BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 9 2005, pages 2112–2113
doi:10.1093/bioinformatics/bti183

Gene expression

# MACAT—microarray chromosome analysis tool

Joern Toedling, Sebastian Schmeier, Matthias Heinig, Benjamin Georgi and Stefan Roepcke*

Freie Universitaet Berlin, Bioinformatics programme and Max Planck Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany

Received on October 4, 2004; revised on November 15, 2004; accepted on November 24, 2004
Advance Access publication November 30, 2004

## ABSTRACT

**Summary:** By linking *differential gene expression* to the chromosomal localization of genes, one can investigate microarray data for characteristic patterns of expression phenomena involving sizeable parts of specific chromosomes. We have implemented a statistical approach for identifying significantly differentially expressed chromosome regions. We demonstrate the applicability of the approach on a publicly available data set on acute lymphocytic leukemia.

**Availability:** The R-package *MACAT* can be obtained from http://www.compdiag.molgen.mpg.de/software/macat.shtml

**Contact:** roepcke@molgen.mpg.de

**Supplementary information:** http://www.compdiag.molgen.mpg.de/software/macat.shtml

Microarray data analysts have defined tumor subtypes by specific gene expression profiles, consisting of genes that show differential expression between subtypes (Yeoh *et al*., 2002). However, tumor subtypes have also been characterized by phenomena involving large chromosomal regions. For instance, Christiansen *et al*. (2004) report on a subtype of acute myeloid leukemia, showing mutations in the *AML1* gene on chromosome 21 along with deletions or loss of chromosome arm 7q. A natural approach to bridging the gap between these two paradigms is to link scoring for differential gene expression to the chromosomal localization of genes. Tumor subtypes can be defined by differential expression patterns affecting sizeable regions of certain chromosomes.

To assist in the identification of significantly differentially expressed chromosome regions, we provide the implementation of a statistical approach. *MACAT* is written in the R statistical programming language and is part of the developmental branch of the popular Bioconductor package (Gentleman *et al*., 2004). We assume normalized expression data, which can be provided as a matrix or expression set in R or as a delimited text file. In a preprocessing step the expression data is integrated with gene location data into one common data format. To date, this step has only been implemented for commercial Affymetrix® oligo-nucleotide microarrays.

For each gene, we compute a statistic denoting the degree of differential expression between two groups of samples. This statistic is the regularized $t$-score introduced in Tusher *et al*. (2001). In essence, it is Student's $t$-statistic augmented by a fudge factor $s_0$ in the denominator, which prevents a high statistic for genes with a low variance. We set $s_0$ to the median over all gene standard deviations, analogous to Tibshirani *et al*. (2002).

The distribution of measured genes is not uniform over the length of the chromosome. Since we want to evaluate differential expression over the whole chromosome, we interpolate the statistic for positions between measured genes. This interpolation, however, does not aim to assign statistics to non-coding regions, but to provide a smooth estimate of differential expression over large chromosomal regions.

The following kernel functions are used for interpolation:

- $k$-nearest neighbor: For every chromosomal coordinate compute the average of the $k$ nearest genes.
- Radial basis function (rbf): For every coordinate compute the average over all genes weighted by distance from the coordinate.
- Base-pair distance: Similar to the $k$-nearest-neighbors, but the average is taken over all genes within a certain radius of the coordinate.

The free parameters of the kernels determine the degree of smoothing. By default, optimal parameter settings are estimated from the data by cross-validation (for details see the package's vignette).

To judge the significance of differential expression, we investigate random permutations of the class labels. To obtain a reliable simulation of the empirical distribution, we suggest observing at least $B \geq 1000$ permutations. For each permutation, the regularized $t$-statistic is computed for each gene. Thus, for each gene we obtain $B$ permutation statistics and consequently an *empirical p-value*, denoting the proportion of the permutation statistics being greater or equal than the gene's actual statistic that is based on the true class labels. The permutation statistics also provide upper and lower significance borders, which are smoothed using the same kernel function as for the original statistics.

Optionally, to judge the significance of differential expression over chromosomal regions, one can instead investigate permutations of the ordering of genes on chromosomes.

Meaningful and concise visualization facilitates a better understanding of both the data and the statistical analysis. *MACAT* includes functions for plotting expression levels and statistical scores versus base-pair coordinate on the chromosome. Regions showing significant differential expression are highlighted in these plots. One can also generate HTML-pages, which contain additional information on genes located within the highlighted chromosomal regions (Fig. 1). For each gene comprehensive annotation, a LocusLink ID,

*To whom correspondence should be addressed.

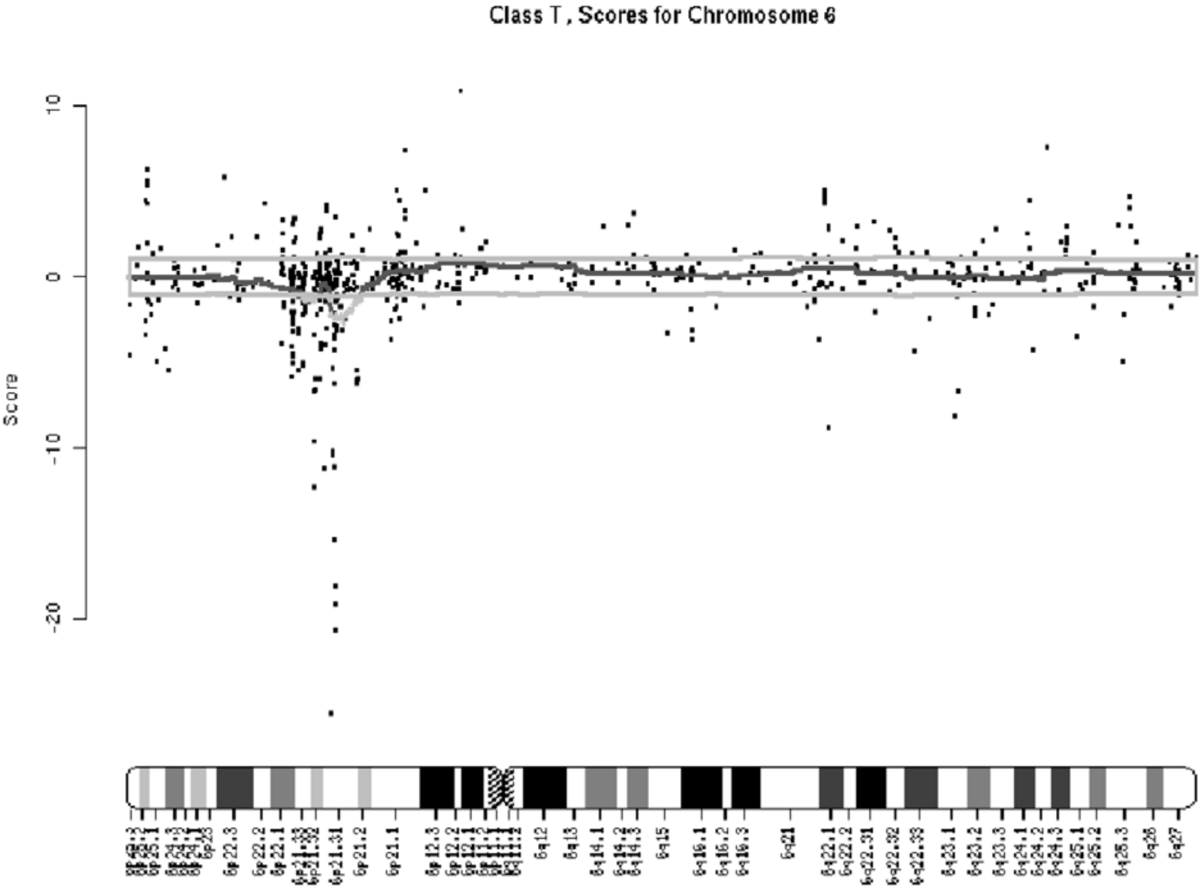

**Fig. 1.** Excerpt of a generated HTML-page for the *T-versus B-lymphocyte ALL* analysis.

with a hyperlink to the NCBI website, and the empirical *p*-value are provided. In addition, *MACAT* includes functions for writing gene expression levels and statistics into text files, which can be used with other programs for further analyses.

As an example, we present the results of an analysis on *T-versus B-lymphocyte ALL* within the publicly available data set described in Yeoh *et al*. (2002). A region on the p-arm of chromosome 6 could be identified as significantly under-expressed (Fig. 1). Among the genes within that region are the MHC class II genes, which are known to be expressed by B-lymphocytes, but not by T-lymphocytes. Since these genes are distributed over a large part of the p-arm of chromosome 6, it makes sense to assume that all genes in this region are significantly less transcribed in T-lymphocytes compared to B-lymphocytes. This gives an indication that chromosomal regions highlighted by our method are indeed biologically meaningful.

The method which we have described can detect significant differential expression for chromosomal regions. However, the reason for the differential expression, be it a mutation, translocation, hypermethylation, loss of heterozygosity, or another event, remains to be investigated.

## REFERENCES

Christiansen,D.H., Andersen,M.K. and Pedersen-Bjergaard,J. (2004) Mutations of *AML1* are common in therapy-related myelodysplasia following therapy with alkylating agents and are significantly associated with deletion or loss of chromosome arm 7q and with subsequent leukemic transformation. *Blood*, **104**, 1474–1481.

Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tusher,V., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to ionizing radiation response. *Proc. Natl Acad. Sci.*, **98**, 5116–5121.

Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.