GenePattern

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.
Authors:	Joshua Gould, Gad Getz, Stefano Monti, Alper Uzun (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. 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 (1). 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 (6).
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 Westfall (7), 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.
18. **k** – If performing a two-sided test or a one-sided test for markers of class zero, the number of permuted scores greater than or equal to the observed score. If testing for markers of class one, then the number of permuted scores less than or equal to the observed score.

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

Parameters:

Name	**Description**
input filename	The input file - .res, .gct, .odf of type Dataset
cls filename	The class file - .cls
confound variable cls filename	The class file containing the confounding variable - .cls If you specify a confounding variable class file, permutations shuffle the phenotype labels only within the subsets defined by that class file. For example, in Lu, Getz, et. al. (2005), to select features that best distinguish tumors from normal samples on all tissue types, tissue type is treated as the confounding variable. In this case, the confounding variable class file lists each tissue type as a phenotype and associates each sample with its tissue type. Consequently, when ComparativeMarkerSelection performs permutations, it shuffles the tumor/normal labels only among samples with the same tissue type.
test direction	The test to perform (up-regulated for class 0, up-regulated for class 1, two-sided). By default, ComparativeMarkerSelection performs the two-sided test. If you are interested in a one-sided test: perform the one-sided test, perform the two-sided test, and consider both sets of results.
test statistic	The statistic to use:

t-test	$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}}$ where μ is the average σ is the standard deviation n is the number of samples
t-test (median)	same as t-test, but uses median rather than average
t-test (min std)	same as t-test, but enforces a minimum value for σ (minimal standard deviation)
t-test (median, min std)	same as t-test, but uses median rather than average and enforces a minimum value for σ (minimal standard deviation)
SNR (signal-to-noise ratio)	$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$ where μ is the average σ is the standard deviation
SNR (median)	same as SNR, but uses median rather than average
SNR (min std)	same as SNR, but enforces a minimum value for σ (minimal

	<table><tr><td></td><td>standard deviation)</td></tr><tr><td>SNR (median, min std)</td><td>same as SNR, but uses median rather than average and enforces a minimum value for σ (minimal standard deviation)</td></tr></table>
min std	Used only if *test statistic* includes the min std option. If σ is less than *min std*, σ is set to *min std.*
number of permutations	The number of permutations to perform (use 0 to calculate asymptotic p-values). The number of permutations you specify depends on the number of hypotheses being tested and the significance level that you want to achieve (3). The greater the number of permutations, the more accurate the p value.
complete	Whether to perform all possible permutations. By default, *complete* is set to no and *number of permutations* determines the number of permutations performed. If you have a **small** number of samples, you might want to perform all possible permutations.
balanced	Whether to perform balanced permutations. By default, *balanced* is set to no and phenotype labels are permuted without regard to the number of samples per phenotype (for example, if your dataset has 12 samples in class 0 and 10 in class 1, any permutation of class 0 has 12 randomly selected samples). If you set *balanced* to yes, phenotype labels are permuted by balancing the number of samples per phenotype (for example, if your dataset has 12 samples in class 0 and 10 in class 1, any permutation of class 0 has an equal number of samples from class 0 and class 1. Balancing samples is rarely necessary, but may be useful if your samples are very unevenly distributed across phenotypes.
random seed	The seed for the random number generator.
smooth p values	Whether to smooth p-values by using the Laplace's Rule of Succession. By default, *smooth p values* is set to yes, which means p-values are always less than 1.0 and greater than 0.0.
phenotype test	Tests to perform when cls file has more than two classes: one-versus-all, all pairs. **Note**: The p-values obtained from the one-versus-all comparison are not fully corrected for multiple hypothesis testing.
output file	The name of the output file

Output Files:

An odf file of type ComparativeMarkerSelection

References:

1. 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.

2. Golub, T., Slonim, D. et al. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science* **286**, 531-537.
3. Good, P. (2000) Permutation Tests: A Practical Guide for Testing Hypotheses, 2nd Ed., New York: Springer-Verlag
4. Gould J., Getz G., Monti S., Reich M., and Mesirov J.P. (2006) Comparative gene marker selection suite. *Bioinformatics* **22**, 1924-1925; doi:10.1093/bioinformatics/btl196.
5. Lu, J., Getz, G., Miska, E., et al. (2005) MicroRNA Expression Profiles Classify Human Cancers. *Nature* **435**, 834-838
6. Storey, J.D. and R. Tibshirani (2003) Statistical significance for genomewide studies. *PNAS*, **100**(16): p. 9440-9445.
7. 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:

Module type:	Gene List Selection
CPU type:	any
OS:	any
Java JVM level:	1.5
Language:	Java, R