

Generalized Random Set Framework for Functional Enrichment Analysis Using Primary Genomics Datasets

Johannes M. Freudenberg¹, Siva Sivaganesan², Mukta Phatak¹, Kaustubh Shinde¹, and Mario Medvedovic^{1,*}

¹ Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA.

² Mathematical Sciences Department, University of Cincinnati, Cincinnati, OH 45221, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Functional enrichment analysis using primary genomics datasets is an emerging approach to complement established methods for functional enrichment based on pre-defined lists of functionally related genes. Currently used methods depend on creating lists of “significant” and “non-significant” genes based on ad-hoc significance cutoffs. This can lead to loss of statistical power and can introduce biases affecting the interpretation of experimental results.

Results: We developed and validated a new statistical framework, Generalized Random Set analysis (GRS), for comparing the genomic signatures in two datasets without the need for gene categorization. In our tests GRS produced correct measures of statistical significance, and it showed dramatic improvement in the statistical power over four existing methods. We also developed a procedure for identifying genes driving the concordance of the genomics profiles and demonstrated a dramatic improvement in functional coherence of genes identified in such analysis.

Availability: GRS can be downloaded as part of the R package CLEAN from <http://ClusterAnalysis.org/>. An online implementation is available at <http://GenomicsPortals.org/>.

Contact: Mario.Medvedovic@uc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Elucidating the functional significance of differences in gene expression, or of patterns of different gene expression regulatory events in the context of existing knowledge about gene function has become the most challenging, but also the most rewarding aspect of the genomics data analysis (Rhodes and Chinnaiyan, 2005). The most commonly used strategy relies on sets of reference lists containing functionally related genes, such as Gene Ontologies (GO) (Ashburner *et al.*, 2000) and KEGG pathways (Kanehisa and Goto, 2000), to identify functional categories enriched by the differentially regulated genes.

Statistical methods and computational procedures for identifying functionally related sets of genes that are associated with new experimental data have been studied in detail (Ackermann and Strimmer, 2009; Newton *et al.*, 2007; Sartor *et al.*, 2009; Subramanian *et al.*, 2005; Tian *et al.*, 2005). Two basic approaches are used to establish statistical significance of such enrichment. The first approach relies on counting the number of genes in a given functional category that are also differentially expressed, and using the Fisher’s exact or Chi-squared tests to establish the statistical significance of such overlaps. The inherent limitation of this type of analysis is the requirement to choose statistical significance cutoff levels for differential expression. This can reduce the statistical power of the analysis (Ackermann and Strimmer, 2009; Newton *et al.*, 2007; Sartor *et al.*, 2009; Subramanian *et al.*, 2005), and different threshold choices may lead to dramatically different enriched categories (Pan *et al.*, 2005). The second approach uses the complete distribution of differential expressions of all genes to identify enriched gene lists, without categorizing them into differentially and non-differentially expressed. Statistically, these methods use either rank-based Kolmogorov-Smirnov-like tests (Subramanian *et al.*, 2005), or the traditional location shift tests (Newton *et al.*, 2007; Sartor *et al.*, 2009; Tian *et al.*, 2005) to identify lists of functionally related genes that are “more differentially expressed” than a randomly drawn list of genes of the same length. Systematic performance assessments established that, in general, “location shift” methods using all data outperform Chi-squared- and Kolmogorov-Smirnov-like methods (Ackermann and Strimmer, 2009; Sartor *et al.*, 2009; Tian *et al.*, 2005).

Comparing a new, experimentally derived gene list to other, pre-defined lists of functionally related genes, while widely used, has its limitations. Despite the large number of such gene lists, they are often not adequate for precisely characterizing functional consequences of experimentally derived genomics profiles. First, functional relationship does not necessary imply co-expression or co-regulation. For example, only a fraction of genes associated with the same GO terms also exhibit coordinated gene expression (Wren, 2009). Consequently, a potentially large portion of genes belonging to a functional category or pathway will never be informative about the association between a genomics profile and the

*To whom correspondence should be addressed.

pathway. Second, lists associated with a biological concept or a process are often incomplete (Pena-Castillo et al., 2008).

The alternative is to use the vast array of public domain genomics datasets to perform functional enrichment analysis by primary genomics data. In this approach, newly generated genomics profiles are directly compared to profiles from a large collection of reference genomics datasets. Functional interpretation of new data is then based on phenotypic characteristics of the reference datasets that have concordant genomics profiles. A number of methods use this approach including EXALT (Yi et al., 2007), EPSA (Tenenbaum et al., 2008), GEMS (Li et al., 2008), and GEM-TREND (Feng et al., 2009). All these methods categorize genes into “significant” and “non-significant” based on a fixed significance threshold (e.g. FDR<0.1) which can result in loss of statistical power as well as biases as is the case in the traditional analysis of gene sets.

Here, we introduce a new statistical framework, Generalized Random Set (GRS) analysis, to assess the concordance in genomics profiles between two datasets. GRS extends the Random Set (RS) method for functional enrichment analysis by gene lists (Newton et al., 2007) and does not require specification of a significance cutoff for neither the query signature nor the reference datasets. To the best of our knowledge, this is the first method without such a requirement. We first describe our approach in detail and then compare our new method to other existing procedures which employ primary genomics data in this way. We find that GRS outperforms the other methods due in part to the loss of statistical power associated with the categorization of genes that is necessary for the other methods. GRS is implemented in CLEAN R package (available at <http://ClusterAnalysis.org/>), and the online version is available through Genomics Portals (<http://GenomicsPortals.org/>) (Shinde et al., 2010).

2 METHODS

2.1 Generalized Random Set Analysis

Suppose that we calculated measures of differential expression and associated statistical significances for a set of genes G and let s_g denote a measure of the level of differential expression for gene g to be used in the analysis. In our case we set

$$s_g = -\log_{10}(p\text{-value}_g) \quad (1)$$

Let p_g be an index variable for the membership of each gene in a specific functional category F . p_g is set to 1 if gene g is the member of the functional category F and it is set to 0 otherwise:

$$p_g = \begin{cases} 1 & \text{if } g \in F \\ 0 & \text{if } g \notin F \end{cases} \quad (2)$$

The Random Set statistics, measuring the overall level of differential expression for genes in F , is defined as the average score for the genes in F .

$$\bar{X} = \frac{\sum_g p_g s_g}{\sum_g p_g} \quad (3)$$

Under the null hypothesis that there is no enrichment of differentially expressed genes among the genes in category F , the RS statistics is approximately distributed as the normal random variable with the mean μ being equal to the average of s_g over all genes and the variance being derived using the simple delta method (Casella and Berger, 2001):

$$\bar{X} \sim N(\mu, \sigma^2), \text{ where}$$

$$\mu = \frac{\sum_g s_g}{|G|}, \text{ and} \quad (4)$$

$$\sigma^2 = \frac{1}{|F|} \left(\frac{|G| - |F|}{|G| - 1} \left(\frac{\sum_g s_g^2}{|G|} - \left(\frac{\sum_g s_g}{|G|} \right)^2 \right) \right)$$

This method (Newton et al., 2007) can be applied to our problem by simply using one genomics datasets to create the functional category F by declaring only genes differentially expressed for some cut-off to be members of the category F .

Now, suppose we have probabilities of differential expression for each gene g in two datasets (p_g^i and p_g^j). These probabilities can be estimated based on the p -values for differential expression (Sellke et al., 2001). To avoid the need for categorizing genes into “significant” and “non-significant”, we propose to compute a statistic E_{12} by replacing the index variable p_g with p_g^i while the score s_g for the other dataset remains the same as defined in equation (1). To make the statistics symmetric with respect to two datasets we also compute the statistics E_{21} in which the scores and probabilities are reversed and define the overall Generalized Random Set (GRS) statistic as the average of the two:

$$E = \frac{E_{12} + E_{21}}{2}, \text{ where} \quad (5)$$

$$E_{ij} = \frac{\sum_g p_g^i s_g^j}{\sum_g p_g^i}$$

We derive the approximate distribution for the test statistic Z_E , a standardized version of E , under the null hypothesis of no concordance between the two datasets:

$$Z_E = \sqrt{|G|} \left(\frac{E - \mu_E}{\sigma_E} \right), Z_E \sim N(0,1), \text{ where}$$

$$\mu_E = \frac{\sum_g s_g^1 + \sum_g s_g^2}{2|G|}, \text{ and} \quad (6)$$

$$\sigma_E^2 = \delta \cdot \Sigma \delta$$

That is, σ_E is approximated using the multivariate delta method (Casella and Berger, 2001) such that $\delta = 1/2 \cdot (\bar{p}_1^{-1}, \bar{s}_2 \bar{p}_1^{-1}, \bar{p}_2^{-1}, \bar{s}_1 \bar{p}_2^{-1})'$ and

Σ is the estimated variance-covariance matrix of the random variable $X = (X_1, X_2, X_3, X_4)'$, $X_1 = p_1 s_2$, $X_2 = p_1$, $X_3 = p_2 s_1$, and $X_4 = p_2$.

Details of the derivation can be found in the supplemental material.

2.2 Estimating Probability of Differential Expression

We approximate the probabilities of differential expression based on the p -values of differential expression (Sellke et al., 2001). The posterior probability of a gene being differentially expressed is the 1-posterior probability of the null hypothesis and can be estimated as:

$$p_{post}^{H_1} = 1 - p_{post}^{H_0} = 1 - \frac{B}{1 + B}, \text{ where} \quad (7)$$

$$B = \begin{cases} -e \cdot p\text{-value} \cdot \log(p\text{-value}) & \text{if } p\text{-value} < e^{-1} \\ 1 & \text{if } p\text{-value} \geq e^{-1} \end{cases} \quad (8)$$

However, equation (8) implies that, a priori, H_0 and H_1 are equally likely. In other words, the proportion m_0 of not differentially expressed genes is assumed to be 0.5. Using the False Discovery Rate approach (Storey and

Tibshirani, 2003) we estimate m_0 from the data and modify equation (7) as follows:

$$P_{post}^{H_1} = 1 - \frac{1}{1 + \left(\frac{m_0}{1 - m_0} B\right)^{-1}} \quad (9)$$

2.3 Identifying genes with concordant patterns

In addition to assessing the existence of the concordant gene expression patterns, it is important to identify genes responsible for this concordance. A straightforward way to rank genes based on the likelihood of concordance is to use the scaled measure E_g of the individual contribution of gene g to the GRS statistics:

$$E_g = \frac{|G|}{2} \left(\frac{P_g^1 s_g^2}{\sum_g P_g^1} + \frac{P_g^2 s_g^1}{\sum_g P_g^2} \right) \quad (10)$$

where $|G|$ is the total number of genes.

We estimate the null distribution of E_g by first randomly re-assigning gene labels in both, the query and the reference signature and then re-computing E_g for each gene. After repeating this procedure n times, the resulting quantiles are averaged in order to describe the E_g null distribution.

2.4 Data preprocessing, differential expression profiles, and gene matching

We use two primary breast cancer gene expression datasets (Miller *et al.*, 2005; Schmidt *et al.*, 2008) and the collection of genome-wide expression experiments systematically assessing small molecule perturbations in vitro (“Connectivity Map”) (Lamb *et al.*, 2006) in the analysis. Raw data is preprocessed using the RMA normalization procedures (Irizarry *et al.*, 2003) and the Entrez Gene based custom CDFs (version 10) (Dai *et al.*, 2005).

Statistical significance of differential expression between two groups of samples (e.g. positive vs. negative Estrogen receptor status (ER+/-)) is assessed using empirical Bayes linear models (Sartor *et al.*, 2006; Smyth, 2004). An average expression signature for a given pair of conditions (e.g. ER+ vs. ER-) is computed as the per-gene average difference of \log_2 expression levels between the conditions.

To match genes across datasets, platform-specific identifiers are first mapped to Entrez gene IDs (Maglott *et al.*, 2005) or, if the datasets are from different species, to Entrez HomoloGene IDs. Next, average p -values (geometric mean) and expression measures (arithmetic mean) for each Entrez ID are computed where necessary, and finally genes are matched across datasets by Entrez ID.

2.5 Other methods to assess gene list concordance

We compared our GRS analysis to four other methods. All four procedures require sub-setting genes into “differentially expressed” and “not differentially expressed” using a statistical significance cut-off (Table 1). The alternative methods are as follows.

EPSA (*Expression-based Pathway Signature Analysis*) (Tenenbaum *et al.*, 2008). EPSA computes the Spearman correlation coefficient between average differential expression levels in the *Query* and the *Reference* data-

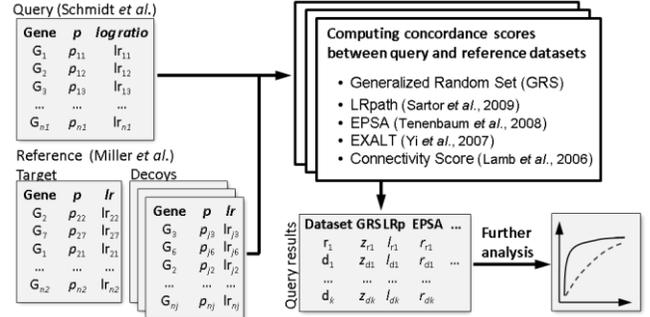


Fig. 1. Computational study for assessing abilities of different methods to distinguish between *Target* and *Decoy* datasets constructed from samples in the Miller (Miller *et al.*, 2005) dataset using the *Query* dataset based on samples from the Schmidt (Schmidt *et al.*, 2008) dataset.

sets for genes with statistically significant differential expression ($FDR < 0.1$) in the *Query* dataset. If none of the genes had statistically significant differences in expression then the score was set to zero.

EXALT (*EXpression signature Analysis Tool*) (Yi *et al.*, 2007). This method computes the Total Identity score by considering genes to be differentially expressed if the FDR (“q-value”) is less than 0.2. Briefly, genes are assigned to either U (up-regulated), D (down-regulated), or X (uncertain) in both *Query* and the *Reference* datasets. The weighted sum of Q-scores (i.e. the $-\log(q\text{-value})$) for discordant genes (U-D, D-U) is then subtracted from the weighted sum of Q-scores for concordant genes (U-U, D-D). Genes assigned to X for either one of the signatures are ignored.

CS (*Connectivity Score*) (Lamb *et al.*, 2006). Differentially expressed genes in the *Query* dataset ($FDR \leq 0.1$) are used as the query list and tagged “up” or “down,” respectively, based on their average expression signature. The Connectivity Score is then computed using the average reference expression signature as the single “treatment instance” to compute a metric based on the Kolmogorov-Smirnov statistic (Lamb *et al.*, 2006).

LRpath (Sartor *et al.*, 2009). This method requires a reference gene list (“functional category”) which we defined as the list of differentially expressed genes in the *Query* dataset ($FDR \leq 0.1$). The LRpath p -value was then computed as described in the original publication by testing the differential expression enrichment in the *Reference* datasets.

2.6 Comparing GRS to other methods

To compare GRS to other methods we test their ability to identify concordant expression profiles defined by the ER status of human primary breast tumors between two independent datasets, as well as expression profiles defined by perturbations that are caused by a number of small molecules (Lamb *et al.*, 2006).

Breast cancer datasets (Miller *et al.*, 2005; Schmidt *et al.*, 2008). For each iteration, we define a *Query* dataset of size $2N$ ($N = 2 \dots 10$) by randomly selecting N samples without replacement among ER+ and N samples among ER- samples in the Schmidt dataset (Schmidt *et al.*, 2008). The *Reference* datasets are constructed using samples from an independent primary breast cancer dataset (Miller *et al.*, 2005) and consist of a single *Target* that is constructed in the same way as the *Query* dataset, and 20 *Decoys* for which two groups of samples of size N are selected randomly from all remaining samples. Using the five different methods (EPSA, EXALT, CS, LRpath, and GRS) we compute respective 21 scores between the *Query* and *Reference* datasets and rank *Reference* datasets based on these scores. We repeat this procedure 500 times for each N (Figure 1).

Connectivity map dataset (Lamb *et al.*, 2006). Here we use the same principle to generate *Query* and *Reference* sets except *Decoy* datasets are based on the true gene expression profiles for perturbants that are different from the *Query* perturbant: For *Query* datasets we randomly draw without

Table 1. Methods included in the comparisons.

Method (Reference)	Similarity measure	Signif. cutoff
CS (Lamb <i>et al.</i> , 2006)	K-S statistic	0.1 FDR (query)
EXALT (Yi <i>et al.</i> , 2007)	Significance score	0.2 FDR (query & ref.)
EPSA (Tenenbaum <i>et al.</i> , 2008)	Spearman correlation	0.1 FDR (query)
LRpath (Sartor <i>et al.</i> , 2009)	Logistic regression	0.1 FDR (reference)
GRS	Random set statistic	Not required

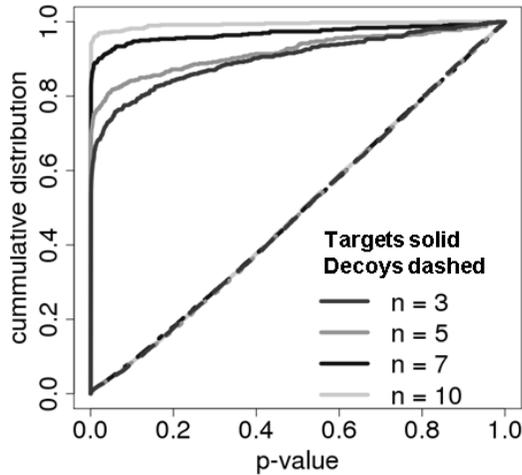


Fig. 2. Cumulative distribution of GRS p-values under the null hypothesis (dashed) and *Query* vs. *Targets* (solid) comparisons using the two breast cancer datasets.

replacement a subset of size N from the pool of samples treated with the same compound (e.g. estrodiol, wortmannin, etc.) and then draw N samples from the pool of corresponding control samples. For *Target* datasets we randomly draw without replacement a subset of size N from remaining pool of samples treated with the same compound and from remaining controls, respectively. For a *Decoy* dataset we draw N samples from a sample pool treated with a compound other than the query compound and N samples from its corresponding controls.

Receiver Operating Characteristics. To compute the True Positive Rate (TPR) we consider only the *Query-Target* pair a true match, any *Query-Decoy* pairing is considered false match. For a fixed sample size, we construct Receiver Operating Characteristics (ROC) curves based on the rankings of each *Query-Reference* pair in 500 trials. For the fixed ranking

threshold t ($t=1, \dots, 20$), True Positive Rates (TPR) are defined as the proportion of trials for which the rank of *Query-Target* score is less than t and False Positive Rates are the average proportion of *Query-Decoy* ranks less than t . In case of ties, ranks are assigned randomly among the tied scores. For each sample size we construct ROC curves by varying t from 1 to 21, and summarize them by calculating the area under each such ROC curves (AUC).

3 RESULTS

We implemented our new approach and the other methods listed in Table 1 and performed a comparison study using expression data from primary human breast cancers (Miller et al., 2005; Schmidt et al., 2008) and the Connectivity Map (Lamb et al., 2006). We then evaluated each method based on their ability to correctly identify a *Target* signature concordant with a *Query* signature among a number of unrelated *Decoy* signatures. This approach allows us to estimate the specificity and sensitivity of each method while mimicking the functional analysis by primary genomics data where the researcher compares a new dataset (*Query*) to a diverse collection of existing genomics datasets (*References*) in order to identify gene signatures that are similar to the *Query*.

3.1 Evaluating GRS measures of statistical significance

To assess how well our null distribution approximates the true null-distribution we plotted the average cumulative distribution of p -values for different sample sizes (Figure 2) under the null hypothesis of no concordance (dashed lines) and the alternative hypothesis (solid lines). The distribution under the null hypothesis was simulated by randomly permuting gene labels in the *Decoy* datasets while leaving the expression data intact. The cumulative distribution of p -values under the null hypothesis fell on the diagonal

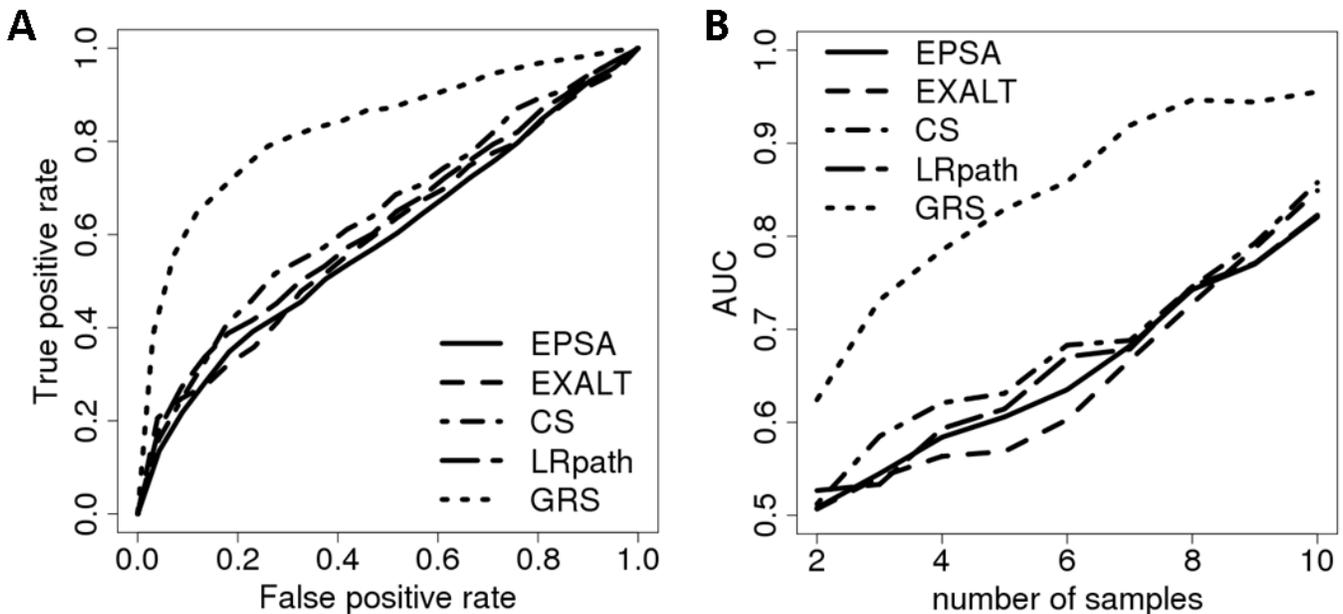


Fig. 3. ROC curves comparing the performance of different methods for the breast cancer data (Miller et al., 2005; Schmidt et al., 2008). **A)** An example of the ROC curves for a fixed samples size ($N=5$). **B)** Areas under the ROC curves for different methods and sample sizes ($N=2,3, \dots, 10$).

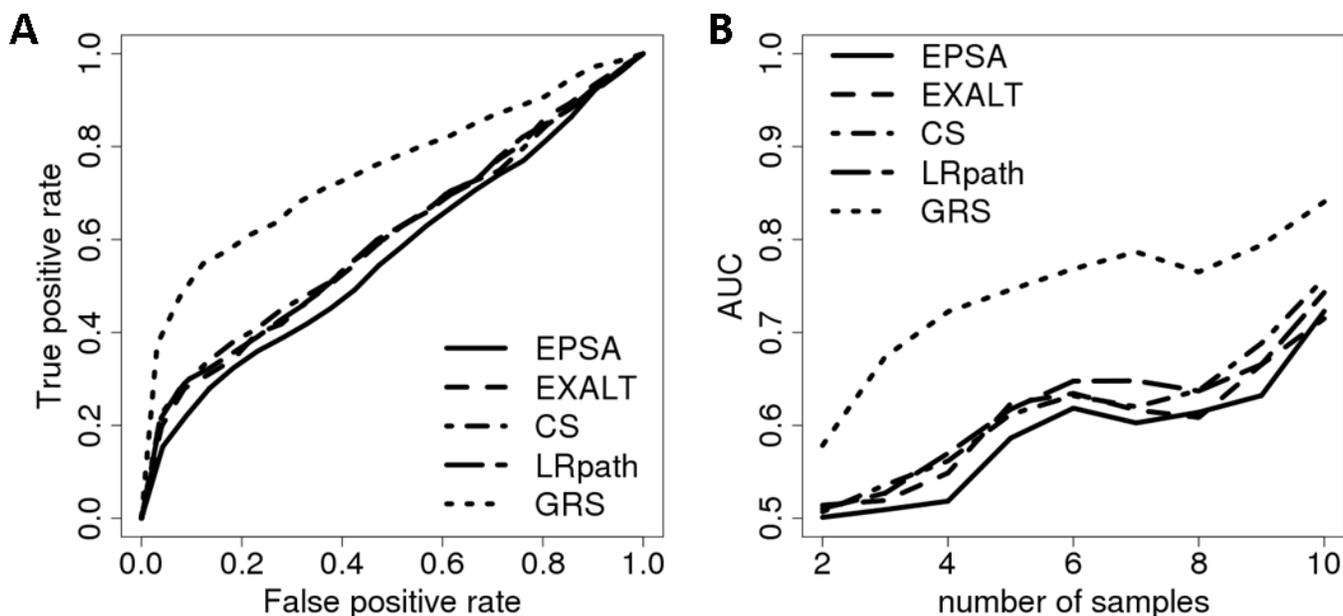


Fig. 4. ROC curves comparing the performance of different methods for the Connectivity Map data (Lamb *et al.*, 2006). **A)** An example of the ROC curves for a fixed samples size ($N=5$). **B)** Areas under the ROC curves for different methods and sample sizes ($N=2,3,\dots,10$).

line indicating perfect control of the Type I error rate. Under the alternative hypothesis (solid lines in Figure 2) p -values were highly enriched by small p -values and enrichments increased with the increased sample size indicating increasing statistical power to detect the concordance.

3.2 Evaluating methods using sub-sampling of breast cancer datasets

To compare our GRS approach to existing methods (EPSA, EXALT, Connectivity Score, and LRpath) assessing gene list concordance, we tested their ability to identify concordant expression profiles defined by the ER status of human primary breast tumors among a number of *Decoys*. The analysis was performed and ROC curves established for a series of sample sizes N Figure 3.A shows the ROC curve for $N=5$. GRS (dashed line) clearly outperforms the other methods producing significantly higher TPRs for any given FPR. ROC curves were summarized for each N by calculating the area under each such curve (Figure 3.B). Our GRS method provided dramatic improvement in precision over four alternative approaches which all rely on designating “differentially expressed” genes. For example, when $N=5$, allowing 10% false positives will, on average, resulted in 70% TPR for our GRS method and only about 40% TPR for other methods (Figure 3.A).

3.3 Evaluating methods using the Connectivity Map

In this case we tested the ability of different methods to identify datasets generated by using the same “perturbant” as used in the *Query* dataset. This time, *Decoy* datasets were not formed based on randomly permuted data, but based on the true gene expression profiles for perturbants different from the *Query* perturbant. As

with the breast cancer data, GRS outperformed alternative methods across all sample sizes, while all methods indicated increased statistical power with increased sample size (Figure 4). All methods performed worse in this setting than in the breast cancer data. This is most likely due to the fact that *Decoys* in the breast cancer data are true decoys since they are based on randomly permuted sample labels, whereas at least some of the decoys generated in the Connectivity Map setting are likely to be similar to the *Target* as multiple perturbants are affecting the same biological processes and hence produce highly concordant transcriptional signatures.

3.4 Identifying genes with concordant patterns

We examined the utility of the gene-level scores (E_g) to identify genes associated with overlapping patterns of expression. As an example, we used two gene expression datasets comparing samples with different proliferation levels in two very distinct biological systems: diets-induced differential proliferation in normal rat mammary epithelium (Medvedovic *et al.*, 2009) and differential proliferation of primary human tumors of different histologic grades (grade 3 vs. grade 1) (Schmidt *et al.*, 2008). The GRS indicated the existence of a strong concordant gene expression signature in the two datasets (p -value $< 10^{-5}$). The functional analysis of individual datasets using LRpath (Sartor *et al.*, 2009) against Gene Ontologies (GO) (Ashburner *et al.*, 2000) indicated the strong enrichment of cell-cycle related genes. Using LRpath with gene-level GRS scores (E_g) as the independent variable we compared the statistical significance of enrichment of top 10 GO terms with results based on individual datasets (Figure 5). All top 10 GO terms were cell-cycle related and there was a dramatic increase in the statistical significance of these GO terms over the analysis of individual datasets. This indicates that GRS analysis was able to effec-

tively identify genes associated with the concordant profiles in the two datasets and accentuate their importance in the functional analysis.

3.5 Example: Diets proliferation signature

To demonstrate how the GRS framework can be used to functionally annotate a newly generated genomics dataset, we continue the example shown in Figure 5 involving diets-induced gene expression changes in normal rat mammary epithelium (Medvedovic *et al.*, 2009). First, we used our web interface (Shinde *et al.*, 2010) to analyze a single reference dataset (query: dataset “BcerDiets”, parameter “Diet”; reference: dataset “GSE11121Entrez”, parameter: “grade 1” vs. “grade 3”). The resulting gene list contained 7,814 common genes where 123 (396) genes had a gene-specific GRS score greater than the 99th (95th) percentile. These genes were highly enriched for proliferation-related pathways which can be viewed by clicking on the corresponding link on the result page. A heatmap for the top 396 genes showing the relative expression levels for the two data sets can be found in supplemental Figure 1. Next, we used the procedures implemented in the R to analyze a collection of 2,980 human reference signatures obtained from GEO (Barrett *et al.*, 2009) and 3,135 signatures derived from the Connectivity Map (Lamb *et al.*, 2006). Both reference sets yielded concordant signatures with highly significant GRS scores that complement the proliferation-related functional annotations found by traditional methods. The GEO reference datasets with the most significant scores include a comparison of mesenchymal and proliferative cells in gliomas (accession GDS1815), breast cancer cells with inactivated FOXM1 transcription factor, and (GDS1477) breast cancer cells affected by 17 β -estradiol (GDS2324) (Supplemental Table 1). The most concordant perturbation signatures are caused by compounds (Etoposide and Methotrexate) that are commonly used in cancer chemotherapies and inhibit DNA replication. The third most significant compound (Monobenzone) also is a cytotoxin which is used for depigmentation therapy but its working mechanism is not yet fully understood (Supplemental Table 2). Reference signatures and the R package CLEAN are available for download at our website (<http://ClusterAnalysis.org/>).

4 DISCUSSION

We have developed a new method to assess the concordance of gene signatures. We compared our new approach to four existing methods evaluating the ability to identify genomics datasets similar to a query dataset based on their respective gene signatures with the goal of using the phenotypic characteristics of the identified reference sets to elucidate underlying molecular functions of the genomic profile in the new dataset. Each of the methods is designed to functionally annotate a newly derived gene signature and all but one (LRpath) was aimed specifically at primary genomics data. None of the existing methods, however, are designed to use the complete signature without requiring the specification of a significance cutoff.

The choice of the optimal cutoff is a difficult problem. A too restrictive cutoff leads to a low number of genes in the signature, particularly in experiments with small sample sizes. Choosing less restrictive cutoffs, on the other hand, often leads to high numbers of genes in the signature which then results in reduced specificity

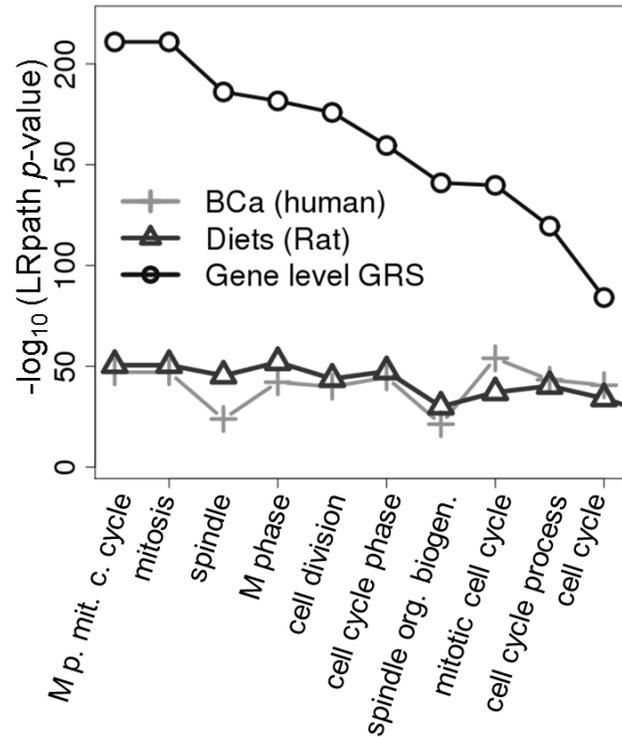


Fig. 5. Precision of functional analysis is greatly improved by combining two proliferation-related datasets.

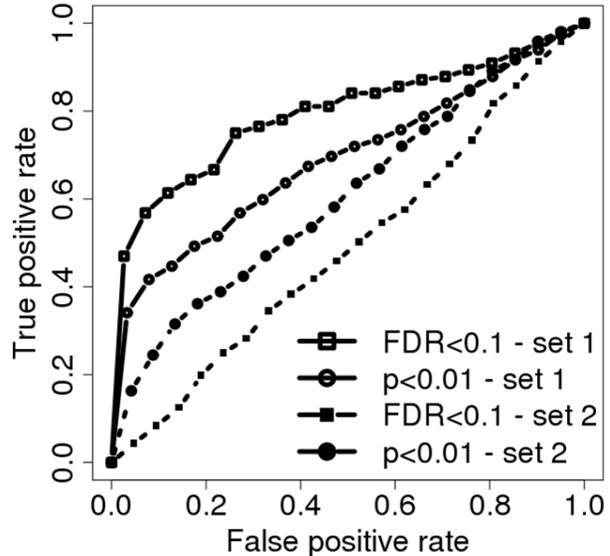


Fig. 6. Effect of significance cut-off choice. In this example we use Connectivity Map data (Lamb *et al.*, 2006) to compute the EPSA score for two sets of *Query* and corresponding *Reference* signatures as described ($N=5$) and compute the respective ROC curves. *Query* set 1 is a set of compound-vehicle comparisons with a relatively large number of genes with low p -values. Set 2 is a more “difficult” set of comparisons where the distribution of p -values for each signature is more uniform. For set 1, the FDR cutoff results in a better ROC curve than for p -value cutoff (upper two curves) while for set 2, the less conservative p -value cutoff produces a better ROC curve than the FDR cutoff (lower two curves).

when computing the concordance measures. To illustrate the impact of the cutoff choice on subsequent concordance analyses, we used the Connectivity Map data (Lamb *et al.*, 2006) to first calculate the EPSA score for two sets of *Query* and corresponding *Reference* signatures as described ($N=5$) each time using a) a stringent significance cutoff (FDR < 0.1) and b) a less conservative one (p -value < 0.01). We then computed the respective ROC curves (Figure 6). *Query* set 1 is a set of compound-vehicle comparisons with a relatively large number of genes with low p -values. Set 2 is a more “difficult” set of comparisons where the distribution of p -values for each signature is more uniform. In the former case, the FDR cutoff was the better choice (i.e. the ROC curve for the FDR cutoff is above the curve for the p -value cutoff) as it resulted in *Query* signatures that were more informative for the concordance assessment. In the latter case however, FDR proved to be too restrictive and the p -value-based cutoff was the better alternative. That is, ROC curves for set 2 overall are worse than for the less “difficult” set 1 but the curve for the p -value cutoff is above the curve for the FDR cutoff in this case (Figure 6).

In an elegant theoretical analysis within their unifying RS framework, Newton (Newton *et al.*, 2007) showed that in certain situations (small number of very significant genes), methods based on categorizing genes can outperform methods that rely on the complete expression profiles in the traditional enrichment analysis by gene lists. While it is possible that this result holds in the context of enrichment by primary data, our procedure indicates that it is probably impossible to *a priori* define optimal significance cutoffs for assessing the concordance of gene signatures. Consequently, one may always be better off using methods that do not rely on defining such cut-offs.

Our method provides additional means to functionally annotate newly generated genomics data exploiting the vast number of datasets publicly available through repositories such as GEO (Barrett *et al.*, 2009). It is readily available as part of the add-on R package CLEAN (Freudenberg *et al.*, 2009) and through our web interface Genomics Portals (Shinde *et al.*, 2010). The online version (<http://GenomicsPortals.org>) facilitates the use of datasets uploaded by the user as well as using any one of the >2,000 primary genomics datasets currently deposited in the Genomics Portals as query and reference datasets.

ACKNOWLEDGEMENTS

Funding: This research was supported by grants from the National Human Genome Research Institute (R01 HG003749), National Library of Medicine (R21 LM009662) and NIEHS Center for Environmental Genetics grant (P30 ES06096).

REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10, 47.
 Ashburner,M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25-29.
 Barrett,T. et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, 37, D885-D890.
 Casella,G. and Berger,R.L. (2001) *Statistical Inference*. Duxbury Press.
 Dai,M. et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucl. Acids Res.*, 33, e175.
 Feng,C. et al. (2009) GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genomics*, 10, 411.

Freudenberg,J.M. et al. (2009) CLEAN: CLustering Enrichment ANalysis. *BMC Bioinformatics*, 10, 234.
 Irizarry,R.A. et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 31, e15.
 Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28, 27-30.
 Lamb,J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313, 1929-1935.
 Li,Y. et al. (2008) Gene expression module-based chemical function similarity search. *Nucleic Acids Res.*, 36, e137.
 Maglott,D. et al. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 33, D54-D58.
 Medvedovic,M. et al. (2009) Influence of fatty acid diets on gene expression in rat mammary epithelial cells. *Physiol Genomics*, 38, 80-88.
 Miller,L.D. et al. (2005) From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS*, 102, 13550-13555.
 Newton,M.A. et al. (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, 1, 85-106.
 Pan,K.H., Lih,C.J. and Cohen,S.N. (2005) Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 8961-8965.
 Pena-Castillo,L. et al. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, 9, S2.
 Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat Genet.*
 Sartor,M.A., Leikauf,G.D. and Medvedovic,M. (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25, 211-217.
 Sartor,M.A. et al. (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 7, 538.
 Schmidt,M. et al. (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, 68, 5405-5413.
 Sellke,T., Bayarri,M.J. and Berger,J. (2001) Calibration of p-values for testing precise null hypothesis. *The American Statistician*, 55, 62-71.
 Shinde,K. et al. (2010) Genomics Portals: integrative web-platform for mining genomics data. *BMC Genomics*, 11, 27.
 Smyth,G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, Article 3.
 Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 9440-9445.
 Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 15545-15550.
 Tenenbaum,J. et al. (2008) Expression-based Pathway Signature Analysis (EPSA): Mining publicly available microarray data for insight into human disease. *BMC Medical Genomics*, 1, 51.
 Tian,L. et al. (2005) Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13544-13549.
 Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, 25, 1694-1701.
 Yi,Y. et al. (2007) Strategy for encoding and comparison of gene expression signatures. *Genome Biology*, 8, R133.