J. Jedediah Smith

Use “Gene Set Analysis: Challenges, Opportunities, and Future Research” to answer the following questions:

1. According to the authors, what are the main benefits of performing gene set analysis as opposed to the single-gene approach?

They can be used to more easily analyze the activity of groups of biologically related genes to help determine which groups are relevant to a phenotype of interested, rather than having to go one gene at a time. It also helps solve many of the shortcomings seen in single-gene approach.

1. Ignoring the math, what is the basic idea behind over-representation analysis (ORA)? Please give me one simple sentence in plain English.

Extended single gene analysis methods to help calculate the initial gene scores that are used to build your contingency table.

1. What are the main assumptions of ORA?

That genes are independent and equally effective in biological process.

1. Ignoring the math, what is the basic idea behind functional class sorting methods (FCS)? Please give me one simple sentence in plain English.

Use all information from an expression matrix to avoid relying on the biologically invalid assumptions of ORA when calculating gene scores.

1. In general terms, what are the possible drawbacks of single variate FCS such as GSEA? Don’t worry whether they have since been addressed or not, just list the major drawbacks that people have been trying to address.

GSEA does not detect differential enrichment of gene sets that have the majority of their genes up-regulated unless the phenotypes are swapped and the procedure is run again. Some GSEA may also ignore highly enriched gene sets based on the size of their set. GSEA has also lead to high enrichment scores from sets clustered around the middle of a sorted list of genes, even when the sets are night associated with the phenotypes under study.

1. In general terms what is the main idea behind multivariate FCS methods such as Globaltest?

They directly calculate gene set scores from expression data and skip the intermediate step of calculating gene scores.

1. In general terms (no math) describe each of the following techniques for determining the significance of a gene set score:
   1. Parametric

The gene set score is proposed. Then by accepting some simplifying assumptions under the null hypothesis, a parametric distribution for the gene set statistic is proposed. The parametric distribution is used to assess the significance of gene set statistics.

* 1. Gene sampling

The significance of a gene set score for a given gene is assessed by comparing it to the scores of randomly assembled sets of genes from the reference set.

* 1. Dynamic programming

Assess the significance of the gene set scores derived from the unweighted Kolmogorov-Smirnov statistic. Though to be faster and more accurate than other methods, but does not extend to as many different gene sets.

1. What are the major drawbacks of each of the following:
   1. Competitive null hypothesis methods

Uses genes as sampling units, ignores correlation between genes within a gene set, and can be severely effected by the inclusion of irreverent genes.

* 1. Self-contained null hypothesis methods

Requires a large number of samples for each phenotypes, a condition not met by many biological experiments.

1. What problem do pathway-topology methods attempt to address?

They try to help quantify the importance of a gene to pathway activity. Could also be used to potentially improve the accuracy of enrichment analysis.

1. According to the authors, what major challenges still face gene set analysis because of the lack of gold standard datasets?

There is a lack of consensus about the best practices. Challenges that face gene set analysis, lick a lack of reproducibility, specificity, or sensitivity are all rooted in the absence of a cold standard dataset.