Using Optimal F-Measure and Random Resampling in Gene Ontology Enrichment Calculations.

Weihao Ge[1,2,†], Zeeshan Fazal[1,3,4,†] and Eric Jakobsson[1,2,5,6,7,8*]

[1]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [2]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [3]Department of Biosciences, COMSATS Institute of Information Technology, Park Road, Tarlai Kalan, Islamabad, Pakistan, [4]Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [5]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [6]Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [7]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, [8]Department of Molecular and Integrative Physiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America.

*Correspondence: jake@illinois.edu

† Equal Contributor

Abstract

Background: A central question in bioinformatics is how to minimize arbitrariness and bias in analysis of patterns of enrichment in data. A prime example of such a question is enrichment of gene ontology (GO) classes in lists of genes. Our paper deals with two issues within this larger question. One is how to calculate the false discovery rate (FDR) within a set of apparently enriched ontologies, and the second how to set that FDR within the context of assessing significance for addressing biological questions, to answer these questions we compare a random resampling method with a commonly used method for assessing FDR, the Benjamini-Hochberg (BH) method. We further develop a heuristic method for evaluating Type II (false negative) errors to enable utilization of F-Measure binary classification theory for distinguishing "significant" from "non-significant" degrees of enrichment.

Results: The results show the preferability and feasibility of random resampling assessment of FDR over the analytical methods with which we compare it. They also show that the reasonableness of any arbitrary threshold depends strongly on the structure of the dataset being tested, suggesting that the less arbitrary method of F-measure optimization to determine significance threshold is preferable.

Conclusion: Therefore, we suggest using F-measure optimization instead of placing an arbitrary threshold to evaluate the significance of Gene Ontology Enrichment results, and using resampling to replace analytical methods

Keywords: Gene Ontology; F-measure; False Discovery Rate; Microarray Data Analysis

Background

Gene Ontology (GO) enrichment analysis is a powerful tool to interpret the biological implications of selected groups of genes. The gene lists from experiments such as microarrays, are gathered into clusters associated with biological attributes, and defined as GO terms [1]). The GO terms are arranged in an acyclic tree structure from more specific to more general descriptions, including biological process (BP), cellular component (CC), and molecular function (MF) [1]. GO aspires to be both a cross-species common language, and means of understanding the uniqueness of each species at in the categories or biological process, location in the cell, and molecular [1]. Each enriched GO term is then evaluated by its significance level, i.e. the probability that the enrichment has not occurred by pure chance.

Enrichment tools have been developed to process large gene lists to generate significantly enriched ontologies. Huang *et.al* (2009) summarizes the tools widely used for GO enrichment [2]. Different tools emphasize different features. Gorilla [3], DAVID [4], g:profiler [5] are web interfaces that integrate functional annotations including GO annotations, disease and pathway databases etc. Blast2GO [6] extends annotation of gene list to non-model organisms by sequence similarity. GO-Miner [7], Babelomics[8], FatiGO[9], GSEA[10], and ErmineJ [11] apply resampling or permutation algorithms on random sets to evaluate the number of false positives in computed gene ontologies associated with test

sets. DAVID [4] and Babelomics [8] introduced level-specific enrichment analysis; that is, not including both parents and children terms. The TopGO algorithms "eliminate" and "parent-child" eliminate or reduce the weight of genes in the enriched children terms when calculating parent term enrichment [12]. TopGO [13] and GOstats [14] provide R-scripted tools for ease of further implementation. Cytoscape plugin in BinGO [15] is associated with output tree graphs.

To calculate raw *p*-values for GO enrichment without multiple hypothesis correction, methods used include hypergeometric distribution, Fisher's distribution, Binomial distribution, or χ^2 distribution [16]. Rivals *et. al*. discussed the relative merits of these methods [16].

Uncorrected *p*-values are subjected to multiple hypothesis correction by the methods of Bonferroni [17], Benjamini-Hochberg (BH) [18], or Benjamini-Yekutieli (BY) [19], or recently, a hierarchical method proposed by Bogmolov *et.al.*[20]. Bonferroni is the most stringent one among these multiple hypothesis correction methods. Benjamini-Hochberg has been widely applied in enrichment tools such as BinGO [15], DAVID[4], GOEAST [21], Gorilla [3], and Babelomics [8], to name a few. The Benjamini-Yekutieli method is included in the GOEAST package [21]. GOssip provides a direct analytical estimation of false positives that compares well with resampling [22]. In random resampling, a null set is constructed by random sampling from the same structured database that the test set enrichment is computed from. Because it is most directly related to the question of how likely it is that an enrichment result may arise by chance, it can be reasonably considered the most reliable method for estimating false positives [22]. Resampling is more computer-intensive than other methods [22], but high-throughput techniques have demonstrated that it is possible to keep resampling time in a reasonable range [7].

In applying all the cited methods and tools, it is common to apply a threshold boundary between "significant enrichment" and "insignificance". Such assignment to one of two classes is an example of a binary classification problem. Often such classifications are made utilizing an optimum F-measure [23]. Rhee, *et.al.* (2008) have suggested application of F-measure optimization to the issue of gene ontology enrichment analysis [24]. In the present work, we present an approach to gene enrichment analysis based on F-measure optimization, which considers both precision and recall and provides a flexible reasonable threshold for data sets depending on user choice as to the relative importance of precision and recall. We also compare a resampling method to the Benjamini-Hochberg and Benjamini-Yekutieli methods for estimation of FDR and use with F-measure optimization. The work suggests that resampling is preferable to analytical methods to estimate FDR, and F-measure optimization is preferable to an arbitrary threshold, in computing enrichment in gene ontology analysis.

Methods

Enrichment Tool

For results reported in this study (described below), the TopGO [13] package is implemented to perform GO enrichment analysis, using the "classic" option. In this option, the hypergeometric test is applied to the input gene list to calculate an uncorrected *p*-value.

FDR Calculation

The empirical resampling and Benjamini-Hochberg (BH) methods are used to estimate the FDR. The *p*-value adjustment using Benjamini-Hochberg is carried out by a function implemented in the R library. http://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html

The resampling method is based on the definition of *p*-value as the probability that an observed level of enrichment might arise purely by chance. To evaluate this probability, we generate several null sets, which are the same size as the test set. The genes in the null sets are randomly sampled from the background/reference list. GO enrichment analysis was carried out on both test set and null set. The average number of enriched results in the null sets would be the false positives. In all the results shown in this paper, 100 null sets were used to compute the average. In the pipeline, available for download in Supplementary material, the number of null sets is an adjustable parameter. The ratio of false positives to total positives is the FDR.

F-measure Optimization

The F-measure is a weighed value of precision and recall (Powers, 2011). The parameter β is chosen based on the research question, whether minimization of type I (false positive) or type II (false negative) error, or balance between the two, is preferred, according to the equation:

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 Precision + Recall} \qquad Equation\ 1$$

The larger the magnitude of β the more the value of F_β is weighted towards recall; the smaller the value of β the more the value of F_β is weighted towards precision. To obtain F-measure, precision and recall are derived from enrichment results. For an analytical method such as BH, we first calculate precision by the equation: $Precision = 1 - FDR$. The true positive is derived by $(True\ Positive) = (Total\ Positive) \cdot Precision$. On the other hand, for the resampling method, the number of enriched terms from random set indicates false positives. Therefore, we first calculate true positive number by:

$$True\ Positive = (Total\ Positive) - (False\ Positive).$$

Then, we calculate the precision:

$$Precision = \frac{True\ Positive}{Total\ Positive}$$

Recall is defined as

$$Recall = \frac{True\ Positve}{Relevant\ Elements}$$

"Relevant Elements" is defined by

$$Relevant\ Elements = True\ Positives + False\ Negatives$$

In the absence of the ability to calculate "False Negatives" directly, we estimate the number of relevant elements as the maximum true positive achieved across the range of possible *p*-values, as shown graphically in Figure 1 for BH method of computing false positives, for a gene list to be described in detail later in the paper. In this figure we plot total positives, false positives (False Discovery Rate x total positives, and true positives (total positives – false positives) vs. uncorrected *p*-value for the entire range of *p*-values from 0 to 1. At very lenient *p*-values the FDR approaches 1, resulting in the true positives approaching 0. It is difficult to evaluate false negatives and thus assign a number for "relevant elements", since a false negative is an object that escaped observation, and thus can't be counted directly. Yet such estimation is

essential to applying F-measure. In our case, if we follow the trajectory of the true positives in Figure 1 as the threshold is relaxed, we see that at very stringent *p*-values all positives are true positives. As the threshold is relaxed further, more false positives are generated, so the total positive and true positive curves start to diverge. At $p = 0.2$ (a far higher value than would ordinarily be used as a cutoff) the true positives reach a maximum, and the number of true positives starts to decline as *p* is further relaxed. We utilize this maximum value as the maximum number of GO categories that can be possibly regarded as enriched in the data set; i.e., the number of relevant elements.

Based on precision and recall at each raw *p*-value cut-off, we can obtain a table and curve of F-measure vs raw *p*-value. Optimizing F-measure provides us a threshold which emphasize precision ($\beta<1$) or recall ($\beta>1$), or balance of both ($\beta=1$). Note that precision and recall are extreme values of F-measure; that is, Precision=F_0 and Recall=F_∞.

Data Sets

Environmental Stress Response (ESR)

First dataset is the Yeast Environmental Stress Response (ESR) data [25], a robust data set for a model organism. The ESR set is list of genes commonly differentially expressed in response to environmental stresses such as heat shock, nutrient depletion, chemical stress, etc. Approximately 300 genes are up-regulated and 600 genes are down-regulated in ESR set [25]. We expect this set to be "well-behaved" (give reasonable results with standard methods of analysis), since the data come from a very well annotated model organism subject to a widely studied experimental intervention.

Alarm Pheromone (AP)

The second data set is comprised of human orthologs to the honey bee Alarm Pheromone set [26]. The Alarm Pheromone set is a list of genes differentially expressed in honey bee brain in response to the chemical alarm pheromone, which is a component of the language by which honey bees communicate with each other. Previous studies have shown that the Alarm Pheromone set is enriched in placental mammal orthologs, compared to other metazoans including non-social insect orthologs [27]. The Alarm Pheromone set is much smaller than the ESR set, with 91 up-regulated genes and 81 down-regulated genes. We expect the AP set to be not so "well-behaved" compared to the ESR set, as we are using model organism orthologs (human) to a non-model organism (honey bee) and the organisms diverged about 600 million years ago.

Random Test Sets

To generate a baseline of the analysis for each data set using different FDR calculation methods, we have applied the pipeline to analyze randomly-generated sets as "test" set inputs, where FDR should equal to 1 for all uncorrected *p*-values.

The BH FDR curves are calculated in the following way: The R program p.adjust is applied to generate a list of analytically calculated FDR (BH) corresponding to uncorrected *p*-values for each "test" sets. Then the lists of FDRs are merged and sorted by uncorrected *p*-values. The FDRs are smoothed by a "sliding window" method: at each uncorrected *p*-value point, the new FDR is the average value of 11 FDRs centered by the uncorrected *p*-value point.

The Resampling FDR curves are calculated in the following way: The output uncorrected *p*-values are binned in steps of 1E-4. The counts below the upper bound of each *p*-value bin for the "test" set enrichment categories are the "Total Positives", and average counts for the null set enrichment categories are the "False Positives". The process is repeated for the multiple "test" sets, and corresponding to each test set, 100 null sets were generated for "False Positive" calculation. Then the number of total and false positives are averaged, respectively. The FDR would be the quotient of the averaged total and false positives. Then, all the FDRs are plotted against the uncorrected *p*-values.

Results

In this section, we present the results of applying our methods to the two previously published sets of data introduced in the Methods section, the ESR set and the human orthologs of the Alarm Pheromone set. For both above data sets, we show the results from analyzing the genes using the biological process (BP) category of the gene ontology.

ESR Set (Environmental Stress Response, yeast)

Benjamini-Hochberg (BH)

Figure 2 shows the results of F-measure optimization on the ESR data based on FDR calculated by Benjamini-Hochberg (BH) method. As expected by their definitions, precision (F_0) decreases with increasing *p*-value while recall increases with increasing *p*-value. $F_{0.5}$ (precision-emphasized), F_1 (precision and recall equally weighted) and F_2 (recall-emphasized) all show relative maxima, providing a rational basis for assigning a threshold for significance. The horizontal scale is extended far enough to visualize the determination of the number of relevant items. In the case of the up-regulated gene set, maximum F_1 occurs at an uncorrected *p*-value close to 0.05. In the case of the down-regulated gene set however, it appears that a more stringent cutoff would be appropriate.

Resampling

Figure 3 shows the results of F-measure optimization on the ESR data using resampling to calculate FDR. The false positives are calculated by average number of GO categories enriched in random sets. All the F-measures optimize at much lower uncorrected *p*-values than do the F-measures calculated by the BH method.

Alarm Pheromone Set (human orthologs)

Benjamini-Hochberg (BH)

Figure 4 shows exactly the corresponding results as Figure 2, this time on the human orthologs to the honey bee alarm pheromone set. F-measures are maximized at much higher thresholds than for the ESR set. The difference in optimal F-measure is largely due to the different shapes of the recall curves. For the ESR set, precision drops significantly more rapidly with increasing uncorrected *p*-value than does the AP set. Therefore, a higher uncorrected *p*-value can be used for the latter set with essentially the same degree of confidence.

Resampling

Figure 5 shows the number of GO categories and F-measures for the alarm pheromone set human orthologs using resampling method. The resampling method have found more false positives than BH, and therefore the precision is much lower than the precision calculated from BH, and the F-measures are optimized at lower uncorrected *p*-values than the F-measures calculated from BH.

From the above Figures 2-5, we can note the stepped structure in the number of enriched GO categories. The stepped structure lies in the fact that the number of genes associated with any GO category, in the test set or reference set, must be an integer with limited number of choices. Therefore, the uncorrected *p*-values calculated would be in a discrete set instead of a continuum. Consequently, the number of positives as a function of *p*-values increases in a stepped way. As a result, the F-measures derived from the number of GO categories have spikes. But as our graphs have demonstrated, the optimal F-measures reflect the different weights on precision and recall despite the spikes.

Comparison of FDR (False Positive) Calculation by Benjamini-Hochberg (BH) and Resampling

In the previous section, we have demonstrated how to use F-measure optimization to obtain a flexible threshold based on requirement of the research problem, whether precision or recall is more heavily weighted. This section demonstrates how the resampling method applies to the F-measure optimization approach by providing an alternative way to estimate FDR. We have plotted the FDR calculated by BH and Resampling of the randomly selected sets that are same size as the test sets. The random "test" sets are selected from the same reference set as the test sets they are compared to. Each random "test" set result is averaged over 50 runs.

Figure 6 shows that for the ESR set, the BH method and resampling estimate similar FDR at low *p*-value. As the threshold increases, the BH method estimates lower false discovery rate, and therefore higher precision, than the resampling method at the same raw *p*-value. For the Alarm Pheromone set, the BH method estimates lower FDR than resampling.

To further evaluate the methods, we carried out multiple runs using random sets as test sets. In this case, the FDR should in principle be 1, for any uncorrected *p*-value. The results of this test are shown in Figure 9a, where for each segment of *p*-values (bin size = 0.0001) we show the mean plus/minus the standard deviation. The Resampling method passes the test on the average, but the results are noisy; and the BH method systematically underestimates FDR. Figure 9b shows that the noise in the Resampling method results in Figure 9a are largely due to the variation in the random "test" sets, and that the noise level in using random resampling for real data is acceptably low.

Threshold comparison summary

In this section, we show the bar graphs (Figures 10 and 11) of number of enriched biological process GO categories associated with ESR and alarm pheromone sets, at different thresholds including the commonly-used FDR < 0.05, optimization of $F_{0.5}$, F_1, and F_2, with BH and resampling methods. This is an alternative representation of data already presented in Figures 2-5.

In Figures 8 and 9, we are comparing the number of enriched GO categories found using thresholds calculated by BH and Resampling. The leftmost bars in each cluster represents FDR under 0.05, and the next three bars are results from flexible thresholds by F-measure optimization.

We see that the most widely used method (BH FDR<0.05) is generally quite stringent. When we weight more on precision by optimizing $F_{0.5}$ using the resampling method, a more stringent threshold is calculated for the alarm pheromone set. Maximizing F_1 would bring back many more terms. Investigation into precision tells us that sometimes the F_1-maximized FDR is too high (near 1) for us to tell apart signal and noise, and might not be a good threshold. However, whether F_1 optimization is reasonable depends on the data set. In the ESR set, where precision and recall can reach the balance where both are reasonable (precision=0.84, recall =0.96 for the up-regulated; precision=0.93, recall=0.94 for the down-regulated), the F_1 optimized threshold is reliable in the sense we can be confident in validity of the large majority of the terms that are returned. On the other hand, for the alarm pheromone set, precision becomes very low when F_1 is optimized (precision=0.46, recall=0.35 for the up-regulated; precision=0.53, recall=0.81 for the down regulated), so the user may wish to use a more stringent threshold. The major point is that the threshold will not be arbitrary, but rather based on the scientist's judgment on the relative biological significance of how much weight to give precision and recall.

Conclusions

In this work, we have addressed two issues with the commonly used methods in the GO enrichment analysis: the arbitrariness of the threshold for significance, and the relationship between resampling vs. Benjamini-Hochberg theory for estimating false discovery rate. For the first part, we introduced optimization of F-measures so that both type I and II errors are considered. Unlike arbitrarily applied threshold of BH FDR<0.05 or raw *p*-value<0.01 for any data set, the F-measure optimization approach provides a flexible threshold appropriate to the nature of the data set and the research question. If the data set is high in noise-to-signal ratio and the penalty for letting in false positive is high, we can choose to optimize F-measures weighing more on precision. If the data set fails to show much enrichment by commonly-applied methods, we can relax the threshold and extract the best information indicated by F-measure optimization. For the second part, we introduced resampling approach using random sets to directly estimate false positives, and consequently derive values of FDR, precision, recall, and F-measures. We believe that for the GO enrichment analysis, a resampling method is more universally applicable than the BH method, because the resampling method is a direct algorithmic representation of the false discovery rate; that is, the likelihood of getting a positive result by pure chance. Thus, results from resampling are independent of the structure of the data set, such as parent-child relationships.

A concern is that, because of the nature of the problem, we were forced to use a heuristic (albeit reasonable) method to estimate the false negatives, essential for calculating recall. We judge that this concern is more than offset by the advantage of enabling the replacement of an arbitrary threshold with F-measure optimization. In the supplementary material, we present our automatic pipeline integrating TopGO with resampling and analyzing functions to carry out the whole process of resampling, enrichment analysis, F-measure calculation, and representing results in tables and figures. The pipeline also includes a GOstats (Falcon and Gentleman, 2007) module for easy analysis of under-represented terms. As demonstrated, the pipeline can also calculate analytical FDR including, but not limited to, the BH method.

In summary, we suggest replacing a fixed *p*-value for assigning a threshold in enrichment calculations with an optimal F-measure, which incorporates the well-established and well-defined concepts of precision and recall.

Abbreviations

GO: gene ontology; FDR: false discovery rate: BH: Benjamini-Hochberg method; BP: biological process; CC: cellular component; MF: molecular function; BY: Benjamini-Yekutieli method; ESR environmental stress response genes; AP: honey bee genes in response to Alarm Pheromone, human orthologs.

Ethics approval and consent to participate

N/A

Consent for publication

N/A

Availability of data and material

Computer codes available in Supplementary Material

Competing interests

The authors declare that they have no competing interests.

Authors contributions

WG and ZF both did parts of the calculation, and worked together to develop the automated pipeline. EJ suggested the overall direction of the work. WG wrote the first draft of the manuscript. All three authors worked on refining the manuscript

Acknowledgements

We gratefully acknowledge useful discussions with especially Dr. Enes Kotil and Santiago Nunez-Coralles and also other members of our research group in the Beckman Institute. Professor Saurabh Sinha of our Computer Science Department provided useful discussions and wise advice.

Funding

We gratefully acknowledge support from R01GM098736 from National Institute of General Medical Science to EJ.

References

[1] Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
[2] Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
[3] Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
[4] Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
[5] Reimand,J. *et al.* (2016) g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, **44**, W83–W89.
[6] Conesa,A. *et al.* (2005) Blast2GO: A universal annotation and visualization tool in functional genomics research. Application note. *Bioinformatics*, **21**, 3674–3676.
[7] Zeeberg,B.R. *et al.* (2005) High-throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, **6**, 168.
[8] Al-Shahrour,F. *et al.* (2006) BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**.
[9] Al-Shahrour,F. *et al.* (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**.
[10] Subramanian,A. *et al.* (2007) GSEA-P: A desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.
[11] Ballouz,S. *et al.* (2016) Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Res.*, gkw957.
[12] Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
[13] Alexa,A. and Rahnenfuhrer,J. (2010) topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. *October*.
[14] Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
[15] Maere,S. *et al.* (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
[16] Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401.
[17] Bland,J.M. and Altman,D.G. (1995) Multiple significance tests: the Bonferroni method. *BMJ*, **310**, 170.
[18] Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 289–300.
[19] Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under depencency. *Ann. Stat.*, **29**, 1165–1188.
[20] Bogomolov, M. *et.al.* (2017) Testing hypotheses on a tree: new error rates and controlling strategies. arXiv preprint arXiv:1705.07529
[21] Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, 36.
[22] Blüthgen,N. *et al.* (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Informatics*, **16**, 106–115.
[23] Powers,D.M.W. (2011) Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.*, **2**, 37–63.
[24] Rhee,S. *et al.* (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.
[25] Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
[26] Alaux,C. *et al.* (2009) Honey bee aggression supports a link between gene regulation and behavioral evolution. *Proc. Natl. Acad. Sci.*, **106**, 15400–15405.
[27] Liu,H. *et al.* (2016) Conservation in Mammals of Genes Associated with Aggression-Related Behavioral Phenotypes in Honey Bees. *PLoS Comput. Biol.*, **12**.

Figures

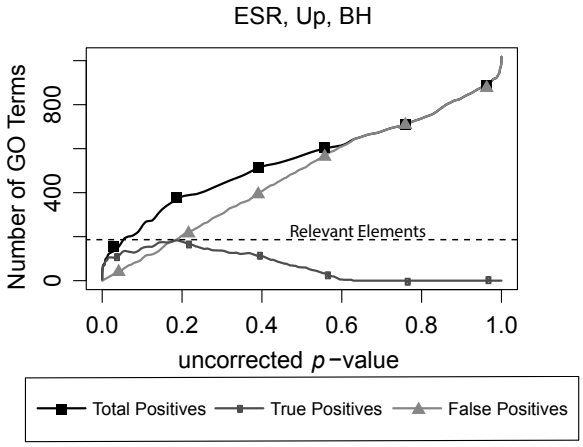

Figure 1. **Number of positives for the yeast environmental stress response (ESR) set over the full range of uncorrected *p*-values from 0 to 1.** "Total Positives" is the number of Biological Process GO categories returned as a function of the *p*-value threshold for significance. "False Positives" is the number of total positives multiplied by the False Discovery Rate as calculated by the Benjamini-Hochberg formulation. "True Positives" is "Total Positives" minus "False Positives". "Relevant Items", necessary to estimate number of false negatives, is estimated as the largest number of true positives computed at any uncorrected *p*-value.

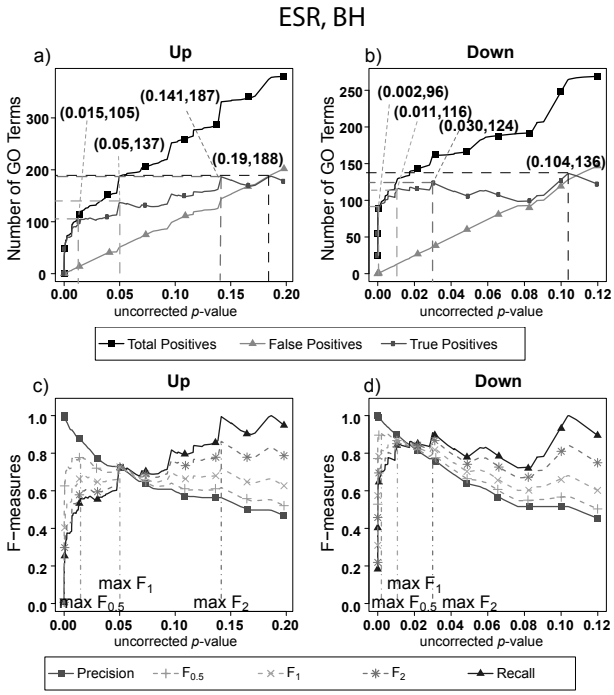

Figure 2. Number of positives and F-measure values for ESR set, BH-estimated FDR. a) Shows the number of enriched biological process Gene Ontology categories as a function of uncorrected p-value, the Benjamini-Hochberg number of false discoveries, and the projected true positives, namely the difference between the total positives and the false positives, for the upregulated ESR gene set. This panel is from the same data set at Figure 1. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the downregulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of relevant items, necessary to calculate recall (and therefore (F-measure)), is approximated by (total positives – false positives) $_{max}$. The *p*-value at which the computed true positives are a maximum is 0.19 for upregulated gene list (a) and at 0.104 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the *p*-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall where we obtain an estimation of relevant elements by maximizing true positive).

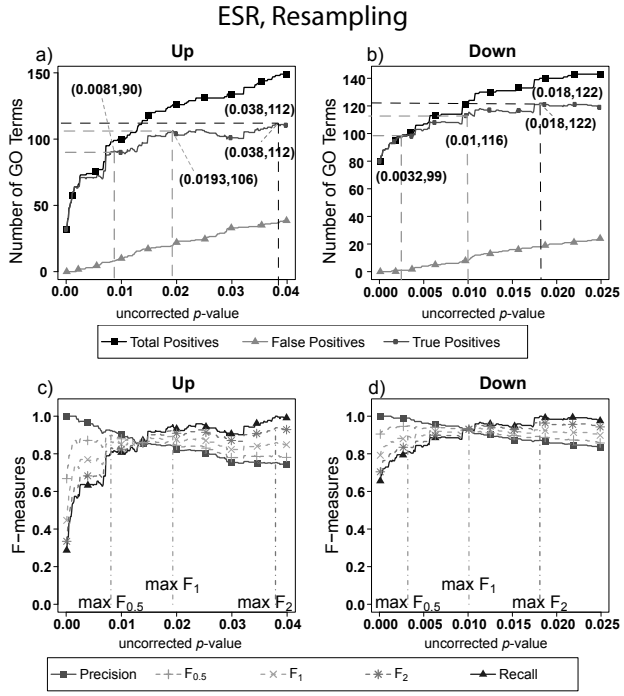

Figure 3. Number of positives and F-measure values for ESR set, Resampling-estimated FDR. A) Shows the number of enriched biological process Gene Ontology categories as a function of uncorrected p-value, the average number of enriched Gene ontology categories from the random set as the false positives, and the projected true positives, namely the difference between the total positives and the false positives, for the up-regulated ESR gene set. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the downregulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of relevant items, necessary to calculate recall (and therefore (F-measure)), is approximated by (total positives – false positives) $_{max}$. The *p*-value at which the computed true positives are a maximum is 0.038 for upregulated gene list (a) and 0.018 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the p-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

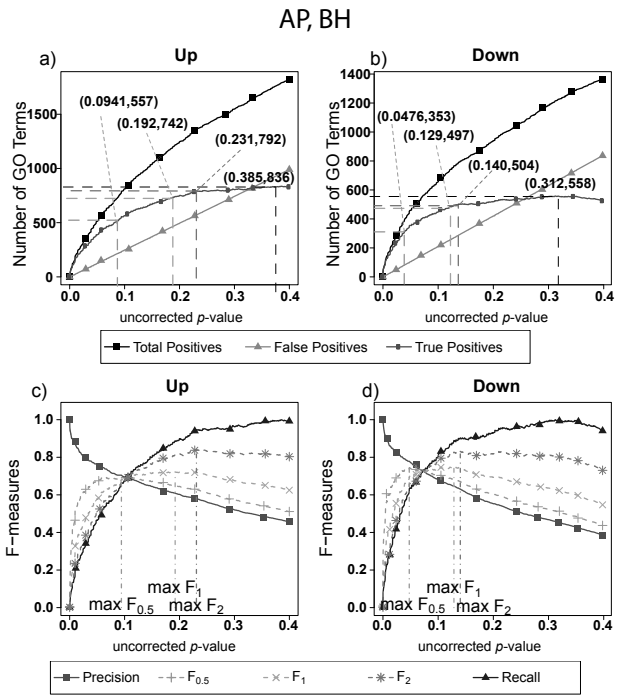

Figure 4. **Number of positives and F-measure values for Alarm Pheromone set, BH-estimated FDR** a) shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the Benjamini-Hochberg number of false discoveries, and the projected true positives, namely the difference between the total positives and the false positives, for the upregulated alarm pheromone human orthologs gene set. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the downregulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of relevant items, necessary to calculate recall (and therefore (F-measure)), is approximated by (total positives – false positives) $_{max}$. The *p*-value at which the computed true positives are a maximum is 0.385 for upregulated gene list (a) and at 0.312 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the p-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall) and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

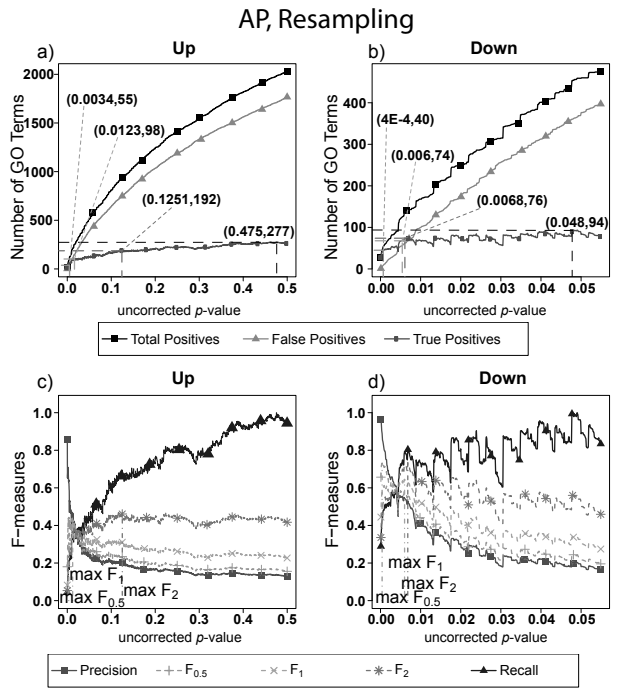

Figure 5. Number of Positives and F-measure values for AP set, Resampling-estimated FDR. The figure shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the average number of enriched Gene ontology categories from the random set as the false positives, and the projected true positives, namely the difference between the total positives and the false positives, for the up-regulated alarm pheromone human orthologs gene set. b) is the same as a) for the down-regulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of relevant items, necessary to calculate recall (and therefore (F-measure)), is approximated by (total positives – false positives) $_{max}$. The *p*-value at which the computed true positives are a maximum is 0.475 for upregulated gene list (a) and at 0.048 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the *p*-value and number of returned GO terms at significant markers, specifically at maximum $F_{0,5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

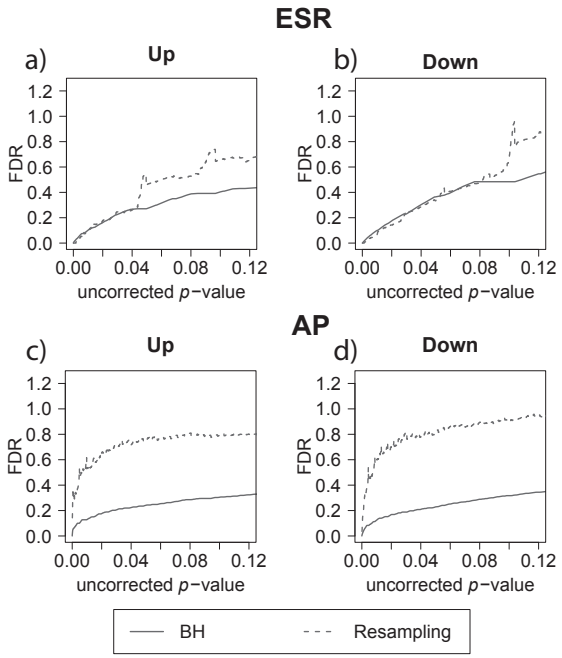

Figure 6. **False discovery rate comparison**. False discovery rate estimated by Benjamini-Hochberg (solid curve) and Resampling (dashed curve) for the ESR set and Alarm Pheromone set. Figure 6 compares the number of false discovery rate calculated by Benjamini-Hochberg (solid) and Resampling (dashed) in each set: a) up-regulated ESR, b) down-regulated ESR, c) up-regulated Alarm Pheromone set, and d) down-regulated Alarm Pheromone set. Generally, resampling has found higher false discovery rate than Benjamini-Hochberg. At low *p*-values, the BH and resampling methods get similar estimation of false discovery rate for the ESR set.

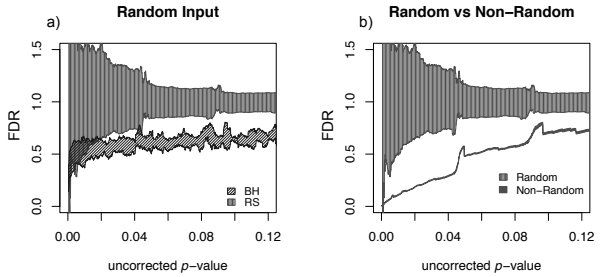

Figure 7. Comparison of different FDR calculation method on accuracy and convergence. a) Comparison of BH and Resampling on random "test" sets. At each *p*-value (*p*-values binned at intervals of .0001), the mean and standard deviation are calculated and plotted as shown. The random test sets consist of 281 yeast genes, against the background of the entire yeast genome. For each of the methods 50 test sets were used and the mean plus/minus standard deviation plotted as shown. Resampling hits the mark on the average but with substantial noise, while BH systematically underestimates FDR. b) Evaluation of resampling convergence on a real data set, ESR upregulated considered in this paper. This set is run against five different ensembles of null sets, each ensemble containing 100 null sets. The mean and standard deviation are plotted and compared to the results from the random test sets. It is seen that the noise of the resampling method on a real data set is acceptable.

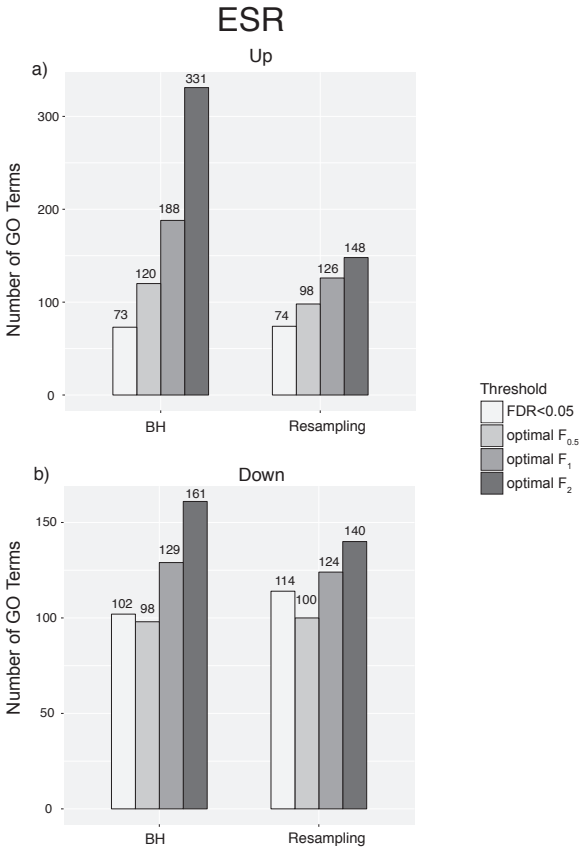

Figure 8. Number of GO categories obtained at different thresholds including: FDR<0.05, $F_{0.5}$ optimization, F_1 optimization, and F_2 optimization, using Benjamini-Hochberg (BH) and Resampling. Calculated FDR for the ESR set, up- and down-regulated. BH-estimated $F_{0.5}$, F_1, and F_2 optimization gives 120, 188, and 331 GO terms for up-regulated ESR set respectively, and 98, 129, 161 terms for down-regulated set respectively. Resampling-estimated $F_{0.5}$, F_1, and F_2 optimization gives 98, 126, and 148 terms for up-regulated ESR set respectively, and 100, 124, 140 terms for down-regulated set respectively. As more emphasis is placed on recall, the threshold would increase and more terms would be recovered. The F-measure optimization thresholds estimated by BH is more relaxed by resampling and consistently brings back more terms. The FDR<0.05 threshold is more stringent for the up-regulated set than $F_{0.5}$ optimization, but less stringent for the down-regulated set than $F_{0.5}$ optimization. For both up- down- regulated sets, the most stringent thresholds given by BH and resampling are close to each other (FDR<0.05 for up-regulated, and $F_{0.5}$ optimization for down-regulated), indicating BH and resampling estimates similar FDR at the most stringent thresholds for the ESR set, but then deviates as thresholds increase.

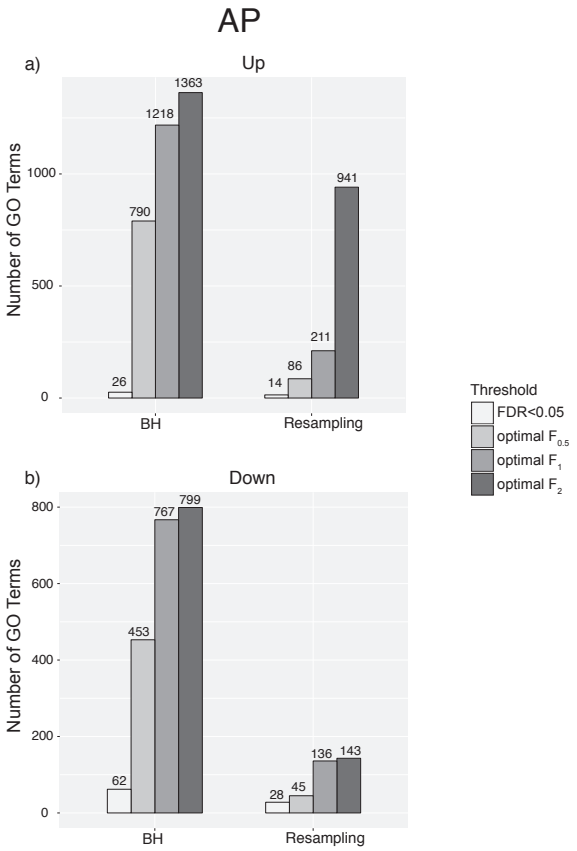

Figure 9. Number of GO categories obtained at different thresholds including: FDR<0.05, $F_{0.5}$ optimization, F_1 optimization, and F_2 optimization, using Benjamini-Hochberg (BH) and Resampling – calculated FDR for the AP set, up- and down-regulated. BH-estimated $F_{0.5}$, F_1, and F_2 optimization gives 790, 1218, and1363 GO terms for up-regulated AP set respectively, and 453, 767, 799 terms for down-regulated set respectively. Resampling-estimated $F_{0.5}$, F_1, and F_2 optimization gives 86, 211, and 941 terms for up-regulated AP set respectively, and 45, 136, 143 terms for down-regulated set respectively. As more emphasis is placed on recall, the threshold would increase and more terms would be recovered. The F-measure optimization thresholds estimated by BH is more relaxed by resampling and consistently brings back more terms. The FDR<0.05 threshold is more stringent for the up-regulated set than $F_{0.5}$ optimization, but less stringent for the down-regulated set than $F_{0.5}$ optimization. Increasing the thresholds to optimize $F_{0.5}$, an F-measure which includes the effect recall but still emphasize on precision, brings in many more terms. Therefore, for the alarm pheromone set a cutoff of FDR<0.05 might leave out too many possible candidates.

Additional Files

Additional file 1 --- pipelinemanual .docx
"A TopGO- and GOstats-based automated pipeline for GO enrichment analysis using F-measure optimization based on resampling and traditional calculation"
This is a word document giving detailed description of how to run the pipeline for resampling or analytical FDR calculation and obtain thresholds maximizing F-measures

Additional file 2 --- pipeline.gz
This file contains source codes of the pipeline and the ESR and AP data sets for demo runs.