



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. **Score** - The value of the test statistic.
4. **Feature P** - The feature-specific p value based on permutation testing.
5. **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.
6. **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.
7. **Bonferroni** - The value of the Bonferroni correction applied to the feature specific p value.
8. **Q Value** - An estimate of the FDR using the procedure developed by Storey and Tibshirani (4).
9. **maxT** - The adjusted *p*-values for the maxT multiple testing procedure described in (5), which provides strong control of the FWER.

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 is T-Test (min |

GenePattern

| | |
|------------------------|--|
| | std) |
| number of permutations | The number of permutations to perform |
| 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 |
| significance booster | Whether to attempt to increase the p-value confidence for significant features |
| theta | Value for removing features when using significance booster |
| output file | The name of the output file |

Return Value:

An odf file of type ComparativeMarkerSelection

References:

1. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression*. Science, 1999. **286**(5439): p. 531-537.
2. Slonim, D.K., et al., *Class Prediction and Discovery Using Gene Expression Data*, in *RECOMB 2000: The Fourth Annual International Conference on Research in Computational Molecular Biology*. 2000: Tokyo, Japan. p. 263-272.
3. Benjamini, Y. and Y. Hochberg, *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): p. 289-300.
4. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies*. PNAS, 2003. **100**(16): p. 9440-9445.
5. Westfall, P.H. and S.S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley Series in Probability and Statistics. 1993, New York: Wiley.

Platform dependencies:

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