GenePattern

ComparativeMarkerSelection Documentation

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, Alper Uzun (Broad Institute), gp-help@broadinstitute.org

Summary

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.

By default ComparativeMarkerSelection expects the data in the input file to **not** be log transformed. Some of the calculations such as the fold are not accurate when log transformed data is provided and not indicated. To indicate that your data is log transformed, be sure to set the "log transformed data" parameter to "yes". Also, ComparativeMarkerSelection requires **at least three** samples per class to run successfully.

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.

References

1. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological*). 1995;57(1):289-300.
2. Golub T, Slonim D, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science.* 1999;286:531-537.
3. Good P. *Permutation Tests: A Practical Guide for Testing Hypotheses*, 2nd Ed. New York: Springer-Verlag. 2000.
4. Gould J, Getz G, Monti S, Reich M, Mesirov JP. Comparative gene marker selection suite. *Bioinformatics.* 2006;22;1924-1925, doi:10.1093/bioinformatics/btl196.
5. Lu J, Getz G, Miska E, et al. MicroRNA Expression Profiles Classify Human Cancers. *Nature.* 2005;435:834-838.
6. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *PNAS.* 2003;100(16):9440-9445.
7. Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment, in *Wiley Series in Probability and Statistics*. New York: Wiley, 1993.

Parameters

input file

The input file - .res, .gct

Note the following constraints:

- If the expression data contains duplicate identifiers, ComparativeMarkerSelection generates the error message: "An error occurred while running the algorithm." The UniquifyLabels module provides one way of handling duplicate identifiers.
- If the expression data contains fewer than three samples per class, ComparativeMarkerSelection appears to complete successfully but test statistic scores are not shown in the results.
- If the expression data contains missing values, ComparativeMarkerSelection completes successfully but does not compute test statistic scores for rows that contain missing values.
- For more information on the RES and GCT file formats, see the File Formats page in the GenePattern documentation: (http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_fileformats)

cls file

The class file - .cls

ComparativeMarkerSelection analyzes two phenotype classes at a time. If the expression data set includes samples from more than two classes, use the *phenotype test* parameter to analyze each class against all others (one-versus-all) or all class pairs (all pairs).

For more information on the CLS file format, see the File Formats page in the GenePattern documentation:
(http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_fileformats)

confounding variable cls file

The class file containing the confounding variable - .cls

If you are studying two variables and your data set contains a third variable that might distort the association between the variables of interest, you can use a confounding variable class file to correct for the affect of the third variable. For example, the data set in Lu, Getz, et. al. (2005) contains tumor and normal samples from different tissue types. When studying the association between the tumor and normal samples, the authors use a confounding variable class file to correct for the effect of the different tissue types.

The phenotype class file identifies the tumor and normal samples:

```
75 2 1 # Normal Tumor 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1
1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
1 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1
```

The confounding variable class file identifies the tissue type of each sample:

```
75 6 1 # colon kidney prostate uterus human-lung breast 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5
5 5 5 5
```

Given these two class files, when performing permutations, ComparativeMarkerSelection shuffles the tumor/normal labels only among samples with the same tissue type.

test direction

By default, ComparativeMarkerSelection performs a two-sided test; that is, the test statistic score is calculated assuming that the differentially expressed gene can be up-regulated in either phenotype class. Optionally, use the *test direction* parameter to specify a one-sided test, where the differentially expressed gene must be up-regulated for class 0 or for class 1.

test statistic

The statistic to use.

t-test	This is the standardized mean difference between the two classes. It is the difference between the mean expression of class 1 and class 2 divided by the variability of expression, which is the square root of the sum of the standard deviation for each class divided by the number of samples in each class. $\frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$ 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	The signal-to-noise ratio is computed by dividing the difference of class means by the sum of their standard deviations. $\frac{m_A - m_B}{s_A + s_B}$ where μ is the average and 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 standard deviation).
SNR (median, min std)	Same as SNR, but uses median rather than average and enforces a minimum value for σ (minimal standard deviation).

Paired t-test	The Paired T-Test can be used to analyze paired samples; for example, samples taken from patients before and after treatment. This test is used when the cross-class differences (e.g. the difference before and after treatment) are expected to be smaller than the within-class differences (e.g. the difference between two patients). For example if you are measuring weight gain in a population of people, the weights may be distributed from 90 lbs. to say 300 lbs. and the weight gain/loss (the paired variable) may be on the order of 0-30 lbs. So the cross-class difference ("before" and "after") is less than the within-class difference (person 1 and person 2). The standard T-Test takes the mean of the difference between classes, the Paired T-Test takes the mean of the differences between pairs: $\frac{\overline{C}_D - m_0}{s_D/\sqrt{N}}$ where the differences between all pairs are calculated and X_D is the average of the differences and s_D the standard deviation. **Note**: For the Paired T-Test, paired samples in the expression data file must be arranged by class, where the first samples in each class are paired, the second samples are paired, and so on. For example, sample pairs A1/B1, A2/B2 and A3/B3 would be ordered in an expression data file as A1, A2, A3, B1, B2, B3. Note that your data must contain the same number of samples in each class in order to use this statistic.

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

ComparativeMarkerSelection uses a permutation test to estimate the significance (p-value) of the test statistic score. The number of permutations you specify depends on the number of hypotheses being tested and the significance level that you want to achieve (3). If the data set includes at least 10 samples per class, use the default value of 10000 permutations to ensure sufficiently accurate p-values.

If the data set includes fewer than 10 samples in any class, permuting the samples cannot give an accurate p-value. Specify a value of 0 permutations to use asymptotic p-values instead. In this case, ComparativeMarkerSelection computes p-values assuming the test statistic scores follow Student's t-distribution (rather than using the *test statistic* to create an empirical distribution of the scores). Asymptotic p-values are calculated using the p-value obtained from the standard independent two-sample t-test.

log transformed data

By default ComparativeMarkerSelection expects the data in the input file to **not** be log transformed. Some of the calculations such as the fold are not accurate when log transformed data is provided and not indicated. To indicate that your data is log transformed, set this parameter to "yes".

complete

When the *complete* parameter is set to yes, ComparativeMarkerSelection ignores the *number of permutations* parameter and computes the p-value based on all possible sample permutations. Use this option only with small data sets, where the number of all possible permutations is less than 1000.

balanced

When the *balanced* parameter is set to yes, ComparativeMarkerSelection requires an equal and even number of samples in each class (e.g. 10 samples in each class, not 11 in each class or 10 in one class and 12 in the other).

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 filename

The name of the output file.

Output Files

1. ODF file

   An .odf file of type ComparativeMarkerSelection. For more information on the ODF file format, see the File Formats page in the GenePattern documentation: (http://www.broadinstitute.org/cancer/software/genepattern/tutorial/gp_fileformats)

Platform Dependencies

Module type:	Gene List Selection
CPU type:	any
OS:	any
Language:	Java (minimum version 1.5), R