Statistical concerns about the GSEA procedure

To the editor:
Mootha *et al.*[1] propose a statistical method (Gene Set Enrichment Analysis; GSEA) to discern changes in expression levels of sets of genes selected *a priori* in transcriptional profiling experiments. Although consideration of groups of genes is an interesting strategy, the proposed test statistic may not necessarily determine "…if the members of a given gene set are enriched among the most differentially expressed genes between two classes"[1].

Situations will probably arise when using GSEA in which genes with the highest values of the difference metric will be ignored solely due to the size of the selected gene sets, unrelated to any biological context of the genes comprising the set. By way of illustration, consider the following hypothetical example. Assume that a given data set consists of three potentially interesting sets of genes S1, S2 and S3, of respective sizes n, $5n$ and $4n$ genes, where n is any integer. Assume also that all of the genes in S1 are ranked higher (*i.e.*, they have greater differences in expression) than the genes in S2, which in turn are ranked higher than the genes in S3. The GSEA procedure yields enrichment scores (ES)[1] of $3n$, $4n$ and 0 for S1, S2 and S3, respectively. The maximum ES[1] is $4n$ and is attributed to S2. S2 will therefore be singled out as the candidate for further investigation over S1, even though S1 comprises the highest ranked genes. This does not seem reasonable, because S2 has been chosen only by virtue of containing a larger number of genes. In other words, GSEA can be at odds with the picture suggested by the gene ranking.

A second observation, using the same illustrative example as above, gives another counterintuitive result. In the absence of a defined third gene set (S3), the ES for S2 = 0 and the ES for S1 remains positive. Therefore, S1, and not S2, is chosen by GSEA, a result opposite to that of the previous scenario. An unusual situation has arisen in which a choice or preference between sets of high ranking is affected simply by the presence or absence of a lower ranking set.

The behavior of GSEA can not be dismissed as one of the usual power issues encountered due to noise in data, small sample size or lack of robustness to model assumptions. The simple example outlined here indicates that the power of the test statistic is sensitive to the *a priori* definition of the hypotheses of interest. These limitations should be clearly understood in applying and interpreting the results of the approach.

Doris Damian[1] & Malka Gorfine[2]

[1]Beyond Genomics, Waltham, Massachusetts, USA. [2]Department of Mathematics and Statistics, Bar Ilan University, Ramat Gan 52900, Israel. Correspondence should be addressed to D.D. (ddamian@beyondgenomics.com).

1. Mootha, V.K. *et al. Nat. Genet.* **34**, 267–273 (2003).

In reply

Our manuscript[1] described Gene Set Enrichment Analysis (GSEA) as "designed to detect subtle but coordinated differences in expression of *a priori* defined sets of functionally related genes." The method requires two inputs: (i) a list of genes that have been ranked according to expression difference between two states and (ii) *a priori* defined gene sets (*e.g.*, pathways), each consisting of members drawn from this list. A gene set then receives an enrichment score (ES) that is a measure of statistical evidence rejecting the null hypothesis that its members are randomly distributed in the ordered list. By definition, the ES is a function of the size of a gene set, the total number of genes in the entire list and the relative ranks of the members of the gene set.

Damian and Gorfine's first comment is that ES can be influenced by the size of a gene set. We completely agree, because in general, statistical significance is a function of two parameters: the estimated magnitude of an effect and the variance in this estimate. Because estimates based on larger numbers of measurements have lower variance than those based on fewer measures, the ES (a measure of statistical significance) may be greater for a set of 100 genes than for a second set of only 5 genes. This can be true if some or all of the 100 genes individually rank lower than the smaller set containing 5 genes. We note that scoring by statistical significance is common; for example, it is standard in genetic linkage analysis to rank regions based on the lod score, which is a measure of statistical significance (not effect size).

In their second example, Damian and Gorfine show that by removing almost half of the lowest-ranking genes in their hypothetical experiment, the ES for gene set S2 falls. The ES falls not simply because of the definition of membership in gene sets (as they claim), but rather because of the selective removal of all genes ranked lower than those in S2. As the members of S2 are now relegated to the bottom of the list, rather than being near the top, this gene set must receive a lower ES. Contrary to Damian and Gorfine's correspondence, the mere presence or absence of gene sets (without changing the underlying list of genes) will not affect the ES of a defined gene set.

Damian and Gorfine conclude by stating that GSEA is sensitive to "*a priori* definition of the hypotheses of interest." We completely agree, as this is the desired behavior of "an analytic technique designed to test *a priori* defined gene sets"[1]. Given that the explicit goal of GSEA is to combine information about functional relationships with measurements of gene expression, it would be quite surprising if the definition of the gene sets had no influence on the results.

*Vamsi K Mootha[1], Mark J Daly[1], Nick Patterson[1], Joel N Hirschhorn[1], Leif C Groop[2] & David Altshuler[1]**

[1]Broad Institute, Harvard University and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Department of Endocrinology, University Hospital MAS, Lund University, Malmo, Sweden. Correspondence should be addressed to V.K.M. (vmootha@broad.mit.edu) or D.A. (altshuler@molbio.mgh.harvard.edu).

1. Mootha, V.K. *et al. Nat. Genet.* **34**, 267–273 (2003).