Q-Value Primer

Michael R. Mehan

The q-value algorithm we use is based on John Storey’s work on false discovery rates. The tool we use is called fdrtool and is written by K. Strimmer. It uses the distribution of the p-values to predict what proportion of the p-values is truly significant. If no significant markers exist between two groups then the p-values for KS-distances between the two groups should be uniformly distributed. Therefore, you will get some significant, i.e., P < 0.05, by chance alone. This is especially true when measuring close to 1000 proteins. You are as likely to get a p-value of .01 as 0.99 if there are no truly significant markers that separate the two groups.

Therefore, the FDR algorithm views the p-value distribution as a mixture of two distributions: the null distribution which is uniform and is derived from all the truly insignificant markers; and the alternative distribution which is derived from the truly significant markers which should be heavily skewed towards low p-values. The algorithm tries to predict what proportion of p-values is from the alternative distribution. If the p-values show no evidence of having any skew towards low p-values, or to say it another way, if the p-values are uniformly distributed, then the q-values will all be 1.0.

Below is a histogram of the p-values for Classifier 3. Notice there is no enrichment of low p-values relative to the rest of the large p-values. The accompanying figure on the right is a Q-Q plot of the p-value distribution with a uniform distribution. The fact that this line is very linear indicates that the p-values are similar to a uniform distribution. I have included the same 2 figures for a study in which we had highly significant p-values on the next page for your reference.



