# Generalized Random Sets for Functional Enrichment Analysis Using Primary Genomics Datasets

**Johannes M. Freudenberg, Siva Sivaganesan, Mukta Phatak, Kaustubh Shinde, Mario Medvedovic**

Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue ML 56, Cincinnati OH 45267, USA

## Introduction

Functional analysis using primary genomics datasets is an emerging approach which allows mining of currently largely underutilized public domain data. It complements established methods for functional enrichment analysis of newly generated genomics data that are based on lists of functionally related genes. Genomics profiles are directly compared to reference profiles from a collection of reference genomics datasets such as the Connectivity Map or the GEO repository. Similarities between the new experimental data and a reference dataset implies an overlap in the function of the underlying regulatory networks. Functional interpretation of new data is then based on phenotypic characteristics of the reference datasets with concordant genomics profiles. Currently used methods for enrichment analysis by primary genomics data depend on creating lists of "significant" and "non-significant" genes derived from ad-hoc significance cutoffs. This often leads to loss of statistical power and can introduce biases affecting the interpretation of experimental results.

We developed a statistical framework for performing functional enrichment analysis using different types of primary genomics datasets to indentify concordant datasets enabling researchers to use publically available data to gain insight into the genomic conditions underlying human disease. We demonstrate the utility of the proposed method and the improvements in statistical power compared to currently available methods.

## Methods

Suppose that we have 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. The Random Set statistics [1] measures the overall level of differential expression for genes in a specific functional category $F$.

Now suppose we have a reference genomics profile instead of $F$. To avoid the need for categorizing genes into "significant" and "non-significant", we propose to compute a statistic $E_{12}$ by replacing the membership index variable with a continuous variable defined by the estimated probability of differential expression for each gene. 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 (Figure 1).

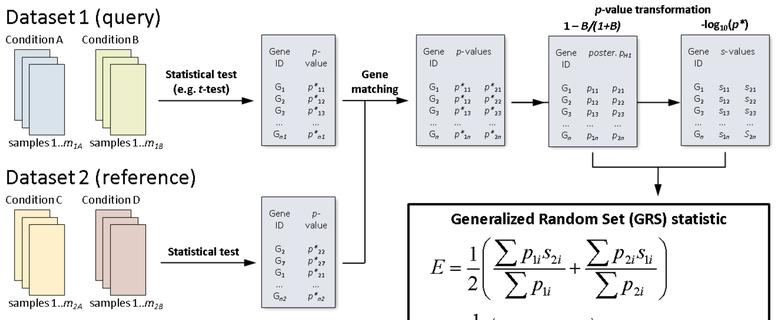



**Figure 1**. Overview of the GRS approach. A genomics profile (query) is compared to reference profiles using the Generalized Random Set (GRS) statistic.

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 where the standard deviation is approximated using the multivariate delta method [2].

## 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 [3,4] and the Connectivity Map [5]. 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 (Figure 2).

**Figure 2**. Computational study for assessing abilities of different methods to distinguish between Target and Decoy datasets.

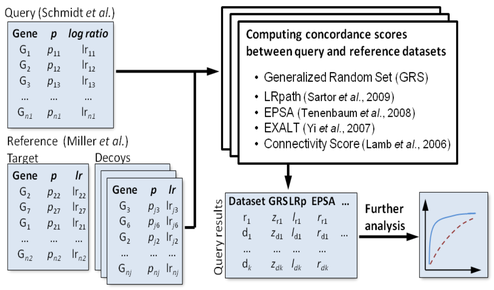



**Table 1**. Other methods to assess the concordance of genomics profiles.

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

GRS (dashed blue line) clearly outperforms the other methods producing significantly higher true positive rates (TPR) for any given false positive rate (FPR) (Figures 3-4). ROC curves were summarized for each $N$ by calculating the area under each such curve. Our GRS method provided dramatic improvement in precision over four alternative approaches which all rely on designating "differentially expressed" genes. For example, for $N$=5, allowing 10% false positives will, on average, result in 70% TPR for our GRS method and only about 40% TPR for the other methods we evaluated (Figure 3.A).

**Figure 3**. ROC curves comparing the performance of different methods for the two primary breast cancer data sets [3,4].

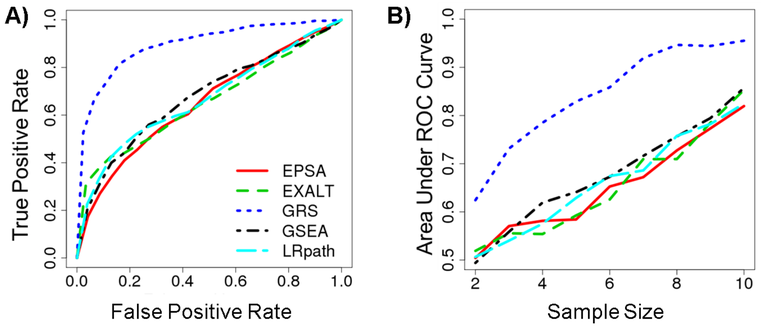
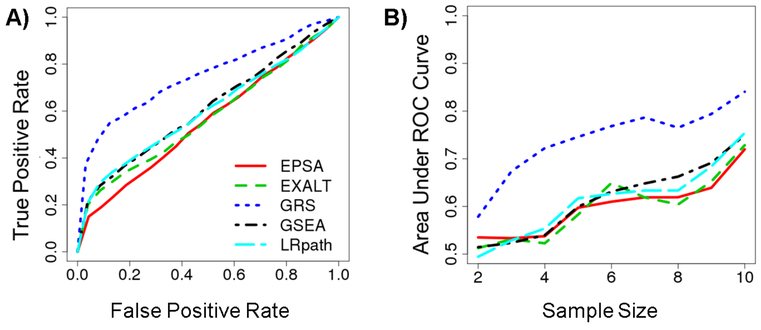



**Figure 4**. ROC curves comparing the performance of different methods for the Connectivity Map data [5].




## Results (cntd.)

We then defined each gene's contribution to the overall GRS score as a gene-level score $E_g$ in order to identify genes associated with the concordant patterns of expression. As an example, we used two datasets comparing samples with different proliferation levels in two very distinct biological systems: diets-induced differential proliferation in normal rat mammary epithelium [6] and differential proliferation of primary human breast tumors of different histologic grades (grade III vs. I) [4]. We compared the statistical significance of enrichment of top 10 GO terms with results based on individual datasets (Figure 5) showing a dramatic increase in the statistical significance of these GO terms over the analysis of individual datasets.

**Figure 5**. Precision of functional analysis is greatly improved by combining proliferation-related datasets using the gene-specific GRS score.

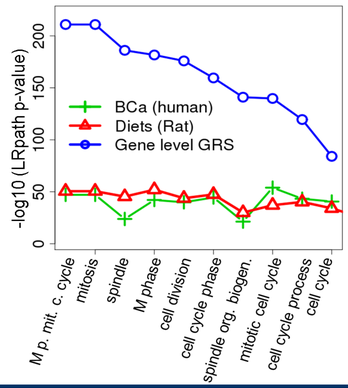
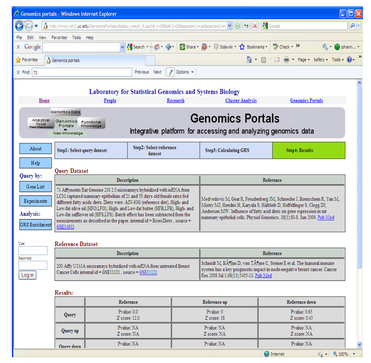



**Figure 6**. GRS is available as part of an R package and as an online tool through our web interface Genomics Portals (http://GenomicsPortals.org/).




## 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 concordant genomics datasets 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. None of these existing methods, however, are designed to use the complete signature without requiring 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. GRS successfully circumvents this problem altogether.

## References

1. Newton, M.A. *et al.* (2007). The Annals of Applied Statistics, 1, 85-106.
2. Casella, G. and Berger, R.L. (2001) Statistical Inference. Duxbury Press.
3. Miller, L.D. *et al.* (2005). PNAS, 102, 13550-13555.
4. Schmidt, M. *et al.* (2008). Cancer Research, 68, 5405-5413..
5. Lamb, J. *et al.* (2006). Science, 313, 1929-1935.
6. Medvedovic, M. *et al.* (2009). Physiological Genomics, 38, 80-88.

## Acknowledgements

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)..