BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 21 2011, pages 2949–2956
doi:10.1093/bioinformatics/btr488

Genome analysis Advance Access publication August 25, 2011

# Finding recurrent copy number alterations preserving within-sample homogeneity

Sandro Morganella[1,2,*], Stefano Maria Pagnotta[1] and Michele Ceccarelli[1,2,*]

[1]Department of Science, University of Sannio, 82100, Benevento and [2]Bioinformatics CORE, BIOGEM s.c.a.r.l., Contrada Camporeale, Ariano Irpino, Italy

Associate Editor: Alex Bateman

## ABSTRACT

**Motivation:** Copy number alterations (CNAs) represent an important component of genetic variation and play a significant role in many human diseases. Development of array comparative genomic hybridization (aCGH) technology has made it possible to identify CNAs. Identification of recurrent CNAs represents the first fundamental step to provide a list of genomic regions which form the basis for further biological investigations. The main problem in recurrent CNAs discovery is related to the need to distinguish between functional changes and random events without pathological relevance. Within-sample homogeneity represents a common feature of copy number profile in cancer, so it can be used as additional source of information to increase the accuracy of the results. Although several algorithms aimed at the identification of recurrent CNAs have been proposed, no attempt of a comprehensive comparison of different approaches has yet been published.

**Results:** We propose a new approach, called Genomic Analysis of Important Alterations (GAIA), to find recurrent CNAs where a statistical hypothesis framework is extended to take into account within-sample homogeneity. Statistical significance and within-sample homogeneity are combined into an iterative procedure to extract the regions that likely are involved in functional changes. Results show that GAIA represents a valid alternative to other proposed approaches. In addition, we perform an accurate comparison by using two real aCGH datasets and a carefully planned simulation study.

**Availability:** GAIA has been implemented as R/Bioconductor package. It can be downloaded from the following page http://bioinformatics.biogem.it/download/gaia

**Contact:** ceccarelli@unisannio.it; morganella@unisannio.it

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on April 22, 2011; revised on July 11, 2011; accepted on August 18, 2011

## 1 INTRODUCTION

Recently, the impact of copy number alterations (CNAs) in human diseases is increasingly being recognized (Albertson *et al*., 2003; Shlien and Malkin, 2009; Taylor *et al*., 2008). CNAs are genomic regions, >1 kb, in which copy number differences are observed between two or more genomes (Feuk *et al*., 2006). aCGH technology allows the measurement of copy number for hundreds of thousands locations (probes) in the genome, where copy number is expressed by the log$R$ Ratio (LRR), which gives an indirect measure of copy number of each probe, computed as the ratio of observed to expected hybridization intensity. The potential of aCGH has been exploited in a large number of publications aimed at the identification of functional genetic mutations involved in cancer (Beroukhim *et al.*, 2010). For example in Astolfi *et al*. (2010) and Venkatachalam *et al*. (2011), the genomes of a cohort of patients are analyzed to extract the CNAs that are common to a significant number of subjects. These analyses are based on the fact that CNAs functionally related to the disease under study (driver alterations) will be present in many of the analyzed genomes (recurrent CNAs); in contrast, random somatic mutations (passenger alterations) are subject-specific and they will be present only in a small number of subjects. We can define a CNA as 'a set of continuous probes that show a high enough evidence to being altered in at least some samples' (Rueda and Diaz-Uriarte, 2010). Of course, rare CNAs also hold a very important role in cancer development (Nagy *et al*., 2004), but most studies based on aCGH are designed to find recurrent CNAs. For this reason, discovery of recurrent CNAs represents a current challenge in *Bioinformatics*.

Several algorithms aimed at the discovery of recurrent CNAs have been proposed. In this work, we refer to this class of algorithm as rCNA-algorithms. rCNA-algorithms differ in many aspects: input data, model, preprocessing step, output and execution time (an overview of rCNA-algorithms is reported in Section 2). Indeed, there is no comprehensive comparison of the different approaches and very few of the published papers report a comparison with other methods. In addition, both Rueda and Diaz-Uriarte (2010) and Shah (2008) point out the importance of carefully planned simulation studies in performance assessment of rCNA-algorithms.

This work has two goals: the description of a novel rCNA-algorithm (GAIA) and the evaluation of rCNA-algorithm performance. GAIA uses a discrete representation of the data to perform a permutation test. A novel iterative procedure taking into account both significance and within-sample homogeneity (homogeneous peel-off) is used to identify the most significant peaks. In Section 3, we present a careful comparison of rCNA-algorithms both on synthetic data and on real aCGH dataset. Synthetic data are generated in agreement to three typical patterns of recurrent CNAs so that quantitative comparison can be performed in a variety of well-defined simulated datasets. We also use two recently published real datasets to evaluate the results of compared algorithms. The first dataset is on colorectal cancer: current knowledge can be used to compare and validate the considered

*To whom correspondence should be addressed.

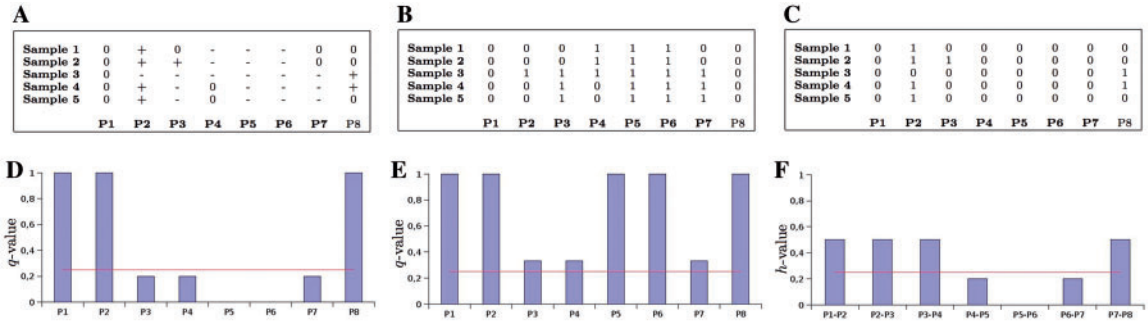


**Fig. 1.** In (**A**), an example of matrix $D$: **0** denotes no alteration, + denotes gain and − denotes loss. In $D$ there are two homogeneous regions from probes P4 to P6 for samples S1, S2 and S3 and from probes P5 to P7 for samples S3, S4 and S5 (blue and green squares, respectively). In (**B**) and (**C**), the matrices $D_L$ and $D_G$ of the matrix in (A) are shown. In (**D**), the $q$-value configuration for $D_L$ before the first iteration of peel-off. In (**E**), the $q$-value configuration for $D_L$ after that the peak P5–P6 has been removed by standard peel-off. In (**F**), the $h$-values for the matrix $D$ in (A). In (D) and (E), red line represents the $q$-value significance threshold ($q_{\text{thr}}$), and in (F) red line represents the $h$-value threshold ($h_{\text{thr}}$).

algorithms. The second dataset refers to gastrointestinal stromal tumor for which well-known aberrant cytobands are used to perform the comparison. In this dataset for each sample, $>1.6$ million of probes are measured, so it can also be used to validate the performance of the algorithms in terms of execution time.

## 2 METHODS

In general, the data analysis process of a large-scale CNA experiment can be described as reported in Supplementary Figure SF4, and GAIA can be summarized as a procedure containing two main steps:

- *Significance testing*: it consists of computing the statistical significance of observed genomic aberrations among various samples at a given site, under the null hypothesis that a given genomic locus is not a site of a recurrent CNA. The distribution of the test statistics under of the null hypothesis (null distribution) is computed by random permutations.
- *Homogeneous peel-off*: the *peel-off* is an iterative procedure aimed at the identification of significant peaks in a region, the selected peaks are iteratively removed by the set of significant locations and the remaining significance values are corrected by FDR (Storey *et al.*, 2004). The procedure continues until no further peak above the significance threshold remains. Here, we propose a novel peel-off procedure taking into account both the statistical significance and the homogeneity within samples.

### 2.1 GAIA

The input data are obtained by an external segmentation procedure having in charge the assignment of a label (i.e. loss, normal and gain) to each probe. The data are arranged as a matrix $D$ of dimensions $N \times M$ ($N$ is the number of samples and $M$ the number of observed probes). Figure 1A shows a simple example of matrix $D$ for a chromosome having eight observed probes and for a dataset composed of five samples. The matrix $D$ is split in two matrices $D_L$ and $D_G$ where the element $d_{ij} \in D_L$ ($D_G$) $i=1,\cdots,N$ and $j=1,\cdots,M$, is set to 1 when a loss (gain) is found in the $j$-th marker of the $i$-th sample, while is set to 0 otherwise. Figures 1B and C show the matrices $D_L$ and $D_G$ of the matrix $D$ reported in Figure 1A.

*Significance testing*: the detection of recurrent CNA sites is performed through a test procedure where the null hypothesis can be intuitively stated as $\mathcal{H}_0$: *the j-th probe is not a site of a recurrent CNA*. More formally, if $\mu$ is the number of CNA in a site and $\mu_0$ is the expected count when no recurrent aberration is present in a site, then $\mathcal{H}_0$: $\mu=\mu_0$ VS $\mathcal{H}_1$: $\mu>\mu_0$.

The adopted test statistic $X$ is the stochastic version of the counts $x_j=\sum_{i=1}^{N} d_{ij}$ of the loss-aberrations observed in the $j$-th site. When the interest is on the gains, the $D_G$ matrix has to replace $D_L$. In both cases, the steps for assessing the significance are the same. The $j$-th probe can be considered as a site of CNA (rejection of $\mathcal{H}_0$) when the corresponding $x_j$ is beyond a critical value $c$ at a given level of significance.

The sampling distribution under the null hypothesis is approximated by adopting a shuffling scheme similar to that of permutation test (Westfall and Young, 1993). The algorithm for the approximation is described by considering the loss-alteration. Let $D_L^{(k)}$ be a matrix obtained by applying a random permutation to each row of $D_L$, and let $\hat{P}^{(k)}(x)$ the empirical probability function estimated on the $x_1^{(k)}$, $x_2^{(k)}$, …, $x_M^{(k)}$, where $x_j^{(k)}=\sum_{i=1}^{N} d_{ij}^{(k)}$. If $K$ is the total number of the $D_L^{(k)}$'s considered, then the approximate distribution of $X$ given $\mathcal{H}_0$ is

$$P(x)=\frac{1}{K}\sum_{k=1}^{K}\hat{P}^{(k)}(x).$$

The decision of rejecting $\mathcal{H}_0$ is based on the $q$-values (Storey *et al.*, 2004), which are a corrected version of the $P$-values

$$p_j=P[X\geq x_j]=\sum_{x=x_j}^{N}P(x)$$

that are computed for each site.

*Homogeneous peel-off*: peel-off is an iterative procedure aimed at the identification of significant peaks of a region. Peel-off uses the $q$-value's computed by the FDR and in each iteration it selects and extracts a new peak and increases to 1 the $p$-value correspondent to the selected probes. In the next iteration, peel-off applies the FDR on this new configuration of $p$-values and finishes if no significant peak exists. A peak is considered significant if its $q$-value is lower than a fixed threshold $q_{\text{thr}}$. In the standard peel-off (Beroukhim *et al.*, 2007), the selected peak is the one having minimum $q$-value. As we will see, the problem of this approach is that it tends to identify just one peak of a region omitting the adjacent significant peaks.

In several biological studies, it was observed that cancer is remarkably homogeneous in its copy number profile (Beroukhim *et al.*, 2009; Sartore-Bianchi *et al.*, 2007; Snijders *et al.*, 2003). For this reason, within-sample homogeneity can represent a source of information allowing to predict CNA regions which are more consistent with the underlying biological phenomenon. Here we propose a new variant of peel-off, called homogeneous, where significant peaks are selected by using the concept of within-sample homogeneity.

Here, as in van de Wiel *et al*. (2007) we define a homogeneous region as a sequence of contiguous probes which for every sample are (almost) constant. Given a matrix $D$, we say that two adjacent probes have a maximum homogeneity when they have the same state, on the other hand two adjacent probes have a medium homogeneity when a probe has a normal state and the other one has an aberrant state (loss or gain) and, finally, two adjacent probes have the minimum homogeneity when they have opposite states. For the matrix $D$ depicted in Figure 1A, in the Sample S2 the pair P4–P5 has maximum homogeneity, the pair P1–P2 has medium homogeneity and the pair P3–P4 has minimum homogeneity. In this way, in agreement with the definition of within-sample homogeneity, we are focusing on the difference in the state of two adjacent probes and not on the state of a single probe. Let $H$ be a matrix $N \times M-1$ where for each sample and for each pair of probes of the matrix $D$ the element $H_{ij}$ measures the degree of homogeneity between the probes $j$ and $j+1$ for the $i$-th sample. In particular, $H_{ij}$ has value 0, 0.5 and 1 for maximum, medium and minimum homogeneity, respectively. From this matrix, we compute the overall homogeneity ($h$-value) as:

$$h_j = \frac{1}{N}\sum_{i=1}^{N} H_{ij}, \quad j = 1, \ldots, M-1 \tag{1}$$

From the configuration of the $h$-values, we can obtain information on the homogeneity of the dataset. In the case in which $h_j = 0$, we know that each sample has the same state for the probes $j$ and $j+1$; in contrast, where we have $h_j = 1$ we know that each sample has opposite states for those probes. Homogeneous peel-off works in the following way: let $l$ and $m$ be the indices representing, respectively, the left and the right boundary of the peak with minimum $q$-value, where $1 \leq l, m \leq M$, we expand the left boundary of the region if:

$$q_{l-1} \leq q_{\text{thr}} \quad \text{AND} \quad h_{l-1} \leq h_{\text{thr}} \tag{2}$$

and we expand the right boundary if:

$$q_{r+1} \leq q_{\text{thr}} \quad \text{AND} \quad h_r \leq h_{\text{thr}} \tag{3}$$

The procedure iteratively expands the boundaries of the region until the above conditions are satisfied. The value of $h_{\text{thr}}$ represents an important parameter for homogeneous peel-off, in fact imposing $h_{\text{thr}} = 0$ the computation follows the same scheme of standard peel-off, in contrast by imposing $h_{\text{thr}} = 1$ the whole significant region will be extracted obtaining many spurious peaks. A value $0 < h_{\text{thr}} < 1$ accounts for the amount of contextual information adopted to measure the homogeneity.

Consider the example in Figure 1A, of course probes P5 and P6 represent two recurrent CNAs, but we can notice that there are two homogeneous regions included between blue and green squares. By using standard peel-off only the peak overlapping probes P5–P6 is extracted. Indeed, in the first iteration we have the $q$-value distribution of Figure 1D and peel-off removes peak P5–P6. At the next iteration, the new $q$-value configuration is the one reported in Figure 1E where no significant peak exists (all probes have a $q$-value greater than the significance threshold of 0.25 represented by the red line) and peel-off ends. If we increase the value of $q_{\text{thr}}$ standard peel-off captures probes P3, P4 and P7 but in this case P3 represents a spurious peak. In other words, in standard peel-off adjacent significant peaks can be extracted by an increase of the significance threshold, but in consequence of this increment spurious peaks are also detected. To overcome this limitation in JISTIC (Sanchez-Garcia *et al*., 2010), a variant of peel-off, called limited, was proposed. In JISTIC, neighboring probes of a significant peak are also considered. It has been demonstrated that limited peel-off allows to obtain more accurate results than standard peel-off, but this approach also suffers from the effect of spurious peaks. For the example depicted in Figure 1A, JISTIC identifies the region P4–P7 but it also extracts the false peak located in P3 (more detail on JISTIC are provided in the following).

In contrast to standard and limited peel-off, GAIA extracts only the region P4–P7. Figure 1F shows the $h$-value configuration for the data of Figure 1A where the red line represents the $h_{\text{thr}}$. GAIA starts from the peak P5–P6 (the peak having minimum $q$-value) and when it tries to expand the detected region on the right side it finds that in P6 the condition in (3) is satisfied and the right boundary of the region is moved in P7. In P7, the condition (3) is false both for the $q$-value and for the $h$-value and GAIA stops the expansion on the right side. In P5, the condition (2) for the left boundary is satisfied and the left boundary is moved in P4, when the condition (2) is tested in P4 it is found true for the $q$-value but false for the $h$-value so the expansion is stopped. In this way, GAIA is able to detect the peak covering the probes P4–P7 omitting the spurious peak in P3.

## 2.2 GISTIC and JISTIC

JISTIC (Sanchez-Garcia *et al*., 2010) is a variant of the previously published rCNA-algorithm GISTIC (Beroukhim *et al*., 2007). These algorithms use smoothed LRRs to compute a statistic (G-score) representing the strength of the aberration of each probe. The computed G-score is compared with the one expected by chance using a permutation test, where significance is corrected by the FDR proposed in Benjamini and Hochberg (1995). JISTIC and GISTIC differ in their peel-off procedure. As explained before, standard peel-off of GISTIC simply extracts the peak of a region with the minimum $q$-value, in contrast, limited peel-off proposed in JISTIC also considers the probes close to the selected peak. In particular, JISTIC decomposes the G-score for a probe in two parts: one representing what remains from the selected peak ($G_r$) and another part that is the complementary of $G_r$ ($G_n$), which considers the independent contribution of the peak. JISTIC considers any reduction in the aberration that is consistent for a fixed number of adjacent probes within a window of fixed size. A new peak is detected if within the window a probes exists for which $G_n$ is lower then a fixed threshold. Sanchez-Garcia *et al*. (2010) demonstrated that JISTIC performs better than GISTIC in terms of specificity and recall, so we chose to use JISTIC instead of GISTIC to perform the comparisons.

Although GAIA is similar to GISTIC and JISTIC, some important differences exist. The first difference is on the input data: JISTIC and GISTIC directly work on the LRRs; in contrast, GAIA works on the list of aberrant regions with the respective labels; it uses a discrete representation of the data. Indeed, another important difference is on the method used to obtain the derivation of the null distribution. Although all algorithms use a permutation test, GISTIC and JISTIC derive a semiexact estimate of this null distribution by using a convolution of histograms, while GAIA explicitly performs the permutations. But of course the most important difference is the peel-off procedure. It is important to notice that the proposed homogeneous peel-off can be applied only if a discrete representation of the data is used. JISTIC is implemented as a platform-independent Java application and it can be easily adapted for working on synthetic and real aCGH data.

## 2.3 GADA

In GADA (Pique-Regi *et al*., 2009), the observed LRR is decomposed into three components: the change in hybridization due to altered copy number, the reference hybridization intensity for non-aberrant probes and the noise component modeled as a zero-mean Gaussian process. The basic assumption of GADA is that the copy number hybridization component is piece wise constant with a small number of regions. GADA uses a sparse Bayesian prior different for each sample and an expectation maximization (EM) algorithm to jointly estimate the model parameters. An hyperparameter is used to control the expected degree of sparseness and in order to efficiently explore solutions with different levels of sparseness, a backward elimination procedure is applied: breakpoints separating two regions are removed if they have a score lower then a fixed threshold. GADA is available as an R package.

## 2.4 cghMCR

cghMCR (Aguirre *et al*., 2004) is a rCNA-algorithm where smoothed data are used to distinguish between normal and altered probes. In particular, segments above the 97th or below the 3rd percentile are considered altered (note that this way a discretized representation of the data is obtained); altered segments are joined if they are either adjacent or separated by a segment $< 500$ kb. cghMCR considers as 'informative' segments $< 20$ Mb and among

them it returns all regions that are found aberrant in at least 75% of samples. This threshold represents a fundamental parameter for the algorithm and the user can change it and obtain different results. Finally, segments separated by just one probe are joined. cghMCR was developed to be applied on a real aCGH dataset of pancreatic adenocarcinoma and it allowed the rediscovery of known aberrant cytobands and the identification of undescribed CNAs. cghMCR is available as an R/Bioconductor package.

### 2.5 Other rCNA-algorithms

In this section, we provide a brief review of rCNA-algorithms. The work of Rouveirol *et al*. (2006) represents one of the first efforts in the development of a rCNA-algorithm. In their work, the authors describe two approaches, called MAR and CMAR, where discrete data profiles are used to search for rectangles delimiting probes having the same aberrant state. Authors used an approach widely applied in Data Mining, in the area of frequent itemset mining. CMAR differs from MAR by the usage of a set of boundary constraints. CGHregions (van de Wiel *et al*., 2007) accepts as inputs discrete labels for each observed probe and by a matrix dimension reduction it identifies patterns of probes that remain (almost) constant. CGHregions also attempts to focus the analysis of aCGH data on the change of the state between adjacent probes and not on the state of a single probe. In this work, the performance of CGHregions are not evaluated because it was not designed for the discovery of recurrent CNAs, but it rather computes homogeneous regions over the subjects. In KC-SMART (Klijn *et al*., 2008), positive and negative LRRs are separately summarized across the samples. In this approach, a flat top Gaussian kernel function is used to perform locally weighted regression and to produce a smoothed estimation of the CNAs. pREC-A and pREC-S (Rueda and Diaz-Uriarte, 2009) are two rCNA-algorithms that, starting from observed LRRs, use a Hidden Markov Model to identify recurrent CNAs. In particular, these approaches compute the joint probabilities of alteration for a sequence of adjacent probes, that is, for each sample they compute the probabilities that the same kind of aberration is found in a subset of consecutive adjacent probes. BSA (Yang *et al*., 2009) uses a Bayesian hierarchical model based on the assumption that copy number observations for different subjects at different genomic positions are independent conditional on the boundaries. BSA sequentially detects the boundaries of aberrant segments and it ends if the average of the highest sample mean of candidate segments is less than a fixed threshold. MPCBS (Zhang *et al*., 2010) enables the detection of CNA from multiple platforms (i.e. Affymetrix, Agilent and Illumina) providing in output a 'multi-platform consensus' score for each probe. Generally, results from multiple platforms are combined after the segmentation; in contrast, in MPBCS the statistical evidence is directly computed across platforms in the segmentation step. CNAnova (Ivakhno and Tavaré, 2010) is based on the assumption that the dataset is divided into a reference set of normal individuals and a set of cancer samples. Based on this assumption and starting from the observed LRRs, it computes the distribution of the $F$- and $t$-statistics by using the one-way analysis of variance (ANOVA). Boundaries of CNA regions are detected by a gradient kernel density estimation and the FDR is used to identify significant regions. DiNAMIC (Walter *et al*., 2011) is a recently published rCNA-algorithm that works either on discrete or on continuous segmented data and implements a cyclic shift procedure to compute the null distribution. DiNAMIC can be applied in the analysis of data from individual chromosome or genome wide and it uses a peel-off procedure to extract the final list of CNAs. Here just JISTIC, GADA and cghMCR will be considered for comparison purposes.

### 2.6 Segmentation of aCGH data

Segmentation represents the first analysis step for many rCNA-algorithms. Aim of a segmentation algorithm is the identification of breakpoints in the genome that identify regions in which probes have a similar copy number profile or share the same state. GAIA needs segmented input data where a discrete label (i.e. loss, normal and gain) is assigned to each region. In order to provide this input we used VEGA, a segmentation algorithm presented in our previous work (Morganella *et al*., 2010). Also JISTIC needs segmented data; in particular, it requires as input a list of regions spanning the whole genome with the respective smoothed mean. For this reason, in the original paper of GISTIC, authors used the GLAD segmentation algorithm (Hupé *et al*., 2004), and hence we used GLAD also for JISTIC. In order to perform a fair comparison between JISTIC and GAIA, we also used VEGA to provide input data for JISTIC. Smoothed data also represent the input for cghMCR. The R/Bioconductor package implementing cghMCR uses a segmentation algorithm known as DNAcopy (Olshen *et al*., 2004).

## 3 RESULTS

For real data, the ground truth of CNA is not available, so in order to perform a qualitative evaluation of rCNA-algorithms we need to extract from biological studies a list of verified CNAs. Although this list can contain CNAs that are not present within the analyzed dataset, it can be efficiently used to have an idea about the results obtained by the rCNA-algorithms. To overcome the limitations occurring on the evaluation performed on real data, we generated synthetic data for which the ground truth is available in order to perform a quantitative assessment.

In order to run the rCNA-algorithms on the considered data, the respective input parameters must be chosen. GAIA has just two main parameters: the significance threshold ($q_{\text{thr}}$) and the homogeneity threshold ($h_{\text{thr}}$) for which we used values 0.25 and 0.12, respectively. This parameter setting has been used both on synthetic and real data. These parameters can be considered as algorithm driven and do not need to be selected on the basis of the dataset. In JISTIC, six parameters must be specified and the default setting suggested by authors has been used to perform all comparisons. On real data, both GADA and cghMCR were used with the suggested parameter configuration, in contrast on synthetic data we perform a parameter tuning in order to obtain the best performance for these algorithms. In the Supplementary Material, more details about the parameter settings are provided.

### 3.1 Synthetic dataset

*Generation strategy*: we generated synthetic data according to three main scenarios (Supplementary Fig. SF1), because they represent the fundamental patterns that are found in real datasets (Rueda and Diaz-Uriarte, 2010), and many other scenarios can be considered as a combination of these three patterns. Scenario I is the simplest case of recurrent CNA regions: there is a recurrent region sharing the same position in all samples. In Scenario II, there are two non-overlapping CNA regions that differ in the kind of aberration and in the fraction of affected subjects (40% of samples show a loss and 60% of samples show a gain). Finally, Scenario III is a hybrid of Scenarios I and II: different aberrations share the same position and each of them affects only a fraction of the subjects.

For each selected Scenario, we simulated a chromosome of 1000 probes considering different CNA widths (100, 200 and 400). In each dataset, both position and kind of the CNA were randomly chosen and the unbiased LRRs for loss, normal and gain were considered to be $log_2(\frac{1}{2})=-1$, $log_2(\frac{2}{2})=0$ and $log_2(\frac{3}{2})=0.58$, respectively. From this unbiased dataset, we generated perturbed datasets by using two different noise models. The first model is an intensity noise affecting the values of unbiased LRRs which is assumed to follow a white Gaussian distribution $\sim\mathcal{N}(0,\sigma)$, while

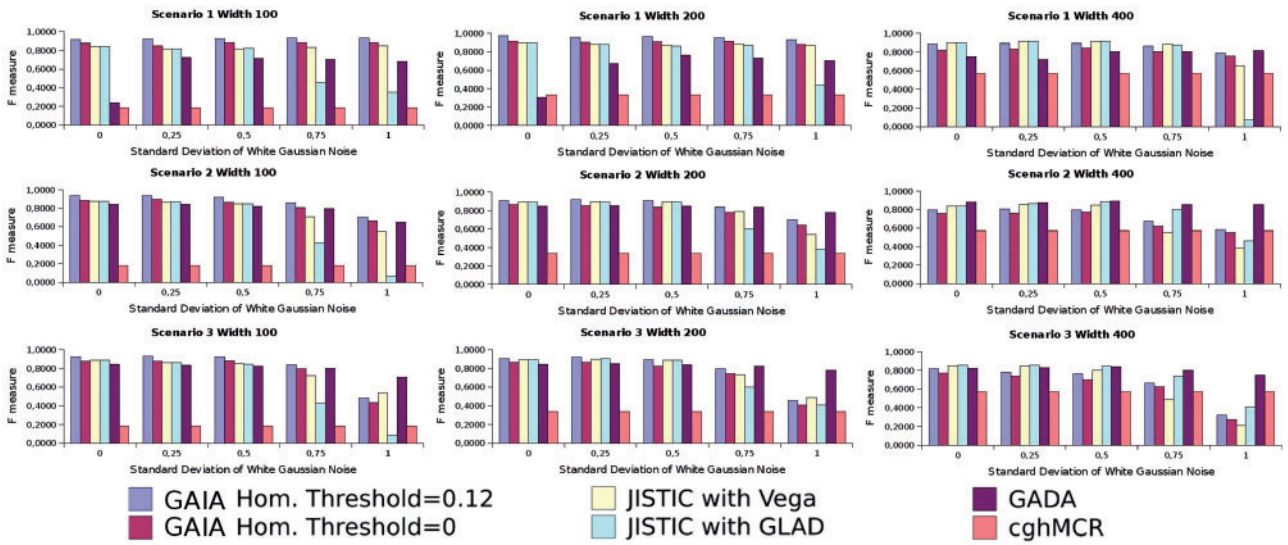


**Fig. 2.** Results on the synthetic dataset perturbed by both intensity and spatial noise. In each chart, $x$-axis is the SD $\sigma$ of the white Gaussian process modeling the intensity noise $\sim\mathcal{N}(0,\sigma)$ and $y$-axis reports the $F$-measure for all compared approaches. The values of $\alpha$ and $\beta$ for the generation of the random resizing and shifting are both set to 0.25

the second model represents a spatial noise that resizes and shifts the location of the CNAs. This kind of perturbation could better model biological noise and errors in the localization of the boundary of the sample aberration induced by segmentation algorithms. Resizing is modeled as multiplicative Gaussian noise $\sim\mathcal{N}(1,\alpha\cdot w)$ and the random shifting of the middle position of the aberration is modeled as a zero mean Gaussian $\sim\mathcal{N}(0,\beta)$ ($w$ is the width of the CNA). We generated two simulated datasets: the first perturbed with only intensity noise, while the second synthetic dataset contains intensity, resizing and shifting perturbations. Typical examples of this second dataset is reported in Supplementary Figure SF2. We can see that the second synthetic dataset is quite realistic and challenging. By varying the value $\sigma$ of the intensity noise (0, 0.25, 0.5, 0.75 and 1) and by using $\alpha=\beta=0.25$, we obtained a total of 2250 different synthetic datasets which were used for comparison purposes.

*Evaluation metrics*: quantitative assessment is based on the number of true positives (TP), false positives (FP) and false negatives (FN). A TP occurs when the rCNA-algorithm correctly detects a CNA for a probe contained into a simulated CNA region; in contrast, if the detected CNA corresponds to a probe outside of the simulated aberrant region, we have an FP. Finally, we have an FN when the rCNA-algorithm does not detect an aberration for probes located into a simulated CNA region. As a single figure of merit, we calculate the $F$-measure representing harmonic mean of precision (exactness) and recall (completeness).

*Results*: Figure 2 shows the $F$-measures obtained from the synthetic dataset perturbed by both intensity and spatial noise, and results are also summarized in Supplementary Table ST2. For GAIA, two different values of $h_{\text{thr}}$ were used: 0 and 0.12, in particular $h_{\text{thr}}=0$ corresponds to standard peel-off. The reported experiments show that in all considered cases GAIA produced more accurate results by using within-sample homogeneity than the standard peel-off. As expected, noise influenced performance of rCNA-algorithms and the impact of noise was more evident in Scenario III where different aberrations are overlapped. All approaches (except for cghMCR) provided good results in all simulated scenarios and the width of CNAs did not influence the respective performance. From the performance of JISTIC, we can note that it produced more accurate results by using VEGA rather than GLAD, and this observation is particularly true for width of 100 and $\sigma$ values of 0.75 and 1. Results also show the ability of GADA in detection of simulated CNAs. Another consideration is on the lower performance of cghMCR, indeed, although we performed a tuning of the parameters, this approach presented important difficulties in the detection of simulated CNAs. In particular, we can notice that performance of cghMCR improved as the width of the aberration increased. This behavior was due to the large amount of FP produced by this algorithm. However, cghMCR has the advantage to be very fast as reported in Section 3.4. The performance trend highlighted on the synthetic dataset perturbed by both noise models was also confirmed by the results on the simpler dataset perturbed only by intensity noise (Supplementary Fig. SF3 and Table ST1).

## 3.2 Results on colorectal cancer

Colorectal cancer (CRC) is one of the most common tumors and many biological studies have been designed to characterize the aberrations involved in this disease. These studies have generated a list of verified CRC alterations that can be used to perform a qualitative analysis of rCNA-algorithms on real aCGH data. In this work, we analyzed a recently published CRC dataset (Venkatachalam *et al*., 2011) extracted from 41 patients who were diagnosed with microsatellite-stable CRC without polyposis. From this dataset, we selected 30 samples that were hybridized on SNP 250k Affymetrix GeneChip arrays (data available in GEO with identifier GSE13429). Raw data were preprocessed by PennCNV tool (Wang *et al*., 2007).

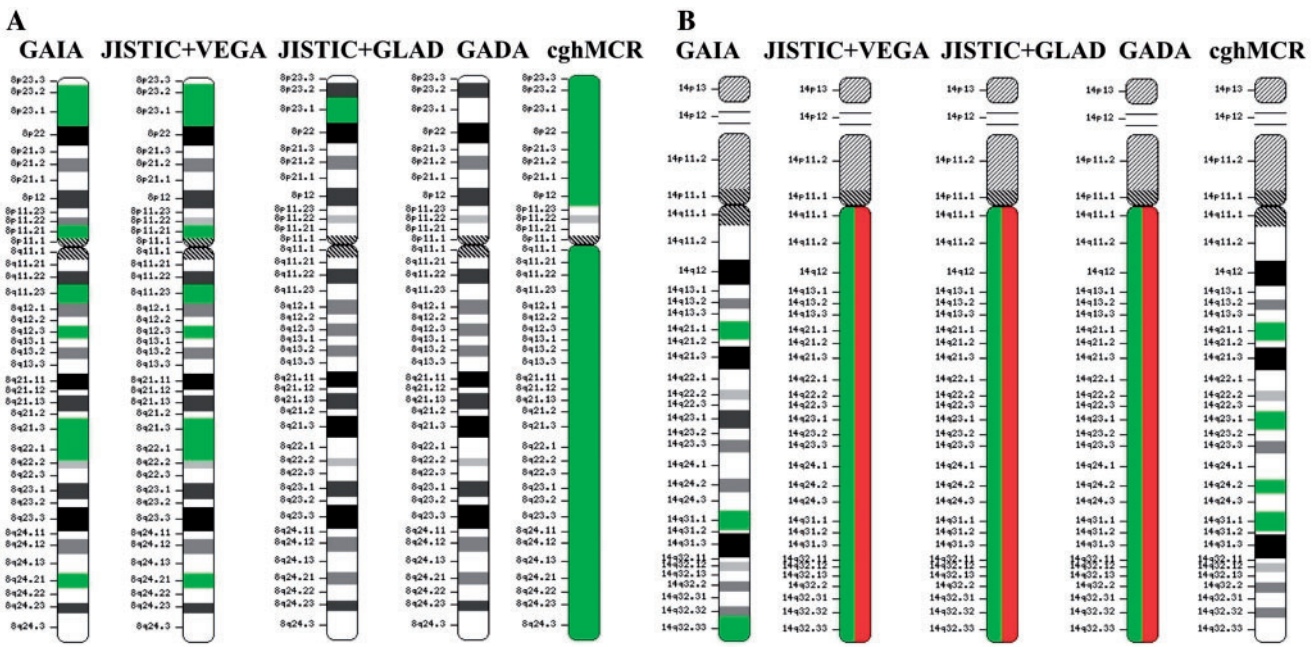


**Fig. 3.** Results on real aCGH data. In (**A**), results for CRC dataset on chromosome 8. In (**B**), results for GIST dataset on chromosome 14. Green and red indicate loss and gain, respectively.

Development of CRC is a multistep process that involves an accumulation of mutations in tumor suppressor genes and oncogenes. One of the earliest trigger genetic event in CRC is the inactivation of the *APC*, a tumor suppressor gene antagonist of *Wnt* signaling pathway (Locker *et al*., 2006). Other evidences in CRC are mutations of the well-known tumor suppressor gene *PTEN* and high-level amplifications at 20q13.2 containing the target gene *STK6* (Nakao *et al*., 2004). Methylation of *CDKN2A* and *P16* were detected in CRC (Goto *et al*., 2009) and loss of *UNC5C* expression was observed in 68% of primary CRC (Bernet *et al*., 2007). The list of all considered aberrations is reported in Supplementary Table ST3a. GAIA and JISTIC + VEGA produced nearly the same number of cytobands (184 and 182, respectively), JISTIC + GLAD computed 39 cytobands, cghMCR detected 587 cytobands with a recurrence threshold of 50% (increasing this value to 60% only six cytobands were obtained and with the default value of 75% no aberrations were found) and GADA returned nine cytobands (Supplementary Table ST3a). From the results in Supplementary Table ST3a, we can note that GADA and JISTIC + GLAD had significant difficulties in the detection of the considered CNAs. In contrast GAIA, JISTIC + VEGA and cghMCR produced a list of aberrations consistent to the biological evidence, but some disagreements with the expected aberrations were found. In particular, for cghMCR some biological contradictions were observed, among them the gain of *PTEN* cytoband may be considered a false evaluation. JISTIC + VEGA was in agreement with the biological evidence in 57% of cases, while both GAIA and cghMCR correctly identified 64% of mutations, but cghMCR produced more than three times the number of cytobands of GAIA. In Figure 3A, the results produced on the chromosome 8 by all considered approaches are reported. This chromosome was chosen because the loss of 8p23.1 is the only mutation detected by all rCNA-algorithms (except for GADA). From this figure, we can notice that both GAIA and JISTIC produced narrow aberrations; in contrast, cghMCR found the whole chromosome as lost and GADA did not detect any aberrations.

### 3.3 Results on gastrointestinal stromal tumor

Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal tumors of the gastrointestinal tract. We used the data published by Astolfi *et al*. (2010) where 25 fresh tissue specimens of GISTs were collected and hybridized by Affymetrix Genome Wide SNP 6.0 (GEO identifier GSE20710). Raw data were preprocessed by PennCNV tool (Wang *et al*., 2007) obtaining the LRR for ∼1.6 million of probes. In order to qualitatively investigate the performance of the compared approaches, as above, we used some well-known cytogenetic aberrations characterizing GIST (Supplementary Table ST3b). In particular, loss of cytobands 14q11.2, 14q32.33, 22q12.2 and 22q13.31 appears to play an important role in the early stage of tumor formation and in late tumor progression (Ässämäki *et al*., 2007; Lasota *et al*., 2005). Cytogenetic losses of 1p36.23 and 9p21.3 have also been related with GISTs (Ässämäki *et al*., 2007; Perrone *et al*., 2005) and gains of 7p11.2 and 12q15 were confirmed by *in situ* hybridization (Tornillo *et al*., 2005).

The first observation is about the number of computed cytobands (Supplementary Table ST3b): GAIA detected 272 cytobands, JISTIC + VEGA detected 577 cytobands, JISTIC + GLAD detected 548 cytobands, cghMCR detected 320 cytobands and GADA detected 698 cytobands. For GADA, the recurrence parameter was set to 65%; with the default value of 75% no significant result was obtained and decreasing this threshold to 50% resulted in the detection of >2.5 billion of aberrant probes. From Supplementary Table ST3b,

we can see that cghMCR and GADA were the only approaches that detected the gain in 12q15, but cghMCR failed to detect all other aberrations except for 9p21.3 and 8q24.21. GADA performed well on GIST dataset but there are several cytobands which were reported both as loss and as gain (9p21.3, 14q11.2, 14q32.33, 7p11.2 and 22q13.31). Also on GIST, the segmentation produced by VEGA allowed JISTIC to obtain more consistent results than GLAD. But JISTIC had the same anomalies of GADA in cytobands 14q11.2, 14q32.33, 22q13.31 and 8q24.21 which were found both in loss and in gain. In Figure 3B, the results produced on the chromosome 14 are reported. From this figure, it is evident that both JISTIC and GADA identified very large aberrations, in contrast GAIA and cghMCR detected narrow CNAs. Another consideration is that JISTIC and GADA found the whole long arm chromosome both in gain and in loss, highlighting the previously described problem.

### 3.4 Computational aspects

The actual implementation of GAIA processes one chromosome at time. In GAIA, the most expensive operation is the computation of the null distribution by the permutation test. Let $N$ be the number of samples and $M$ the number of observed probes, GAIA requires a time $\mathcal{O}(NMK)$ where $K$ is the number of performed permutations. JISTIC uses the same scheme of GAIA, but it needs to compute for each sample and for each probe the respective G-score: this operation requires a linear time of $\mathcal{O}(NM)$. Moreover, in JISTIC the null distribution is computed by a convolution of histograms that requires a time of $\mathcal{O}(MH^2)$, $H$ is the number of bins of the histogram. So the overall time required by JISTIC is $\mathcal{O}(M(N+H^2))$. In cghMCR, a linear number of steps is required $\mathcal{O}(NM)$, indeed, it simply performs a counting of the number of alteration for each probe. Finally, the time required by GADA depends on the convergence of the EM algorithm used to estimate the parameters of the model. Each iteration of EM can be performed in linear time and the overall complexity is $\mathcal{O}(NMT)$ where $T$ is the number of iterations to reach the convergence.

A significant improvement in the performance of GAIA, with a complexity of $\mathcal{O}(NK)$, can be obtained by considering an approximation of the null distribution. In particular, after computing the frequency of alteration for each sample $\theta_j$, we can simulate a matrix with dimension $N \times K$ where each column is a vector in which the element $j$ has a probability for drawing 1 equal to $\theta_j$. This asymptotically approximates the original simulation for large $M$ and $K$.

Supplementary Table ST4 reports the execution times for the compared algorithms. An important consideration is the fact that reported times depend on the respective software packages and, of course, a careful implementation (e.g. by using *ad hoc* data structures) can notably improve the performance of algorithms. Reported execution time of cghMCR is related to the counting of the aberrations, and it does not contain neither the time required by the segmentation (that can be very long) nor the time required to produce a formatted output. Execution time of GADA is improved by calling, from R, a function written in C. In JISTIC, the generation of the needed input matrix file from smoothed data is not considered for calculation of execution time. Finally, reported execution time of GAIA includes both data loading and formatted output file construction. From Supplementary Table ST4, the computational improvement induced by the approximation of the null distribution is evidenced.

## 4 DISCUSSION

GAIA uses within-sample homogeneity to obtain results consistent with the biological nature of copy number profiles and the reported results suggested that this heuristic can improve the accuracy. Extensive comparison on simulated data allowed the observation of the behavior of the considered algorithms. GADA seemed to be particularly robust with respect to noise. GAIA and JISTIC had comparable performance with a slight advantage for GAIA and finally less accurate performance were obtained by cghMCR. Analysis of two real aCGH datasets suggested an important observation: approaches directly working on LRR were strongly affected by the chip resolution. In particular, both for JISTIC and GADA the number of identified regions increased with increasing resolution. In contrast, GAIA, which uses a discrete representation of the data, resulted in a stable estimate of the number of computed regions with respect to the chip resolution. Results on synthetic and real aCGH data showed that cghMCR produced several spurious peaks as also observed in Rueda and Diaz-Uriarte (2010). Another important aspect pointed out by our analysis is on the execution time. cghMCR was very fast but it simply performs a counting of the number of alteration for each probe. Execution time of GADA is related to the convergence of the EM algorithm used to estimate the parameters of the model. Convergence is slower in high-resolution scenarios, but its execution time seemed to be good. Finally, both GAIA and JISTIC use a conservative permutation model to compute the distribution of the null hypothesis. In its approximated version, GAIA seems to be faster than JISTIC. But GAIA had also an acceptable computational time (1 h in the largest dataset) when it explicitly performs the permutations. We also highlighted the fact that the list of computed CNAs must be integrated with information measuring the strength of evidence of aberrations (i.e $q$-value). Indeed, this information can be used to resolve ambiguous results (this is the case of JISTIC and GADA on GIST dataset where regions were found both in loss and in gain) and to choose CNAs representing the target of further biological investigations.

Although our methodology can be subject to further improvements, our opinion is that GAIA is a valid alternative to other rCNA-algorithms, especially for the analysis of high-resolution aCGH data. Indeed, the performed analysis showed that GAIA can be particularly useful for high-resolution data which have been already segmented (as for example TCGA CNA data). Indeed, we report a better stability with respect to the number of detected regions in different resolution scenarios. In other words, the amount of false positives produced is not related to the resolution, and this contrasts noticeably with the rCNA-algorithms considered for comparison. In addition, when GAIA used an approximated approach to compute the null distribution, the processing of large amount of data is faster than other approaches based on significance analysis. We also pointed out the fact that, of course, accurate tuning of the input parameters can significantly improve the results of the rCNA-algorithms. But often, this tuning involves many parameters and their perfect combination is very hard to reach. Therefore, another very important advantage of GAIA is that in contrast to other considered rCNA-algorithms, it requires just two input parameters. Finally, although in literature discretization

of LRR data has been criticized because of the potential loss of information it entails, the results presented here show that if we feel confident (as we were for VEGA, but in general this occurs in high-resolution scenario) on the region labeling performed by the segmentation algorithm, we can produce more accurate list of CNAs than algorithms working directly on the LRR which have the problem that smoothed LRRs might not be comparable across the arrays.

## 5 CONCLUSION

Recurrent CNAs represent a very important source of information about genetic diseases like cancer, and rCNA-algorithms can be used to find significant associated aberrations. In this work, a new rCNA algorithm has been presented and a performance analysis of rCNA-algorithms was performed. Results showed that joint analysis of all available samples without a preprocessing segmentation step (as in GADA) can improve robustness respect to noise and that within-sample homogeneity can be used to increase the accuracy of the results. In addition, we point out two interesting strong points of GAIA: its ability in working on high-resolution data and its parameter setting that is very straightforward.

## ACKNOWLEDGEMENT

We would like thank the reviewers for their detailed comments which were useful to improve the manuscript. One of the reviewers suggested to use the approximation of the distribution which improved the computational performance of the developed algorithm.

*Funding*: MiUR (Ministero dell'Università e della Ricerca) under (grant PRIN2008-20085CH22F).

*Conflict of Interest*: none declared.

## REFERENCES

Aguirre,A.J. *et al*. (2004) High–resolution characterization of the pancreatic adenocarcinoma genome. *Proc. Natl Acad. Sci. USA*, **101**, 9067–9072.

Albertson,D.G. *et al*. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Ässämäki,R. *et al*. (2007) Array comparative genomic hybridization analysis of chromosomal imbalances and their target genes in gastrointestinal stromal tumors. *Genes Chromosomes Cancer*, **46**, 564–576.

Astolfi,A. *et al*. (2010) A molecular portrait of gastrointestinal stromal tumors: an integrative analysis of gene expression profiling and high-resolution genomic copy number. *Lab. Investig.*, **90**, 1285–1294.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bernet,A. *et al*. (2007) Inactivation of the UNC5C Netrin-1 receptor is associated with tumor progression in colorectal malignancies. *Gastroenterology*, **133**, 2045–2049.

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.

Beroukhim,R. *et al*. (2009) Patterns of gene expression and copy-number alterations in VHL disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res.*, **69**, 4674–4681.

Beroukhim,R. *et al*. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Feuk,L. *et al*. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Goto,T. *et al*. (2009) Aberrant methylation of the p16 gene is frequently detected in advanced colorectal cancer. *Anticancer Res.*, **29**, 275–277.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Ivakhno,S. and Tavaré,S. (2010) CNAnova: a new approach for finding recurrent copy number abnormalities in cancer SNP microarray data. *Bioinformatics*, **26**, 1395–1402.

Klijn,C. *et al*. (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, **36**, e13.

Lasota,J. *et al*. (2005) Loss of heterozygosity on chromosome 22q in gastrointestinal stromal tumors (GISTs): a study on 50 cases. *Lab. Investig.,*, **85**, 237–247.

Locker,G.Y. *et al*. (2006) The I1307K APC polymorphism in Ashkenazi Jews with colorectal cancer: clinical and pathologic features. *Cancer Genet. Cytogenet.*, **169**, 33–38.

Morganella,S. *et al*. (2010) VEGA: variational segmentation for copy number detection. *Bioinformatics*, **26**, 3020–3027.

Nagy,R. *et al*. (2004) Highly penetrant hereditary cancer syndromes. *Oncogene*, **23**, 6445–6470.

Nakao,K. *et al*. (2004) High resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics,* **5**, 557–572.

Perrone,F. *et al*. (2005) 9p21 locus analysis in high-risk gastrointestinal stromal tumors characterized for c-kit and platelet- derived growth factor receptor gene alterations. *Cancer*, **4**, 159–169.

Pique-Regi,R. *et al*. (2009) Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics*, **25**, 1223–1230.

Rouveirol,C. *et al*. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

Rueda,O.M. and Diaz-Uriarte,R. (2009) Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics*, **10**, 308.

Rueda,O.M. and Diaz-Uriarte,R. (2010) Finding recurrent copy number alteration regions: a review of methods. *Curr. Bioinformatics*, **5**, 1–17.

Sanchez-Garcia,F. *et al*. (2010) JISTIC: identification of significant targets in Cancer. *BMC Bioinformatics*, **11**, 189.

Sartore-Bianchi,A. *et al*. (2007) Epidermal growth factor receptor gene copy number and clinical outcome of metastatic colorectal cancer treated with panitumumab. *J. Clin. Oncol.* **25**, 3228–3245.

Shah,S.P. (2008) Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet. Genome Res.,* **123**, 343–351.

Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 62.

Snijders,A.M. *et al*. (2003) Genome-wide- array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in Fallopian tube carcinoma. *Oncogene*, **22**, 4281–4286.

Storey,J.D. *et al*. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc.*, **66**, 187–205.

Taylor,B.S. *et al*. (2008) Functional copy-number alterations in cancer. *PLoS One*, **3**, e3179.

Tornillo,L. *et al*. (2005) Array comparative genomic hybridization analysis of chromosomal imbalances and their target genes in gastrointestinal stromal tumors (GIST). *Lab. Investig.*, **85**, 921–931.

van de Wiel,M.A. *et al*. (2007) CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informat.*, **3**, 55–63.

Venkatachalam,R. *et al*. (2011) Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int. J. Cancer*, **129**, 1635–1642.

Walter,V. *et al*. (2011) DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. *Bioinformatics*, **27**, 678–685.

Wang,K. *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.

Westfall,P.H. and Young,S.S. (1993) Resampling-based multiple testing: examples and methods for pvalue adjustment. John Wiley, New York.

Yang,L. *et al*. (2009) A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics* **25**, 1669–1679.

Zhang,N.R. *et al*. (2010) Joint estimation of DNA copy number from multiple platforms. *Bioinformatics*, **26**, 153–160.