BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4084–4091
doi:10.1093/bioinformatics/bti677

Genome analysis

A comparison study: applying segmentation to array CGH data for downstream analyses

Hanni Willenbrock[1] and Jane Fridlyand[2,*]

[1]Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark and [2]Department of Epidemiology and Biostatistics, University of California at San Francisco, 2340 Sutter Street, N224, San Francisco, CA 94143, USA

Received on July 5, 2005; revised on September 8, 2005; accepted on September 12, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: Array comparative genomic hybridization (CGH) allows detection and mapping of copy number of DNA segments. A challenge is to make inferences about the copy number structure of the genome. Several statistical methods have been proposed to determine genomic segments with different copy number levels. However, to date, no comprehensive comparison of various characteristics of these methods exists. Moreover, the segmentation results have not been utilized in downstream analyses.

Results: We describe a comparison of three popular and publicly available methods for the analysis of array CGH data and we demonstrate how segmentation results may be utilized in the downstream analyses such as testing and classification, yielding higher power and prediction accuracy. Since the methods operate on individual chromosomes, we also propose a novel procedure for merging segments across the genome, which results in an interpretable set of copy number levels, and thus facilitate identification of copy number alterations in each genome.

Availability: http://www.bioconductor.org

Contact: jfridlyand@cc.ucsf.edu

Supplementary Information: http://www.cbs.dtu.dk/~hanni/aCGH/

1 INTRODUCTION

Development of solid tumors is associated with acquisition of complex genetic alterations. The particular types of genomic derangement seen in tumors reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantages. Comparative genomic hybridization (CGH) is a technique by which it is possible to detect and map genetic changes that involve gain or loss of segments of genomic DNA. Microarray formats of CGH provide copy number information at thousands of locations distributed throughout the genome. For a review of existing array platforms see Pinkel and Albertson (2005).

Genomic profiles greatly vary in their complexity. Depending on the instability present in the tumor and the selection environment, tumor cells may acquire alterations ranging from large segments with single copy number alterations to narrow homozygous deletions or high level amplifications. In many tumors the magnitude of measurable changes is reduced because the cell population is heterogeneous, thus frequently containing a significant proportion of normal cells. For a given genomic profile, the initial computational step is commonly referred to as segmentation and it involves reliable identification of locations with copy number transitions or breakpoints. An example of how a genomic profile may look is illustrated in Figure 1(A and B). Downstream analyses involve classifying the samples and finding copy number alterations that are associated with known biological markers. Thus, additional opportunities arise in the analysis of array CGH data compared with the established analyses of gene expression microarrays. In particular, one has to make efficient use of the physical dependency of nearby clones.

Several segmentation methods have been proposed for partitioning clones into sets with the same copy number. Performances of a hidden Markov models (HMM) approach (Fridlyand *et al*., 2004), a non-parametric change-point method (DNAcopy) (Olshen *et al*., 2004) and a Gaussian model-based approach (GLAD) (Hupe *et al*., 2004) are compared in this article and these approaches are described in the Methods section in detail. Additional segmentation methods involve building hierarchical clustering-style trees along each chromosome (CLAC) (Wang *et al*., 2005), using a penalized likelihood criterion to estimate breakpoints (Picard *et al*., 2005) or applying an expectation–maximization-based method (Myers *et al*., 2004). Other proposals include a Bayesian model that uses parameterized prior distributions and a prior-less maximum a posteriori (MAP) technique to estimate the underlying model (Daruwala *et al*., 2004), a wavelet approach (Hsu *et al*., 2005) and use of a genetic local search algorithm to identify potential breakpoints and perform data smoothing (Jong *et al*., 2004).

To date, most proposed segmentation methods have been evaluated on a simple simulation model and/or a small set of karyotyped Coriell cell lines containing a limited spectrum of one-copy number alterations. Some approaches to simulate array CGH data were to randomly and uniformly select breakpoints throughout the genome (Daruwala *et al*., 2004); assign loss, normal or gain according to a fixed probability transition matrix (Hupe *et al*., 2004) or to draw lengths of segments from a theoretical distribution and then assign either normal or one-copy gain (Hsu *et al*., 2005). Some additional variations have been used to make the simulation resemble real data, e.g. adding a trend parameter (Olshen *et al*., 2004) or simply adding random Gaussian noise to karyotyped Coriell cell lines (Fridlyand *et al*., 2004). However, many of these simulations produce unrealistically simple array CGH data involving few copy number

*To whom correspondence should be addressed.

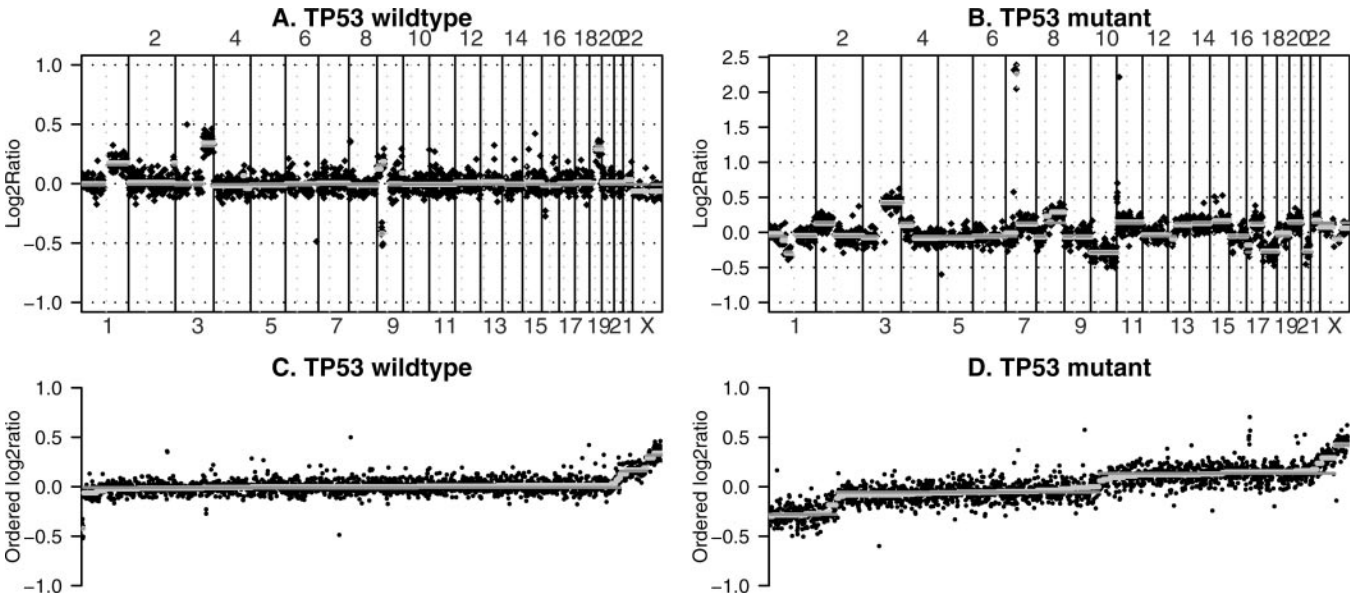

Fig. 1. (**A** and **B**) Genomic profiles for oral cancer samples segmented by DNAcopy and merged by MergeLevels. The observed $\log_2$-ratios are ordered according to their physical mapping along the genome. (**C** and **D**) Discretized $\log_2$-ratios by segmentation and merging. Log_2-ratios are sorted according to predicted $\log_2$-ratios. Observed $\log_2$-ratios are shown in black. Log_2-ratios predicted by DNAcopy are shown in light grey and $\log_2$-ratios following the application of MergeLevels are shown in dark grey. The merged profiles yield better interpretability.

changes. Moreover, until recently, no formal comparisons had been made among proposed algorithms except for Hsu *et al*. (2005) who compare their method with a previous method in terms of its breakpoint detection ability. A very recent paper (Lai *et al*., 2005) describes an extensive study that compares the ability of a large number of methods to assign copy number alterations. However, they did not specifically examine the behavior of aberrations at the boundaries and their simulation model does not lead to sufficiently complicated genomic profiles. With the explosion of interest in copy number microarrays and published computational approaches, there is a need for establishing a standard for systematic comparison of computational segmentation approaches. Here we create a simulation schema that generates genomic profiles of comparable complexity with real life data. This is achieved by resampling segments from a large set of primary tumors. We use the simulated data to compare three original published segmentation methods that were chosen on the basis of free access and ability to output appropriate and comparable segmentation information.

All available methods operate on individual chromosomes. Thus, as a result of segmentation, profiles are partitioned into numerous copy number levels with varying means. This presents a problem when identifying regions of gain or loss. It is currently done on a clone-by-clone basis either by considering normal range using normal–normal hybridizations (Veltman *et al*., 2003; Wang *et al*., 2005) or by estimating the level of experimental noise for a given profile and considering all clones with values outside x times standard deviation range to be altered (Hodgson *et al*., 2001; Nakao *et al*., 2004) where x is frequently set to 3. In this paper, we present a novel level-merging algorithm. The merging step does not compromise on the detection accuracy of the breakpoints and is indispensable as it allows us to identify a genomic base level, if present, and thereby easily assigns regions of copy number gain and loss to characterize individual genomes in terms of the number of copy number levels and to describe regions with respect to their relative copy number level.

Similarly, the physical positions of clones are ignored when identifying regions where the copy number is significantly associated with a phenotype of interest, e.g. a cancer subtype. A standard approach to the problem is to individually test each clone for the association on a 'clone-by-clone' basis. In this paper, we evaluate the benefits of segmenting data before performing downstream analyses and introduce a novel idea of segmenting test statistics to identify entire genomic regions of interest, facilitating the interpretation of results. Thus we compare the downstream analyses such as testing and classification using simulated and real datasets by applying clone-by-clone and region-based approaches.

This paper is organized as follows: in the Methods section, we provide details on the three methods under comparison and a novel level-merging algorithm. We also present novel approaches to incorporate segmentation into downstream analyses such as genome-wide testing and gain/loss detection. The simulation model and the primary tumor dataset are described in the Study Design section. In the Simulation Results section, we compare the ability of the three segmentation methods to detect breakpoints, identify altered regions and detect copy number associations with a phenotype of interest. In the Real Data Example section, we show a case study using a primary tumor array CGH dataset. Finally, in the Discussion and Conclusion section, we discuss the limitations of the study and future work.

2 METHODS

The methods to be compared are available for the R statistical language from Bioconductor (http://www.bioconductor.org/) and have copy number level assignments as their main output.

aCGH. This package contains a HMM-based method that assigns clones to underlying states with constant copy number, thus allowing for determination of breakpoints. It fits an unsupervised HMM in which any state is reachable from any other state. The state emission distributions are Gaussian with state-specific means and fixed variance. The re-estimation is done with a backward–forward algorithm. For a given number of states (k), the initialization is performed using k-means partitioning and transition probabilities are set to be proportional to the copy number distance between the pair of states. The number of states, k, is selected using a model selection criterion, e.g. Akaike information criterion (AIC) (Fridlyand *et al*., 2004). The HMM outputs two types of segmented values: predicted and smoothed $\log_2$-ratios, where the predicted values are state medians and smoothed values are state medians weighted by the estimated probability of being in each state. Here, we use aCGH version 1.1.4 and refer to the method as 'HMM'.

DNAcopy. This entirely non-parametric method is based on circular binary segmentation (CBS), which is a modification of a change-point approach allowing for tertiary splits by connecting the two chromosomal ends. It splits the chromosomes into contiguous regions of equal copy number by modeling discrete copy number gains and losses. Using a permutation reference distribution, it bypasses parametric modeling of the data for assessing significance of the proposed splits (Olshen *et al*., 2004). The model selection is done in the forward way by repeatedly splitting each contiguous segment until no significant splits are found. As predicted values, DNAcopy outputs mean $\log_2$-ratios of each predicted segment. Here, we use DNAcopy version 1.1.0 and we refer to the method as 'DNAcopy'.

GLAD. This Gaussian-based approach detects chromosomal breakpoints by estimating a piecewise constant function that is based on adaptive weights smoothing (AWS). A local constant Gaussian regression model $Y_i = \theta(X_i) + \varepsilon_i$ is considered where the ε_i are independently and identically distributed as $N(0, \sigma^2)$, and $\theta(X_i)$ is a piecewise constant function, where the disjoint regions and the total number of regions are unknown. AWS is based on local-likelihood modeling and is an iterative algorithm that, around every location X_i, finds the maximal possible neighborhood in which the θ parameter is constant (Hupe *et al*., 2004). GLAD contains a procedure for merging segmented levels by iteratively removing excessive breakpoints and subsequently cluster segments across chromosomes to assign levels of copy number gain and loss (Hupe *et al*., 2004). We use the median of the original $\log_2$-ratios of each initial predicted level as unmerged GLAD data; and the median of the original $\log_2$-ratios for each predicted cluster as the GLADmerge values. Since we used GLAD version 1.0.1, it was modified slightly to optimize its performance in our comparison study (see Supplementary information for details). We refer to this method as 'GLAD' and 'GLADmerge' for its level-merging procedure.

2.1 Merging of the copy number levels

As an alternative to model-based GLADmerge, which is not easily combined with other segmentation methods, we propose the following novel method (referred to as 'MergeLevels') for merging copy number levels across the genome. The method merges two segmented levels if the distributions of the $\log_2$-ratios of the clones mapped to those segments are not significantly different or if the predicted level values are closer than a dynamically determined threshold. The algorithm performs the following steps: (1) Order distances between predicted levels using copy number scale ($2^{\text{level value}}$), where level value is the predicted value of the segment. (2) Starting from the smallest distance, test whether two levels should be merged according to either of two criteria: (a) Wilcoxon rank sum test P-value >1e−04 between observed values for two states or (b) distance less than a given threshold. States with <3 clones in each may only be merged based on the threshold criterion (b). (3) After a successful merge, steps 1 and 2 are repeated until no two adjacent levels can be merged. (4) Steps 1–3 are repeated for increasing thresholds. (5) For each threshold, we use Ansari–Bradley 2-sample test (Bauer, 1972) to determine whether the distribution of the current residuals (current merged values minus observed $\log_2$-ratios) is significantly different from the distribution of the original residuals (original segmented values minus observed $\log_2$-ratios). (6) Optimal threshold is chosen as the largest threshold where the Ansari–Bradley P-value >0.05, i.e. where two types of residuals do not differ significantly. The Ansari–Bradley and Wilcoxon rank sum test significance thresholds were chosen based on an independent simulation dataset. See Supplementary information for details.

2.2 Using segmentation results for identifying regions of gain and loss, testing and classification

We test the application of segmentation followed by merging for identification of copy number alterations by defining the level of no alteration as the level with predicted $\log_2$-ratio closest to 0. Thus, all clones belonging to the remaining segments are either gained or lost. For comparison, we estimate experimental variability as the median absolute deviation (MAD) of difference between the observed and predicted $\log_2$-ratios and define threshold for determining gain and loss as three times MAD (factor of 2.5 is used in real data example).

We also introduce a region-based method for copy number association studies, which allows us to compute test statistics for entire regions rather than for individual clones. Student's t-test (equal variance) was used as a test statistic. For multiple testing corrections, we use a permutation-based single-step maxT procedure to control the family wise error rate (FWER) (Westfall and Young, 1993). Thus, the reference distribution was estimated by repeatedly permuting a phenotype with respect to the copy number data, re-computing relevant statistics and recording a permutation absolute maximum. A total of 100 permutations were used for simulation data and 1000 for primary tumor data. Adjusted P-values were derived by comparing an observed statistic with the distribution of the permutation maxima. The significance was declared at maxT adjusted P-value <0.05. Finally, we investigated whether using segmented values improved prediction accuracy for a phenotype predictor (e.g. *TP53* mutational status). For simplicity we used a linear discriminant analysis classifier with diagonal covariance matrix (DLDA) which has previously demonstrated very good performance in microarray studies (Dudoit *et al*., 2002). Performance was assessed using leave-one-out cross-validation for a varying number of input variables which were ranked by their F-statistic within each cross-validation.

3 STUDY DESIGN

3.1 Simulation model

The ratio profiles for array CGH data were simulated to emulate the complexity of real tumor profiles. To accomplish that, we segmented a primary breast tumor dataset of 145 samples (Chin, K. *et al*., unpublished data) using DNAcopy and randomly sampled copy number levels from the empirical distribution of segment mean values, where mean values were binned into the intervals less than −0.4 (0 copies), between −0.2, and −0.4 (one copy), between −0.2 and 0.2 (2 copies), >0.2 but <0.4, (three copies), between 0.4 and 0.6 (four copies) and >0.6 (five copies). Note that defined intervals enrich for more extreme copy number changes and are not intended to present a realistic $\log_2$-ratio-copy number relationship but rather were constructed to increase complexity of the simulated genomes allowing for higher copy number diversity.

The lengths for normal levels (copy number 2) were assigned by sampling from the empirical length distribution of levels falling into the [−0.2, 0.2] bin. Similarly, we assigned lengths to the altered segments by sampling from the length distribution for segments with levels outside that bin, i.e. altered segments, without distinguishing among length distributions with different copy numbers. Thus, the 'true' breakpoints were derived and recorded. Each sample was assumed to be diploid and was assigned a proportion of

tumor cells (P_t), which was drawn from a uniform distribution between 0.3 and 0.7 to resemble the proportion of tumor cells often seen in tumor biopsies and to incorporate this into our simulation model in a controlled way. Consequently, the expected $\log_2$-ratio for each clone was computed as $\log_2[(c \times P_t + 2^*(1 - P_t))/2]$ where c was the assigned copy number.

Finally, Gaussian noise of mean 0 and varying variance were added to each sample. Appropriateness of the Gaussian distribution has previously been demonstrated using samples with limited number of alterations (Hodgson *et al*., 2001). Since hybridization quality and thus experimental variability of the samples may vary greatly, a sample-specific variance was added to each profile by drawing a standard deviation from a uniform distribution with range between 0.1 and 0.2. This variability reflects what is typically observed in the lower quality examples of UCSF BAC array hybridizations (data not shown). A total of 500 samples with 20 chromosomes containing 100 clones each were simulated with lengths of the edge segments truncated. This simulation was used to compare sensitivity and specificity of the three segmentation methods with regard to the breakpoint detection, to compare the two level-merging algorithms and to evaluate merging-based approach for identification of copy number alterations.

We created a different set of simulations to emulate real datasets with samples from two tumor classes. These datasets were used to specifically test whether the segmentation approach was more powerful than a univariate clone-by-clone approach for testing of copy number associations with a phenotype. For this simulation, we created 500 datasets each consisting of 20 samples drawn at random from either of two genome templates constructed as described previously with a few exceptions. Without loss of generality the length of each genomic profile was reduced to 500 clones placed on just one chromosome and each sample was only assigned a probability of 0.7 of having a given aberration (i.e. in all samples ∼30% of segments with copy number gains or losses were re-assigned a normal copy number of 2). Because the proportion of segments with copy number changes in each sample was decreased thereby, we doubled the probability of drawing altered segments from the copy number/segment length distribution. Segments with differences in copy number between the two classes were recorded. On average, each dataset had 211 clones in such segments.

3.2 Breakpoint detection and merging

We compared the sensitivity and false discovery rate (FDR) of HMM, GLAD and DNAcopy to detect and correctly locate breakpoints for originally predicted segments as well as merged segments. Here, the sensitivity is the proportion of true breakpoints that were identified, whereas the FDR is the proportion of falsely predicted breakpoints among the predicted ones. Additionally, MergeLevels and GLADmerge were compared based on the precision of their predicted values relative to expected $\log_2$-ratios and the accuracy of identifying altered clones. We also considered all pairwise combinations of the clones and determined the proportion of clone pairs that were incorrectly assigned to the same or different states, referred to as discordant pairs.

3.3 Copy number association study: testing for differential copy number

A standard approach to identifying genomic regions associated with a particular phenotype, e.g. a cancer subtype, is to individually test each clone for an association, i.e. on a 'clone-by-clone' basis. Here, comparisons were done between the standard and the 'region-based' approaches which included performing *t*-tests either on segmented $\log_2$-ratios or on the observed $\log_2$-ratios followed by segmenting the resulting statistic. Here, for HMM the segmented values corresponded to the HMM-smoothed values (weighted means of the state means). The performance of the methods was evaluated by sensitivity and specificity using a multiple testing corrected significance threshold and by comparing ROC curves.

3.4 Application to primary tumor data

Real array CGH data from BAC arrays with formalin-fixed primary oral squamous cell carcinomas (SCCs) (Snijders *et al*., 2005) were reanalyzed using the approaches introduced in this manuscript. The dataset consisted of 14 *TP53* mutant samples and 61 wildtype samples. The scientific question of interest was the comparison of genomic features between *TP53* mutant and *TP53* wildtype tumor samples. *TP53* status was determined by sequencing. Based on the methods' comparative performance assessment on simulated data, we chose to apply DNAcopy to the tumor data followed by merging with MergeLevels. The two tumor types were compared in terms of their overall genomic instability measured using the total number of breakpoints in each genome. We also assigned gain and loss status to each clone using threshold and segmentation-based methods, and displayed an example of a typical disagreement between the two approaches. Furthermore, we tested for copy number associations with phenotypes using clone-by-clone and region-based approaches. Finally, we build a predictor of the *TP53* phenotype and demonstrate that providing segmented data as an input to a classifier greatly improves prediction accuracy estimated using leave-one-out cross-validation error rate.

4 SIMULATION RESULTS

4.1 Breakpoint identification and merging

From the output of each method, it is possible to infer predicted breakpoints. These were compared with the location of known breakpoints for the simulated data (15 breakpoints per sample on average). Figure 2 illustrates how the methods perform with regard to breakpoint detection at the correct position ($w = 0$) or with an offset (localization error) of one or two clones, $w = \{1,2\}$, within which a predicted breakpoint was assigned as correctly identified. As expected, the sensitivity increased while FDR decreased with larger accepted offsets. By merging, some true breakpoints were removed and consequently sensitivity decreased slightly. Since many excessive breakpoints were removed as well, the FDR greatly decreased, especially for HMM and GLAD.

Of the compared methods, DNAcopy was most sensitive while having the lowest FDR (*P*-value <2.2e−16, paired Wilcoxon rank sum test). GLAD was least sensitive and HMM had the highest FDR. Not surprisingly, both merging procedures decreased FDR while reducing sensitivity for DNAcopy and GLAD (*P*-value <2.2e − 16, paired Wilcoxon rank sum test). MergeLevels was less aggressive than GLADmerge in removing breakpoints resulting not only in higher sensitivity but also in higher FDR. Notice that DNAcopy is very sensitive and has a low FDR when applied alone, and thus benefits least from merging with regard to breakpoints. As an example, when accepting an offset of two clones, DNAcopy has a median sensitivity of 88% while having a median FDR of 6%.

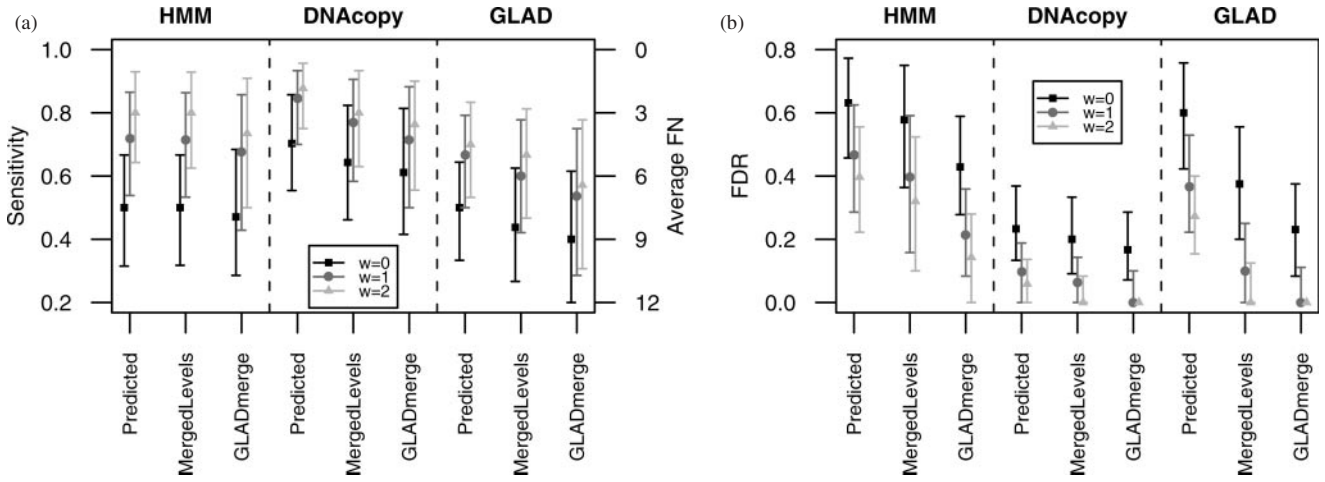

Fig. 2. Results from simulation identifying breakpoints using either HMM, DNAcopy or GLAD or after removal of excessive breakpoints by MergeLevels or GLADmerge following segmentation. (**a**) It shows the median sensitivity and corresponding average number of false negatives (FN.) (**b**) FDR for breakpoint detection with error bars depicting the interquartile range is shown. Breakpoints were classified as correctly identified at its exact location ($w = 0$) or if within an offset of 1–2 clones ($w = 1$–2) of a correct breakpoint.

This corresponds to 1.8 missed breakpoints on average and 0.8 false breakpoints. Both HMM and GLAD had significantly more trouble identifying precise breakpoint locations than DNAcopy based on examination of the offset for predicted breakpoints. The comparative performance between methods was independent of the magnitude of the signal/noise ratio defined as the ratio of the proportion of the tumor cells to the variability of noise (P_t/sd), i.e. DNAcopy consistently performed the best while GLAD was least sensitive and HMM had the highest FDR (see Supplementary information).

Additional studies indicated that the comparative performance did not change when introducing a larger proportion of smaller segments in the simulated data using empirical length distributions generated by either HMM or GLAD using the same primary breast tumor dataset as for DNAcopy. However, further examination of the spatial resolution of the three segmentation methods revealed that HMM had the greatest power to detect the shortest segments with DNAcopy surpassing HMM for longer segments. However, DNAcopy had by far the lowest FDR for all segment lengths. GLAD consistently performed worse than the other two methods except for the detection of the longest segments (see Supplementary Information for details).

The merging step allows us to identify segments on different chromosomes corresponding to the same copy number. As an example, Figure 3 shows simulated data overlaid with known $\log_2$-ratios and with either *HMM* segmented $\log_2$-ratios before merging (A) or after application of MergeLevels (B). For this example, merging clarifies the genomic profile and is able to correctly identify the base (no change) level as well as other copy number levels. This is also true for most other samples (see Supplementary Figure S1). Note that for highly aberrant genomes, such a base level does not exist and it is not possible to infer gain and loss reliably.

To verify that merging performed reasonably, four different measures were considered: (1) sum of squared (SSQ) distance; (2) MAD between predicted $\log_2$-ratios and known $\log_2$-ratios; (3) accuracy of assigning copy number gain and loss and (4) the proportion of discordant pairs (Table 1). Here, the SSQ distance and MAD were calculated with respect to the residuals between the observed

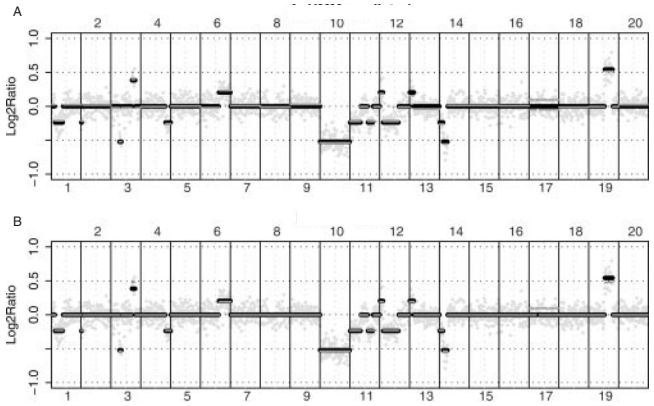

Fig. 3. An example of simulated array CGH data with 100 clones on each of 20 chromosomes. The figure shows simulated $\log_2$-ratios in light grey, ordered by position and chromosome. 'True' $\log_2$-ratios were recorded from the simulations prior to the addition of Gaussian noise and are overlaid in black. (**A**) Predicted or merged $\log_2$-ratio levels are overlaid in dark grey for *HMM* predicted $\log_2$-ratios before merging and (**B**) after applying MergeLevels. Merging brings predicted values closer to their true copy numbers.

(predicted, merged) values and the expected $\log_2$-ratios computed as a function of the copy number and the proportion of the tumor cells. All the clones with the true copy number not equal to 2 were considered to be 'altered' and the 'accuracy' was calculated as the proportion of the clones correctly assigned to altered or unaltered states. To calculate the proportion of the discordant pairs, all pairwise combinations of the clones were considered and the proportion of clone pairs that were incorrectly assigned to the same or different copy number levels, referred to as discordant pairs was determined. Segmentation alone improved all four measures and both types of merging further decreased MAD, and as expected, further increased the accuracy of assigning copy number alterations and dramatically decreased the proportion of discordant clone pairs. No significant difference was observed between the performance of MergeLevels

Table 1. Result using four difference performance measures for the array CGH analyses

	Original	Predicted	MergeLevels	GLADmerge
SSQ distance	47.38	5.08	4.88	7.25
MAD	0.104	0.015	0.0044	0.0047
Accuracy	0.93	0.93	0.97	0.98
Proportion of discordant pairs	—	0.73	0.04	0.04

Median of each performance measure for original $\log_2$-ratios, HMM predicted $\log_2$-ratios and HMM predicted $\log_2$-ratios merged by MergeLevels or by GLADmerge. Results are based on 500 simulated samples. SSQ and MAD are calculated with respect to the residuals between the observed (predicted, merged) values and the expected $\log_2$-ratios. Accuracy refers to the proportion of correctly assigned copy number alterations. The proportion of discordant pairs is the proportion of clone pairs that were incorrectly assigned to the same or different states relative to their true state.

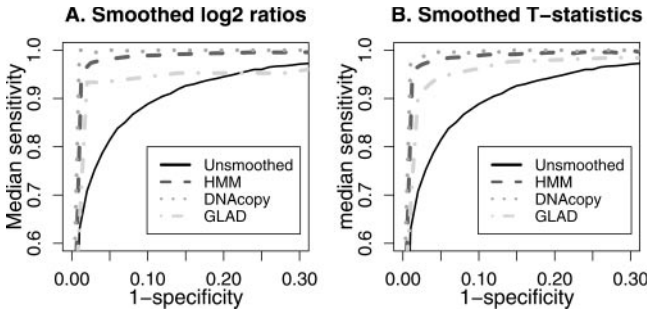

Fig. 4. ROC curve illustrating the results from the copy number association power study. For varying thresholds, it shows the sensitivities versus '1-specificity' (false positive rate). Results are based on 500 simulations and binned median sensitivities are depicted. (**A**) *T*-statistics based on segmented $\log_2$-ratios. (**B**) Segmented *T*-statistics based on raw $\log_2$-ratios.

and GLADmerge except for the SSQ distance where the application of GLADmerge resulted in significantly larger squared error compared with those obtained when only applying segmentation. Thus, while some merging is beneficial—'over-merging' may occur, which is also reflected in the sensitivity/specificity trade-off in Figure 2.

The same four measures were used to assess the benefits from merging DNAcopy and GLAD segmented data and similar overall results were obtained. Moreover, to ascertain that our results and conclusions were not an artifact of our data simulation model or the DNAcopy segmentation results for determination of the empirical length distribution used in our simulation model, a second set of simulated data was generated using the model for high-rearrangement profiles as described by Hupe *et al*. (2004) without their outlier addition. Their model led to much simpler datasets than the data simulated using our model, and consequently improved results for all methods. However, the comparative performance of the three methods was similar (see Supplementary information for details).

4.2 Copy number association power study: testing

We tested copy number associations of individual clones and genomic segments with the simulated binary phenotype, by testing whether a clone had a significantly different $\log_2$-ratio in samples from one subgroup (class 1 template) as compared with the $\log_2$-ratio for the same clone in samples from the other subgroup (class 2 template). Thus, we assessed the sensitivity and specificity of the clone-by-clone approach and the region-based approaches. The latter used segmented $\log_2$-ratios or segmented test statistics as described in Methods. For segmented test statistics, all clones assigned to the same segment would have the same test statistics. Here, the sensitivity is the proportion of known differential clones that were identified, while the specificity is the proportion of known non-differential clones identified as such.

ROC curves were used to evaluate the power to detect associations of the genomic alterations with a phenotype. Thus, in Figure 4, we plotted the median sensitivity over all datasets for small binned intervals of '1-specificity' corresponding to a sequence of different significance thresholds. It shows a combined ROC curve based on results from all 500 simulations. Application of any of the three methods resulted in greatly improved performance, which is evident by a higher sensitivity for any given specificity. Both region-based approaches are superior to the clone-by-clone (original) approach for all three segmentation methods with DNAcopy performing significantly better than HMM and GLAD (see also Supplementary Figure S8). For HMM and GLAD, the family wise multiple testing cutoff was often driven by single extreme values. The levels were predicted correctly in most cases, but the cutoff derived from the maxT reference permutation distribution was too conservative, resulting in many distinct segments being classified as non-differential. We refer to Westfall and Young (1993) and the Supplementary information for details on the permutation-based single-step maxT procedure to control the FWER. Alternatively, when applying a gFWER(k)-controlling single-step common-cutoff augmentation procedure to define significance thresholds, the sensitivity increased greatly, especially for HMM and GLAD, whereas the specificity only decreased slightly (see Supplementary information for details).

5 REAL DATA EXAMPLE: ORAL SQUAMOUS CELL CARCINOMA

Experimental data are inherently variable and segmentation involves bias/variance trade-off. We used DNAcopy and MergeLevels to re-analyze 75 oral SCC samples from a recently published study (Snijders *et al*., 2005) and demonstrated how segmentation may improve the analysis. The aim was to quantitatively compare the *TP53* mutant and wildtype tumor samples in terms of their genomic instability as measured by the number of breakpoints, to identify specific genomic regions associated with the *TP53* mutation and to use copy number data to predict mutation status of tumor samples.

Figure 1A and B illustrates profiles of a wildtype and a mutant sample showing original $\log_2$-ratios overlaid by segmented and merged $\log_2$-ratios. Figure 1C and D shows the effect of segmenting and merging, with merged and original $\log_2$-ratios sorted according to the values of predicted levels. A median of 17 and 28 breakpoints were identified in *TP53* wildtype and mutant samples, respectively. Thus, *TP53* mutant tumors were significantly more unstable genomically (*P*-value <0.03). Similar to simulations, merging only removed a small number of breakpoints for DNAcopy (final median of breakpoints 16 and 24, respectively).

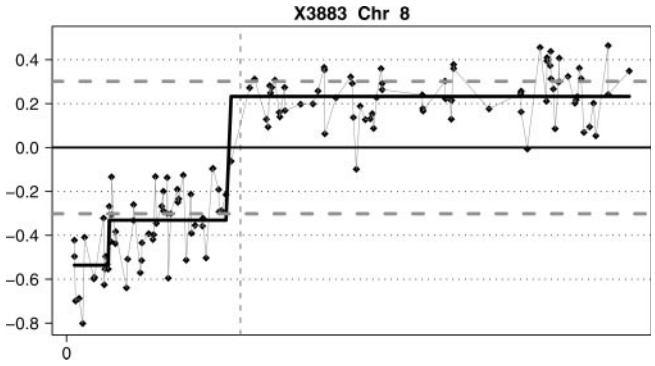

Fig. 5. Identification of gained and lost clones using threshold-based and region-based approaches. A threshold for calling aberrations is indicated by dashed horizontal line at −0.31 and 0.31. The solid curves indicate the segmented values. The baseline is at 0, thus all clones are altered according to the region-based approach with only small proportion of clones altered with the threshold-based method.

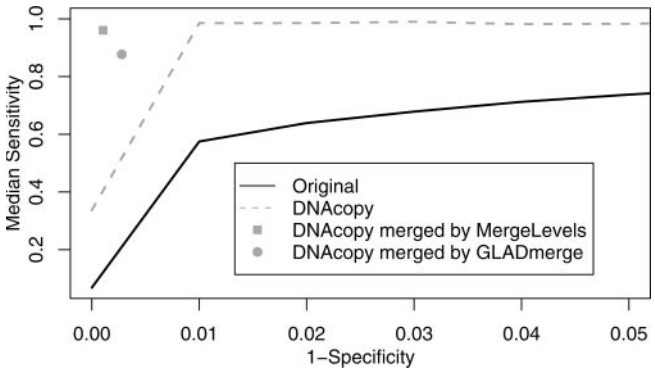

Fig. 6. Simulation study results: ROC curve of calling gains and losses are shown for DNAcopy for varying $\log_2$-ratio thresholds. Median sensitivity based on 500 simulated samples is shown for bins of '1 minus specificity'. Dots for merged results are shown for median sensitivity and median '1 minus specificity' for MergeLevels and GLADmerge.

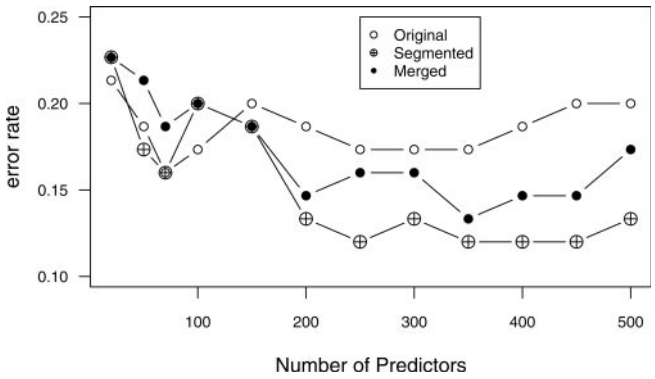

Fig. 7. Misclassification error rate for DLDA classifier using original or segmented and merged data with an increasing number of variables re-selected at each leave-one-out cross-validation step.

Now recall that assigning alterations could be done either on a clone-by-clone basis by drawing a genome-specific threshold or by using merged segments. The difference between the proportions of autosomal clones declared to be altered was dramatic between these two approaches: median value of 5 versus 33%, respectively (see Supplementary Figure S11). The large difference arose partly because of significant heterogeneity of the SCC samples combined with high experimental noise for paraffin-embedded tumors such as the samples in the SCC study. For these, a threshold-based approach is likely to miss many clones within real alterations. For instance, if the threshold is near a true copy number level, half of the clones with that copy number will be incorrectly declared unaltered. Figure 5 demonstrates the threshold-based and segmentation/merging-based methods for calling alterations. The dashed horizontal lines indicate the tumor-specific threshold. Thus, only clones above and below this threshold would be assigned an altered state. However, following segmentation and merging, assignment of the breakpoints agreed with the segmentation done using visual inspection and all clones on this chromosome can be assigned to an altered state. Of course, the threshold for the first method may be decreased; however, this would occur at the expense of a higher false positive rate as illustrated in Figure 6 (note the figure is based on results from breakpoint simulation study). This figure shows an ROC curve for assigning alterations on a clone-by-clone basis using original $\log_2$-ratios or those from a DNAcopy segmentation, and compares it with the results obtained by applying each of the level-merging algorithms. Segmentation by DNAcopy alone improves the results significantly; however, the merging approach is far superior to any threshold for the clone-by-clone approach illustrated by points to the left of both ROC curves.

Next clones with significant differences in copy numbers between *TP53* mutants and wildtype samples were identified (see Supplementary Figure S12 for resulting *t*-statistics). Only 29 clones were significantly differential for original $\log_2$-ratios. Using segmented $\log_2$-ratios for testing 66 clones were found to be significant, and when using segmented *t*-statistics a total of 139 clones were identified as differential. These 139 clones were concentrated in segments on chromosome 8p, 8q, 11q and 18q. Compared to the 29 clones originally identified, only 4 were missing. They corresponded to a single clone on chromosome 2, and a small cluster of 3 clones separated by single non-significant clones on chromosome 10. Thus, the segments picked by region-based approaches produced more biologically meaningful results than traditional univariate testing method. Note that segmentation of the test statistic outputs entire regions of interest and thus eases the interpretation of the results.

Finally, to investigate whether noise reduction via segmentation would allow for more accurate classification, we constructed a predictor for *TP53* mutants versus wildtypes based on observed $\log_2$-ratios, predicted segmented $\log_2$-ratios, or segmented and merged $\log_2$-ratios as input to the classifier. Figure 7 illustrates the resulting error rate curve and demonstrates that segmentation decreases prediction error rate, while use of merged data result in inferior results compared with use of segmented data alone. However, this was to be expected as we have observed that merging occasionally removed a true breakpoint. It is also possible that DLDA classifier is a suboptimal choice for the merged data which is discretized.

6 DISCUSSION AND CONCLUSION

Numerous methods have been proposed for segmentation of array CGH data, thus allowing for identification of copy number

transitions. However, no comprehensive comparison or even basic evaluation of the performance of the proposed methods in terms of their breakpoint detection ability has been attempted; nor have the segmentation results been utilized in downstream analyses. Here, we have presented a realistic simulation study comparing three popular algorithms designed to segment array CGH data. Moreover, we have evaluated a novel merging algorithm linking segmentation output to downstream analyses. Finally, we have proposed a region-based testing algorithm and demonstrated its superior performance.

Our results have indicated that segmentation by any of the three methods aids downstream analyses of array CGH data. Of the methods under comparison, DNAcopy has the best operational characteristics in terms of its sensitivity and FDR for breakpoint detection. However, it should be noted that it is not able to identify single clone aberrations. While our comparison was limited to only three methods, albeit widely used, our study sets an example as a reference point for evaluating future algorithms. Also, our simulation model successfully emulates the complexity of real array CGH data. Moreover, our results agree well with the recently published results by Lai *et al*. (2005), where they used a limited number of simple data simulations to demonstrate that DNAcopy generally performed better than GLAD and HMM with regard to detection of copy number alterations. Their results also indicated that HMM performed the best for small aberrations given a sufficient signal/noise ratio and GLAD did better than HMM for wider aberrations.

Merging of the resulting segments is of paramount importance in downstream use of the segmentation results. This aspect of the analysis has been largely ignored up to now except for a post-processing procedure in GLAD. We have introduced a novel merging algorithm and evaluated its performance against the existing one obtaining comparable results. We have also demonstrated that level-merging improves gain/loss detection, quantification of genomic instability for a tumor and assignment of clones to the same copy number classes. However, small reductions in sensitivity brought on by merging may hurt some downstream analyses such as testing and classification since these analyses are very sensitive to the removal of even a few true breakpoints. Ideally, a merging step could be incorporated into the initial segmentation.

Currently, identifying regions with differential copy number is done using the same approaches as in transcriptional microarray studies without special consideration for known physical dependence. We have introduced a novel method for identifying such regions, which explicitly uses segmentation results. The new approach delivers great improvements in detection power as demonstrated by our analysis.

In this paper we have demonstrated the superior performance of DNAcopy. However, an HMM approach is adaptable to perform a whole genome fit by doing constrained optimization of the segment means and variances across the entire genome, and thus consistently improving its performance with more observations. Moreover, in problems where simultaneous inferences need to be made, e.g. copy number and methylation, it may be of an advantage to use more model-based approaches such as an HMM and its extensions. Several papers on this have already been published (e.g. see Zhao *et al*., 2004) and we are continuing to work on evaluating and extending exciting methods to such problems.

ACKNOWLEDGEMENTS

The authors wish to thank Donna Albertson, Adam Olshen and E. S. Venkatraman for many useful discussions and Dan Pinkel for his ideas and critical reading of this manuscript. The authors would also like to acknowledge Peter Dimitrov for his assistance with implementation of the aCGH package. Finally, thanks to the anonymous referees for their useful comments. This work was partially supported by the grant NCI P50 CA58207 (J.F.) and The Danish Center for Scientific Computing and The Danish Technical Research Council (H.W0).

Conflict of Interest: none declared.

REFERENCES

Bauer,D.F. (1972) Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc*., **67**, 687–690.

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

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal*., **90**, 132.

Hodgson,G. *et al*. (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet*., **29**, 459–464.

Hsu,L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

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

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

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

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.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37** (suppl), S11–S17.

Snijders,A.M. *et al*. (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24**, 2432–4242.

Veltman,J.A. *et al*. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*., **63**, 2872–2880.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. Wiley, NY.

Zhao,X. *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*., **64**, 3060–3071.