Statistical tests for detecting differential expression in pair-wise experiments

1. Statistical tests that do not require replicates

Given a target protein (*X*), the spectral counts from a pair-wise experiment are arranged in a two-way table:

	Condition 1	Condition 2	Total
Spectral count for a target protein *X*	x_1	x_2	x
Spectral count for any other protein	y_1	y_2	y
Total spectral count across all proteins	n_1	n_2	n

The proportion of spectral count for target protein *X* is $r_1 = x_1 / n_1$ under Condition 1 and $r_2 = x_2 / n_2$ under Condition2. Testing whether a target protein *X* is differentially expressed under the two conditions is equivalent to testing the statistical significance of the difference between the two proportions. Note that the G test, the Fisher's exact, and the AC test do not require replicates, thus the spectral count data for these tests are pooled for replicated experiments.

1.1 The *G* test

The *G* test[1] tests the goodness of fit of the observed spectral count values to their expectationed ones based on the null hypothesis of independence. Given the two-way table, the *G* test is calculated based on a multinomial distribution:

$$G = 2 \times (x_1 \ln x_1 + x_2 \ln x_2 + y_1 \ln y_1 + y_2 \ln y_2 + n \ln n - n_1 \ln n_1 - n_2 \ln n_2 - x \ln x - y \ln y).$$

The *G* statistics approximately follows the χ^2 distribution with one degree of freedom. This approximation can lead to a higher type I error than the intended level. William's correction (*w*)[1] can be used for adjustment:

$$w = 1 + \frac{\left(\frac{n}{n_1} + \frac{n}{n_2} - 1\right)\left(\frac{n}{x} + \frac{n}{y} - 1\right)}{6n}.$$

The adjusted *G* statistics is then defined as:

$$G_{adj} = G/w.$$

1.2 The Fisher's exact test

The Fisher's exact test[2] assumes that the row totals and column totals are fixed in the two-way table. Hence, any one entry in the table completely determines the others. If the spectral count for protein *X* under Condition 1 is denoted as *K*, then the probability of obtaining the value x_1 for *K*, given the row and column margins, could be calculated based on the hypergeometric distribution:

$$p(K = x_1) = \frac{n_1!n_2!x!y!}{n!x_1!x_2!y_1!y_2!}.$$

If $x_1/n_1 > x_2/n_2$, then $\sum_{i=x_1}^{n_1} p(K = i)$ could be used in a one-sided test.

On the other hand, if $x_1/n_1 < x_2/n_2$, then $\sum_{i=0}^{x_1} p(K = i)$ could be used in a one-sided test.

1.3 The AC test

The AC test[3] calculates the conditional probability of finding x_2 spectral count in Condition 2 given the fact that x_1 spectral count has been observed in Condition 1:

$$p(x_2 \mid x_1) = \frac{(n_2/n_1)^{x_2}(x_1 + x_2)!}{x_1!x_2!(1 + n_2/n_1)^{(x_1+x_2+1)}}.$$

If $x_2/n_2 < x_1/n_1$, then $\sum_{i=0}^{x_2} p(i \mid x_1)$ could be used in a one-sided test.

On the other hand, if $x_2/n_2 > x_1/n_1$, then $\sum_{i=0}^{x_1} p(i \mid x_2)$ could be used in a one-sided test.

2. Statistical tests that require replicates

The spectral count data for replicates are kept separately in these tests. The global normalization is applied to normalize the data before using the tests. Specifically, for a target protein (X), its spectral count in each experiment is divided by the average spectral count in that experiment across all the proteins so that the global average count is the same across all experiments. The normalized spectral counts are used for the tests.

2.1 The *t*-test

Assuming a target protein X has an average spectral count of $Mean_1$ across the n_1 replicates $\{x1_1, x1_2, \ldots x1_{n1}\}$ under Condition 1, and an average spectral count of $Mean_2$ across the n_2 replicates $\{x2_1, x2_2, \ldots x2_{n2}\}$ under Condition 2, the t statistic[4] is:

$$t = \frac{Mean_1 - Mean_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

where

$$\sigma_1^2 = \frac{\sum_{i=1}^{n_1}(x1_i - Mean_1)^2}{n_1 - 1} \text{ and } \sigma_2^2 = \frac{\sum_{i=1}^{n_2}(x2_i - Mean_2)^2}{n_2 - 1}.$$

2.2 The LPE test

The LPE test[5] is modified from the *t*-test. It pools proteins with similar counts by percentile intervals and fits a smooth local regression curve to the variance estimates on

the percentiles. The averages in the *t*-test are replaced by medians. The variances estimated from a single protein are replaced by those estimated from the pooled values. The LPE test implemented in bioconductor (http://www.bioconductor.org) is used for this study.

References:

1. Sokal, R. R.; Rohlf, F. J., *Biometry*. 3 ed.; W.H. Freeman and Company: New York, 1995.
2. Fisher, R. A., *Statistical methods for research workers*. Oliver & Boyd: Edinburgh, 1925.
3. Audic, S.; Claverie, J. M., The significance of digital gene expression profiles. *Genome Res* **1997,** 7, (10), 986-95.
4. Student, On the error of counting with a haemacytometer. *Biometrika* **1907,** 5, 351-360.
5. Jain, N.; Thatte, J.; Braciale, T.; Ley, K.; O'Connell, M.; Lee, J. K., Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **2003,** 19, (15), 1945-51.