BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 24 2010, pages 3020–3027
doi:10.1093/bioinformatics/btq586

Genome analysis

Advance Access publication October 19, 2010

VEGA: variational segmentation for copy number detection

Sandro Morganella[1,2], Luigi Cerulo[1,2], Giuseppe Viglietto[2,3] and Michele Ceccarelli[1,2,*]

[1]Department of Biological and Environmental Studies, University of Sannio, Via Port'Arsa 11, 82100, Benevento, [2]Bioinformatics Core, IRGS Istituto di Ricerche Genetiche "G. Salvatore", BioGeM s.c.a r.l. Ariano Irpino (AV) and [3]Dipartimento di Medicina Sperimentale e Clinica G Salvatore; Università Magna Graecia, Catanzaro, Italy

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Genomic copy number (CN) information is useful to study genetic traits of many diseases. Using array comparative genomic hybridization (aCGH), researchers are able to measure the copy number of thousands of DNA loci at the same time. Therefore, a current challenge in bioinformatics is the development of efficient algorithms to detect the map of aberrant chromosomal regions.

Methods: We describe an approach for the segmentation of copy number aCGH data. Variational estimator for genomic aberrations (VEGA) adopt a variational model used in image segmentation. The optimal segmentation is modeled as the minimum of an energy functional encompassing both the quality of interpolation of the data and the complexity of the solution measured by the length of the boundaries between segmented regions. This solution is obtained by a region growing process where the stop condition is completely data driven.

Results: VEGA is compared with three algorithms that represent the state of the art in CN segmentation. Performance assessment is made both on synthetic and real data. Synthetic data simulate different noise conditions. Results on these data show the robustness with respect to noise of variational models and the accuracy of VEGA in terms of recall and precision. Eight mantle cell lymphoma cell lines and two samples of glioblastoma multiforme are used to evaluate the behavior of VEGA on real biological data. Comparison between results and current biological knowledge shows the ability of the proposed method in detecting known chromosomal aberrations.

Availability: VEGA has been implemented in R and is available at the address http://www.dsba.unisannio.it/Members/ceccarelli/vega in the section Download.

Contact: ceccarelli@unisannio.it

Supplementary information: Supplementary information is available at *Bioinformatics* online.

Received on July 22, 2010 ; revised on October 11, 2010 ; accepted on October 12, 2010

1 INTRODUCTION

Recent biological studies show the close relationship between chromosomal regions aberrant in copy number (CN) and diseases like tumor (Beroukhim *et al*., 2010; Harada *et al*., 2008; Zhao *et al*., 2003) and mental retardation (Fan *et al*., 2007; Sebat *et al*., 2007). High resolution CN estimation makes use of comparative genomic hybridization arrays. DNA from a test sample and normal reference sample are labeled differentially, using different fluorophores, and hybridized to several thousand probes. The ratio of the fluorescence intensity of the test to that of the reference DNA is then calculated, to measure the CN changes for a particular location in the genome. In particular, the log$R$ ratio (LLR) gives an indirect measure of CN of each probe by plotting the ratio of observed to expected hybridization intensity. After a microarray has been constructed and hybridized, the corresponding image can be acquired and additional analysis steps can be used both to reduce noise and to increase the statistical confidence of the observations (generally each probe is spotted in several copies) (Khojasteh *et al*., 2005). The output of this process is a list of LRR values with the respective genomic positions. The so called *segmentation* algorithms use this list of measurements to compute the features (breakpoint positions and kind of mutation) of the aberrant regions along the genome. The availability of efficient segmentation algorithms plays an important role in the mapping of aberrant chromosomal regions. Accurate aberration mapping can be used to extract new insight on the mechanisms leading to genetic diseases. For this reason in literature, several segmentation approaches have been proposed.

In likelihood function-based approaches (Jong *et al*., 2003, 2004; Myers *et al*., 2004), breakpoint positions are estimated by using a maximum likelihood criterion in which penalty terms are used to limit the complexity of the solution. Often penalty terms are controlled by weights which can be chosen adaptively to the data (Hupe *et al*., 2004; Picard *et al*., 2005). An interesting likelihood function-based approach is DNACopy (Olshen *et al*., 2004) which is based on a modification of the original binary segmentation proposed by Sen and Srivastava (1975). DNACopy segments the chromosome into contiguous regions of equal CN using a non-parametric permutation reference distribution which takes in account the effect of noise. In particular, the authors model CN data as a sequence of random variables and the maximum likelihood is used recursively to look for change points where adjacent random variables have a different distribution function. In addition, DNACopy uses a pruning algorithm to control the number of regions. Willenbrock and Fridlyand (2005) and Lai *et al*. (2005) showed that DNACopy algorithm performs well in terms of sensitivity and false discovery rate both on synthetic and real data. Other statistical models frequently used for CN segmentation are Bayesian approaches and Hidden Markov models. In the Bayesian framework, the prior distributions are combined with some posterior distribution functions to construct the most plausible hypothesis concerning the data segmentation (Daruwala *et al*., 2004; Pique-Regi *et al*., 2008). In Hidden Markov model-based approaches, hidden states represent the underlying CN of probes. In Fridlyand *et al*.

*To whom correspondence should be addressed.

(2004), the model is characterized in terms of three parameters: the initial state probability, the transition probability and the collection of Gaussian emission probability functions defined within each state. SMAP (Andersson *et al*., 2008) is a recent approach based on the discrete-index hidden Markov model where a maximum *a posteriori* approach is used to split the chromosome into regions. The authors adapt the maximum *a posteriori* approach so that user-defined priori informations can be integrated within their model for limiting noise influence (modeled by a Gaussian distribution).

In recent literature, variational-based approaches are emerging to deal with the CN segmentation problem (Nilsson *et al*., 2008, 2009). The works of Mumford and Shah (1989) and Rudin *et al*. (1992) are the pioneers of discontinuity-adaptive variational models which have been successfully applied in a wide variety of problems. The original model, based on the Total Variation norm (Giusti, 1984), was proposed to recover image corrupted by noise preserving important image features such as object edges (Ceccarelli, 2007a). Afterwards, discontinuity-adaptive variational models have been applied in many different research areas, such as texture segmentation (Vese *et al*., 2002), medical image analysis (Ceccarelli *et al*., 2007b) and shape identification in 'synthetic-aperture radar' imagery (Redding *et al*., 1999). Variational models are based on the minimization of a functional controlling the similarity between the computed segmentation and the observed image, penalizing at the same time complex solutions. The complexity of the solution is controlled by the *scale parameter* (also called *regularization parameter*). As the regularization parameter increases less regions are computed, therefore the choice of a good regularization parameter is a common open question in many variational data analysis algorithms.

Discontinuity-adaptive variational models are very skillful in segmentation of piecewise constant (PWC) images. In PWC images, the pixels belonging to the same object have the same intensity, but noise changes image features and the segmentation task becomes more difficult. This situation is very similar to aCGH data, for which segmentation errors are due to noise that shifts LRR values. A recent work (Nilsson *et al*., 2009) presents a CN segmentation algorithm based on the total variation minimization process proposed by Rudin *et al*. (1992).

Here, we propose a new segmentation algorithm (VEGA) based on the Mumford and Shah variational model (Mumford and Shah, 1989). The PWC assumption is used to define a functional considering both accuracy and parsimony of the boundaries. The segmentation problem is put as a minimization problem of an energy functional encompassing both the quality of interpolation of the data by a piecewise constant function and the 'complexity' of the solution measured by the length of the boundaries between segmented regions. It is well known that the resulting energy functional is non-convex and can have many local minima. In order to efficiently compute a solution here we adopt a greedy steepest-descent algorithm based on a pyramidal multiscale approach. The resulting algorithm belongs to the class of region growing segmentation algorithms similar to that proposed in (Koepfler *et al*., 1994). In addition, we propose a data-driven heuristics for the computation of a suitable regularization parameter. The use of a variational segmentation approach for CN variation estimation has also been proposed in (Nilsson *et al*., 2009), but there are some significant differences with the method proposed here. First, VEGA is based on the Mumford and Shah model while Ultrasome (Nilsson *et al*., 2008, 2009) adopts the Rudin's model (Rudin *et al*., 1992). Moreover, VEGA performs the minimization of the energy functional with a bottom-up approach, by a sequence of successive merging of smaller regions into larger ones. This leads to a greedy multiscale algorithm driven by an increasing series of values of the regularization parameter similar to the image segmentation approach proposed by Koepfler *et al*. (1994). Whereas, in (Nilsson *et al*., 2008) a dynamic programming approach is used, by choosing a fixed value of the regularization parameter.

In order to validate our approach, we choose as comparison methods DNACopy (a likelihood function-based approach whose good performance have been demonstrated), SMAP (a recent statistical model-based approach) and Ultrasome (that uses a variational model as in VEGA). Results are compared both on synthetic and on real biological data. Both SMAP (version 1.12.0) and DNACopy (version 1.16.0) are available as Bioconductor R packages, while Ultrasome (version 2.0) is available in a command line version (for Windows and Linux) and a graphical user interface version (for Windows).

2 METHODS

2.1 Mumford and Shah model

The basic idea of the Mumford and Shah model (Mumford and Shah, 1989) is the so-called piecewise smooth model. Given an observed signal u_o defined on the domain Ω, we can model u_o by a partition of Ω into a set of disjoint connected components Ω_i, with

$$\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \Omega_n$$

in such a way that the signal u_0 varies smoothly within each Ω_i and it varies discontinuously across the boundaries between different Ω_i. The set of points on the boundary between the Ω_i is denoted as Γ. This means that the segmentation problem is put as a problem of optimal piecewise smooth approximation, i.e. we look for an approximation u of u_0 whose restrictions to the regions Ω_i are smooth. The search of the optimal piecewise smooth approximation of u_0 can be cast into the minimization of the following functional:

$$E(u,\Gamma) = \alpha \int_{\Omega} (u-u_o)^2 dx\, dy + \int_{\Omega\setminus\Gamma} |\bigtriangledown u|^2 dx\, dy + \lambda|\Gamma| \qquad (1)$$

where α and λ are two non-negative parameters weighting the different terms in the energy: the first term requires that u approximates u_o, the second term takes in account the variability of u within each connected component Ω_i and the third term penalizes complex solutions in terms of the length of the boundaries $|\Gamma|$. Smaller values of E are associated with better solutions (u,Γ) for the observed signal u_o. A special case of Equation (1) is obtained when the approximation u of the signal u_o is considered to be a piecewise constant function (u constant within each connected components Ω_i). For this case, Mumford and Shah (1989) proposed the so-called *piecewise constant Mumford-Shah model*:

$$E(u,\Gamma) = \sum_i \int_{\Omega_i} (u_o-u_i)^2 dx\, dy + \lambda|\Gamma| \qquad (2)$$

It is easy to show that, given a fixed set of boundaries Γ, in order to have a minimum the variables u_i should be set as the mean of u_o within of each connected component Ω_i.

In this work, we adopt the Mumford and Shah model defined in (2) for segmenting CN data. The role of the two terms of (2) is important, the first term can be considered as the error in the approximation of u by a constant function within each region and the second term as a penalty to complex segmentations consisting of many small regions with irregular boundaries. Therefore, this kind of functional is a compromise between the accuracy of the approximation within each region and parsimony of the boundaries.

The resulting segmentation depends on the scale parameter λ, indeed it determines the amount of regions of the computed segmentation: when λ is small many boundaries are allowed so the resulting segmentation will be *fine*, while as λ increases the segmentation will be coarser and coarser.

2.2 The proposed approach

2.2.1 The model Let $D \in \mathbb{R}^n$ the data vector containing n LRR probes of a chromosome where the observations are ordered by the respective genomic position. We define a segmentation S of D as a set of ordered positions (breakpoints) $b_1,\ldots,b_{M+1}$ partitioning D into M connected regions $\mathbf{R}=\{R_1,\ldots,R_M\}$. The region R_i is identified by the indexes in $[b_i,b_{i+1})$ with $i=1,\ldots,M$ and where the breakpoints b_1 and b_{M+1} are fixed to the values 1 and $n+1$, respectively. We use the one-dimensional version of the piecewise constant Mumford and Shah functional, in this case the length of the boundaries between regions has no influence on the segmentation, and the second term of (2) reduces to the number of regions, denoted here as M.

$$E(u,\Gamma)=\sum_i \int_{\Omega_i} (u_o-u_i)^2 dx\, dy+\lambda M \tag{3}$$

Given the n-dimensional data D, it is easy to show that the optimal segmentation must be chosen among the 2^n possible solutions. In genomic data, we have a resolution that provides tens of thousands of observations, so brute force algorithms cannot be applied and suitable solutions must be found by using heuristic strategies. Here, we use a greedy procedure as explained below.

2.2.2 Minimization process The minimization of (3) is carried out by a region growing process with small regions progressively merged to create larger ones leading to a pyramidal algorithm going from finer segmentations to coarser segmentations. Given two adjacent regions R_i and R_j, it can be shown that the reduction of the energy (3) after the merging of these two regions is given by:

$$E(u,\Gamma\setminus R_i\cup R_j)-E(u,\Gamma))=\frac{|R_i|\,|R_j|}{|R_i|+|R_j|}||u_i-u_j||^2-\lambda \tag{4}$$

where $R_i,R_j\in\mathbf{R}$ with $i\neq j$, $|R_i|$ and u_i are the length and LRR mean value of the i-th region, respectively. $||\cdot||$ represents the L_2 norm. Following a greedy procedure, we start with a segmentation having n regions each for each LRR measure, then at each step we choose as the next pair of regions to be merged that producing the maximum decrease of the energy functional. For a fixed value of λ, the algorithm iteratively computes the pair of adjacent regions for which (4) is as negative as possible, if such pair of regions exists. If no such pair of regions exists, then the value of λ is increased. The resulting method is therefore considered a multiscale algorithm Koepfler *et al*. (1994) since the value of λ represents the *scale*, as λ grows the segmentation gets coarser. The algorithm stops when the maximum value of the scale parameter is reached. The sequence of values of this parameter is called λ-schedule by analogy with temperature schedule of simulated annealing.

2.2.3 λ-schedule selection The λ values determine the quality of the final segmentation so the choice of the λ-schedule is very important for final segmentation. Here, we propose a dynamic λ-schedule selection where the λ stop value is computed considering both the sequences of λ values and the data variability. We modify the *Full λ-Schedule Segmentation* approach proposed by Redding *et al*. (1999) so that it can work on one-dimensional spaces. In particular from Equation (4) we associate to each breakpoint b_i with $i=2,\ldots,M$ (b_1 and b_{M+1} breakpoints are fixed) the value:

$$\hat{\lambda}_i=\frac{|R_{i-1}|\,|R_i|}{|R_{i-1}|+|R_i|}||u_{i-1}-u_i||^2 \tag{5}$$

Note that $\hat{\lambda}_i$ represents the cost required for merging the regions R_{i-1} and R_i. The i-th breakpoint can be deleted (and the adjacent region are merged) if $\hat{\lambda}_i<\lambda$. If more regions remain and no $\hat{\lambda}_i$ agrees the previous inequality, then there is no merging that decreases the functional and a new scale parameter, λ, must be chosen. Here, we select the next scale parameter value as the smallest $\hat{\lambda}_i$ and adding to this value a positive constant ϵ close to zero. By using this update rule, new region merges are allowed, so the region growing process will be composed by fine merging operations.

2.2.4 Optimal λ selection A critical point for many variational approaches is the selection of the stop condition, i.e. the selection of the last value of λ for which the output solution is obtained. Different values of the regularization parameter can be associated to suitable segmentations. In order to provide a stopping criterion, we use a modified version of the minimum variance method which was first studied by Otsu (1979) in image segmentation. Otsu (1979) proposed a stopping criterion based on the maximization of the ratio $\sigma^2(B)/\sigma^2(T)$ where $\sigma^2(B)$ and $\sigma^2(T)$ are the between class-variance and the total variance, respectively. $\sigma^2(B)$ measures the variability between different segmented regions and it can be calculated considering only the adjacent regions.

Our idea starts from the consideration that the $\hat{\lambda}_i$ measures the degree of compatibility between two adjacent regions, so we can use $\hat{\lambda}_i$ to estimate the variability between adjacent regions (similarly to the between class-variance). Therefore, measuring the distance between two consecutive λ-schedule values ($\triangle\lambda=\lambda_{l+1}-\lambda_l$), we can obtain information on the consistency of the segmentation. In particular, small $\triangle\lambda$ values indicate a merging of compatible regions, in contrast the more $\triangle\lambda$ grows the more we deviate from the compatibility between adjacent regions. In order to take also in account the total variability of the data we use the standard deviation (SD), ν, computed chromosome by chromosome, and the proposed stop criterion is:

$$\triangle\lambda=\lambda_{l+1}-\lambda_l\leq\beta\nu \tag{6}$$

where β is a positive constant. We tested different values for β and we obtained the best performance by using $\beta=0.5$. In absence of further prior information, this value allows to obtain segmentation results consistent to the data taking also in account the complexity of the solution. All results reported in this article were obtained by using $\beta=0.5$. The detailed VEGA algorithm is summarized in pseudocode in the Supplementary Material.

2.2.5 Computational complexity If we have n probes then the starting segmentation will have n regions corresponding with $n+1$ breakpoints that have to be maintained in order to efficiently extract the minimum at step 8 of the Algorithm 1 (Supplementary Material). Here, we use a priority queue based on a heap data structure (Cormen *et al*., 2009). Therefore, step 4 requires $O(n)$, step 8 requires $O(1)$, whereas steps 11 and 12 require $O(\log n)$ each. Considering that the cycle in steps 7–18 is repeated at maximum n times, the complexity of the algorithm VEGA is $O(n\log n)$.

2.2.6 Region labeling The region growing process described above produces a segmentation S composed by M regions $\mathbf{R}=\{R1,\ldots,R_M\}$ each represented by the value u_i, the mean of the observations contained in the i-th region (with $i=1,\ldots,M$). In order to assign a label to each region indicating if this region correspond to a normal, loss or gain aberration we use the rule proposed by Willenbrock and Fridlyand (2005): a region R_i is labeled as loss if $\mu_i<-0.2$, while it is labeled as a gain if $\mu_i>0.2$ and for values $-0.2\leq\mu_i\leq0.2$ the region is considered normal.

3 RESULTS

In order to evaluate the performance of our approach, we use both synthetic and real data. We compare our results with the ones obtained by DNACopy (Olshen *et al*., 2004), Ultrasome (Nilsson *et al*., 2009) and SMAP (Andersson *et al*., 2008). Among the considered methods just SMAP provides a label assignment for the segmented regions, while Ultrasome and DNACopy (similar to VEGA) provide the LRR mean value of each region. Therefore, when the quantitative analysis requires for each region a discrete

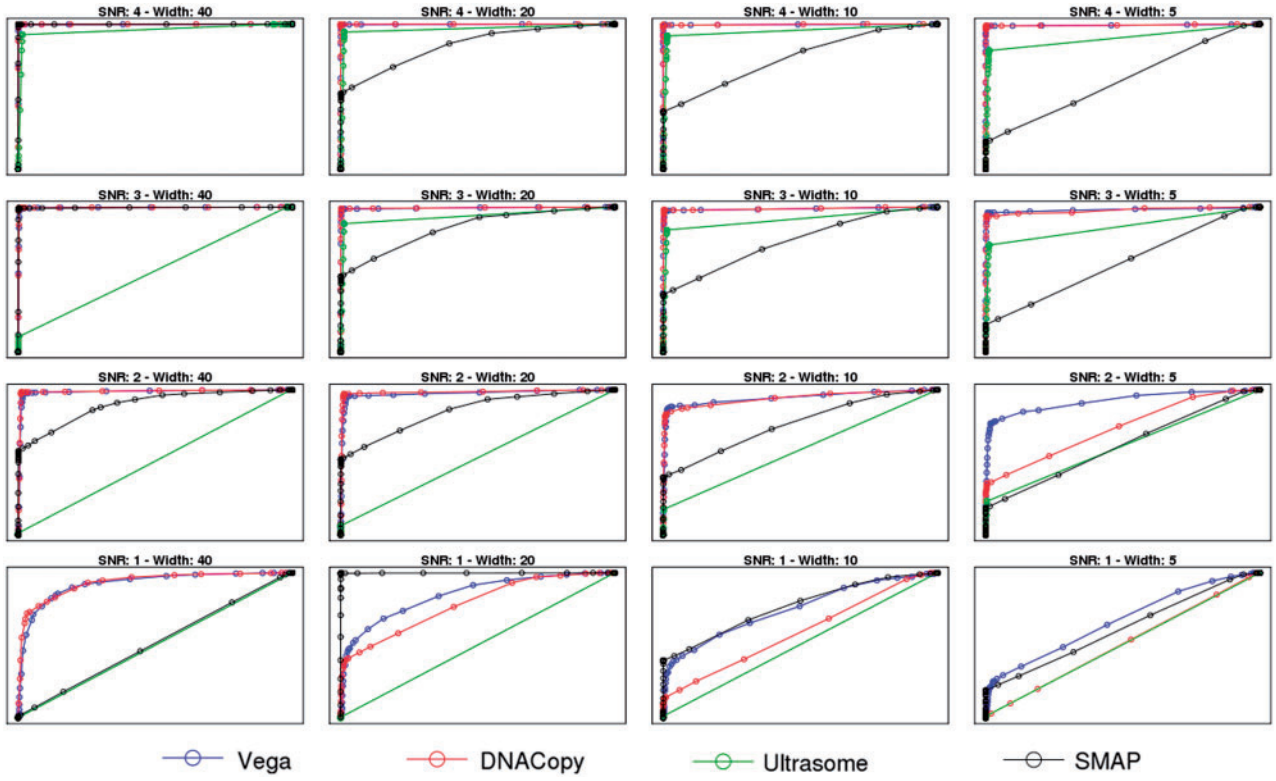


Fig. 1. ROC curves at different aberration widths and SNR. The *x*-axis is the FPR and the *y*-axis is the TPR. The curves were generated by measuring the TPR and FPR on the simulated data published in Lai *et al*. (2005). Blue is VEGA, red is DNACopy (Olshen *et al*., 2004), green is Ultrasome (Nilsson *et al*., 2009) and black is SMAP (Andersson *et al*., 2008).

label (loss, normal or gain), we apply to Ultrasome and DNACopy the same label assignment rules used in our approach (see Section 2.2.6), while when the fitted LRR in each region is required we use for SMAP the mean value of all LRRs contained in the region. Algorithms ran with their default parameters, except for SMAP for which parameters were chosen by the user.

3.1 Results on synthetic data

3.1.1 Evaluation metric For a quantitative evaluation of the performance we used the Precision, the Recall and their harmonic mean (*F*-measure) which are three widely used statistical classifications. The Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness. Both measures are defined by using the concepts of true positive (TP), false positive (FP) and false negative (FN). A TP represents a perfect correspondence between the computed label and the true ground, a FP occurs when the algorithm detects a mutation that is not present in the true ground and we have a FN evaluation when a mutation in the true ground is not detected by the algorithm. Algorithm accuracy has also been reported by calculating the receiver operating characteristic (ROC) curve as described in Lai *et al*. (2005) where the true positive rate (TPR) was defined as the number of probes inside the aberration whose fitted values are above the threshold level divided by the number of probes in the aberration and the false positive rate (FPR) was defined as the number of probes outside the aberration whose fitted values are above the threshold level divided by the total number of probes outside the aberration.

3.1.2 Validation on Lai et al. synthetic data Given that several different segmentation methods have been published in literature, this makes it difficult to have a fair comparison between all of them. Here, as in Magi *et al*. (2010), we address this problem by using already available synthetic data previously published in Lai *et al*. (2005). About a dozen methods have been benchmarked using this dataset, and we use it to evaluate the performance of VEGA. Moreover, Lai *et al*. (2005) found DNACopy as the method performing consistently well on both synthetic and real data, so we include the performance of DNACopy in our comparison.

The synthetic dataset proposed by Lai *et al*. (2005) simulates four aberration widths (5, 10, 20 and 40) in different noise conditions (signal-to-noise ratio SNR of 1, 2, 3 and 4). For each aberration width and SNR, 100 artificial chromosomes of 100 probes were generated. As shown in Figure 1, in high SNR conditions (SNR values of 3 and 4) VEGA results are very similar to the ones of DNACopy, while for low SNRs and small aberration widths (lower right panels) VEGA appears to perform better than DNACopy. By analyzing SMAP's performance we can notice how it is strongly influenced by the aberration width and for SNR of 1 its performance is not simple to be interpreted. Eventually this may be due to the used parameter setting. For Ultrasome, we can notice that although it has acceptable performance for high SNR levels (4 and 3), it has some problems in low SNR scenarios.

Table 1. Results on synthetic data for the considered approaches

σ	VEGA			Ultrasome			DNACopy			SMAP		
	Precision	**Recall**	***F*-measure**	**Precision**	**Recall**	***F*-measure**	**Precision**	**Recall**	***F*-measure**	**Precision**	**Recall**	***F*-measure**
0.00	**1.000**	**1.000**	**1.000**	0.937	0.926	0.931	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
0.10	**1.000**	**1.000**	**1.000**	0.938	0.926	0.932	0.999	**1.000**	0.999	**1.000**	**1.000**	**1.000**
0.20	0.998	**0.999**	**0.999**	0.938	0.920	0.929	0.996	**0.999**	0.997	**0.999**	**0.999**	**0.999**
0.30	0.968	0.978	0.973	0.933	0.898	0.915	0.960	0.977	0.968	**0.985**	**0.982**	**0.983**
0.40	0.876	0.928	0.902	0.912	0.852	0.881	0.860	**0.931**	0.894	**0.931**	0.892	**0.911**
0.50	0.825	**0.891**	0.856	**0.892**	0.824	**0.857**	0.789	0.881	0.832	0.869	0.771	0.817
0.60	0.679	0.828	0.746	**0.819**	0.748	**0.782**	0.643	0.780	0.705	0.516	**0.900**	0.656
0.70	0.586	0.804	0.678	**0.762**	0.681	**0.719**	0.581	0.730	0.647	0.306	**0.985**	0.467
0.80	0.556	0.792	0.653	**0.728**	0.669	**0.697**	0.545	0.711	0.617	0.287	**1.000**	0.446
0.90	0.487	0.787	0.601	**0.654**	0.623	**0.638**	0.493	0.644	0.558	0.270	**0.998**	0.425
1.00	0.440	0.793	0.566	**0.607**	0.578	**0.592**	0.460	0.521	0.489	0.275	**0.991**	0.430
Mean	0.765	0.891	**0.816**	**0.829**	0.786	0.807	0.757	0.834	0.791	0.676	**0.956**	0.739

σ represents the SD of the white Gaussian noise. Bold represents the best performance for each noise level.

3.1.3 Validation on the proposed synthetic data Although the dataset proposed by Lai *et al*. (2005) allows an extensive comparison of segmentation algorithms it is far from real aCGH data where, for each chromosome, thousands of probes are observed and multiple gain and loss mutations are expected. In order to overcome these limitations we created an artificial dataset simulating the profiles of three chromosomes having different size of 500, 750 and 1000 probes (50 profiles for each chromosome were generated). In each sample, a set of mutations were randomly inserted with size ranging from 11 to 25 where both position and class (loss or gain) of each aberration was randomly chosen. In particular, 10 mutations (with size ranging from 11 to 20) were inserted within the data with 500 observations and 15 aberrations (with size ranging from 11 to 25) were inserted within the samples having 750 and 1000 probes. The model used for the generation of the synthetic profiles follows:

$$d_m = x_m + \varepsilon_m \tag{7}$$

where d_m is the observed LRR for the probe m which is the sum of two components, the first one is the real LRR x_m for this probe and the second one, ε_m is the noise component corrupting the data. The true component x_m can assume values $log_2(\frac{1}{2}) = -1$, $log_2(\frac{2}{2}) = 0$ and $log_2(\frac{3}{2}) = 0.58$ for loss, normal and gain, respectively. The noise component is modeled as a white Gaussian process $\varepsilon_m \sim \mathcal{N}(0, \sigma)$. Performance assessment is made by varying for each sample the value of σ from 0 to 1 with step 0.1.

In Table 1, Recall, Precision and *F*-measure of each considered method are reported with respect to the noise SD σ.

Observing the results obtained for the different chromosome sizes (Supplementary Figs S2–S4), we note that the number of probes does not affect the performance of the compared algorithms. Analysing the trends of the considered approaches with respect to the noise level σ (Table 1 and Fig. S2), we note that SMAP works well until the value of σ is not greater than 0.5, after this value its performance tends to drop down. Also DNACopy's performance is negatively affected by the noise, but the trend is less steep than SMAP. In contrast, both VEGA and Ultrasome results are not drastically affected by the values of σ but Ultrasome does not compute the correct segmentation in no-noise condition ($\sigma = 0$), this is caused by the number FN and FP that produce a Precision of 0.937 and a Recall of 0.926. This first analysis demonstrates that variational approaches tend to be more robust to noise than the other considered ones. These considerations are also supported from the ROC curves and the corresponding area under the curves (AUCs) reported in Supplementary Figures S4–S11 where in addition we can see that these algorithms are more skillful at detecting losses than gains.

Figure 2 shows the results of the compared algorithms on a simulated chromosome profile of 1000 probes for σ noise values of 0 (A) , 0.5 (B) and 1 (C). In these Figures, black lines represent the true profiles and the red lines show the profiles computed by the algorithms. Comparing computed and true profiles we can note that in absence of noise ($\sigma = 0$), all algorithms provide the correct segmentation, while for $\sigma = 0.5$ VEGA and SMAP compute more consistent profiles. Finally in high noise condition ($\sigma = 1$), only VEGA and Ultrasome seem to provide an acceptable segmentation.

3.2 Results on real data

3.2.1 Glioblastoma Multiforme (GBM) data GBM is a particular malignant and aggressive type of brain tumor. Several chromosomal mutations have been observed in glioma as gain on chromosome 7 and losses of chromosomes 10, 13 and 22. We used two samples (GBM31 and GBM29) representing primary GBMs in the glioma data from Bredel *et al*. (2005) which were used in Lai *et al*. (2005). In sample GBM31, chromosome 13 is characterized by a large region of loss (Fig. 3A). Results show that VEGA, DNACopy and SMAP identified the expected deletion while Ultrasome did not detect this mutation. For the chromosome 7 of the sample GBM29, three small amplifications with high amplitude around EGFR locus are expected (Fig. 3B). Among the compared algorithms, SMAP is the only one that did not find the gains, DNACopy combined the first two amplifications together while both VEGA and Ultrasome detected all three mutations.

3.2.2 Mantle Cell Lymphoma (MCL) data MCL is an aggressive lymphoma with median patient survival times of $\sim$3 years. We used eight MCL cell lines (Granta-519, HBL-2, JVM-2, NCEB-1, REC-1, SP49, UPN-1 and Z138C) previously analyzed by DeLeeuw *et al*. (2004). This analysis was performed by SMRT aCGH containing 97299 elements, representing 32433 overlapping genomic segments spanning the entire human genome (Ishkanian *et al*., 2004). DeLeeuw *et al*. (2004) provided a comprehensive list of cytobands with the respective biological references for these cell lines. In

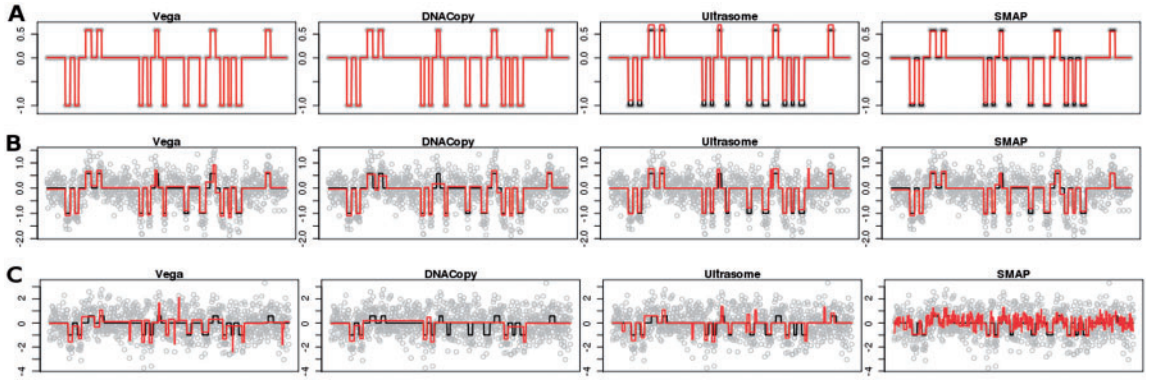


Fig. 2. Segmentation results on a simulated chromosome profile of 1000 probes. Red line reports the segmentation computed by the algorithm, black line is the true ground. In **(A)** the results on the simulated data without the noise component. In **(B)** the results on the simulated data with Gaussian noise, zero mean and SD equal to 0.5. In **(C)** the results on the simulated data with Gaussian noise, zero mean and SD equal to 1.

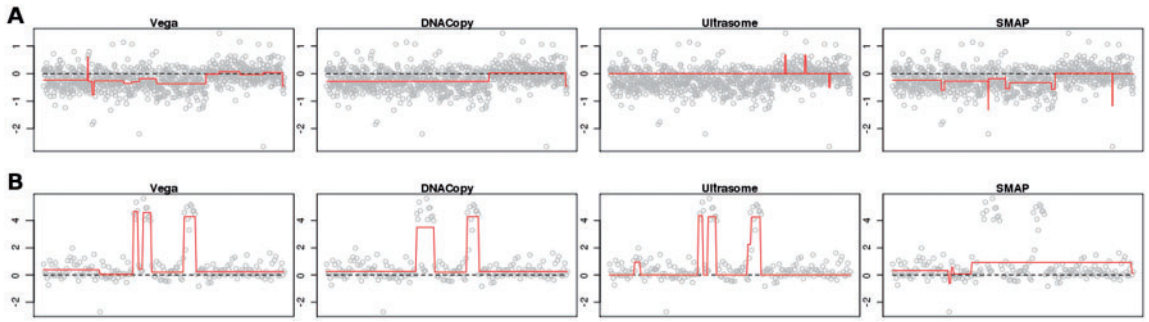


Fig. 3. Segmentation results on Glioblastoma Multiforme (GBM) data published in Bredel *et al*. (2005). In **(A)** the results on the chromosome 13 of the sample GBM31 are shown. This chromosome has a partial loss of low magnitude which is detected by VEGA, DNACopy and SMAP, only Ultrasome does not detect this mutation. In **(B)** the results around EGFR (chromosome 7) of the sample GBM29 are shown. This chromosomal region has three amplifications which are properly detected by VEGA, DNACopy and Ultrasome, only SMAP does not segment these aberrations.

addition, they confirmed the deletion of 13q14.3 cytoband in Granta-519 by using fluorescence in situ hybridization (FISH) analysis. This list was used as our 'ground truth' to assess the performance of VEGA, Ultrasome, DNACopy and SMAP (Fig. 4).

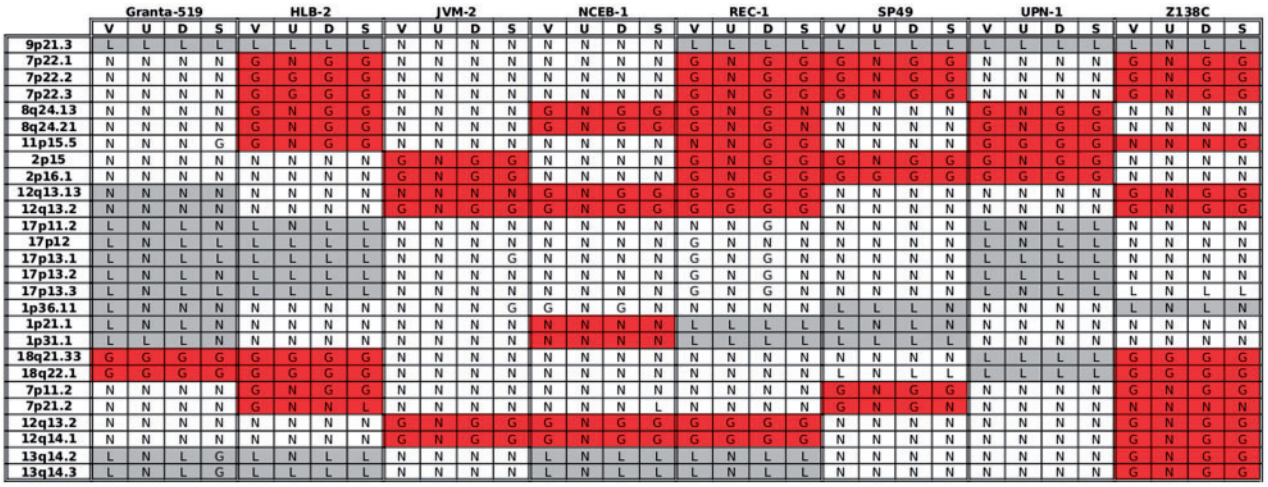
By using this 'ground truth' we can see that Ultrasome did not identify some interesting mutations such as the gain of 8q24.13-8q24.21 [this cytoband contains the well-characterized *MYC* oncogene that is reported to be overexpressed in MCL (Hofmann *et al*., 2001)] and the loss of 13q14.2-13q14.3 which has been validated for Granta-519 by using FISH analysis. Moreover, for this last chromosomal region SMAP found a gain which is in contradiction with respect to FISH analysis. Loss of 9p21.3 was confirmed by all approaches and only Ultrasome considered this region as normal in Z138. In Granta-519, the loss of 1p36.11 was detected only by VEGA, while SMAP did not find a loss for this cytoband for Granta-519, SP49 and Z138. By analyzing Granta-519, we can see that VEGA and DNACopy agreed with DeLeeuw's analysis for all reported cytobands except for the mutation in 12q13.13-12q13.2 for which no method provided a loss. In contrast, both Ultrasome and SMAP had some problems in region detection for this cell line.

As a further a qualitative analysis of VEGA, Ultrasome, DNACopy and SMAP approaches, we analyzed the chromosome 8 of NCEB-1 for which a gain covering the whole long branch is expected (DeLeeuw *et al*., 2004; Martinez-Climent *et al*., 2001; Rinaldi *et al*., 2005). In Supplementary Figure S12, the segmented regions for this chromosome are shown, and this figure shows that Ultrasome detected very few and small aberrant regions on the long branch of this chromosome, SMAP provided a very 'jagged' segmentation, while VEGA and DNACopy computed the expected amplification, but DNACopy detected an unexpected normal region.

Supplementary Table S1 shows the number of aberrant regions obtained by the different approaches. From this table, we note that SMAP computed the maximum amount of regions (97.06), on the contrary Ultrasome provided a mean of 25 aberrant regions for cell line, in the middle we find VEGA and DNACopy algorithms with a similar mean of aberrant regions (85.94 and 86.75, respectively).

4 DISCUSSION AND CONCLUSIONS

In this article, we present VEGA: a new approach for DNA copy number segmentation. VEGA uses a variational-based method for image segmentation, known as Mumford and Shah model. This model is modified so that it can be efficiently applied in CN segmentation. A greedy multiscale region growing process is used to find the solution. The region growing process follows a λ-schedule selection that provides an efficient scheme to control the quality of



	Granta-519				HLB-2				JVM-2				NCEB-1				REC-1				SP49				UPN-1				Z138C			
	V	U	D	S	V	U	D	S	V	U	D	S	V	U	D	S	V	U	D	S	V	U	D	S	V	U	D	S	V	U	D	S
9p21.3	L	L	L	L	L	L	L	L	N	N	N	N	N	N	N	N	L	L	L	L	L	L	L	L	L	L	L	L	L	N	L	L
7p22.1	N	N	N	N	G	N	G	G	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	N	N	N	N	G	N	G	G
7p22.2	N	N	N	N	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	N	N	N	N	G	N	G	G
7p22.3	N	N	N	N	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	N	N	N	N	G	N	G	G
8q24.13	N	N	N	N	G	N	G	G	N	N	N	N	G	N	G	G	G	N	G	N	N	N	N	N	G	N	G	G	N	N	N	N
8q24.21	N	N	N	N	G	N	G	G	N	N	N	N	G	N	G	G	G	N	G	N	N	N	N	N	G	N	G	G	N	N	N	N
11p15.5	N	N	N	G	G	N	G	G	N	N	N	N	N	N	N	N	N	N	G	G	N	N	N	N	G	G	G	G	N	N	N	G
2p15	N	N	N	N	N	N	N	N	G	N	G	G	N	N	N	N	G	N	G	G	G	N	G	G	G	N	G	G	N	N	N	N
2p16.1	N	N	N	N	N	N	N	N	G	N	G	G	N	N	N	N	G	N	G	G	G	G	G	G	G	G	G	G	N	N	N	N
12q13.13	N	N	N	N	N	N	N	N	N	N	N	N	G	N	G	G	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G
12q13.2	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G
17p11.2	L	N	L	N	L	N	L	L	N	N	N	N	N	N	N	N	N	N	G	N	N	N	N	N	L	N	L	L	N	N	N	N
17p12	L	N	L	L	L	L	L	L	N	N	N	N	N	N	N	N	G	N	N	N	N	N	N	N	L	N	L	L	N	N	N	N
17p13.1	L	N	L	L	L	L	L	L	N	N	N	G	N	N	N	N	G	N	G	N	N	N	N	N	L	L	L	L	N	N	N	N
17p13.2	L	N	L	N	L	L	L	L	N	N	N	N	N	N	N	N	G	N	G	N	N	N	N	N	L	L	L	L	N	N	N	N
17p13.3	L	N	L	L	L	L	L	L	N	N	N	N	N	N	N	N	G	N	G	N	N	N	N	N	L	N	L	L	L	N	L	L
1p36.11	L	N	N	N	N	N	N	N	N	N	N	G	G	N	G	N	N	N	N	N	L	L	L	N	N	N	N	N	L	N	L	N
1p21.1	L	N	L	N	N	N	N	N	N	N	N	N	N	N	N	N	L	L	L	L	L	N	L	N	N	N	N	N	N	N	N	N
1p31.1	L	L	L	N	N	N	N	N	N	N	N	N	N	N	N	N	L	L	L	L	L	L	L	L	N	N	N	N	N	N	N	N
18q21.33	G	G	G	G	G	G	G	G	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	L	L	L	L	G	G	G	G
18q22.1	G	G	G	G	G	G	G	G	N	N	N	N	N	N	N	N	N	N	N	N	L	N	L	L	L	L	L	L	G	G	G	G
7p11.2	N	N	N	N	G	N	G	G	N	N	N	N	N	N	N	N	N	N	N	N	G	N	G	G	N	N	N	N	G	N	G	G
7p21.2	N	N	N	N	G	N	N	L	N	N	N	N	N	N	N	L	N	N	N	N	G	N	G	N	N	N	N	N	N	N	N	N
12q13.2	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G
12q14.1	N	N	N	N	N	N	N	N	G	N	G	G	G	N	G	G	G	G	G	G	N	N	N	N	N	N	N	N	G	N	G	G
13q14.2	L	N	L	G	L	N	L	L	N	N	N	N	L	N	L	L	L	N	L	L	N	N	N	N	N	N	N	N	G	N	G	G
13q14.3	L	N	L	G	L	L	L	L	N	N	N	N	L	N	L	L	L	N	L	L	N	N	N	N	N	N	N	N	G	N	G	G

Fig. 4. This figure shows the cytobands reported in DeLeeuw *et al*. (2004) with the respective mutations: grey and red boxes indicate loss and gain, respectively. For each cytoband the results of VEGA (V), Ultrasome (U), DNACopy (D) and SMAP (S) are reported. L indicates loss, N indicates normal and G indicates gain.

the segmentation. The optimal scale parameter value is computed in an automatic way considering both the λ-schedule and the data variability. VEGA algorithm works separately on each chromosome and it has computational complexity of $O(nlogn)$ where n is the number of observed probes for the considered chromosome.

We compare VEGA with three approaches that can be considered the state of the art in CN segmentation: DNACopy (a likelihood function-based approach), SMAP (an algorithm based on discrete hidden Markov model) and Ultrasome (an algorithm that as VEGA uses a variational-based model). Results on synthetic data show the expected robustness to noise of variational-based approaches (VEGA and Ultrasome), and the performance of VEGA in all noise conditions. DNACopy and SMAP have similar performance for low noise levels but their performance may be significantly affected by noise. In order to assess the performance of the compared approaches on real data, we use eight MCL cell lines and two samples of glioblastoma multiforme. The 'true ground' for this data is composed by a list of cytobands which are extracted from current biological knowledge. Results on real data show the ability of VEGA in segmentation of well-characterized mutations and they also confirm the properties of the DNACopy algorithm. In contrast, real data analysis reveals some limitations of SMAP and Ultrasome. In addition, both synthetic and real results show important differences between VEGA and Ultrasome performance. These differences are mainly due to: the underlying model (VEGA uses Mumford and Shah model while Ultrasome uses Rudin model), the minimization process (VEGA uses a region growing process while Ultrasome uses a top-down approach) and the selection of the scale parameters used in our segmentation process. We think that SMAP and Ultrasome performance can be improved by a tuning of their input parameters. The problem is that these algorithms require a significant number of arguments and the tuning process can be hard for non-expert users.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their constructive comments.

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

Conflict of Interest: none declared.

REFERENCES

Andersson,R. *et al*. (2008) A segmental maximum a posteriori approach to genome-wide copy number profiling. *Bioinformatics*, **24**, 751–758.

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

Bredel,M. *et al*. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res.*, **65**, 4088–4096.

Ceccarelli,M. (2007a) A finite Markov random field approach to fast edge-preserving image recovery. *Image Vis. Comput.*, **25**, 792–804.

Ceccarelli,M. *et al*. (2007b) Automatic measurement of the intima-media thickness with active contour based image segmentation. In *IEEE International Workshop on Medical Measurement and Applications. MEMEA '07*, IEEE, Washington, DC, pp. 321–331.

Cormen,T.H. *et al*. (2009) *Introduction to Algorithms*, 3rd edn. MIT Press, Cambridge, MA.

Daruwala,R.S. *et al*. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl Acad. Sci.*, **101**, 16292–16297.

DeLeeuw,R.J. *et al* (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum. Mol. Genet.*, **13**, 1827–1837.

Fan,Y.S. *et al*. (2007) Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. *Hum. Mutat.*, **28**, 1124–1132.

Fridlyand,J. *et al*. (2004) Hidden Markov Models approach to the analysis of array CGH data. *J. Multivariate Anal.*, **90**, 132–153.

Giusti,E. (1984) *Minimal Surfaces and Functions of Bounded Variation*. Birkhauser, Basel, CH.

Harada,T. *et al*. (2008) Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene*, **27**, 1951–1960.

Hofmann,W.K. *et al*. (2001) Altered apoptosis pathways in mantle cell lymphoma detected by oligonucleotide microarray. *Blood*, **98**, 787–794.

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

Ishkanian,A.S. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genetics*, **36**, 299–303.

Jong,K. *et al*. (2003) Chromosomal breakpoint detection in human cancer. Vol. 2611 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 54–65.

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Khojasteh,M. *et al*. (2005) A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics,* **6**.

Koepfler,G. *et al*. (1994) A multiscale algorithm for image segmentation by variational method. *SIAM J. Numer. Anal.*, **31**, 282–299.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Magi,A. *et al*. (2010) A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics*, **11**, 265–280.

Martinez-Climent,J.A. *et al*. (2001) Loss of a novel tumor suppressor gene locus at chromosome 8p is associated with leukemic mantle cell lymphoma. *Blood*, **98**, 3479–3482.

Mumford,D. and Shah,J. (1989) Optimal Approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.*, **41**, 577–684.

Myers,C.L. *et al*. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Nilsson,B. *et al*. (2008) An improved method for detecting and delineating genomic regions with altered gene expression in cancer. *Genome Biol.*, **9**, R13.

Nilsson,B. *et al*. (2009) Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*, **25**, 1078–1079.

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

Otsu,N. (1979) A threshold selection method from gray-level histograms. *Sys. Man Cybern. IEEE Trans.*, **9**, 62–66.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**.

Pique-Regi,R. *et al*. (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, **24**, 309–318.

Redding,N.J. *et al*., (1999) An efficient algorithm for Mumford-Shah segmentation and its application to SAR imagery. In *Procedings Conference on Digital Image Computing Techniques and Applications (DICTA'99)*, pp. 35–41.

Rinaldi,A. *et al*. (2005) Genomic and expression profiling identifies the B-cell associated tyrosine kinase Syk as a possible therapeutic target in mantle cell lymphoma. *Br. J. Haematol.*, **132**, 303–316.

Rudin,L. *et al*. (1992) Nonlinear total variation based noise removal algorithms. *Physic. D*, **60**, 259–268.

Sebat,J. *et al*. (2007) Strong association of de novo copy number mutations with Autism. *Science*, **316**, 445–449.

Sen,A. and Srivastava,M.S. (1975) On tests for detecting change in mean. *Ann. Stat.*, **3**, 98–108.

Vese,L.A. *et al*., (2002) A multiphase level set framework for image segmentation using the Mumford and Shah Model. *Int. J. Comput. Vis.*, **50**, 271–293.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

Zhao,X. *et al*. (2003) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.*, **64**, 3060–3071.