Summary: Using gene correlation data to identify enriched GO Term biological processes

Colin Sheehan

Gene Ontology terms are classifications representing a molecular function, biological process, or cellular component, which are composed of a collection of elements falling under that classification. For instance, a term like "glucose transmembrane transport" will include experimentally-supported genes involved with this term, such as SLC2A6. Further, there can be very broad GO terms like "metabolic process" that contain multiple daughter terms like, "glycosylation." GO terms are heavily used within bioinformatics to predict the involvement of particular biological processes within large sets of data.

Often, gene correlations are used as evidence for the involvement of a particular gene of interest with a biological process; however, analyzing the overall correlation from a set of multiple genes involved in that process may represent a more powerful and unbiased approach. One might expect that if a particular gene is systemically involved in the up-regulation or down-regulation of a process, there would be an overall bias in the correlation coefficient for the genes in that GO term. These results could then be visually represented as a "tree plot" whereby genes are ranked according to their probability value along the x-axis. We might expect a completely random set of data to have coefficients distributed around zero, forming the image of a rotated tree (figure 1).

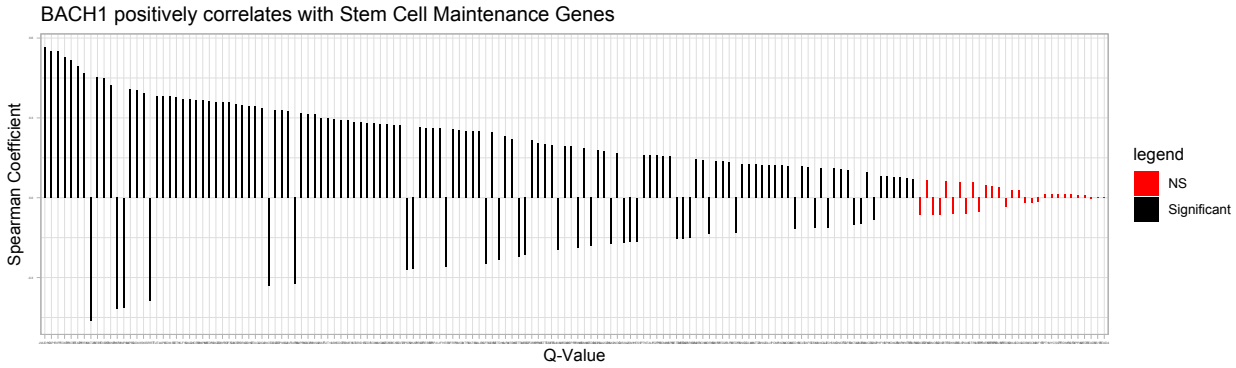

Figure 1: Example of tree plot formed by correlation between BACH1 and the "Stem Cell Maintenance" GO Term. Note the systemic bias towards positive correlations might suggest BACH1 is involved in the up-regulation of this process.

Although this plot is a visually appealing form of data representation, given its ease of interpretation, it is unclear how statistically significant this data is without the assignment of a p-value. One possible solution to this problem is to run multiple simulations involving sets of randomized genes, and taking the mean of the multiple correlations (Figure 2). From this, we can treat the resulting means like a probability distribution to test the following H_0 with real data—there is no difference between this distribution and a distribution produced by randomly selected genes. The mean was initially chosen as the statistic value in order to be weighted by the values of each coefficient, as opposed to merely the number of positive versus total correlation coefficients in the set. However, it is not ideal because it may artificially lower the value towards zero due to insignificant correlations included in the list; therefore, future work will seek to revise a better measure. An example of the resulting data from one simulation is shown in figure 3.

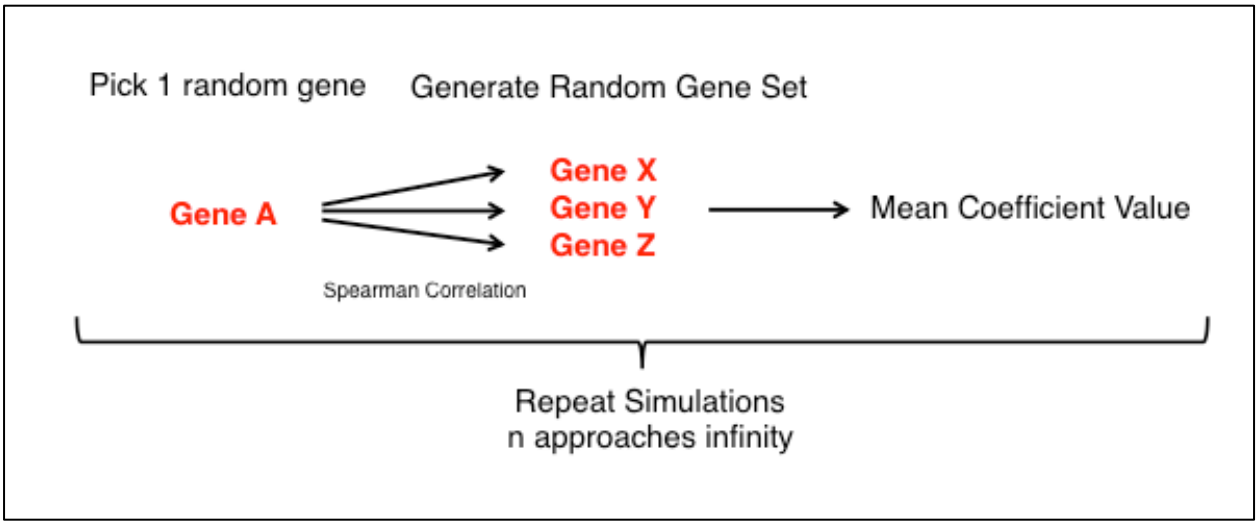

Figure 2: Schema of simulations for generating "probability distribution"

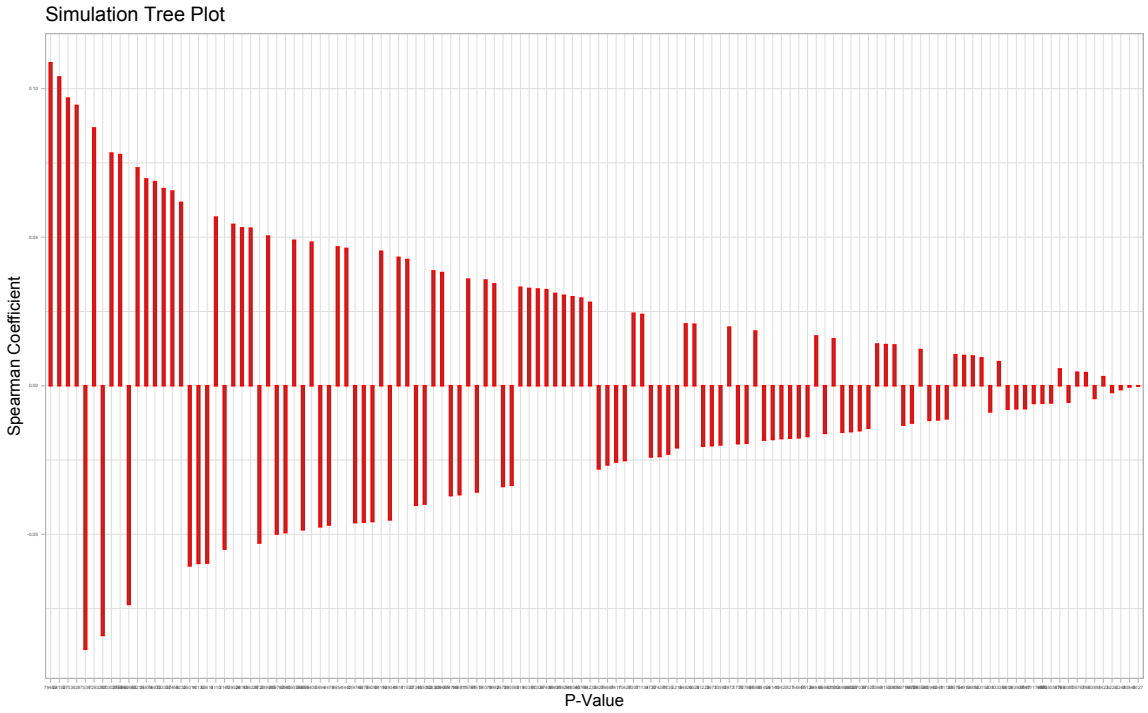

Figure 3: Example of tree plot generated by one simulation

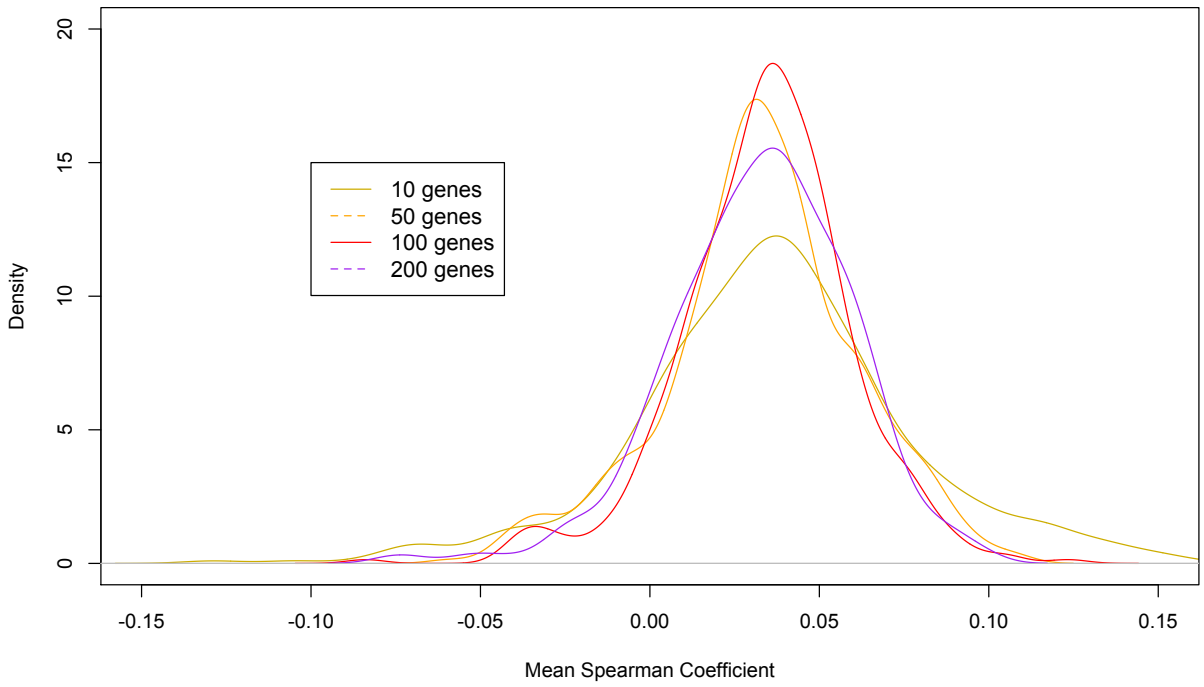

Figure 4: Density function of the resulting simulations, using different sizes for the gene set

To determine if the size of the gene set had influence on the behavior of the distribution, 500 simulations were run using sets including 10, 50, 100, or 200 genes. Note that the resulting density functions were all fairly similar regardless of the set size, although becomes wider at small gene sets (10 genes). Interestingly, the distribution is NOT centered on zero; instead, it is slightly shifted to the right, indicating that positive correlations are a slightly more common within the TCGA tumor data.

To test the efficacy of this method, I looked for the involvement of particular genes of interest within a list of 30 different metabolic GO terms containing between 50 and 300 genes in size. The "probability-value" was calculated as the area under the curve between positive/negative infinity and the mean spearman coefficient generated for that GO term, using the density function generated by the 200 gene set simulations. An example of this is shown with OCT4 in figure 5, where no significant hits were uncovered. One can see that all the red points (Metabolic GO Terms) are closely situated under the density function from the simulations. This was then repeated using BACH1, and several interesting hits were identified, in order of statistical significance: positive regulation of ROS metabolic process, positive regulation of carbohydrate metabolic process, electron transport chain, and the top hit was OXPHOS (Figure 6, 8). Given the previous work that has identified BACH1 as a significant regulator of OXPHOS, this validated these methods as a useful tool for identifying biological processes involved with a particular gene of interest. Examples of the plots generated by some of these GO terms are shown in figures 7 (non-significant terms) and 8 (significant terms).

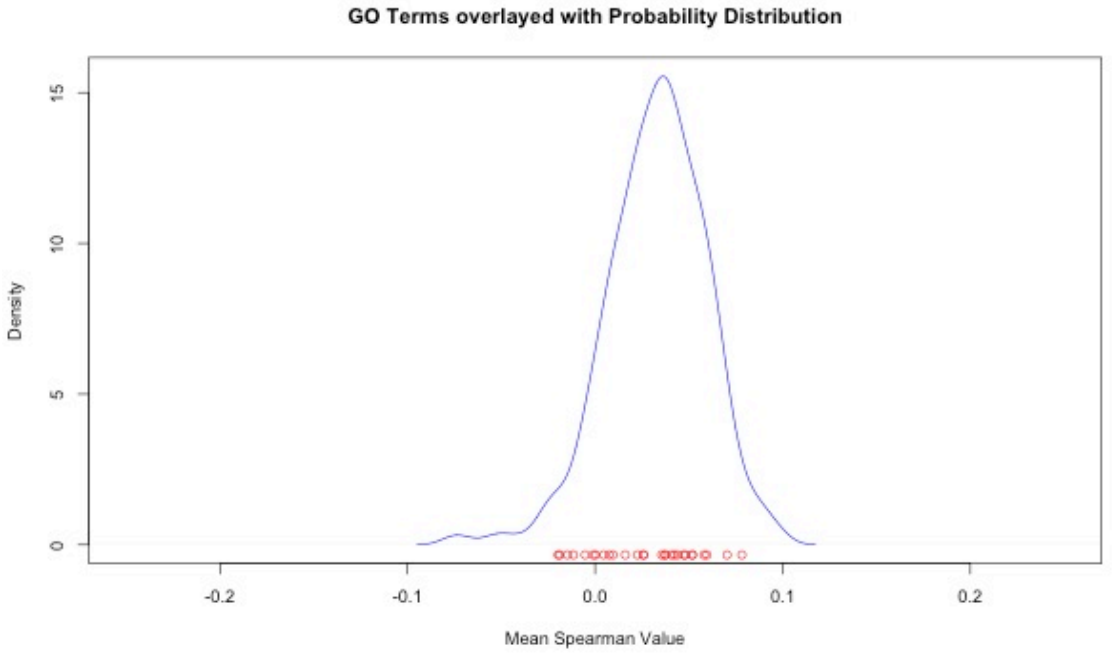

Figure 5: No Significant hits for metabolic GO terms with OCT4.

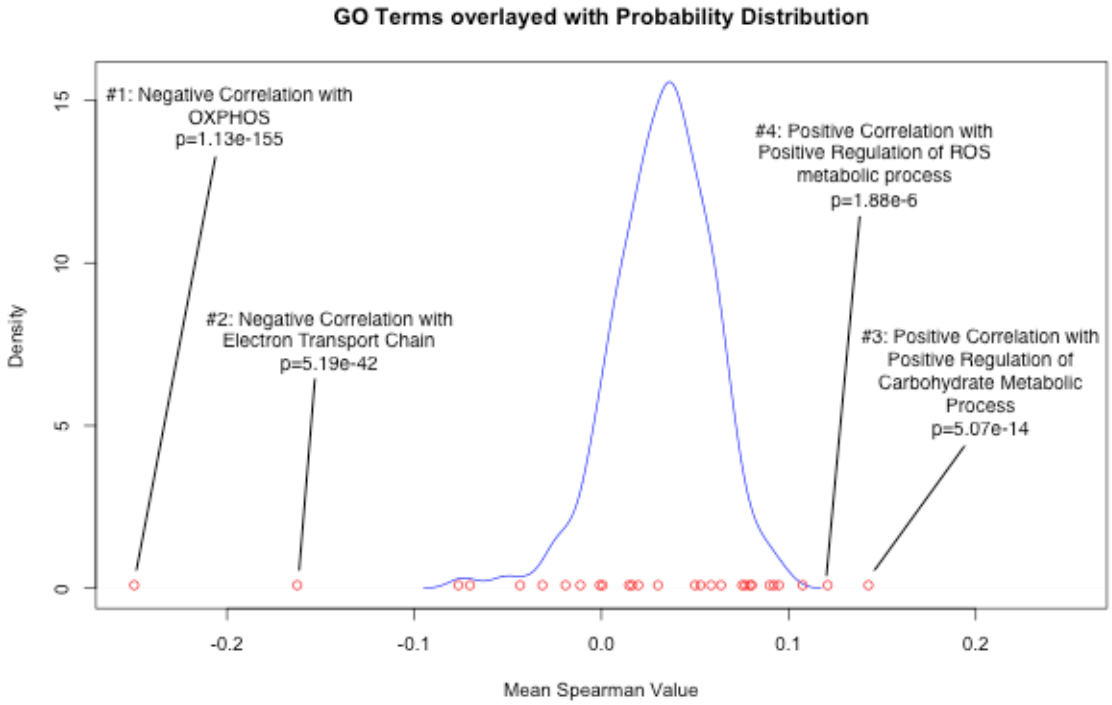

Figure 6: Multiple hits for metabolic GO terms seen with BACH1.

Shown in figure 7 are examples of plots generated by some of the insignificant GO terms of the BACH1 analyses, such as lipid catabolic process and membrane biosynthesis.

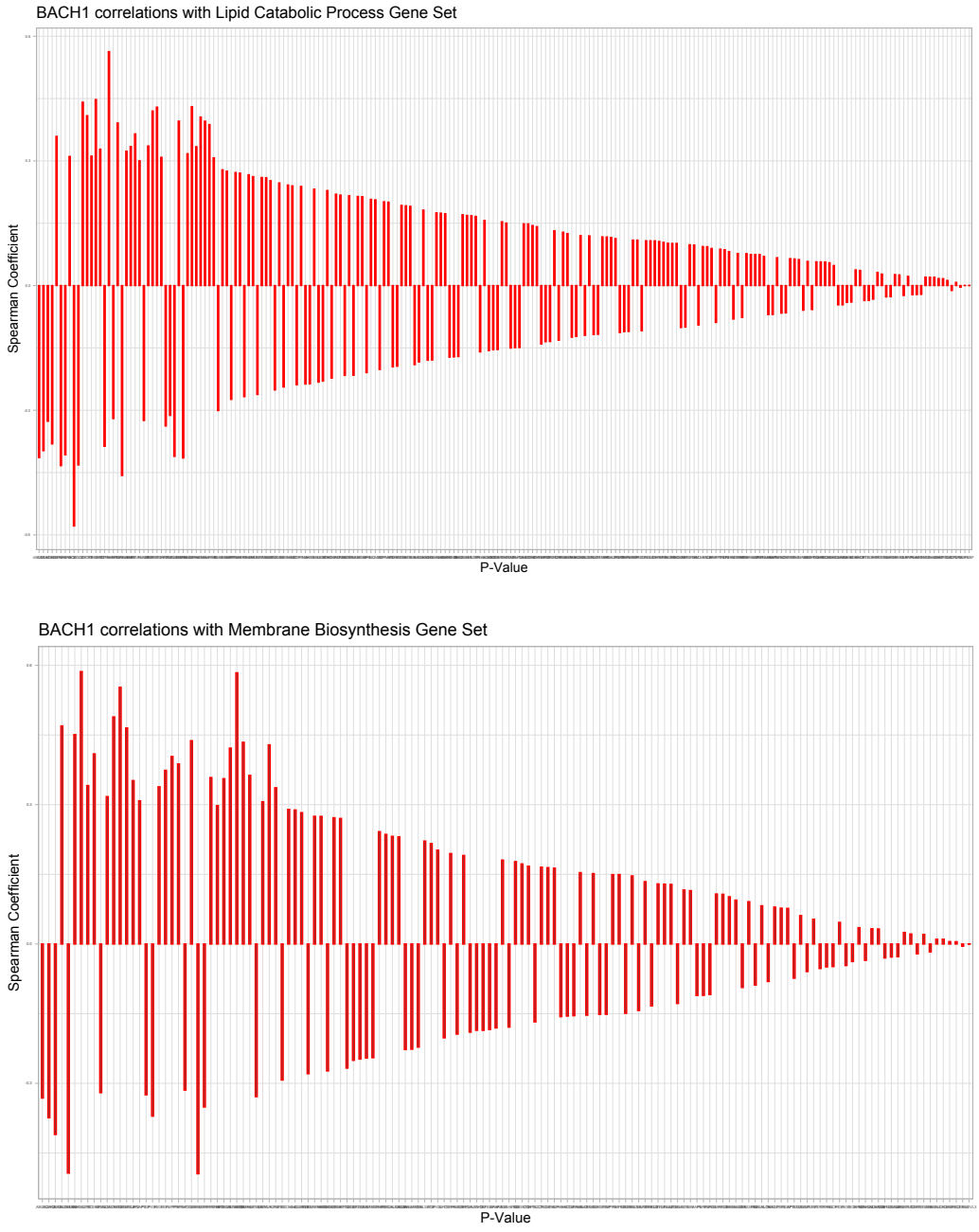

Figure 7: Examples of Insignificant hits from the BACH1 data, note the apparent random distribution of coefficients around zero. Shown above is lipid catabolic process (p=0.44), and below is membrane biosynthesis (p=0.26).

Figure 8: Significant hits of Metabolic GO Terms with BACH1. Shown above is the fourth hit, positive regulation of ROS metabolic process gene set (p=1.88e-6), and below is the top hit, OXPHOS (p=1.1e-155).