**Full Title:**

FRY: A Fast Approximation to ROAST Gene Set Test with Mean Aggregated Set Statistics

**Authors**

Göknur Giner (a,b) and Gordon K. Smyth (a,c)

**Affiliations**

1. Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville VIC 3052, Australia
2. Department of Medical Biology, The University of Melbourne, Parkville VIC 3010, Australia
3. Department of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3052, Australia

***Abstract***

Gene set tests are often used in differential expression analyses to explore the behaviour of a group of related genes. This is useful for identifying large-scale co-regulation of genes belonging to the same biological process or molecular pathway.

One of the most flexible and powerful gene set tests is the ROAST method in the limma R package. ROAST uses residual space rotation as a sort of continuous version of sample permutation. Like permutation tests, it protects against false positives caused by correlations between genes in the set. Unlike permutation tests, it can be used with complex experimental design and with small numbers of replicates. It is the only gene set test method that is able to analyse complex “gene expression signatures” that incorporate information about both up and down regulated genes simultaneously.

ROAST works well for individual expression signatures but has limitations when applied to large collections of gene sets, such as the Broad Institute’s Molecular Signature Database with over 8000 gene sets. In particular, the p-value resolution is limited by the number of rotations that are done for each set. This makes it impossible to obtain very small p-values and hence to distinguish the top-ranking pathways from a large collection. As with permutation tests, the p-values for each set may vary from run to run.

This talk presents Fry, a very fast approximation to the complete ROAST method. Fry approximates the limiting p-value that would be obtained by performing a very large number of rotations with ROAST. Fry preserves most of the advantages of ROAST but also provides exact high-resolution p-values very quickly. In particular, it can distinguish the most significant sets in large collections and to yield statistically significant results after adjustment for multiple testing. This makes it an ideal tool for large-scale pathway analysis.

Another important consideration in gene set tests is the possible unbiased or incorrect estimation of P-values due to the correlation among genes in the same set or dependence structure between different sets.

**What will make your presentation engaging, enlightening and enjoyable?**

This talk addresses one of the most commonly used bioinformatics concepts. Gene set testing tools have been developed since the early 2000s and become crucial for better understanding of the biological processes. For this reason, our recently developed gene set testing tool will engage the bioinformatics audience, and the talk will be enlightening by presenting the statistical approach to this. To make my talk enjoyable, I will be highlighting the essential parts of my work clearly and demonstrate the idea behind it on a simulated data set with visual representations.