BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 21 2006, pages 2658–2666
doi:10.1093/bioinformatics/btl455

Gene expression

# A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions

A. Callegaro, D. Basso and S. Bicciato*

Department of Chemical Process Engineering, University of Padua, Via Marzolo 9, I-35131 Padua, Italy

Received on February 11, 2006; revised and accepted on August 18, 2006
Advance Access publication September 1, 2006
Associate Editor: John Quackenbush

## ABSTRACT

**Motivation:** The systematic integration of expression profiles and other types of gene information, such as chromosomal localization, ontological annotations and sequence characteristics, still represents a challenge in the gene expression arena. In particular, the analysis of transcriptional data in context of the physical location of genes in a genome appears promising in detecting chromosomal regions with transcriptional imbalances often characterizing cancer.

**Results:** A computational tool named locally adaptive statistical procedure (LAP), which incorporates transcriptional data and structural information for the identification of differentially expressed chromosomal regions, is described. LAP accounts for variations in the distance between genes and in gene density by smoothing standard statistics on gene position before testing the significance of their differential levels of gene expression. The procedure smoothes parameters and computes *p*-values locally to account for the complex structure of the genome and to more precisely estimate the differential expression of chromosomal regions. The application of LAP to three independent sets of raw expression data allowed identifying differentially expressed regions that are directly involved in known chromosomal aberrations characteristic of tumors.

**Availability:** Functions in R for implementing the LAP method are available at http://www.dpci.unipd.it/Bioeng/Publications/LAP.htm

**Contact:** silvio.bicciato@unipd.it

**Supplementary Information:** http://www.dpci.unipd.it/Bioeng/Publications/LAP.htm

## INTRODUCTION

The exploration of all genes at once, in a systematic fashion, still represents a sort of revolution that is shifting molecular biology from a reductionistic, hypothesis-driven approach towards deciphering the mechanisms underlying gene transcription and regulation. These technological advances have been accompanied by the development of bioinformatic methods for the analysis and interpretation of an overwhelming mass of biological data. The common objective of all these methods is to identify statistically relevant genes sharing particular profiles from huge matrices bearing values for thousands of molecules. Although hundreds of studies have fully demonstrated the relevancy of genomic approaches in analyzing the transcriptional status of different physiopathological conditions, it is becoming increasingly clear that one of the next challenges in the gene expression arena is the systematic integration of expression profiles with other types of gene information, such as chromosomal localization, ontological annotations and sequence characteristics. In particular, analyzing gene expression data in context of the physical location of genes in a genome should be effective in detecting those chromosomal regions with transcriptional imbalances often characterizing cancer.

Several functional genomics studies suggested a relationship between genomic structural abnormalities and expression imbalances (under- or over-expression), and identified groups of physically contiguous genes characterized by similar, coordinated transcriptional profiles. The pioneering work of Caron and coworkers (Caron *et al*., 2001) illustrated how whole chromosome views can reveal a higher order organization of the genome with most chromosomes presenting large regions of highly expressed genes, called RIDGEs (regions of increased gene expression), interspersed with regions where gene expression is low. This research is the basis for the Human Transcriptome Map (http://bioinfo.amc.uva.nl/HTMseq/controller) where genome sequences have been integrated with mRNA expression profiles obtained from many different tissues using Serial Analysis of Gene Expression (SAGE) libraries and oligonucleotide microarrays. Comparative screenings of Transcriptome Map data highlighted numerous clusters of overexpressed and physically close genes implicated in cancer or tissue-specific pathologies (Caron *et al*., 2001; Versteeg *et al*., 2003). As a consequence, gene expression profiles have been analyzed in a chromosomal context to identify amplified or deleted chromosomal regions containing genes related to tumor initiation and progression (Lu *et al*., 2001; Crawley and Furge, 2002; Husing *et al*., 2003; Zhou *et al*., 2003; Furge *et al*., 2004; Cromer *et al*., 2004; Masayesva *et al*., 2004; Midorikawa *et al*., 2004; Reyal *et al*., 2005). Lately, genotyping studies have revealed that there is a considerable influence of copy number on gene expression patterns, and that organizing gene-expression data by genomic location and scanning for regions with significantly modulated gene-expression signals may reveal chromosomal amplifications and deletions (Pollack *et al*., 2002; Hyman *et al*., 2002; Heidenblad *et al*., 2005).

Several computational approaches have been adopted to identify chromosomal regions of increased or decreased expression from transcriptional data (Caron *et al*., 2001; Crawley and Furge, 2002; Pollack *et al*., 2002; Kano *et al*., 2003; Husing *et al*., 2003; Zhou *et al*., 2003, 2005; Toedling *et al*., 2005). All these methods score differentially expressed genes using standard statistics and then scan an array-based gene map using windows of

*To whom correspondence should be addressed.

fixed length or containing a pre-selected number of genes. In particular, the R-package called MACAT (Toedling *et al.*, 2005, http://www.compdiag.molgen.mpg.de/software/macat.shtml) links differential gene expression data to the chromosomal location of genes, interpolating a regularized *t*-score for distances between measured genes. The parameters of the kernel smoothing functions are estimated and optimized from the data through a cross-validation procedure, and statistics are smoothed over a moving window of fixed length. This strategy implicitly assumes that the distance between genes and/or the gene density are constant within chromosomes. To overcome this limitation, Levin and coworkers (Levin *et al.*, 2005) developed a model-based method that accounts for variations in gene distance, density and sequence characteristics (e.g. GC content). The statistical significance of chromosomal regions with differential gene expression is determined in comparison with a theoretical null distribution given the assumptions that the expression of each gene is normally distributed, the distances between genes are distributed as independent and identical exponential random variables, and the lengths of genes are negligible in comparison with the distances between them.

The purpose of this work is to present a non-parametric model-free statistical method, named locally adaptive statistical procedure (LAP), for the identification of differentially expressed chromosomal regions, which accounts for variations in gene distance and density. The method is based on the computation of a standard statistic (e.g. SAM *t*-statistic) as a measure of the difference in gene expression patterns between groups of samples, assessed on high-density microarrays. Once the statistical scores have been calculated, probes (or probe sets for Affymetrix arrays) are re-annotated in terms of Entrez Gene IDs and the statistics are sorted, on each chromosome, according to the chromosomal coordinate (in base pairs) of the corresponding gene. For each chromosome, the statistic is locally smoothed using non-parametric estimation of regression function over the positional coordinate. Differentially expressed regions are identified using a permutation procedure. In particular, gene positions are randomly shuffled and the randomly generated statistics are smoothed to generate the null smoothed distribution. This empirical null distribution is finally used to estimate the *q*-value measure of significance.

The LAP procedure has been tested on three public datasets (Virtaneva *et al.*, 2001; Ross *et al.*, 2003; Nutt *et al.*, 2003) to assess the correspondence between differentially expressed chromosomal regions and known chromosomal aberrations such as trisomies, translocations and deletions.

In this article, the Methods section gives the computational details of the statistics in the various response cases, of the smoothing algorithm and of the permutation test to identify differentially expressed chromosomal regions. The Results section presents the application of LAP to the analysis of publicly available expression data. Additional tables and figures are available in the Supplementary information section (denoted as _SI throughout the text). The Results section also includes a comparison of the proposed procedure with MACAT and Levin's approach. Finally, Conclusions summarizes the main characteristics of the LAP procedure and discusses the proposed approach in comparison with other methods for the identification of differentially expressed chromosomal regions.

## METHODS

LAP procedure consists of three main steps: (1) computation of a statistic for ranking probes in order of strength of the evidence for differential expression; (2) smoothing of the statistic after sorting the statistical scores according to the chromosomal position of the corresponding genes and (3) application of a permutation test to identify differentially expressed chromosomal regions.

### Statistic

Let **X** be the matrix of normalized expression levels $x_{ij}$ for gene $i$ in sample $j$ ($i = 1, 2, \ldots, G$; $j = 1, 2, \ldots, n$) and **Y** a response vector $y_j$ ($j = 1, 2, \ldots, n$) for the $n$ samples. The statistic $d_i$ is based on the ratio of change in gene expression $r_i$ to the standard deviation in the dataset $s_i$ for each probe set $i$, as defined by Tusher *et al.* (2001) in the SAM method:

$$d_i = \frac{r_i}{s_i + s_0}, \tag{1}$$

where the estimates of gene-specific variance over repeated measurements are stabilized by a fudge factor $s_0$ [see Tusher *et al.* (2001) and SAM technical manual for details].

Since the chromosomal analysis can be applied to different data types (e.g. two- and multi-class problems, paired data, quantitative responses, time course experiments, survival analyses), the quantities $r_i$ and $s_i$ have different formulations in different experimental designs.

Specifically, in the two-class unpaired case, $y_j = 1$ or $y_j = 2$. Considering $C_k$ the set of indices for the $n_k$ samples in group $k$, $C_k = \{j : y_j = k,\ k = 1, 2\}$, $\bar{x}_{i1} = \sum_{j \in C_1} x_{ij}/n_1$ and $\bar{x}_{i2} = \sum_{j \in C_2} x_{ij}/n_2$, then $r_i$ and $s_i$ can be computed as follows:

$$r_i = \bar{x}_{i2} - \bar{x}_{i1} \tag{2}$$

$$s_i = \left[\frac{(1/n_1 + 1/n_2)\left\{\sum_{j \in C_1} (x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C_2} (x_{ij} - \bar{x}_{i2})^2\right\}}{(n_1 + n_2 - 2)}\right]^{1/2} \tag{3}$$

and $d_i$ represents a two-sample *t*-like statistic with variance stabilization.

In the case of a multi-class response vector, $y_j \in \{1, 2, \ldots, K\}$, given the $C_k$ indices of observations in class $k$, the $n_k$ samples in $C_k$, $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$ and $\bar{x}_i = \sum_j x_{ij}/n$, $r_i$ and $s_i$ have the following formulations:

$$r_i = \left[\left\{\frac{\sum n_k}{\prod n_k}\right\} \sum_{k=1}^{K} n_k (\bar{x}_{ik} - \bar{x}_i)^2\right]^{1/2} \tag{4}$$

$$s_i = \left[\frac{1}{\sum (n_k - 1)} \left(\sum \frac{1}{n_k}\right) \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2\right]^{1/2} \tag{5}$$

with $d_i$ taking the form of an $F$-like statistic with variance stabilization.

In the case of censored survival data, $y_j = (t_j, \Delta_j)$ where $t_j$ is time and $\Delta_j = 1$ if the observation is a death while $\Delta_j = 0$ if the observation is censored. Considering $D$ the set of indices of the $K$ unique death times $z_1, z_2, \ldots, z_K$, $R_1, R_2, \ldots, R_K$ the indices of the observations at risk at these unique death times, i.e. $R_k = \{l : t_l \geq z_k\}$, $m_k$ the number of observations in $R_K$, $d_k$ the number of deaths at time $z_k$, $x^*_{ik} = \sum_{t_j = z_j} x_{ij}$ and $\bar{x}_{ik} = \sum_{j \in R_K} x_{ij}/m_k$, then $r_i$ and $s_i$ are defined as follows:

$$r_i = \sum_k \left[x^*_{ik} - d_k \bar{x}_{ik}\right] \tag{6}$$

$$s_i = \left[\sum_k \left(\frac{d_k}{m_k}\right) \sum_{j \in R_k} (x_{ij} - \bar{x}_{ik})^2\right]^{1/2} \tag{7}$$

and $d_i$ represents a Cox-like statistic with variance stabilization.

All the SAM statistics can be computed using the R package samr freely available at http://www-stat.stanford.edu/~tibs/SAM/.

## Smoothing

To link statistical scores $d_i$ and chromosomal locations, probes are re-annotated in terms of Entrez Gene ID using the annotate package of Bioconductor and statistics are sorted, on each chromosome, according to the physical coordinate (in base pairs) of the corresponding gene. Since each locus must be characterized by a single value of the statistic, in the case of multiple probes mapping to the same locus (e.g. probe set redundancy in Affymetrix arrays), the chromosomal locus is assigned the statistic of the probe with the highest absolute value of $d_i$. To verify whether this choice introduces any bias or affects the permutation procedure, the issue of redundancy introduced by replicate probes on the array has been also addressed by assigning to each locus the mean (Crawley and Furge, 2002) or the median statistics of multiple probes mapping to the same gene.

Once sorted, the $d_i$ statistics are smoothed over the chromosomal coordinate, obtaining for each locus link a smoothed statistic $S_i$. The smoothing process allows estimating the differential expression of chromosomal regions, taking into account that the distribution of investigated genes is not uniform over the genome. Specifically, the LAP procedure interpolates the statistics between investigated chromosomal positions using a non-parametric estimation of regression functions based on kernel regression estimators and automatically adapted local plug-in bandwidth (Herrmann, 1997). As described by Toedling and coworkers (Toedling *et al*., 2005), smoothing of the statistic can be formally stated as a non-parametric regression problem where the $d_i$ score is to be estimated over the chromosomal coordinate. Non-parametric regression problems can be approached using various methods, as kernel smoothing, orthogonal series, spline functions or wavelets. A critical issue in selecting the regression strategy is represented by the procedure for adapting the smoothing parameters. Indeed, the smoothing parameters, e.g. the bandwidth, can be adapted globally or locally (Herrmann, 1997).

In particular, the LAP procedure is based on a local variable bandwidth kernel estimator, the lokerns function, developed by Herrmann (1997) and coded in the R package lokern freely available at http://www.sourcekeg.co.uk/cran/src/contrib/Descriptions/lokern.html. Lokern comprises a function which automatically estimates the optimal bandwidths iteratively, adapting a bandwidth to the regression function and minimizing the asymptotic mean squared error. Polynomial kernels and boundary kernels are used with a fast and stable updating algorithm for kernel regression estimation. Theoretical aspects and mathematical formulation of lokerns can be found in Herrmann (1997).

## Permutation test

The identification of differentially expressed chromosomal regions as the result of a system perturbation in a microarray experiment can be formally stated as a hypothesis-testing problem in which a defined statistic is used to rank transcripts in order of evidence against the null hypothesis. In general, data obtained from microarray studies do not support asymptotic hypothesis testing. Nevertheless, nominal type I error can be controlled using statistical tests based on empirically constructed null distributions. Thus, a permutation scheme is used to identify differentially expressed regions under the assumption that each gene has a unique neighborhood and that the corresponding smoothed statistic is not comparable with any statistic smoothed in other regions of the genome. Specifically, the $G$ statistic values $d_i$ are first randomly assigned to $G$ chromosomal locations through permutations and then, for each permutation, smoothed over the chromosomal coordinate. The permutation process over $B$ random assignments allows defining the null smoothed statistic $S_i^{0b}$ for gene $i$ ($b = 1, \ldots, B$). The significance of the differentially expressed genes, i.e. the $p$-value $p_i$ for gene $i$, is computed as the probability that the random null statistic $S_i^{0b}$ exceeds the observed statistic $S_i$ over $B$ permutations:

$$p_i = \frac{\#\{b : |S_i^{0b}| \geq |S_i|, b = 1, \ldots, B\}}{B} \tag{8}$$

The $p$-value of Equation (8) has the peculiarity to be local since the observed smoothed statistic $S_i$ is compared only with null statistics $S_i^{0b}$ smoothed on the same neighborhood of the gene $i$. Indeed, during the permutation process, the chromosomal position is conserved while the statistics are randomly shuffled. Thus, observed and null statistics are smoothed and compared exactly over the same region, taking into account variations in the gene distances and in gene density.

Once the distribution of empirical $p$-values has been generated, the $q$-value is used to identify differentially expressed chromosomal regions (Storey and Tibshirani, 2004). $Q$-values allow quantifying significance in light of thousands of simultaneous tests and can be calculated, directly from the $p$-values of Equation (8), using R qvalue package (http://faculty.washington.edu/~jstorey/qvalue/).

# RESULTS

The LAP procedure has been applied to three public datasets with three different types of response variables: (1) two-class unpaired comparison (acute myeloid leukemia; Virtaneva *et al.,* 2001), (2) multi-class response (pediatric acute lymphoblastic leukemia; Ross *et al*., 2003) and (3) censored survival data (malignant gliomas; Nutt *et al*., 2003).

## Case study 1: two-class unpaired data

Virtaneva *et al*. (2001) used Affymetrix HuGeneFL oligonucleotide microarrays to study global gene expression in acute myeloid leukemia patients with trisomy of chromosome 8 (AML+8) as the sole chromosomal abnormality. The expression profiles of $n = 10$ AML+8 patients were compared with those of $n = 10$ AML patients with normal cytogenetics (AML-CN).

Intensity levels were generated from publicly available CEL files (http://thinker.med.ohio-state.edu/aml/index.html) using the robust multi-array analysis (RMA) procedure described by Irizarry and colleagues (Irizarry *et al*., 2003). After annotation, probe sets without chromosomal location information and those referring to chromosomes X and Y were filtered out. Differential expressions between AML+8 and AML-CN groups were calculated using the regularized $t$-statistic as defined in Equations (1)–(3) and each locus was assigned the score of the probe set with the highest score absolute value. This preprocessing step resulted in 5313 unique IDs from the initial 7129 probe sets. Application of the smoothing procedure generated the smoothed scores of Figure 1a. In detail, dots indicate values of the statistic $d_i$ and traces represent the smoothed statistics $S_i$ for each chromosome, along the chromosomal coordinate. The null statistic was defined through $B = 10\,000$ permutations of the statistic values $d_i$ (i.e. randomly assigning the scores to the 5313 loci) and then, for each permutation, smoothed over the chromosomal coordinate. Finally, differentially expressed chromosomal regions were identified using the $q$-value calculated from the distribution of empirical $p$-values [Equation (8)]. Setting a $q$-value threshold of 0.05 allowed the identification of 44 overexpressed genes in AML+8 samples (Fig. 1b and Table 1_SI) and located on chromosome 8 (cytobands 8q112–q13 and 8q22; $p$-values ranging from $4.8 \times 10^{-9}$ to $2.1 \times 10^{-2}$ in DAVID functional annotation) as expected, given the analyzed specimens (Virtaneva *et al*., 2001). Similar distributions of the smoothed statistics and the same regions on chromosome 8 (e.g. 8q11–q13 and 8q22) were obtained assigning to each locus the highest, median or mean absolute value of $d_i$ calculated over multiple probes (Figs 1_SI and 2_SI). In the whole genome plot of Figure 1b, the red perpendicular lines represent the exact chromosomal locations and orientations of the 44 overexpressed

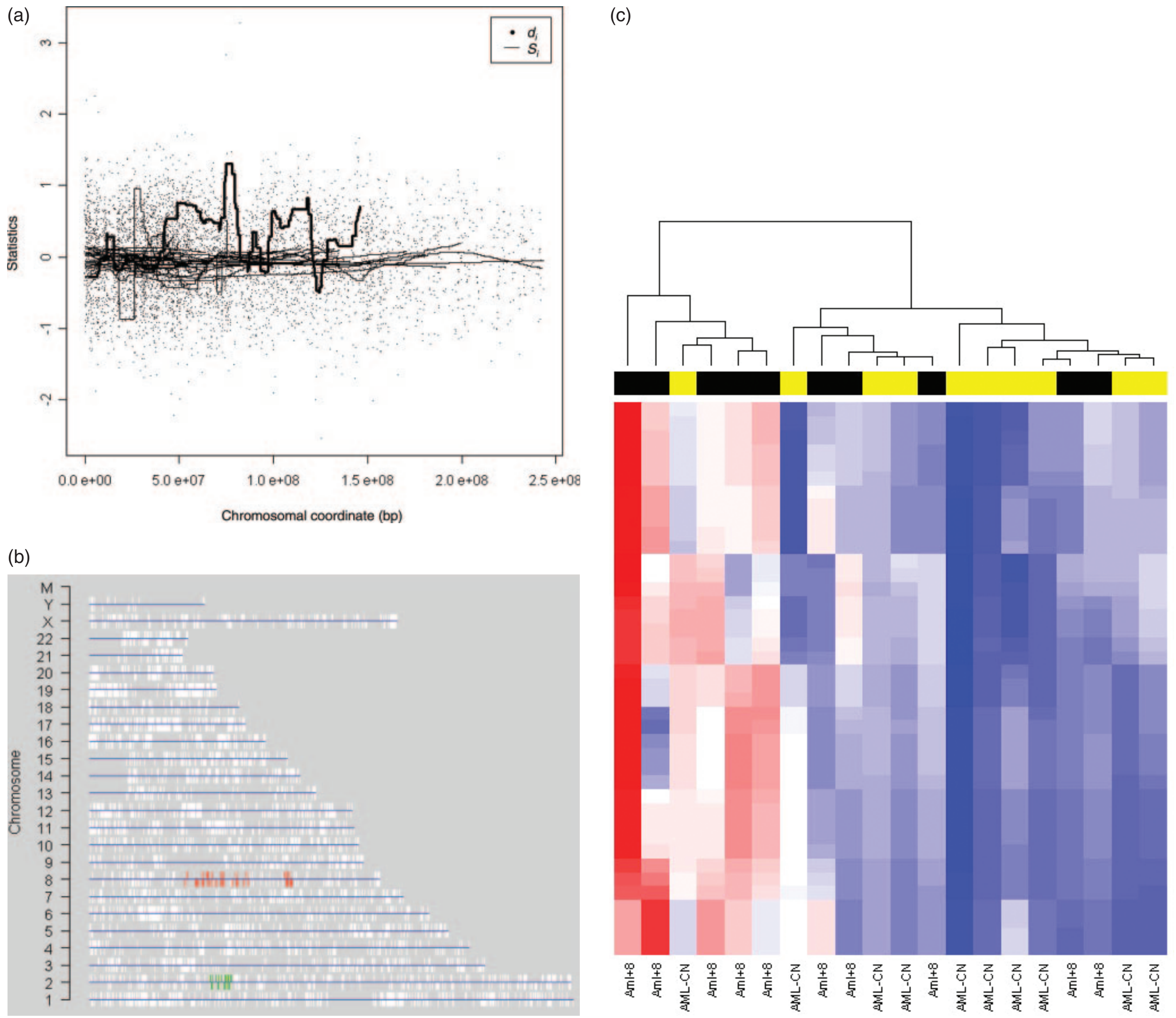

**Fig. 1.** Comparison of gene expression profiles of AML patients with trisomy 8 and AML patients with normal cytogenetics (Case study 1, two-class unpaired data). (**a**) Values of the statistic $d_i$ and of the smoothed statistics $S_i$ along the chromosomal coordinate in base pairs. Dots represent the values of the statistic $d_i$, lines the values of the smoothed statistics $S_i$ for each chromosome and the bold line highlights the smoothed statistics of chromosome 8. (**b**) Whole genome plot of the differentially expressed genes at a $q$-value < 0.05. The red perpendicular lines represent the exact chromosomal locations and orientations of the 44 overexpressed genes, the white bars indicate locations and orientations of all probe sets in the microarray, and the green lines the location of a cluster of 27 downregulated probes. The positions for both the sense and antisense strands are expressed in numbers of base pairs measured from the p (5′ end of the sense strand) to q (3′ end of the sense strand) arms; upper and lower bars stand for genes on sense and antisense strands, respectively. (**c**) Unsupervised hierarchical cluster analysis of the smoothed expression profiles of genes located in the differentially expressed regions of chromosome 8. Euclidean metric and average-linkage were used as distance measure and linkage method, respectively. Black and yellow labeled columns represent AML samples with trisomy 8 and with normal cytogenetics, respectively.

genes, the white bars indicate locations and orientations of all probe sets in the HuGeneFL array and the green lines represent a small chromosomal region of downregulation. The positions for both sense and antisense strands are expressed in numbers of base pairs measured from the p (5′ end of the sense strand) to q (3′ end of the sense strand) arms; upper and lower bars stand for genes on the sense and antisense strands, respectively (see geneplotter package from Bioconductor for details).

To investigate the presence of a chromosomal aberration at the level of the single specimen, the gene expression data were smoothed independently for each sample over the chromosome with the identified expression imbalance and the smoothed signals were used to generate a heatmap (Fig. 1c). Hierarchical clustering of the smoothed data for genes on chromosome 8 segregated the 20 samples into 2 major clusters, one mostly composed of AML+8 specimens with overexpressed regions on chromosome 8 and the other grouping AML+8 and AML-CN samples with lower expression levels of chromosome 8 significant genes. Thus, LAP analysis allowed identifying regions with transcriptional modulation related to trisomy of chromosome 8 even in a heterogeneous group of AML+8 samples. This heterogeneity penalizes methods based on sample permutations, such as SAM (Tusher *et al.*, 2001) and MACAT (Toedling *et al.*, 2004), while it does not affect LAP whose permutation scheme is applied on gene location. Indeed, SAM did not identify any statistically significant genes at a *q*-value of 0.05, and MACAT did not select any differentially expressed chromosomal region (Fig. 3_SI and Table 2_SI; kNN-kernel for interpolation between the scores, optimal parameter settings determined by cross-validation and 10 000 random permutations of the class labels). Differently, the version of MACAT with permutations of genes on chromosomes identified a total of 697 modulated probes scattered along the chromosomes, including up- and down-regulated genes on chromosome 8 (Fig. 4_SI and Table 2_SI; kNN-kernel for interpolation between the scores, optimal parameter settings determined by cross-validation, and 10 000 random permutations of gene positions). Finally, Levin's model-based procedure (Levin *et al.*, 2005) identified only few up-regulated probes on chromosome 8 and several modulated probes on other chromosomes (Fig. 5_SI, *p*-value <0.05). As such, LAP outperformed other approaches in detecting the increase of gene expression on chromosome 8 previously described by Virtaneva *et al.*, 2001 in AML+8 samples.

## Case study 2: multi-class response

Pediatric acute lymphoblastic leukemia (ALL) is a heterogeneous disease with subtypes that differ markedly in their cellular and molecular characteristics as well as in their response to therapy (Ross *et al.*, 2003). Nowadays, treatment protocols achieve an overall survival rate of 70–80% optimizing the intensity of therapy on the patient's risk of relapse. As such, the accurate assignment of patients to risk groups is a critical parameter for establishing the correct therapeutical intervention. Ross and coworkers classified ALL patients into prognostic groups on the basis of expression profiles assessed with Affymetrix HG-U133A microarrays (Ross *et al.*, 2003).

The LAP procedure was applied to a subset of $n = 104$ ALL samples derived from the dataset published by Ross and coworkers (Ross *et al.*, 2003). The subset comprised 14 patients with T-cell lineage ALL (T-ALL) and 90 patients with 5 distinct subtypes of B-cell lineage ALL. Among the latter, 15 patients had a *t*(9;22)(*BCR-ABL*) translocation, 18 had *t*(1;19)(*E2A-PBX1*) and 20 had *t*(12;21)(*TEL-AML1*); moreover, 20 patients had a rearrangement of the *MLL* gene on cytoband 11q23 and 17 had a hyperdiploid karyotype (>50 chromosomes). Intensity levels were generated from publicly available CEL files (http://www.stjuderesearch.org/data/ALL3/) using the RMA procedure (Irizarry *et al.*, 2003). After annotation, probe sets without chromosomal location information and those on chromosomes X and Y were filtered out. Differential expressions among the six karyotypes were calculated using the multi-class statistic as defined in Equations (1), (4) and (5) and each locus was assigned the score of the probe set with the highest score value. This preprocessing step resulted in a total of 11 613 unique IDs. Application of the smoothing procedure generated the smoothed scores of Figure 2a, where dots indicate values of the statistic $d_i$ and traces represent the smoothed statistics $S_i$ along the chromosomal coordinate. As in Case study 1, the null statistic was defined through $B = 10\,000$ permutations of the statistic values $d_i$ and then, for each permutation, smoothed over the chromosomal coordinate. Differentially expressed chromosomal regions were identified using the *q*-value. Setting a *q*-value threshold of 0.01 allowed the identification of 159 differentially expressed genes located on the cytobands q22–q24 of chromosome 1 and centered around gene *PBX1* (Fig. 2b and Table 3_SI). The expression levels of the 159 genes in the identified chromosomal region were smoothed independently on each sample over chromosome 1 and used to inspect each ALL specimen (Fig. 2c). The smoothed expression levels of genes located on cytobands 1q22–1q24 divided the 104 samples into 2 major groups, one comprising 16 of the 18 *t*(1;19)(*E2A-PBX1*) specimens with an up-regulation of 1q22–1q24 genes and the other composed of samples with a lower expression signal in the 1q22–1q24 region. This result indicates that, among the most common ALL chromosomal aberrations, the *t*(1;19) translocation has the highest impact on gene transcription. However, it is currently not understood how the *t*(1;19) translocation leads to up-regulation of region 1q22–1q24 and, as such, the identified features could help shed light on this chromosomal aberration. Since this region is centered on gene *PBX1*, it is possible that most of the analyzed samples with the *t*(1;19) translocation present not only a chromosomal re-arrangement but also a gain of a whole chromosome segment (Paulsson *et al.*, 2005).

## Case study 3: censored survival data

The LAP was finally applied to censored survival data to identify differentially expressed chromosomal regions from gene expression profiles of tumor samples with different clinical outcomes. Specifically, LAP was applied to a set of $n = 50$ primary brain tumors (28 glioblastomas and 22 anaplastic oligodendrogliomas) sampled and assessed on Affymetrix U95Av2 GeneChips before therapy (Nutt *et al.*, 2003). Survival data for all samples were available in terms of months from date of initial diagnosis to death event or, for living patients, to the last follow-up.

Intensity levels were generated from publicly available raw data (http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=82) using the RMA procedure (Irizarry *et al.*, 2003). After annotation, probe sets without chromosomal location information and those on chromosomes X and Y were filtered out. Association with survival time was computed using the score of Equation (1) with the regularized Cox-statistic described in Equations (6) and (7). For each genetic locus, only the probe set with the highest statistic absolute value was further analyzed (8035 unique gene IDs). Application of the smoothing procedure generated the smoothed scores of Figure 3a, where dots indicate values of the statistic $d_i$ and traces represent the smoothed statistics $S_i$ along the chromosomal coordinate. As in the previous cases, the null statistic was defined permuting $B = 10\,000$ times the statistic values $d_i$ and then, for each permutation, smoothing over the chromosomal coordinate. Differentially expressed chromosomal regions were identified setting the *q*-value threshold to zero. This allowed the identification of a large up-regulated region on the p arm of chromosome 1 and a down-regulated region corresponding to cytoband q23–24 of chromosome 10 (Fig. 3b and Table 4_SI), for a total of 170 unique genes. Given the statistic as defined in Equations (1), (6) and (7), an up-regulated chromosomal region is associated with patients with a

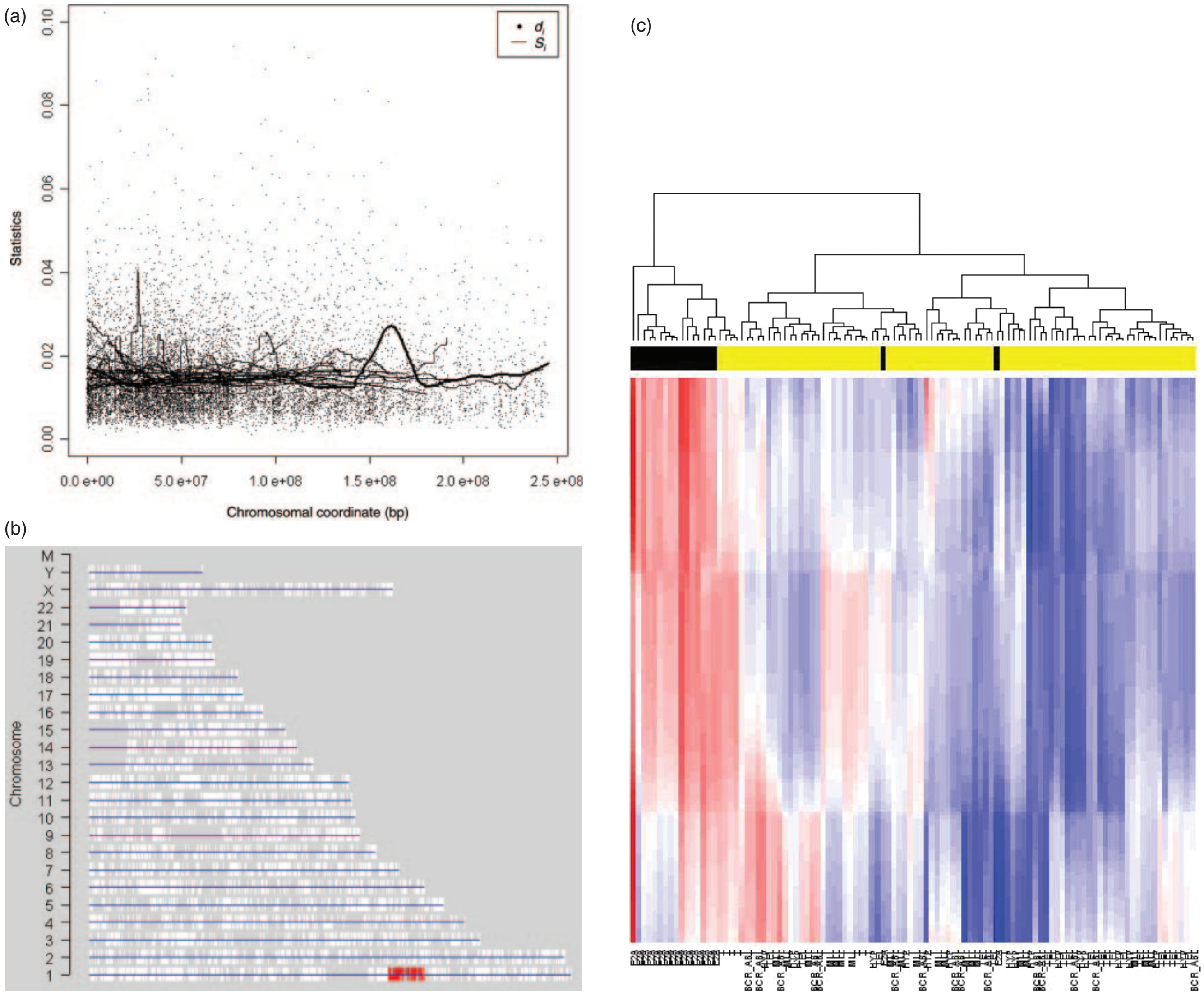

**Fig. 2.** Analysis of gene expression profiles in six groups of pediatric ALL samples (Case study 2, multi-class response). (**a**) Values of the statistic $d_i$ and of the smoothed statistics $S_i$ along the chromosomal coordinate in base pairs. Dots represent the values of the statistic $d_i$ and lines the values of the smoothed statistics $S_i$ for each chromosome. The bold line highlights the smoothed statistics of chromosome 1. (**b**) Whole genome plot of the differentially expressed genes at a $q$-value < 0.01. The red perpendicular lines represent the exact chromosomal locations and orientations the 159 differentially expressed genes, while the white bars indicate locations and orientations of all probe sets in the microarray. The positions for both the sense and antisense strands are expressed in numbers of base pairs measured from the p (5′ end of the sense strand) to q (3′ end of the sense strand) arms; upper and lower bars stand for genes on the sense and antisense strands, respectively. (**c**) Unsupervised hierarchical cluster analysis of the smoothed expression profiles of genes located in the differentially expressed regions of chromosome 1. Euclidean metric and average-linkage were used as distance measure and linkage method, respectively. Black and yellow column labels represent samples with and without the *t*(1;19)(*E2A-PBX1*) translocation, respectively.

higher risk of death and, conversely, down-regulated genes characterize patients with a better outcome. Cluster analysis of the smoothed gene expression data from the p arm of chromosome 1 (156 genes) partitioned the samples into two main groups, one composed of 15 patients with a lower signal pattern (Fig. 3c, cluster A) and another comprising 35 samples with an over-expression of the region 1p35–1p21 (Fig. 3c, cluster B). Survival analysis of these two groups with opposite transcriptional patterns on chromosome 1 (Fig. 4) revealed that patients of cluster A (down-regulation of 1p35–1p21) had significantly better survival (log-rank $p$-value = 0.00197) than patients in cluster B (up-regulation of 1p35–1p21). This result is in agreement with the observation that deletion of the short arm of chromosome 1 (1p) is strongly associated with increased chemosensitivity and is considered a favorable prognostic factor in glioma tumors (Idbaih *et al*., 2005; Sasaki *et al*., 2002).

## DISCUSSION

Although several annotation schemes have been successfully applied to identify chromosomal regions enriched with differentially expressed genes from microarray data, only a few methods have been developed to integrate transcriptional and structural

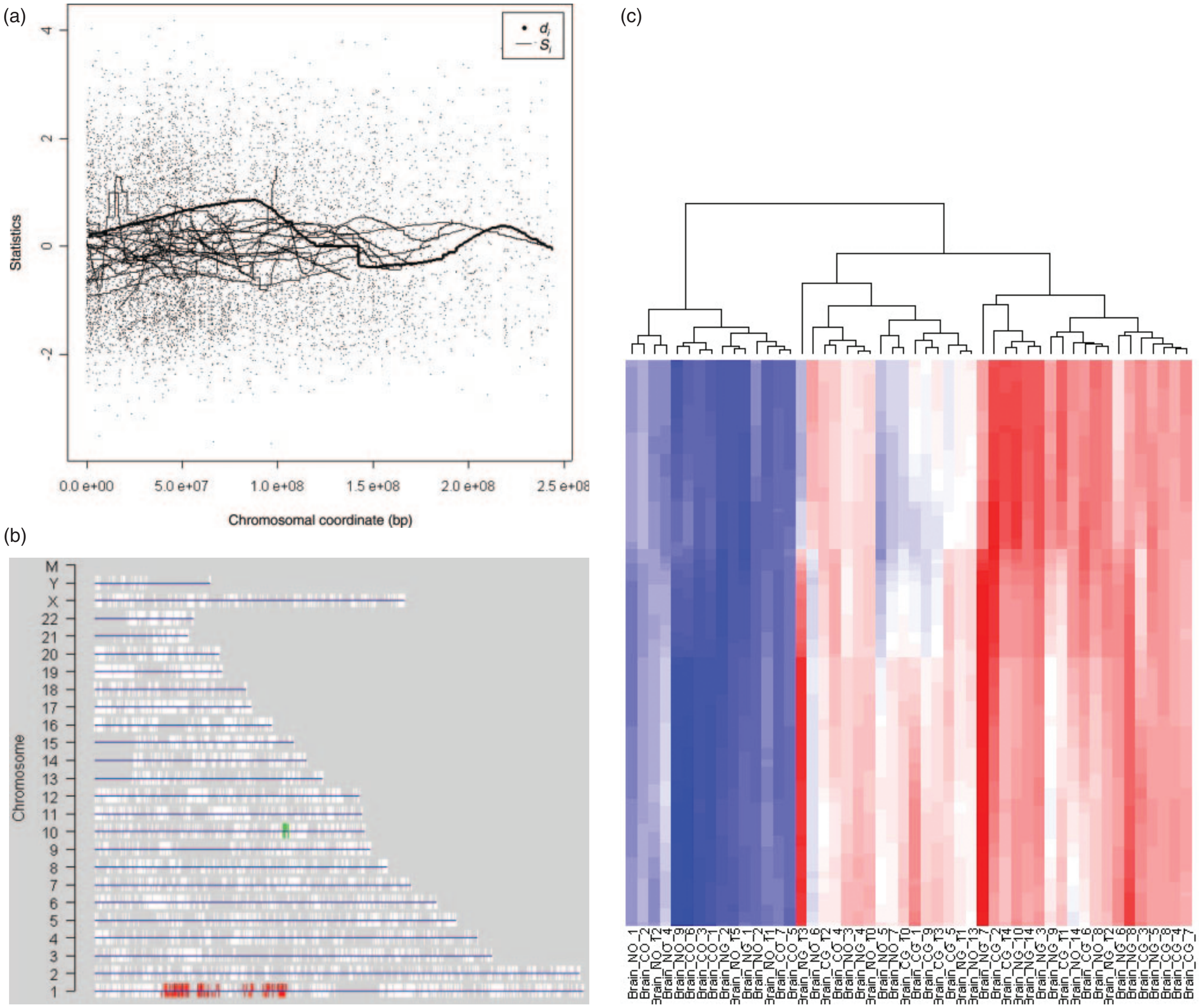

**Fig. 3.** Analysis of gene expression profiles in 50 primary brain tumors (Case study 3, censored survival data). (**a**) Values of the statistic $d_i$ and of the smoothed statistics $S_i$ along the chromosomal coordinate in base pairs. Dots represent the values of the statistic $d_i$ and lines the values of the smoothed statistics $S_i$ for each chromosome. The bold line highlights the smoothed statistics of chromosome 1. (**b**) Whole genome plot of the differentially expressed genes at a $q$-value $= 0$. The colored perpendicular lines represent the exact chromosomal locations, orientations, and up- (red) or down-regulation (green) states of the 170 differentially expressed genes, while the white bars indicate locations and orientations of all probe sets in the microarray. Positions for both the sense and antisense strands are expressed in numbers of base pairs measured from the p (5′ end of the sense strand) to q (3′ end of the sense strand) arms; upper and lower bars stand for genes on sense and antisense strands, respectively. (**c**) Unsupervised hierarchical clustering of the smoothed expression profiles of genes located in the differentially expressed regions of chromosome 1 (156 genes). Euclidean metric and average-linkage were used as distance measure and linkage method, respectively. Cluster analysis divided samples into two main groups: a group with down-regulated smoothed expression levels (cluster A, 15 samples) and a group with up-regulated smoothed expression levels (cluster B, 35 samples).

information before or during the data-analysis process (Crawley and Furge, 2002; Pollack *et al*., 2002; Kano *et al*., 2003; Husing *et al*., 2003; Zhou *et al*., 2003, 2005; Toedling *et al*., 2005). In this context, the LAP is a non-parametric model-free statistical method to detect differential expression over large chromosomal regions. The notions about gene location and local density are incorporated with the differential expression data, interpolating the observed and null statistics on gene positions. The smoothing parameters and measures of significance (e.g. $p$-values) are adapted and computed locally, thus accounting for genomic complexity and more precisely estimating differentially expressed chromosomal regions. The main characteristics of the LAP procedure are (1) the use of a locally adaptive smoothing function, (2) the application of a permutation scheme of statistic values over the whole genome, (3) the computation of local $p$-values and (4) the control of multiplicity through $q$-value calculation. Specifically, a kernel regression estimator and automatically adapted local plug-in bandwidth are used to locally adapt the smoothing parameters. Although

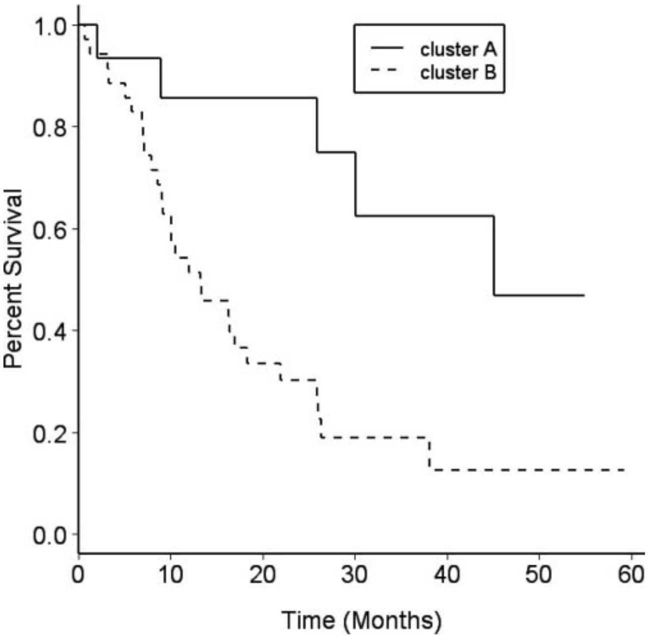

**Fig. 4.** Case study 3: Kaplan–Meier estimates of overall survival for the groups of samples identified by unsupervised hierarchical clustering of the smoothed gene expression data from the p arm of chromosome 1. Patients forming cluster A (down-regulation of 1p35–1p21) had a significantly better survival (log-rank $p$-value = 0.00197) than patients in cluster B (up-regulation of 1p35–1p21).

**Table 1.** Comparison among LAP, MACAT, and Levin's methods in terms of Type I and Type II errors in the detection of differentially expressed regions at various confidence thresholds

| DNA status | CNAG copy number | Type I error (%) | Type II error (%) |
|---|---|---|---|
| LAP $q$-value < 0.01 | | | |
| Gain | >2 | | 59.14 |
| Normal | 2 | 25.35 | |
| Loss | <2 | | 49.15 |
| LAP $q$-value < 0.05 | | | |
| Gain | >2 | | 32.95 |
| Normal | 2 | 52.85 | |
| Loss | <2 | | 20.82 |
| Levin's $p$-value < 0.01 | | | |
| Gain | >2 | | 96.62 |
| Normal | 2 | 5.57 | |
| Loss | <2 | | 99.88 |
| Levin's $p$-value < 0.05 | | | |
| Gain | >2 | | 61.46 |
| Normal | 2 | 45.33 | |
| Loss | <2 | | 63.92 |
| MACAT sample permutation | | | |
| Gain | >2 | | 91.51 |
| Normal | 2 | 13.77 | |
| Loss | <2 | | 92.01 |
| MACAT gene permutation | | | |
| Gain | >2 | | 93.06 |
| Normal | 2 | 11.43 | |
| Loss | <2 | | 93.70 |

suffering from a sub-optimal asymptotic behavior, local variable bandwidth kernel estimators are still competitive, at least in practical terms, over new and more complex methods, such as locally adaptive wavelets. Permutation of statistics over the whole set of genomic loci allows using the information contained in all chromosomes and identifying significant chromosomal regions even in the case of heterogeneous groups. In this case, methods which use a standard permutation scheme of sample labels are indeed penalized, as exemplified by the analysis of Case study 1. Permuting a quantity over thousands of positions implies a huge computational charge. Nevertheless, results from LAP are quite stable with $B > 10\,000$. The local computation of $p$-values permits taking into account the complex structure of the genome in the neighborhood of each gene. Indeed, each observed smoothed statistic $S_i$ is compared only with null statistics smoothed exactly on the same neighborhood of gene $i$.

The proposed method can analyze gene expression data from the most common experimental designs (e.g. two-class comparison, multi-class response, censored survival data, time-course experiment) and in principle can be applied to signals obtained from any high-throughput platform (e.g. copy number data from genotyping arrays). Using a statistic based on criteria such as the difference between the maximum and the minimum expression levels for each gene, the standard deviation, or the interquartile range, it is also possible to perform unsupervised analyses and screen single-class experimental designs to identify regions of increased gene expression common to an entire set of samples.

As in MACAT (Toedling *et al*., 2005), LAP involves the interpolation of a statistic for distances between investigated genes but it does not assume that gene distances and densities are constant within chromosomes. Moreover, the smoothing functions and the $p$-value computations are different in the two methods, and the statistic of MACAT has not been implemented for multi-class and survival analyses. Differently from the model-based scan statistic proposed by Levin and coworkers (Levin *et al*., 2005), LAP is model-free and does not depend on the validity of any assumption about the distributions of expression signal and gene distances. Finally, LAP uses the $q$-value to quantify the significance of modulated regions in light of thousands of simultaneous tests, while MACAT relies on quantile thresholds without any multiplicity control. Moreover, Levin's approach uses pre-defined thresholds of a $z$-score distribution to identify differentially expressed probes (i.e. neither stabilizing the variance nor performing any statistical test) and then computes $p$-values comparing distances of these modulated probes to a theoretical distribution of gene distances over the chromosomes.

The application of the LAP to three independent sets of raw expression data allowed the identification of regions of differential expression that are directly involved in the chromosomal aberrations known to be characteristic of hematological and solid tumors (copy number gain or loss). Recently, the method has been applied to integrate DNA copy number variations obtained from Affymetrix Mapping 100k GeneChips with gene expression data derived from expression arrays (Cifola *et al*., 2006). In the context of a research project focused on the identification of clinical biomarkers of renal cell carcinoma, LAP, MACAT and Levin's approach have been applied to asses the global effect of gene dosage on transcriptional levels in this type of epithelial tumor (see Supplementary

Information). The availability of copy number and gene expression data from samples of Caki-1, a renal carcinoma cell line, allowed a quantitative comparison among the three different methods in terms of Type I and Type II errors in the detection of differentially expressed regions, assuming all gains and losses are known based on copy number analysis (Table 1). In Table 1, Type I and Type II errors are calculated as the number false positives (probes with normal DNA status detected as transcriptionally modulated) and false negatives (probes with gain or loss detected as transcriptionally unchanged), respectively. Results from the comparisons indicate that LAP, although overestimating the amplitude of regions with gene expression imbalance, outperforms both Levin's approach and MACAT in detecting the chromosomal bands where a change in copy number induces a variation of the gene expression profile (see Supplementary information).

## ACKNOWLEDGEMENTS

The authors acknowledge V. Matarese for editing and critically reviewing this manuscript, C. Battaglia (Department of Sciences and Biomedical Technologies, University of Milano) for comparisons with DNA copy number data and the reviewers of the manuscript for helpful comments. This work was supported by grants from the Italian Ministry of University and Research (FIRB RBNE01TZZ8 and RBAU01935A) and by the OncoSuisse Collaborative Cancer Research Project OCS 01517–02–2004.

*Conflict of Interest*: none declared.

## REFERENCES

Caron,H. *et al*. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.

Cifola,I. *et al*. (2006) Integration of whole-genome SNP mapping and transcriptional data in the human metastatic renal carcinoma Caki-1 cell line. *BMC Genomics*, (submitted).

Crawley,J.J. and Furge,KA. (2002) Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biol*., **3**, RESEARCH0075.

Cromer,A. *et al*. (2004) Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis. *Oncogene*, **23**, 2484–2498.

Furge,K.A. *et al*. (2004) Robust classification of renal cell carcinoma based on gene expression data and predicted cytogenetic profiles. *Cancer Res*., **64**, 4117–4121.

Herrmann,E. (1997) Local bandwidth choice in kernel regression estimation. *J. Graphic. Comput. Statist*., **6**, 35–54.

Heidenblad,M. *et al*. (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene*, **24**, 1794–1801.

Husing,J. *et al*. (2003) Combining DNA expression with positional information to detect functional silencing of chromosomal regions. *Bioinformatics*, **19**, 2335–2342.

Hyman,E. *et al*. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res*., **62**, 6240–6245.

Idbaih,A. *et al*. (2005) Two types of chromosome 1p losses with opposite significance in gliomas. *Ann. Neurol*., **58**, 483–487.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graphic. Statist*., **5**, 299–314.

Kano,M. *et al*. (2003) Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions. *Physiol. Genomics*, **13**, 31–46.

Lercher,M.J. *et al*. (2003) A unification of mosaic structures in the human genome. *Hum. Mol. Genet*., **12**, 2411–2415.

Levin,A.M. *et al*. (2005) A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics*, **21**, 2867–2874.

Lu,Y.J. *et al*. (2001) Comparative expressed sequence hybridization to chromosomes for tumor classification and identification of genomic regions of differential gene expression. *Proc. Natl Acad. Sci. USA*, **98**, 9197–9202.

Masayesva,B.G. *et al*. (2004) Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. *Proc. Natl Acad. Sci. USA*, **101**, 8715–8720.

Midorikawa,Y. *et al*. (2004) Distinct chromosomal bias of gene expression signatures in the progression of hepatocellular carcinoma. *Cancer Res*., **64**, 7263–7270.

Nutt,C.L. *et al*. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*., **63**, 1602–1607.

Paulsson,K. *et al*. (2005) Formation of der(19)t(1;19)(q23;p13) in acute lymphoblastic leukemia. *Genes Chromosomes Cancer*, **42**, 144–148.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Reyal,F. *et al*. (2005) Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes—a study of 130 invasive ductal breast carcinomas. *Cancer Res*., **65**, 1376–1383.

Ross,M.E. *et al*. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.

Sasaki,H. *et al*. (2002) Histopathological-molecular genetic correlations in referral pathologist-diagnosed low-grade 'oligodendroglioma'. *J. Neuropathol. Exp. Neurol*., **61**, 58–63.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide experiments. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Toedling,J. *et al*. (2005) MACAT—microarray chromosome analysis tool. *Bioinformatics*, **21**, 2112–2113.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Versteeg,R. *et al*. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*., **13**, 1998–2004.

Virtaneva,K. *et al*. (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl Acad. Sci. USA*, **98**, 1124–1129.

Zhou,Y. *et al*. (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res*., **63**, 5781–5784.

Zhou,X. *et al*. (2005) Identification of discrete chromosomal deletion by binary recursive partitioning of microarray differential expression data. *J. Med. Genet*., **42**, 416–419.