1989. **84**(407): p. 780-786.
52. Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Plans.* Report EFS NSF 163, 1980.

53. Benjamini, Y., et al., *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 2001. **125**(1-2): p. 279-84.
54. Reiner, A., D. Yekutieli, and Y. Benjamini, *Identifying differentially expressed genes using false discovery rate controlling procedures.* Bioinformatics, 2003. **19**(3): p. 368-75.
55. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
56. Ackermann, M. and K. Strimmer, *A general modular framework for gene set enrichment analysis.* BMC Bioinformatics, 2009. **10**(1): p. 47.
57. Jiang, Z. and R. Gentleman, *Extensions to gene set enrichment.* Bioinformatics, 2006. **23**(3): p. 306-313.
58. Langsrud, Ø., *Rotation tests.* Statistics and Computing, 2005. **15**(1): p. 53-60.
59. Gatti, D.M., et al., *Heading Down the Wrong Pathway: on the Influence of Correlation within Gene Sets.* BMC Genomics, 2010. **11**(1): p. 574.
60. Breslin, T., P. Edén, and M. Krogh, *Comparing functional annotation analyses with Catmap.* BMC Bioinformatics, 2004. **5**(1): p. 193.
61. Dorum, G., et al., *Rotation testing in gene set enrichment analysis for small direct comparison experiments.* Stat Appl Genet Mol Biol, 2009. **8**: p. Article34.
62. le Cessie, S. and H.C. van Houwelingen, *Testing the Fit of a Regression Model Via Score Tests in Random Effects Models.* Biometrics, 1995. **51**(2): p. 600-614.
63. Houwing-Duistermaat, J.J., et al., *Testing Familial Aggregation.* Biometrics, 1995. **51**(4): p. 1292-1301.
64. Kost, J.T. and M.P. McDermott, *Combining dependent P-values.* Statistics & Probability Letters, 2002. **60**(2): p. 183-190.
65. Brown, M.B., *400: A Method for Combining Non-Independent, One-Sided Tests of Significance.* Biometrics, 1975. **31**(4): p. 987-992.
66. Andrieu, C., et al., *An Introduction to MCMC for Machine Learning.* Machine Learning, 2003. **50**(1): p. 5-43.
67. Sengupta, S., et al., *Genome-Wide Expression Profiling Reveals EBV-Associated Inhibition of MHC Class I Expression in Nasopharyngeal Carcinoma.* Cancer Research, 2006. **66**(16): p. 7999.
68. Newton, M.A., et al., *Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis.* Ann. Appl. Stat., 2007. **1**(1): p. 85-106.
69. R. Schott, J., *Some tests for the equality of covariance matrices.* Journal of Statistical Planning and Inference, 2001. **94**(1): p. 25-36.
70. Tsai, C.-A. and J.J. Chen, *Multivariate analysis of variance test for gene set analysis.* Bioinformatics, 2009. **25**(7): p. 897-903.
71. Gamage, J., T. Mathew, and S. Weerahandi, *Generalized p-values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA.* Journal of Multivariate Analysis, 2004. **88**(1): p. 177-189.
72. Krishnamoorthy, K. and J. Yu, *Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem.* Statistics & Probability Letters, 2004. **66**(2): p. 161-169.

73. ChatrAryamontri, A., et al., *The BioGRID interaction database: 2017 update.* Nucleic Acids Res, 2017. **45**(D1): p. D369-d379.
74. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
75. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
76. Chatraryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic acids research, 2007. **35**(Database issue): p. D572-D574.
77. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic acids research, 2004. **32**(Database issue): p. D452-D455.
78. Joy, M.P., et al., *High-betweenness proteins in the yeast protein interaction network.* Journal of biomedicine & biotechnology, 2005. **2005**(2): p. 96-103.
79. Shojaie, A. and G. Michailidis, *Network enrichment analysis in complex experiments.* Stat Appl Genet Mol Biol, 2010. **9**: p. Article22.
80. Shojaie, A. and G. Michailidis, *Analysis of gene sets based on the underlying regulatory network.* Journal of computational biology : a journal of computational molecular cell biology, 2009. **16**(3): p. 407-426.
81. Searle, S.R., *Linear Models. John Wiley & Sons, Inc., New York-London-Sydney-Toronto 1971. XXI, 532 S. $9.50.* Biometrische Zeitschrift, 1974. **16**(1): p. 78-79.
82. Chang, K., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nature Genetics, 2013. **45**(10): p. 1113-1120.
83. Levine, D.A., et al., *Integrated genomic characterization of endometrial carcinoma.* Nature, 2013. **497**(7447): p. 67-73.
84. Network, T.C.G.A.R., *Comprehensive molecular characterization of clear cell renal cell carcinoma.* Nature, 2013. **499**(7456): p. 43-9.
85. Geistlinger, L., et al., *Towards a gold standard for benchmarking gene set enrichment analysis.* bioRxiv, 2019: p. 674267.
86. Kośla, K., et al., *A Novel Set of WNT Pathway Effectors as a Predictive Marker of Uterine Corpus Endometrial Carcinoma–Study Based on Weighted Co-expression Matrices.* Frontiers in Oncology, 2019. **9**(360).
87. Larson, J.L. and A.B. Owen, *Moment based gene set tests.* BMC Bioinformatics, 2015. **16**: p. 132.
88. Izzi, V., et al., *Pan-Cancer analysis of the expression and regulation of matrisome genes across 32 tumor types.* Matrix Biology Plus, 2019. **1**: p. 100004.
89. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic acids research, 2016. **45**(D1): p. D353-D361.
90. Nabi, S., et al., *Renal cell carcinoma: a review of biology and pathophysiology.* F1000Research, 2018. **7**: p. 307-307.
91. Liu, Q., et al., *Comparative evaluation of gene-set analysis methods.* BMC Bioinformatics, 2007. **8**(1): p. 431.
92. Hung, J.-H., et al., *Gene set enrichment analysis: performance evaluation and usage guidelines.* Briefings in Bioinformatics, 2012. **13**(3): p. 281-291.

93. Tarca, A.L., G. Bhatti, and R. Romero, *A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity.* PloS one, 2013. **8**: p. e79217-e79217.
94. Rahman, M., et al., *Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results.* Bioinformatics, 2015. **31**(22): p. 3666-3672.
95. Rappaport, N., et al., *Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect.* BioMedical Engineering OnLine, 2017. **16**(1): p. 72.
96. Phipson, B. and K. Smyth Gordon, *Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn*, in *Statistical Applications in Genetics and Molecular Biology*. 2010.
97. Ren, Y., et al., *Predicting mechanism of action of cellular perturbations with pathway activity signatures.* bioRxiv, 2019: p. 705228.
98. Korotkevich, G., V. Sukhov, and A. Sergushichev, *Fast gene set enrichment analysis.* bioRxiv, 2019: p. 060012.
99. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.