

## **FDR-FET – an optimizing gene set enrichment analysis method**

**Rui-Ru Ji<sup>1§</sup>, Karl-Heinz Ott<sup>1</sup>, Roumyana Yordanova<sup>1</sup>, and Robert E.**

**Bruccoleri<sup>2§</sup>**

<sup>1</sup>Applied Genomics, Research and Development, Bristol-Myers Squibb, 311

Pennington-Rocky Hill Road, Pennington, NJ 08534

<sup>2</sup>Congenomics, P.O. Box 1422, Glastonbury, CT 06033

<sup>§</sup>Corresponding authors:

Rui-Ru Ji

Mail Stop 3A0.06

311 Pennington-Rocky Hill Road

Pennington, NJ 08534

Tel: 609-818-6036

Fax: 609-818-3100

Email: [ruiuji@gmail.com](mailto:ruiuji@gmail.com)

Robert E. Bruccoleri

Congenomics

P.O. Box 1422

Glastonbury, CT 06033

Tel: 609-902-8419

Email: [bruc@acm.org](mailto:bruc@acm.org)

## **Abstract**

Gene set enrichment analysis (GSEA) for analyzing large profiling and screening experiments can reveal unifying biological schemes based on previously accumulated knowledge represented as “gene sets”. Most of the existing implementations use a fixed fold-change or  $P$  value cut-off to generate regulated gene lists. However, the threshold selection in most cases is arbitrary and has significant effect on the test outcome and interpretation of the experiment.

We developed a new GSEA method, FDR-FET, which dynamically optimizes the threshold choice and improves sensitivity and selectivity of GSEA. The procedure translates experimental results into a series of regulated gene lists at multiple false discovery rate (FDR) cut-offs and computes the  $P$  value of the overrepresentation of a gene set using a Fisher’s exact test (FET) in each of these gene lists. The lowest  $P$  value is retained to represent the significance of the gene set. We also implement improved methods to define a more relevant global reference set for the FET.

We demonstrate the validity of the method using a published microarray study of three HIV protease inhibitors and compare the results to those from other popular GSEA algorithms. Our results show that combining FDR with multiple cut-offs allows us to control the error while retaining genes that increase information content. We conclude that FDR-FET can selectively identify significant affected biological processes. Our method can be used for any user generated gene lists in the area of transcriptome, proteome and other biological and scientific applications.

**Keywords:** gene set enrichment analysis, false discovery rate, Fisher's exact test, microarray profiling, HIV protease inhibitor

## Introduction

Expression profiling analysis usually begins with the generation of gene lists ranked by fold-changes or  $P$  values. Interpretation of the gene lists can be facilitated by analytical approaches such as gene set enrichment analysis (GSEA)<sup>1</sup>, which utilizes priori constructed reference gene sets that groups genes by classifiers such as biological function or chromosome location (Ackermann and Strimmer 2009). This type of analyses can help identify the underlying biological mechanisms and increase the statistical power by reducing the dimensionality of the problem.

The general framework and methodology of GSEA approaches have been thoroughly analyzed and discussed<sup>2,3</sup>. These methods can be classified as either self-contained or competitive based on the definition of the null hypothesis. A self-contained test compares a gene set to a fixed standard and is not dependent on genes outside of the set. These methods make use of the raw expression data, some of them are based on logistic regression models while others utilize Hotelling's  $T^2$ -tests or the more general MANOVA (multivariate analysis of variance) models<sup>4,5</sup>. By contrast, a competitive test compares the differential expression of a gene set to that of its complement. The majority of these methods examines whether regulated genes are overrepresented in a given gene set by a test of independence in a two by two contingency table, where the test statistic can be constructed based on  $\chi^2$ , hypergeometric, or binomial distribution<sup>6</sup>. A strict fold-change or  $P$  value cut-off is needed to obtain the regulated gene list, however the choice of the cut-off is often arbitrary and can have significant influence on the test outcome and, subsequently, the interpretation of an experiment<sup>7,8</sup>. Alternatively, methods that utilize the whole vector of  $P$  values or fold-changes have been developed<sup>9,10</sup>. For example, PAGE (Parametric Analysis of Gene-

set Enrichment) implements a computationally efficient solution based on the Central Limit Theorem to define an enrichment probability<sup>10</sup>.

## Implementation

We have implemented a new GSEA method, FDR-FET, which was first described by Ji et al.<sup>11</sup> in their transcriptional profiling study of compound dose responses. The current implementation extends the original method and provides options to choose the reference set (i.e. “gene universe”).

FDR-FET automatically optimizes the cut-off criterion for a gene list (L) under investigation using a False Discovery Rate (FDR) procedure that employs a series of linearly increasing critical values<sup>12</sup> and has been shown to control the FDR at pre-specified levels for independent test statistics<sup>13</sup>. Rather than employing a single FDR criterion that would represent an arbitrary limitation of the analysis, we calculate a series of regulated gene lists ( $l_i$ , where  $l_i \subset L$ ,  $1 \leq i \leq 35$ ) corresponding to FDR cut-off values from 1% to 35% (default; or per user specified) in 1% increments.

We denote the gene set collection as S. The overlap between  $l_i$  and a gene set  $s$  of interest ( $s \subset S$ ) is examined using a Fisher’s exact test (FET). We utilize the right test that evaluates the significance of positive association between two lists, i.e. an enrichment of elements of list A (e.g.  $l_i$ ) in list B (e.g.  $s$ ) or *vice versa*<sup>14</sup>. For each  $s$ , there are as many as 35 FETs to be performed by default and the most significant  $P$  value is retained. This procedure is repeated for each gene set  $s$  in S.

We have implemented FDR-FET as a Perl module (Bio::FdrFet) with C inline codes. The module expects: A) gene sets S consisting of gene IDs and associated classifiers; and B) gene list L consisting of unique gene IDs and associated  $P$  values from a study of interest. We also provide an executable program that uses this module and reads

two input files containing these datasets. The Perl module will evaluate each gene set  $s$  and output detailed analysis information such as best  $P$  value, odds ratio and the corresponding FDR cut-off, numbers in the contingency table, and genes in the overlap (between  $s$  and the  $l_i$  with the best  $P$  value), etc. The C inline code of the Perl module is a slightly modified implementation of the FET code found in R<sup>15</sup> that is based on an elegant computation of binomial coefficients<sup>16</sup>. The test data in the module contains the GO pathways and gene  $P$  values used in the example in the next section.

Additional options were provided to deal more rigorously with the choice of reference set that has a major influence on the  $P$  value. We allow four options for the reference set: 1) genes in L ('Genes'); 2) union of genes in L and S ('Union'); 3) intersection of genes in L and S ('Intersection'); 4) user specified arbitrary number ('User'). In particular, choice 3) excludes genes with unknown classification from being counted as negative matches, which may be an issue with  $P$  value calculations. Details of how to use the Perl module can be found by searching for 'Bio::FdrFet' in the CPAN Search website (<http://search.cpan.org/>).

## Results and Discussion

Here we demonstrate the performance of FDR-FET from three perspectives. First, we assessed the selectivity and sensitivity of the method. Second, we compared FDR-FET to other GSEA methods. Since FDR-FET takes  $P$  values as input and does not differentiate the directions of gene regulation, we chose two popular implementations of the same category: a simple FET and PAGE. Third, we compared the results generated from different reference set options.

In general, the sensitivity of GSEA analysis can be improved by removal of background noise, which can have strong impact on the FDR result, through removing the bottom  $n$  percentile of low intensity probes or probes flagged as “absent”, or similar. Consolidation of probes onto the gene level is also recommended to improve independence of measures, which is one assumption of FET (Goeman and Bühlmann 2007). For example, Affymetrix probesets can be consolidated by associating each gene with the most significant  $P$  value among all probesets for the gene.

Alternatively, one can utilize the updated probeset definitions, which have been shown to improve the precision and accuracy of microarray data analysis<sup>17,18</sup>.

We utilized a microarray dataset from a published study on the cellular effects of three HIV protease inhibitors<sup>19</sup>. It is well known that patients taking protease inhibitor drugs to treat HIV-AIDS often develop a lipodystrophy-like syndrome such as hyperlipidemia, peripheral lipoatrophy and central fat accumulation<sup>20</sup>. Parker et al.<sup>19</sup> have shown that protease inhibitors could induce gene expression changes indicative of dysregulation of lipid metabolism, endoplasmic reticulum stress, and metabolic

disturbance. These results are consistent with clinical observations and provide basis for a molecular mechanism for the pathophysiology of protease inhibitor-induced lipodystrophy.

The probeset level expression data was generated using the MAS 5.0 algorithm with quantile normalization<sup>21</sup> and the 20% lowest expressed probesets were removed. A one-way ANOVA with respect to the “drug treatment” factor was performed to generate the sorted gene list by *P* values. We utilized gene sets from both Gene Ontology<sup>22</sup> and KEGG<sup>23</sup>.

### **Validation of FDR-FET**

To demonstrate the sensitivity and selectivity of FDR-FET, we generated 1000 randomized gene lists while retaining the same set of *P* values from the ANOVA. We ran FDR-FET on each of these gene lists using reference set option 1 (i.e. ‘Genes’) and maximal FDR at 35% for every gene set in KEGG. The 95<sup>th</sup> and 99<sup>th</sup> percentiles of the negative log of *P* values were calculated for every gene set, these values are found to center around 1.9 and 2.6, respectively (Figure 1). As expected, no gene set shows any large deviation from the others. By contrast, the *P* values generated from the real dataset exhibits a non-uniform distribution with only a few highly significant gene sets. Importantly, the top three gene sets with the largest separations from the 99<sup>th</sup> percentiles are the targets of HIV protease inhibitors: Aminoacyl-tRNA biosynthesis (KEGG:hsa00970), Biosynthesis of steroids (KEGG:hsa00100), and Glycolysis/Gluconeogenesis (KEGG:hsa00010).

### **Comparison of FDR-FET with a simple FET test**

Many of the existing GSEA implementations are based on FET with a fixed  $P$  value or fold change cut-off. To compare the performance of FDR-FET, which employs a flexible cut-off criterion, with that of a typical GSEA analysis, we analyzed the regulated gene list generated with an arbitrary FDR cut-off (35%). Table 1 contains the ten most significant gene set hits calculated by FDR-FET using reference set option 1 (i.e. 'Genes') and maximal FDR at 35%. This list includes all the established major targets of the HIV protease inhibitors (lipid metabolism, amino acid metabolism, gluconeogenesis, and endoplasmic reticulum). By contrast, when a single arbitrary FDR cut-off (35%) is used, the effect on gluconeogenesis associated with the pathophysiology of protease inhibitors is missed. Moreover, as depicted in Figure 2, the  $P$  values for three representative gene sets reach the maximal significance at different FDR cut-offs, demonstrating that the utilization of a flexible cut-off criterion indeed maximizes the signal to noise ratio of a gene list for individual gene sets.

### **Comparison of FDR-FET with PAGE**

The PAGE analysis was performed using the whole vector of  $P$  values from the ANOVA analysis as input. Since PAGE is based on the Central Limit Theorem that requires gene sets to be sufficiently large, we only examined those gene sets with sizes equal to or larger than ten. The negative log of  $P$  values for three gene sets (GO:0006418, GO:0004812, KEGG:hsa00970) are set to 20 since they all have a  $P$  value of zero by the PAGE analysis. Again we could identify all the major targets of HIV protease inhibitors in the top ten gene set hits from PAGE output (Supplementary1). Interestingly, the results from FDR-FET and PAGE show high concordance despite the fundamental difference in their underlining methodologies

(Figure 3). Using a gene set  $-\log P$  value cut-off of 3, PAGE identified 76 significant affected gene sets whereas FDR-FET identified 79, among which 63 are shared between the two methods. In particular, the two top ten hit lists have eight gene sets in common.

Since PAGE is a parametric test, it is generally more liable to gene outliers. In another word, a gene (or a few) with a sufficiently large fold change may lead to significant testing result for the gene set of which the gene is a member. For instance, GO:0008652 and GO:0000049 have highly significant  $P$  values by PAGE but only modest  $P$  values by FDR-FET (Figure 3). A close examination of the genes annotated to these two gene sets reveals that both contain a couple of genes with extremely low  $P$  values from the ANOVA test (Supplementary2). By contrast, genes in FET-based methods have equal weight and the  $P$  value reflects the gene set enrichment in the regulated gene list, true to the name of GSEA. There are areas where FDR-FET and PAGE can complement each other. For example, FDR-FET is more robust when the gene set size is small when PAGE can not produce a reliable  $P$  value. On the other hand, incomplete gene annotation may affect FET-based methods more than PAGE since lack of knowledge is counted as ‘true negative’ in the contingency table.

### **Comparison of different reference set options**

When the biological experiment is performed using a focused gene array (i.e. a subset of genes from a genome), yet the whole genome is used as the reference set, the number of “true negative” is inflated, leading to unrealistic small  $P$  values in GSEA outputs. Therefore one must evaluate what is (close to) the true “universe” for an enrichment analysis. We have introduced new options to address this issue:

- "Genes": all genes tested are counted in the GSEA calculation assuming that the gene sets are universally representing the genome universe.
- "Intersection" can be used when the gene sets are selected to represent a restricted universe, e.g. signaling pathways. In this case, only genes that are present in at least one of the signaling pathways are counted.
- 'Union' represents the general case by which any genes are counted once they are present in either the regulated gene list or the gene sets ("genome as reference set").

In options 'Genes' and "Union" annotated and unannotated genes are both counted in the reference set while in option 'Intersection', genes are only counted when they are annotated in at least one of the gene sets. Table 2 contains the ten most significant gene set hits by the option 'Genes' and the corresponding  $P$  values and ranks by options 'Union' and 'Intersection' calculated using maximal FDR at 35%. All three options identified the main HIV protease inhibitor targets, present in the top tens except for gluconeogenesis, which is ranked 12th in result generated from the 'Union' option. Using a gene set  $-\log P$  value cut-off of 3, the options 'Genes' and 'Intersection' identified similar numbers of affected gene sets, namely, 79 and 73, respectively, among which 71 gene sets are shared between the two hit lists. By contrast, the 'Union' option identified 96 gene sets, of which 21 is unique to this option and appears to be non-specific and unrelated to the drug effects upon close examination, suggesting a possible loss of selectivity with this option (Supplementary1). The effect of 'Intersection' becomes more apparent when smaller gene sets are used. The  $P$  values and the order of the hits are altered when considering

smaller reference sets (Supplementary 1 and Supplementary3). By selecting an appropriate reference set we can enhance the sensitivity and selectivity and reduce the number of spurious hits.

## **Conclusions**

In summary, the employment of FDR and multiple cut-offs provides statistical rigor with additional flexibility: the gene list size is dynamically adjusted so that genes that increase information content are retained yet the addition of noise is limited. This methodology can be applied to results from divergent experiments (e.g. hit lists from expression profiling and proteomics studies) as often found in chemogenomics and systems biology approaches.

## **Authors' contributions**

RRJ: method conception and development of Fdr-Fet, data analysis, writing of manuscript; KHO: expert advice of method, conception of altered reference sets, data analysis, writing of manuscript; RY: expert advice of method, data analysis, revision of manuscript; REB: implementation and testing of Fdr-Fet, revision of manuscript.

All authors read and approved the final manuscript.

## **Acknowledgements**

This work was supported by Bristol-Myers Squibb, the past and current employer of the authors.

## References

1. Allison DB, Cui X, Page GP, Sabripour M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7:55-65.
2. Ackermann M, Strimmer K. 2009. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47.
3. Goeman JJ, Bühlmann P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23:980-7.
4. Sartor MA, Leikauf GD, Medvedovic M. 2009. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211-7.
5. Ucar D, Neuhaus I, Ross-MacDonald P, Tilford C, et al. 2007. Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, 23(20):2716-24.
6. Khatri P, Drăghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21:3587-95.
7. Breitling R, Amtmann A, Herzyk P. 2004. Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5:34.
8. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. 2005. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, 21:2988-93.
9. Luo W, Friedman MS, Shedden K, Hankenson KD, et al. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161.

10. Kim SY, Volsky DJ. 2005. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6:144.
11. Ji RR, de Silva H, Jin Y, Bruccoleri RE, et al. 2009. Transcriptional profiling of the dose response: a more powerful approach for characterizing drug activities. *PLoS Comput Biol*, 5(9):e1000512.
12. Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751-754.
13. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser*, B57:289-300.
14. Agresti A. 1992. A survey of exact inference for contingency tables. *Statist Sci*, 7:131-153.
15. R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.  
<http://www.r-project.org>.
16. Loader C. 2000. Fast and Accurate Computation of Binomial Probabilities.  
<http://projects.scipy.org/scipy/raw-attachment/ticket/620/loader2000Fast.pdf>.
17. Dai M, Wang P, Boyd AD, Kostov G, et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33(20):e175.
18. Sandberg R, Larsson O. 2007. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8:48.
19. Parker RA, Flint OP, Mulvey R, Elosua C, et al. 2005. Endoplasmic reticulum stress links dyslipidemia to inhibition of proteasome activity and glucose transport by HIV protease inhibitors. *Mol Pharmacol*, 67:1909-19.

20. Calza L, Manfredi R, Chiodo F. 2004. Dyslipidaemia associated with antiretroviral therapy in HIV-infected patients. *J Antimicrob Chemother* 53:10-4.
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185-93.
22. Ashburner M, Ball CA, Blake JA, Botstein D, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25-9.
23. Kanehisa M, Araki M, Goto S, Hattori M, et al. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36:D480-D484.

## Figure legends

### **Figure 1: Performance assessment of FDR-FET using simulated datasets.**

*P* values are calculated for gene sets from the KEGG for each of the 1000 randomized gene lists using FDR-FET (with the option ‘Genes’ and maximal FDR=35%). The 95th (red, squares) and 99th (green, triangles) percentiles of the *P* values are calculated for each of the gene sets. Gene sets are ordered by their *P* values calculated from the real dataset (blue, diamonds). The top three gene sets (highlighted in red circles) with the largest separations from the 99th percentiles are the targets of HIV protease inhibitors: Aminoacyl-tRNA biosynthesis (KEGG:hsa00970), Biosynthesis of steroids (KEGG:hsa00100), and Glycolysis/Gluconeogenesis (KEGG:hsa00010).

### **Figure 2: The impact of cut-off criterion on gene set analysis result.**

The influence of the FDR cut-off on the size of regulated gene list (bars, right axis) and on the significance of selected gene sets (calculated with the option ‘Genes’) for the HIV protease inhibitor experiment: Endoplasmic reticulum (GO:0005783; red, circles); Lipid biosynthetic process (GO:0008610; green, triangles); and Glycolysis/Gluconeogenesis (KEGG:hsa00010; orange, diamonds). The highlighted data points indicate the maximal *P* values (labelled) for the respective hits in the gene sets.

### **Figure 3: Comparison of the analysis result of FDR-FET to that of PAGE.**

*P* values are calculated for gene sets from the Gene Ontology and KEGG for the HIV protease inhibitor experiment using FDR-FET (with the option ‘Genes’ and maximal

FDR=35%) and PAGE (using the whole vector of gene  $P$  values as input). Gene sets of size equal to or larger than ten are included in the plot.