ComparativeMarkerSelection Documentation

Module name:	ComparativeMarkerSelection
Description:	Computes significance values for features using several metrics, including FDR(BH), Q Value, FWER, Feature-Specific P-Value, and Bonferroni.
Author:	Joshua Gould, Gad Getz, Stefano Monti (Broad Institute) gp-help@broad.mit.edu

The ComparativeMarkerSelection module includes several approaches to determine the features that are most closely correlated with a class template and the significance of that correlation. If the input class template has more than two classes, than a one-versus-all comparison is performed for each class. Note that the p-values obtained from the one-versus-all comparison are not fully corrected for multiple hypothesis testing. The module outputs a file containing the following columns:

1. ***Rank*** - The rank of the feature within the dataset based on the value of the test statistic. If a two-sided p-value is computed, the rank is with respect to the absolute value of the statistic.
2. ***Feature*** - The feature name.
3. ***Description*** – The description of the feature.
4. ***Score*** - The value of the test statistic.
5. ***Feature P*** - The feature-specific p-value based on permutation testing.
6. ***Feature P Low*** - *The* estimated lower bound for the feature p-value.
7. **Feature P High** -The estimated upper bound for the feature p-value.
8. ***FDR (BH)*** - An estimate of the false discovery rate by the Benjamini and Hochberg procedure (3). The FDR is the expected proportion of erroneous rejections among all rejections.
9. ***Q Value*** - An estimate of the FDR using the procedure developed by Storey and Tibshirani (4).
10. ***Bonferroni*** - The value of the Bonferroni correction applied to the feature specific p-value.
11. **maxT** - The adjusted *p*-values for the maxT multiple testing procedure described in (5), which provides strong control of the FWER.
12. ***FWER*** **(Family Wise Error Rate)** - the probability of at least one null hypothesis/feature having a score better than or equal to the observed one. This measure is not feature-specific.
13. **Fold Change** – The class zero mean divided by the class one mean.
14. **Class Zero Mean** – The class zero mean.
15. **Class Zero Standard Deviation** – The class zero standard deviation.
16. **Class One Mean** – The class one mean.
17. **Class One Standard Deviation** – The class one standard deviation.

The results from the ComparativeMarkerSelection algorithm can be viewed with the ComparativeMarkerSelectionViewer.

Parameters:

Name	Description

input filename	The input file - .res, .gct, .odf type=Dataset
cls filename	The class file - .cls
confound variable cls filename	The class file containing the confounding variable - .cls
test direction	The test to perform (up-regulated for class 0, up-regulated for class 1, two-sided)
test statistic	The statistic to use
min std	The minimum standard deviation if test statistic includes min std option
number of permutations	The number of permutations to perform (use 0 to calculate asymptotic p-values)
complete	Whether to perform all possible permutations
balanced	Whether to perform balanced permutations
random seed	The seed of the random number generator
smooth p values	Whether to smooth p-values
phenotype test	Tests to perform when cls file has more than two classes (Note: not fully corrected for multiple hypothesis testing
output file	The name of the output file

Return Value:

An odf file of type ComparativeMarkerSelection

References:

- Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. **57**(1): p. 289-300.
- Golub, T., Slonim, D. et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science* **286**, 531-537.
- Good, P. (1994) Permutation Tests: A Practical Guide for Testing Hypotheses, New York: Springer-Verlag
- Lu, J., Getz, G., Miska, E., et al. (2005) MicroRNA Expression Profiles Classify Human Cancers. *Nature* **435**, 834-838
- Storey, J.D. and R. Tibshirani (2003) Statistical significance for genomewide studies. *PNAS*, **100**(16): p. 9440-9445.
- Westfall, P.H. and S. S. Young (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. *Wiley Series in Probability and Statistics*. New York: Wiley.

Platform dependencies:

Task type: Gene List Selection
CPU type: any
OS: any
Java JVM level: 1.4
Language: Java, R