

A practical efficient approach in high throughput screening: using FDR and fold change

Xian-Jin Xie (✉ xian-jin.xie@utsouthwestern.edu)

University of Texas Southwestern Medical Center

Angelique Whitehurst

University of Texas Southwestern Medical Center

Michael White

University of Texas Southwestern Medical Center

Method Article

Keywords: false discovery rate, fold change, high-throughput screening

Posted Date: April 16th, 2007

DOI: <https://doi.org/10.1038/nprot.2007.188>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Introduction

High throughput screening continues to present a common statistical challenge as the number of statistical comparisons vastly exceeds that of the biological replicates. Statisticians have developed methods for such large data sets, including the Bonferroni adjustment, FDR¹ and the recent optimal discovery procedure (ODP)². We detail here the statistical protocol employed for a genome-wide RNAi-based synthetic lethal screening study³. A straightforward combination of both FDR control and fold change criteria identifies a highly reproducible list of hits. Preliminary comparison between this readily implemented method and other recently developed methods (such as ODP's) shows good properties for this practical approach⁴.

Reagents

Equipment

SAS or other statistical software

Procedure

We compared cell viability under two experimental conditions (paclitaxel treated or vehicle treated) for all genes available in the genome-scale RNAi library³. The analysis was performed in triplicate under each condition. The step by step statistical procedures we used are as follows: 1. Collect the raw luminescence measurement from each well of the high throughput screen and record as a numeric value. Information about well location in the plate and plate number should also be recorded. 2. Normalize numeric luminescence values to internal reference control samples (cells with no siRNA in wells) on each plate to allow for plate-to-plate comparisons. The value of each experimental well should be divided by the reference well of the plate. 3. For each gene, perform a two sample t-test (with pooled variance) to determine whether there is a significant difference between the mean values under the two experimental conditions. Record a P -value. We performed this test using both GeneSpring and SAS. The results were almost identical with only minor decimal rounding differences. While we used a two sample t-test, other commonly used statistics, such as S - (SAM), U - (Mann-Whitney) and M - statistics may also be used for computing the P -values for each gene. In situations where there are multiple experimental conditions, ANOVA-type of analysis can also be performed to calculate P -values. 4. With the P -values generated in step 3 (we had 20,960), perform Benjamini-Hochberg's method to control the false discovery rate (FDR)¹; again, we recommend you use the standard operations in GeneSpring and SAS. The essence of this method is to inflate the raw P -values based on their rank in the distribution of all the P -values. Let $P(i)$ = the P -value of gene i , let i = rank of $P(i)$ in the distribution, let m = total number of comparisons (i.e. genes in the genome, in our case, $m=20,960$) and let q^* = false

discovery rate, the FDR for the i -th gene $FDR(i) = P(i) / (m/i)$. Since we pre-specified FDR criterion as $FDR \leq 0.05$, genes with a FDR less or equal to 0.05 were selected into list A. 5. In order to take the magnitude of response into account, sort the viability ratios (e.g. fold change: $\text{mean}_{\text{paclitaxel}} / \text{mean}_{\text{carrier}}$) in ascending order and select the genes with a fold change among the lowest 2.5 percentile of the fold change distribution into list B. 6. Include genes that were identified by both list A (so for our study, $FDR \leq 0.05$) and list B (fold change among the lowest 2.5%) in a high priority hit list C. Genes selected in list C can be sorted either using FDR or fold change for further validation and functional tests. Note: In our “high confidence” list C, we observed high extent of enrichment of proteasome subunits and of Gamma-TuRC subunits. Assuming hypergeometric distribution, the probability of obtaining this high enrichment by chance is close to zero. This further confirms the validity of list C.

Timing

Within hours

Critical Steps

Step 4 uses Benjamini-Hochberg’s false discovery rate instead of a Bonferroni adjustment. This is because the Bonferroni adjustment is far too conservative in our experimental setting, where the number of comparisons is considerably large. We used 5% FDR in our study, however relaxing this cutoff, for example to 10%, may reduce false negatives and return a larger list of potential hits. Step 6 combines the strength of FDR and fold change in creating a highly reproducible joint hit list. This is especially advantageous for reducing false positives.

Anticipated Results

Since the resulting list of “hits” has two criteria: a low FDR cutoff (q^*) and a percentile cutoff of the fold change distribution (q), the reproducibility of these “hits” is expected to be very high ($FDR \leq q^*$). In experiments with over tens of thousands of comparisons such as genome-wide screening, this stringent protocol is important in identifying highly reproducible “hits” if small q^* is specified along with a small percentile criterion q . However, this method may not render a very low false negative rate (FNR) if such criteria are so chosen. In our genome-wide RNAi screening study, this method serves extremely well as our primary goal is to identify a short highly reproducible list of “hits” for further functional analysis. Further improvement includes taking the known correlations among genes into consideration, and controlling FDR and FNR simultaneously when the study goal is not solely a short highly reproducible hit list.

References

1. Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society. B* 57(1):289
2. Storey JD, Dai JY,

and Leek JT. (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, in press. 3. Whitehurst AW, Bodemann BO, Cardenas J, Ferguson D, Girard L, Payton M, Minna JD, Michnoff C, Hao W, Roth MG, Xie X-J, White MA (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature*, in press. 4. Cao J, Xie X-J, Whitehurst A, White M (2007) A Bayesian mixture model for high throughput screening (unpublished study)

Acknowledgements

We thank Aihua Bian, MPH for her helpful technical assistance.