Comparative Gene Marker Selection Suite

Joshua Gould*, Gad Getz, Stefano Monti, Michael Reich, Jill P. Mesirov

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

ABSTRACT

Motivation: An important step in analyzing expression profiles from microarray data is to identify genes that can discriminate between distinct classes of samples. Many statistical approaches for assigning significance values to genes have been developed. The Comparative Marker Selection suite consists of three modules that allow users to apply and compare different methods of computing significance for each marker gene, a viewer to assess the results, and a tool to create derivative datasets and marker lists based on user-defined significance criteria.

Availability: The Comparative Marker Selection application suite is freely available as a GenePattern module. The GenePattern analysis environment is freely available at <http://www.broad.mit.edu/genepattern>.

Contact: jgould@broad.mit.edu

1 INTRODUCTION

When analyzing genome-wide transcription profiles from microarray data, the first step is often to identify genes that can discriminate between distinct classes of samples (usually defined by a phenotype, such as tumor or normal). This process is commonly referred to as marker (or feature) selection. Many statistical approaches have been developed to estimate the significance of marker genes. The Comparative Marker Selection suite implements many of these approaches, and provides visualization tools for easy comparison of the results that are generated. The suite consists of three modules in the GenePattern (Reich et al., 2006) analysis environment and includes (1) an analytic module that computes the statistical significance of each gene, and includes several methods of correcting for multiple hypothesis testing, (2) a visualization module to aid in the evaluation of the analytic module results, and (3) a utility module to create derivative datasets and marker lists based on user-defined significance criteria. Complete documentation for GenePattern is provided on the GenePattern website. Additionally, each module in the GenePattern environment contains detailed documentation. We also provide default settings for all module parameters.

2 ANALYTIC MODULE

The analytic module takes as input a data set of expression profiles from samples belonging to two phenotypes. Either two-channel (cDNA) or absolute value (Affymetrix) data can be used as input to the module and any missing values can be imputed (see FAQ on the GenePattern website). If a dataset contains multiple phenotypes, there is the option to perform all pairwise comparisons or all one-versus-all comparisons. A test statistic (e.g., t-test) is chosen to assess the differential expression between the two classes of samples. Note that technical and biological replicates are handled the same way as independent samples. The significance (nominal p-value) of marker genes is computed using a permutation test, which is a commonly used method for assessing the significance of marker genes. Permutation tests have the advantage of not assuming a parametric underlying distribution of expression values, and importantly they preserve gene-gene correlations which affect some measurements of significance. To construct a distribution of the test statistic, under the null hypothesis of no differential expression, phenotype labels are randomly re-assigned to samples and the test statistic is recomputed for the relabeled data set. This procedure is repeated for a given number of relabelings to yield the empirical null. It should be noted that the total number of possible exhaustive permutations is a function of the number of samples in each class (for example, given 20 samples, 10 of each class, the total number of distinct permutations would be); we suggest a minimum of 10 samples per class. In cases where the number of permutations is insufficient to estimate a significant p-value, the module provides the option of computing asymptotic p-values based on the t-test. In addition to reporting the nominal p-value, we also report the estimated 95% confidence intervals for the nominal p-value to assess p-value accuracy. We also include an option to perform all possible relabelings to obtain exact p-values.

Selecting class markers is a particular instance of the general multiple hypothesis testing (MHT) problem. Since several thousand hypotheses are usually tested at once, the nominal p-values have to be corrected to account for the increased number of potential false positives. For example, if we test 20,000 genes for differential expression, a nominal p-value threshold of 0.01 would only ensure that the expected number of false positives is less than 200 (0.01*20,000).

One approach to adjusting for multiple hypothesis testing is to control the false discovery rate (FDR), the expected fraction of false positives among all genes reported as significant. In most cases, controlling the FDR is sufficient because the purpose of most microarray investigations is to generate hypotheses for further study. The FDR cut-off level controls the fraction of false leads that the user is willing to tolerate. We include two methods for computing the FDR: the BH procedure developed by Benjamini and Hochberg (Benjamini and Hochberg, 1995) and the q-value method of Storey and Tibshirani (Storey and Tibshirani, 2003). The BH procedure gives a more conservative estimate of the FDR than the q-value. The q-value attempts to gain its extra power from estimating π_0 ($0 \leq \pi_0 \leq 1$), the fraction of true null hypotheses among all tested hypotheses (the BH procedure always assumes $\pi_0 = 1$). When this fraction is large, which is the case for many microarray experiments in which we test thousands of genes of which we expect very few to be differentially expressed, little advantage is gained by using the q-value. However, if we expect a large proportion of the tested genes to be differentially expressed, the q-value might allow for the reduction of false negatives when compared with the BH method. We include a plot of the π_0 estimate versus the tuning parameter λ to evaluate the accuracy of the final π_0 estimate.

*To whom correspondence should be addressed.

When a more conservative approach is required, we suggest controlling the family-wise error rate (FWER), the probability of having at least one false positive. For example, the FWER may be preferred when further investigation of any false positive is costly. We include three methods for controlling the FWER. The Bonferroni method is the most conservative, followed by the empirical FWER, and the maxT procedure (Westfall and Young, 1993).

Sometimes data contains extraneous variables that are not accounted for in the design of the experiment and that can distort the results. For example, selecting markers that distinguish tumor from non-tumor samples might lead to incorrect results if some of the samples are male and some are female. We provide the option to control for these confounding phenotypes by providing a restricted permutations (Good, 2000; Lu, Getz, Miska et al., 2005) option, in which the class labels are shuffled only within each confounding phenotype. In the example in which gender is the confounding phenotype, we can restrict the permutations so that female non-tumor labels are only permuted with female tumor labels and male non-tumor labels are only shuffled with male tumor labels.

VISUALIZATION AND UTILITY MODULES

The visualization portion of the suite provides a framework for assessing the results from the analytic module. The module includes interactive histograms used to determine null distributions for each measure of significance. Additionally a plot of the test statistic rank versus the test statistic is included, which is useful for visualizing the number of features that are upregulated in each class (see Figure 1). Pairwise comparison of different significance measures can be plotted to help users assess the relative stringency of the selected hypothesis rejection criteria. Users can visually inspect the profiles of each feature across each sample in a heat map or expression profile format.

Users can view features that pass selected filtering criteria and create derivative data sets and features lists from these filtered features. For example, a user can extract all genes that have a q-value less than 0.1 and save the corresponding dataset and marker list. We include a utility module to automate this function. All the plots in the viewer are dynamically updated to include only the features that pass the selected filtering criteria.

The annotation of features is provided by two mechanisms. Affymetrix probe set identifiers can be interactively annotated from genomic databases such as GenBank, UniGene, SwissProt, LocusLink, and Gene Ontology using the GeneCruiser Web service (Liefeld et al., 2005). Users can also enter their own annotations of features and view these annotations via a color-coding mechanism.

ACKNOWLEDGEMENTS

GenePattern is supported by funding from the National Institutes of Health. The authors wish to thank the following members of the Cancer Program at the Broad Institute: Todd Golub, Ted Liefeld, and David Twomey.

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

Liefeld. T. et al. (*2005)* GeneCruiser: A web service for the annotation of microarray data. *Bioinformatics Advance Access DOI 10.1093/bioinformatics/bti587.*

Lu, J., Getz, G., Miska, E., et al. (2005) MicroRNA Expression Profiles Classify Human Cancers. *Nature* **435**, 834-838

Reich, M., et al. (2006) GenePattern 2.0, *Nature Genetics* (in press)

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.

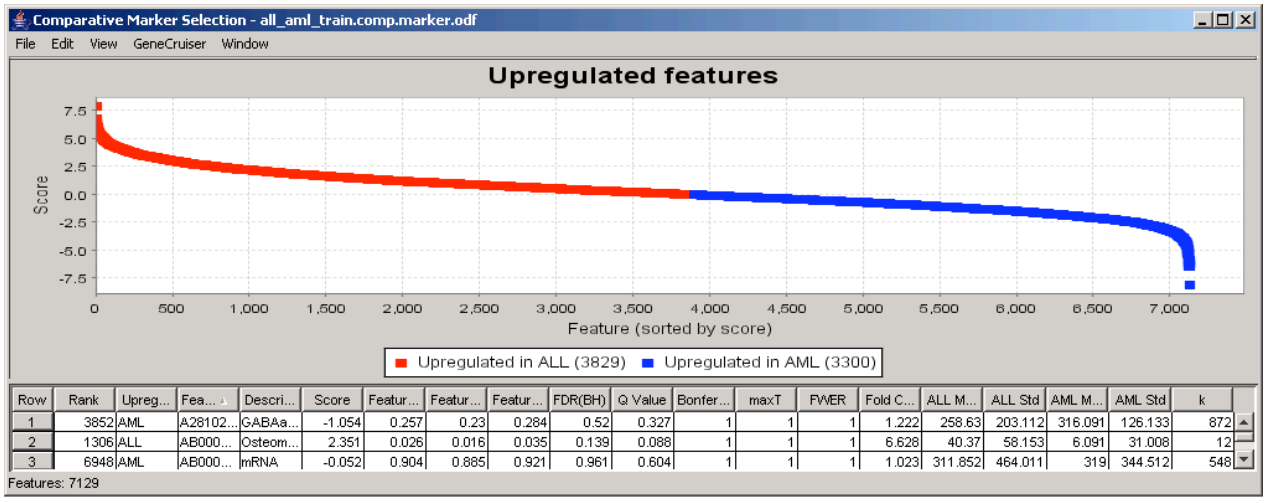

Fig. 1. Plot of test statistic score and table of significance values for data in Golub & Slonim et al., 1999.