BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 18 2012, pages 2357–2365
doi:10.1093/bioinformatics/bts448

Genome analysis Advance Access publication July 13, 2012

A regression model for estimating DNA copy number applied to capture sequencing data

Guillem J. Rigaill[1,2,3], Sidney Cadot[1], Roelof J.C. Kluin[4], Zheng Xue[5], Rene Bernards[2,5], Ian J. Majewski[5] and Lodewyk F.A. Wessels[1,2,*]

[1]Department of Bioinformatics and Statistics and [2]Cancer Systems Biology Center (CSBC), The Netherlands Cancer Institute, Amsterdam, The Netherlands, [3]Unité de Recherche en Génomique Végétale (URGV) INRA-CNRS-Université d'Evry Val d'Essonne, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France, [4]Central Microarray Facility and [5]Department of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Target enrichment, also referred to as DNA capture, provides an effective way to focus sequencing efforts on a genomic region of interest. Capture data are typically used to detect single-nucleotide variants. It can also be used to detect copy number alterations, which is particularly useful in the context of cancer, where such changes occur frequently. In copy number analysis, it is a common practice to determine log-ratios between test and control samples, but this approach results in a loss of information as it disregards the total coverage or intensity at a locus.

Results: We modeled the coverage or intensity of the test sample as a linear function of the control sample. This regression approach is able to deal with regions that are completely deleted, which are problematic for methods that use log-ratios. To demonstrate the utility of our approach, we used capture data to determine copy number for a set of 600 genes in a panel of nine breast cancer cell lines. We found high concordance between our results and those generated using a single-nucleotide polymorphsim genotyping platform. When we compared our results with other log-ratio-based methods, including ExomeCNV, we found that our approach produced better overall correlation with SNP data.

Availability: The algorithm is implemented in C and R and the code can be downloaded from http://bioinformatics.nki.nl/ocs/

Contact: l.wessels@nki.nl

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

Received on December 13, 2011; revised on July 7, 2012; accepted on July 9, 2012

1 INTRODUCTION

The majority of the human genome is composed of repetitive sequences and intergenic regions. Over the past 5 years, a number of hybridization-based approaches have been developed to enable selective enrichment of target DNA sequences, allowing for high coverage sequencing of protein coding genes (Albert *et al*., 2007; Gnirke *et al*., 2009; Hodges *et al*., 2007; Okou *et al*., 2007). Capture sequencing has been used successfully to identify variants underlying rare genetic disorders and to find genes that are recurrently mutated in cancer (Harbour *et al*., 2010; Jones *et al*., 2010; Ng *et al*., 2010; Varela *et al*., 2011). Although capture sequencing has traditionally been used to identify SNVs and small insertions and deletions, it is now clear that it can also be used to derive gene copy number (Sathirapongsasuti *et al*., 2011). This approach has tremendous potential, as detecting recurrent copy number alterations (CNAs) is a powerful way to identify oncogenes and tumor suppressor genes (Beroukhim *et al*., 2010).

Here we present a new statistical approach for the detection of CNAs and illustrate its performances on capture sequencing data. Instead of modeling the ratio or log-ratio between the test and control sample coverage, we modeled the test sample coverage as a linear function of the control sample coverage (a proportionality model). Using the outcome of this model, we segment the data into regions where the proportionality between the test coverage and control coverage is constant, which should equate to regions with similar copy number. An exact algorithm was used to recover the maximum-likelihood (ML) segmentation and the number of segments was selected using a modified Bayesian Information Criterion (BIC) that is adapted for segmentation problems (Zhang and Siegmund, 2007).

To validate our method, we determined copy number from capture sequencing data derived from nine breast cancer cell lines that had previously been analyzed with single nucleotide polymorphism (SNP) genotyping arrays (Forbes *et al*., 2010). One major challenge in determining copy number from sequencing data is that the level of coverage varies greatly between different regions. Using a proportionality model overcomes this problem and produces far fewer outliers than approaches that use the ratio between a test and control sample. The proportionality model is also able to handle homozygous deletions, which are problematic for methods which use log-ratios (Deng, 2011; Lonigro *et al*., 2011; Sathirapongsasuti *et al*., 2011).

Overall, we found that the proportionality model (PropSeg, for proportionality based segmentation) outperformed ratio and log-ratio-based methods for copy number determination, produced the best correlation to SNP array data and the lowest number of false-positive breakpoints with an equivalent recall. In the following sections, we describe the details of the statistical model and the associated inference scheme.

*To whom correspondence should be addressed.
The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

1.1 Some modeling considerations

The general purpose of our model is to infer CNAs from capture sequencing data collected for a test sample. Rather than inferring the copy number for each and every bait, we look for segments (sets of consecutive baits) along the genome that seem to have the same copy number. This process involves two key steps that inform one another: (i) segmenting the genome into an optimal number of regions such that the copy number in every segment is approximately constant (i.e. identifying breakpoints) and (ii) exploiting the fact that when this segmentation is available, many positions in the same segment have approximately the same copy number which provides more robust estimates.

1.1.1 A need for normalization A simple approach to estimate the copy number at a given position on the genome is to consider the coverage of the test sample as an estimator. However, in capture sequencing, this is not an efficient approach as some baits tend to capture more reads than others, which is vividly illustrated in Figure 1A and B. In Figure 1A, we plot the measurements from two different normal germline samples (normals) using the same bait library. The copy number in germline samples should assume a single constant value across all positions, excluding the sex chromosomes. We observed a large range of coverage values, which confirmed that the sequence composition of the baits has a significant influence on sequence coverage obtained after capture (Gnirke *et al.*, 2009). Although there was a large range in coverage between different baits, these values were highly correlated between different normal controls. Using three different normal samples, the pair-wise comparison always produced a Pearson correlation >0.977. The high level of reproducibility allows us to normalize for capture efficiency bias, but this depends on having a good estimate of the coverage in normal samples. To this end, we took the average of several normal samples as reference. This averaging will smooth copy number polymorphisms present in the reference samples. However, our ability to detect cancer-related alterations will remain unchanged.

1.1.2 Modeling bait order rather than the distance between baits The density between baits is extremely variable; they may be as close as a few base pairs (bp) or be separated by many millions. In copy number analysis, some have proposed to model the genomic distance between two data points arguing that one is more likely to find a break between distant genomic positions than between proximal positions (Marioni *et al.*, 2006). This is probably true across a large collection of samples when assuming that breaks are more or less evenly distributed across the whole genome. For an individual sample this is less applicable, as true breaks are at given positions and cannot be considered as random events. Moreover, the fact that breaks are more likely to occur between distant data points would increase the likelihood of finding breaks in regions of the genome where we have very few baits and diminish our chances of finding breaks in regions where we have many baits (where we should have more power to precisely locate breaks). In terms of visualization of the data, considering the order of the baits rather than the genomic location arguably gives a better impression of the quality of the data (Fig. 1C and D). For these reasons, we only consider the order of the baits and not the genomic distance between them.

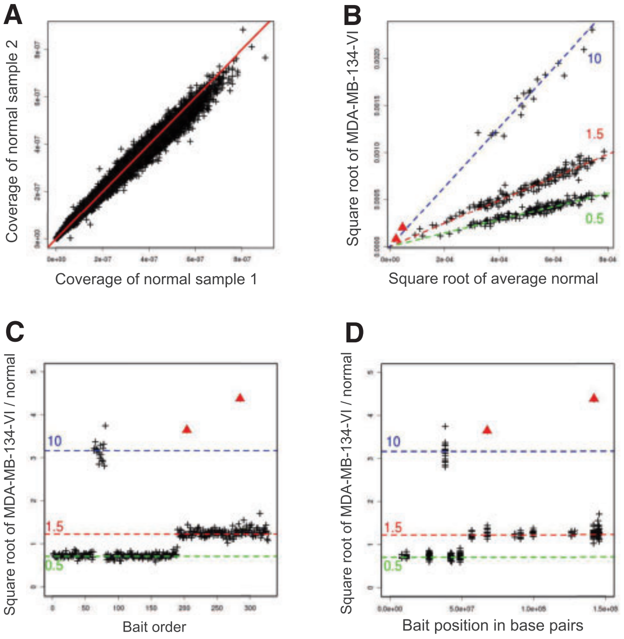

Fig. 1. Capture efficiency bias. **(A)** This graph represents the measured proportion of reads per bp (i.e. the number of reads on the bait divided by the total number of reads and divided by the length of the bait) of two different normal samples using the same library. The 95% confidence interval for this correlation is [0.989, 0.990]. **(B)** This graph represents the square root of the average proportion of reads per bp for all normal samples against that for the MDA-MB-134-VI tumor cell line for Chromosome 8. Panels **C** and **D** represent the square root of the ratio between the MDA-MB-134-VI cell line coverage and the average normal coverage as a function of the order of the baits on the chromosome (Panel C) and as a function of the chromosomal position of the baits in bp (Panel D). In Panels B D, the blue, red and green dotted lines represent a ratio of, respectively, 10, $\frac{3}{2}$ and $\frac{1}{2}$ and the two red triangles represent isolated baits with unexpectedly high ratios due to very low coverage in the normal samples

1.1.3 Proportionality, ratio, difference and log-ratio The ultimate goal of copy number analysis is to identify those regions of the genome where the copy number is constant. The simplest approach to this problem is to summarize the information for every bait by taking the ratio between test and normal sample coverage and then looking for changes in the ratio along the genome. However, the ratio approach results in a loss of absolute coverage values. For example, if we observe 100 reads in the test sample and 50 in the normal sample it is much more likely to represent a real difference in copy number, when compared with a value of 2 reads in the test sample and 1 read in the normal, although they have the same ratio (2:1). The ratio approach is also plagued by outliers; for baits with low normal coverage, a small variation in the test sample leads to a drastic change in the ratio. The two baits shown as red triangles in Figure 1B–D illustrate this problem on real data.

A log-ratio approach is also commonly used for array CGH data and was also recently proposed to derive copy number from exome sequencing data (Deng, 2011; Lonigro *et al.*, 2011; Sathirapongsasuti *et al.*, 2011). As for the ratio approach, using the log-ratio approach for sequencing data results in a loss of absolute coverage values. Furthermore, the log-ratio is not defined for regions that have no coverage. This is an important distinction, because regions that have no coverage may indicate a homozygous deletion.

To limit the influence of outliers, some have proposed thresholding to remove baits that have low coverage (Sathirapongsasuti *et al.*, 2011). ExomeCNV applies a strict cutoff in which every base should be read 15 times (this corresponds to a coverage of 0:3 (15/50) if the reads are 50 bp long; we divide by 50 as ExomeCNV counts every bp on a read whereas we count a complete read only once.). All baits below this cutoff are discarded from the analysis. This is not desirable, as it limits the sensitivity to detect deletions. In an extreme case, a region with an average of 100 reads in the control and zero reads in a tumor will be overlooked. To illustrate this point, we considered the E-cadherin gene (*CDH1*), which is partially deleted in the HCC2218 cell line (this deletion is also detected using SNP 6.0 arrays). In HCC2218 cells, seven consecutive baits in *CDH1* have zero reads and two additional baits have fewer than three reads (Fig. 2). Thus, for the smallest possible threshold (0) seven baits are discarded by ExomeCNV. Importantly, six of these seven baits have >497 reads in normal samples. For any threshold >0.0055 coverage, the two remaining baits are also discarded. This threshold roughly corresponds to a cutoff of 0.275 in ExomeCNV as the read length is 50. Thus, for its standard cutoff ExomeCNV cannot detect this alteration. This clearly illustrates that discarding baits with zero reads results in a loss of useful information and also suggests that choosing an optimal threshold is not trivial. (Also see Supplementary Figure S3, which shows the percentage of reads retained as a function of a fixed threshold.)

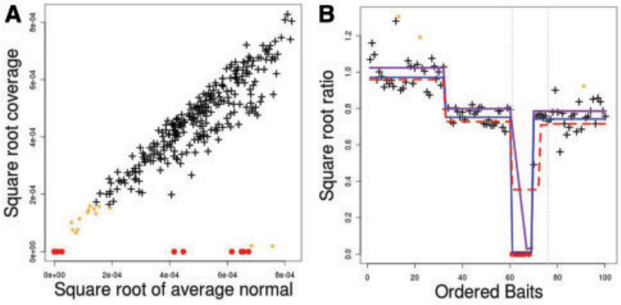

Fig. 2. Normal average coverage and HCC1187 tumor cell line coverage. **(A)** This graph represents the square root of the average proportion of reads per bp for all normal samples against that for the HCC1187 tumor cell line for Chromosome 16. **(B)** The square root of the ratio between the HCC1187 cell line coverage and the average normal coverage as a function of the order of the baits on the Chromosome 16 around the CDH1 gene. Data points depicted by red symbols are those discarded by ExomeCNV with a cutoff of 0 (the lowest possible cutoff). Data points depicted by orange symbols are those discarded by ExomeCNV with a cutoff of 15 (corresponding to a coverage of 0.3). The purple line represents the estimate of the copy number ratio with ExomeCNV with a threshold of 0. The blue line represents our estimate of the copy number ratio. The red dashed line represents the log-ratio estimated using CBS on SNP 6.0 data. The gray vertical dotted lines delimits the CDH1 gene.

Another frequently used solution to overcome the log(0) problem is to add a small constant ($\beta > 0$) to both the control (x) and test counts (Y) and segment the quantity $\log(Y+\beta/x+\beta)$. However, in a region where the copy number ratio is constant $Y/x=\alpha$, the corresponding value of $\log(Y+\beta/x+\beta)=\log(\alpha x+\beta/x+\beta)$ is clearly only constant when $\alpha=1$ or when $\alpha\neq 1$ and x is constant in that region. Thus, from a modeling perspective, looking for regions with constant $\log(Y+\beta/x+\beta)$ is not a desirable approach.

One could also think of taking the difference between the test coverage and the normal coverage as a measure of copy number variation. However, it is quite clear that this difference is not necessarily constant in a region where the underlying copy number ratio is constant. Hence, this model is not suitable either.

To avoid the loss of information and the outlier problems, we model the test coverage as a function of the normal coverage. To be specific, we assume that within a segment, the test coverage can be predicted efficiently by multiplying the normal coverage by a proportionality coefficient. In the remainder of the article, we will demonstrate the superior properties of this approach, both from a theoretical perspective as well as in terms of performance on real data.

2 METHODS

2.1 Capture protocol

DNA was extracted from a panel of nine cell lines including MDA-MB-453, HCC2218, MDA-MB-134-VI, ZR-75-30, OCUB-M, MDA-MB-468, HCC1187, BT-474 and CAMA-1. Paired-end (PE) fragment libraries were prepared using a genomic DNA library preparation kit (Illumina). Briefly, 3 μg of DNA was fragmented to an average size of 150 bp using a Covaris S220 sonicator. DNA fragments were end-repaired and A-tailed, linkers were added by ligation and the fragments were enriched by polymerase chain reaction (PCR). Libraries were amplified with six cycles of PCR using Herculase II polymerase and P5/P7 primers supplied in the SureSelect kit (Agilent). The libraries were hybridized to the SureSelect Human Kinome bait library according to the manufacturer's protocol (Agilent). After stringent washing, the captured DNA was amplified with an additional 12 cycles of PCR and an index was incorporated to facilitate pooling. The size and concentration of captured DNA was assessed using a 2100 Bioanalyzer (Agilent). Enriched libraries were pooled so that six samples were included in each lane. Sequencing was performed on a GAIIx (Illumina) using a 2×50 bp PE protocol. All reads were filtered for quality and aligned to the human genome (GRCh37/hg19) using BWA. Only unique pairs with mapping quality >20 were retained for allele calling and coverage assessment.

2.2 SNP 6.0 data from the Sanger

Affymetrix SNP6.0 data were obtained from the Cancer Cell Line Project, part of the Catalogue of Somatic Mutations in Cancer (Forbes *et al.*, 2010). The analysis and interpretation of the primary data are solely our own.

2.3 Statistical model

In this section, we describe the details of our proportionality statistical model (**P**). We consider one test sample and one chromosome at a time with n baits ordered along the genome. For each bait, $i \in \{1, \dots, n\}$, we determine

- Y_i, the number of reads mapping to bait i in the test sample divided by the total number of reads for the test sample and divided by the length of the bait.
- x_i, the number of reads mapping to bait i in the normal sample divided by the total number of reads for all normal samples and also divided by the length of the bait.

In what follows, we assume the x_i to be fixed (as opposed to random). We can make this assumption as the x_i are obtained as the average of several reference samples and are thus known with a higher precision than the Y_i.

We denote the number of segments by K. The breakpoints are denoted by $\tau_0, \tau_1, \dots, \tau_K$ with the convention that $\tau_0 = 0$ and $\tau_K = n$. The k-th segment, r_k is defined as $\{\tau_{k-1}+1, \dots, \tau_k\}$. The set of all possible segmentations in K segments is defined as $\mathcal{M}_K$. A particular segmentation of $\mathcal{M}_K$ is denoted by m and it consists of K segments denoted by r. The number and positions of breakpoints are unknown and need to be inferred from the data. In each of these segments, the proportionality ratio (α_r) between the test coverage (Y_i) and the normal coverage (x_i) is expected to be constant. Figure 1 gives an example of the relation between the test and normal count in the case where there is some copy number variation. We propose the following statistical model (**P**) to model this behavior:

$$\forall i \in r \qquad Y_i = \alpha_r x_i + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d and } \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

A slightly more complex model is to consider an error-in-variable regression. In this model, the test measurement is a random variable and is thus denoted with a capitalized symbol (X_i). This model can be written as

$$\forall i \in r \qquad Y_i = \alpha_r x'_i + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d and } \varepsilon_i \sim \mathcal{N}(0, \sigma_1^2),$$

$$\forall i \in r \qquad X_i = x'_i + \varepsilon'_i, \quad \varepsilon'_i \text{ i.i.d and } \varepsilon\prime_i \sim \mathcal{N}(0, \sigma_2^2).$$

All ε_i and $\varepsilon\prime_i$ are independant. For the sake of simplicity, we do not consider this model in the rest of this article. From a modeling perspective, this error-in-variable regression is more appropriate for cases where only a single test sample is available.

2.3.1 Recovering the ML segmentations in 1 up to K_{max} segments If the positions of the segments $(\tau_0, \tau_1, \dots \tau_K)$ are known, inference on the proportionality coefficients (α_r) is straightforward. Specifically, for a given segment r, and by using ML estimation, we obtain

$$\hat{\alpha}_{\mathbf{P},r} = \frac{\sum_{i\in r} x_i Y_i}{\sum_{i \in r} x_i^2} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{\sum_{r\in m}\sum_{i\in r}(Y_i - \hat{\alpha}_{\mathbf{P},r}x_i)^2}{n}.$$

Inferring the number of segments as well as their positions is a more difficult issue, as the number of possible segmentation is very large. There are $\binom{k-1}{n-1}$ possible ways to segment a sequence of n points in k segments. Even for a small datasets with $n = 100$ and for a relatively small number of segments $k = 10$ this is already quite a large number as $\binom{99}{9} = 1.73.10^{12}$. Thus, a brute force strategy is prohibitive.

The best segmentation in k segments is the segmentation with the smallest estimated variance. In other words, we look for the segmentation m such that $\sum_{r\in m}\sum_{i\in r}(Y_i - \hat{\alpha}_{\mathbf{P},r}x_i)^2$ is minimal. If we define the cost of a segment $c(r)$ as $c(r) = \sum_{i\in r}(Y_i - \hat{\alpha}_{\mathbf{P},r}x_i)^2$, the cost of a segmentation $C(m)$ can be written as the sum of the costs of its segments: $C(m) = \sum_{r\in m} c(r)$. From this, we obtain that any sub-segmentation of an optimal segmentation is also optimal and can be derived in a $\Theta(Kn^2)$ time and $\Theta(Kn)$ space dynamic programming algorithm following the work of Guédon (2008).

2.3.2 Choosing the number of segments Once we have recovered the best ML segmentations in 1 up to $K_{\max}$ segments, we are still faced with the task of choosing the 'optimal' segmentation. Just considering the likelihood is not sufficient as the likelihood increases with the number of breakpoints and one will always select the solution with $K_{\max}$ segments. Many criteria have been proposed to achieve a trade-off between model complexity (here the number of segments) and goodness of fit. However, model selection criteria such as BIC (Schwarz, 1978) are not well adapted for segmentation problems, since in segmentation problems, the model complexity is exponential with the number of data points (see Birgé and Massart, 2006) resulting in a tendency to overestimate the number of segments (Zhang and Siegmund, 2007). In addition, application of the BIC to segmentation problems is not theoretically justified because it assumes that the likelihood is three times differentiable with respect to the parameters (Lebarbier and Mary-Huard, 2007; Rigaill *et al.*, 2011; Zhang and Siegmund, 2007). In segmentation problems, this is obviously not the case as the positions of breakpoints are discrete. Thus, we decided to use the modified BIC criterion (Zhang and Siegmund, 2007), which is specifically adapted for segmentation problems. To be specific, we used the formula given in Theorem 2 of Zhang and Siegmund (2007) and replaced SS_{all}, SS_{wg} and SS_{bg} by their counterpart in our proportionality model (the segmentation in the mean model and our proportionality model have the exact same number of parameters).

2.4 Bias and variance of estimators

We assessed the bias and variance of our estimators of the copy number ratio for a segment using either the proportionality or the ratio model (**R**). The ratio model can be written as follows:

$$\forall i \in r \qquad Y_i/x_i = \alpha_r + \varepsilon'_i, \quad \varepsilon'_i \text{ i.i.d} \quad \varepsilon'_i \sim \mathcal{N}(0, \sigma'^2).$$

Under the ratio model, for a given segment r, the ML estimator of α_r is

$$\hat{\alpha}_{\mathbf{R},r} = \frac{1}{n_r}\sum_{i\in r} Y_i/x_i,$$

where n_r is the number of baits in segment r. We compared the biaises and variances of $\hat{\alpha}_{\mathbf{R},r}$ and $\hat{\alpha}_{\mathbf{P},r}$ under the ratio (**R**) and proportionality (**P**) models. None of the following results depend on the assumption of normality of the error.

Using the linearity of the expectation, under both models $\hat{\alpha}_{\mathbf{P},r}$ and $\hat{\alpha}_{\mathbf{R},r}$ are unbiased, meaning that $E(\hat{\alpha}_{\mathbf{P},r}) = E(\hat{\alpha}_{\mathbf{R},r}) = \alpha_r$. Under **P**, using the fact that ε_i are independant, we get that

$$V_{\mathbf{P}}(\hat{\alpha}_{\mathbf{P},r}) = \sigma^2 / \sum_{i\in r} x_i^2 \qquad \text{and} \qquad V_{\mathbf{P}}(\hat{\alpha}_{\mathbf{R},r}) = \left(\frac{\sigma}{n_r}\right)^2 \sum_{i\in r}(1/x_i)^2.$$

As the arithmetic mean is always higher than the harmonic mean, we obtain that under **P**, $V(\hat{\alpha}_{\mathbf{P},r})$ is always smaller than $V(\hat{\alpha}_{\mathbf{R},r})$. Assuming that **R** is true and using the fact that ε'_i are independant, we get that

$$V_{\mathbf{R}}(\hat{\alpha}_{\mathbf{P},r}) = \sigma'^2 \frac{\sum_{i\in r} x_i^4}{(\sum_{i\in r} x_i^2)^2} \qquad \text{and} \qquad V_{\mathbf{R}}(\hat{\alpha}_{\mathbf{R},r}) = \frac{\sigma\prime^2}{n_r}$$

Expanding $\sum_{i\in r}\sum_{j\in r}(x_i^2 - x_j^2)^2 \geq 0$ we get

$$\sum_{i\in r}\frac{x_i^2}{\sum_{i\in r} x_i^2} x_i^2 \geq \sum_{i\in r} x_i^2/n_r.$$

Thus, under **R**, $V(\hat{\alpha}_{\mathbf{R},r})$ is always smaller than $V(\hat{\alpha}_{\mathbf{P},r})$.

Summarizing, if the proportionality model is true, the proportionality estimator is better than the ratio estimator and vice versa. However, what is more interesting is to compare improvement of the proportionality estimator over the ratio estimator under **P** and the improvement of the ratio estimator over the proportionality estimator under **R**. In this scenario, we see a clear advantage for the proportionality estimator and present empirical results on this topic in the 'Results' section.

2.5 Comparison to other methods

We compared our approach with a recently proposed method for copy number variation (CNV) detection from exome sequencing data (ExomeCNV; Sathirapongsasuti *et al.*, 2011) which segments the log-ratio between the test and normal coverage. All parameters of ExomeCNV were set to standard values. We tested the standard cutoff of 15 and also a cutoff of 0: the minimum cutoff one can choose as it is required to have strictly positive values for the log transformation. We also compared our approach with the segmentation of the ratio using CBS (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007) and CGHseg (Picard *et al.*, 2005, 2007, 2011). We ran CBS and CGHseg on the coverage ratio and the square root of the ratio with or without a smoothing for obvious outliers using the 'smooth.CNA' function from the DNAcopy package (as recommended in the DNAcopy package). In total, we thus have four variants of CBS and four variants of CGHseg. Finally, we ran our approach on the raw coverage data (Y_i and x_i) and also on the square root-transformed coverage ($\sqrt{Y_i}$ and $\sqrt{x_i}$). We used the square root transform as it represents a simple way to stabilize the variance. Overall, we tested 12 different methods: (i) cbs: cbs on the raw ratio coverage; (ii) cbs-out: cbs on the raw ratio with the smoothing of outliers; (iii) cbs-sqrt: cbs on the square root-transformed ratio coverage; (iv) cbs-sqrt-out: cbs on the square root-transformed ratio with the smoothing of outliers; (v) cghseg: cghseg on the raw ratio coverage; (vi) cghseg-out: cghseg on the raw ratio with the smoothing of outliers; (vii) cghseg-sqrt: cghseg on the square root-transformed ratio coverage; (viii) cghseg-sqrt-out: cghseg on the square root-transformed ratio with the smoothing of outliers; (ix) exomecnv-15: ExomeCNV with a threshold of 15; (x) exomecnv-0: ExomeCNV with the lowest threshold of 0; (xi) propseg: our approach on raw test and normal coverage and (xii) propseg-sqrt: our approach on square rooted test and normal coverage. For all methods, we started with the same table with the number of read per baits or the number of read per baits normalized for bait length (i.e. per base) for normal and test samples.

2.5.1 Comparison to Affymetrix SNP 6.0 data To validate our approach, we compared the predicted copy number ratio to results obtained using SNP genotyping (Affymetrix SNP 6.0). For the comparison, we used the SNP 6.0 data to construct pseudo SNP 6.0 bait profiles for every bait position (S_i). More specifically, for each cell line, if we could find one or more probes matching baits *i*, we set the pseudo bait count to the average of these probes. Otherwise, we set the pseudo count to the intensity of the closest SNP 6.0 probe. For the capture data, we consider the 12 different approaches described above as well as the following baseline approaches for estimating copy number: (1) the raw test sample coverage; (2) the raw and square root-transformed ratio between test and normal coverage, with and without smoothing the most obvious outliers. This resulted in 17 profiles derived from the capture data. We then computed the Euclidean distance between the pseudo SNP 6.0 bait profiles and each of these 17 profiles for the nine different cell lines. We also ran a simple regression model to assess the Pearson correlation, as well as the slope and origin between the pseudo SNP 6.0 bait profiles and the obtained bait profiles. A desirable approach should obtain a small Euclidean distance, a high correlation, an intercept of zero and unit slope. We also segmented the SNP 6.0 data using CBS after the removal of outliers and assessed the ability, in terms of precision and recall, of the different approaches to recover change points found with SNP 6.0 data. CBS is already used by exomecnv-0, exomecnv-1, cbs, cbs-out, cbs-sqrt and cbs-out-sqrt so this evaluation criterion is likely to favor these approaches.

3 RESULTS

3.1 Computational time

10 723 baits were targeted in our experiemnts, with a maximum of 1157 baits per chromosome. The run-time for the dynamic programming algorithm for both the raw and square root-transformed coverage over all samples with a 2.53 Ghz laptop with 4 GB of RAM was 228s. This results in an average time per sample with either the raw or square root-transformed coverage of $228/(18/2) = 7.1$s.

3.2 Proportionality and ratio estimators

We have already shown (Section 2.4) that both models are unbiased estimators. In addition, we have shown (by comparing the variances of the estimators) that, under the ratio model (**R**), the ratio estimator is better than the proportionality estimator and vice versa. We compared improvement of the proportionality estimator over the ratio estimator under the proportionality model (**P**) and the improvement of the ratio estimator over the proportionality estimator under **R**.

To this end, we picked, at random, segments along the genome and computed the ratio between the variance of the ratio and proportionality estimators under **P**. For the same segments, we also computed the ratio between the variance of the proportionality and ratio estimators under **R**. The results are depicted in Figure S4 of Supplementary Materials. In summary, the relative variance of $\hat{\alpha}_{\mathbf{R},r}$ and $\hat{\alpha}_{\mathbf{P},r}$ under $\mathbf{P}$:$V_{\mathbf{P}}(\hat{\alpha}_{\mathbf{R},r})/V_{M1}(\hat{\alpha}_{\mathbf{P},r})$ is usually much larger than the relative variance of $\hat{\alpha}_{\mathbf{P},r}$ and $\hat{\alpha}_{\mathbf{R},r}$ under **R**. When the true model is **R**, the proportionality estimator is off, but is it off by a dramatically smaller margin than the margin by which the ratio estimator is off in the case when **P** is correct. More precisely, $V_{\mathbf{P}}(\hat{\alpha}_{\mathbf{R},r})/V_{\mathbf{P}}(\hat{\alpha}_{\mathbf{P},r})$ varies from 1 to roughly 10^5 whereas $V_{\mathbf{R}}(\hat{\alpha}_{\mathbf{P},r})/V_{\mathbf{R}}(\hat{\alpha}_{\mathbf{R},r})$ only varies from 1 to 5. Furthermore, the advantage of the proportionality estimator was found to be higher than the advantage of the ratio estimator in 228 596 out of 230 000 (99.4%) of the simulations. In the end no model is perfect, but it is worthwhile to consider the huge gains in accuracy that can be obtained with the proportionality estimator, especially when the corresponding model is closer to reality.

3.3 Comparison to other methods

3.3.1 Outliers For all samples in both cell line datasets, we ran 12 different methods. We then flagged segments with fewer than three baits and assessed the coverage of these baits in normal samples. We call these segments with fewer than three baits outliers. In addition, to aid the comparison between methods, we excluded all baits that are discarded by ExomeCNV due to a zero count. Figure 3 shows that for Dataset 2 the ratio approach using CBS or CGHseg tends to call baits with very low-intensity values (x_i) as outliers. Indeed, outliers detected by these methods have a mean normal coverage significantly smaller than the mean normal coverage of all baits denoted by 'All' in Figure 3 ($P < 0.0013$ using a Student *t*-test or Wilcoxon test for all pair-wise comparisons).

Looking at the median in Figure 3, it would seem that the 16 outlier-baits of exomecnv-0 are not biased toward low coverage baits and the difference between all baits and that the

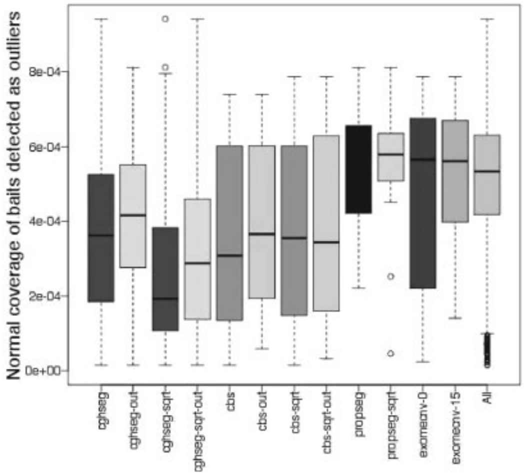

Fig. 3. Comparison of approaches based on outlier detection. This graph represents, for the different approaches (x-axis) box-plots of the square root-transformed mean normal coverage of baits detected as outliers (y-axis). A full description of the different approaches can be found in Section 2.5. The last box-plot represents the mean normal coverage of the 10 723 baits

outlier-baits detected by exomecnv-0 are not significantly different from the other best performing methods based on a t-test or Wilcoxon test. However, a closer look reveals that the boxplot is highly asymetric and that the first quartile is quite low. To assess the importance of this skewing, we resampled 16 baits out of the complete set of baits, computed their mean and repeated this 1 million times. We obtained a mean smaller than the mean of exomecnv-0 in 6465 of the million resamplings. This demonstrate the outlier-baits of exomecnv-0 are biased toward low normal coverage.

The three remaining methods propseg, propseg-sqrt and exomecnv-15 are not significantly different from the mean normal coverage (using either a t-test, wilcoxon test or our permutation scheme). Importantly, propseg and propseg-sqrt do not remove any low coverage baits of the analysis whereas exomecnv-15 does.

3.3.2 Comparison with SNP 6.0 data We established pseudo copy number profiles from SNP 6.0 data for each cell line (detailed in Section 2), and used these profiles to assess the performance of each method for detecting CNVs from capture data. Figure 4 shows, for one cell line, the capture data segmented with propseg-sqrt compare with the result obtained on SNP 6.0 data using CBS.

We then computed the mean Euclidean distance between the different capture approaches and the estimate derived from SNP 6.0 data. The results of ExomeCNV were transformed using the function $x \rightarrow \sqrt{2^x}$. This transformation drastically reduced the Euclidean distance between SNP 6.0 data and ExomeCNV results and also improved the correlation. The results with or without the transformation $x \rightarrow 2^x$ can be found in Supplementary Figures 1 and 2. The mean Euclidean distance of propseg-sqrt was found to be smaller than all other approaches on average (Fig. 5A and B). These differences were found to be significantly different from other methods (Wilcoxon test), except for cbs-sqrt, cbs-sqrt-out and propseg. The Pearson correlation between propseg-sqrt and SNP 6.0 data was higher than all other approaches (Fig. 5C and D). These differences were found to be significant except for cbs-sqrt-out and exomecnv-0. The intercept of the regression of the propseg-sqrt result on the SNP 6.0 data was smaller and closer to zero than all other approches. These differences were found to be significant except for cbs-out and cbs-sqrt-out. The slope of the regression of the propseg-sqrt result on the SNP 6.0 data was larger and closer

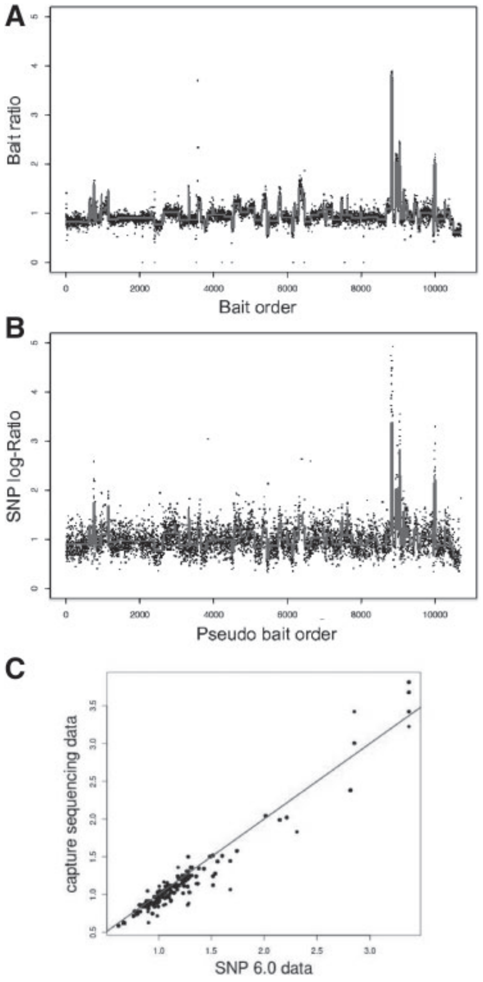

Fig. 4. Comparison of capture sequencing and SNP 6.0 data. **(A)** This graph represents the square root-transformed coverage ratio obtained form capture data for the BT-474 cell line along the whole genome. Baits are ordered according to their chromosomal and genomic position. The thick, grey line represents the segmentation and estimated proportionality coefficient for each segment with propseg-sqrt. **(B)** This graph represent the log-ratio obtained from Affymetrix SNP 6.0 for the BT-474 cell line along the whole genome at the positions of the baits. The thick, grey line represents the segmentation obtained by CBS on the complete SNP 6.0 profile. In total, 119 change points are detected using the bait data and 203 using the SNP 6.0 data. Of these, 89 are common, i.e. lie on the same bait position. **(C)** This graph represents the DNA copy number ratio estimated with propseg-sqrt for the BT474 cell line as a function of the CBS estimated copy number ratio using Affymetrix SNP 6.0 data

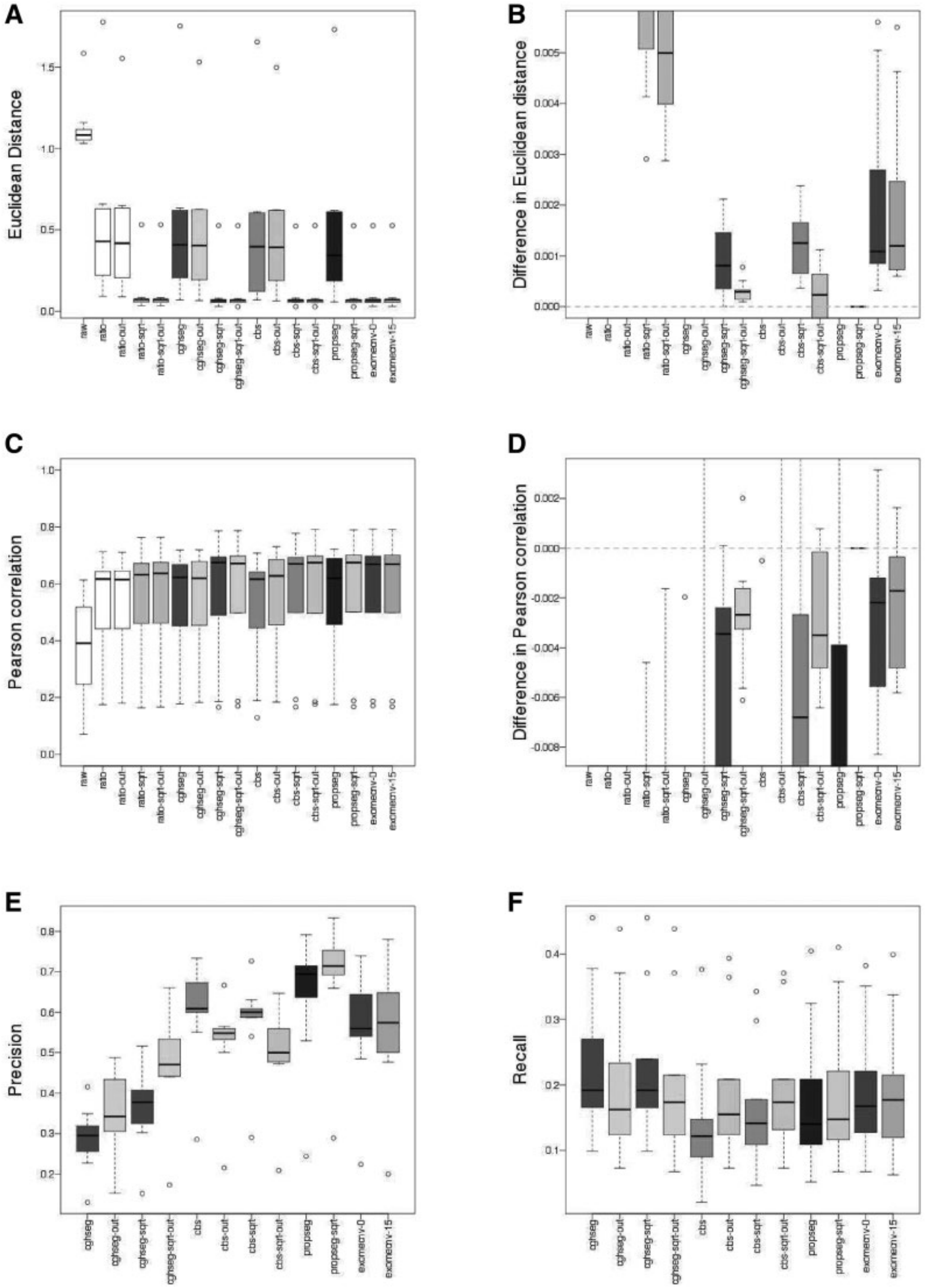

Fig. 5. Comparison of approaches based on Euclidean distance, Pearson correlation precision and recall. Box-plots of the Euclidean distance **(A)** and the Pearson correlation **(C)** between capture data and the raw SNP 6.0 data for all approaches. Box-plots of the differences in Euclidean distance **(B)** and Pearson correlation **(D)** between propseg-sqrt and all other approaches. Box-plots of the precision **(E)** and recall **(F)** of the different approches with respect to the CNAs identified using CBS on the SNP 6.0 data. A full description of the different approaches can be found in the Section 2.5. The white box-plots represent a number of baseline benchmarks. These are, the raw test sample coverage (raw), the raw ratio with or without the smoothing of obvious outliers (ratio and ratio-out). The square root-transformed ratio with or without the smoothing of obvious outliers (ratio-sqrt and ratio-sqrt-out) are represented by the fourth and fifth boxplots from the left, respectively.

to one than all other approaches. These differences were found to be significant for all approaches.

Finally, we looked at the ability of the different methods to recover breaks identified with SNP 6.0 data using CBS (Fig. 5E and F). We found that propseg-sqrt had, on average, the highest precision (number of true positives as a proportion of predicted CNAs). These difference were found to be significant except when compared with propseg. The improvement of propseg-sqrt over cbs-sqrt-out, exomecnv-0 and exomecnv-15 is 18.4, 12.7 and 12.4%, respectively.

We found that only cghseg, cghseg-sqrt and cghseg-out achieved a significantly better recall than propseg-sqrt. The average improvements in terms of recall are 4.1, 0.9 and 3.5%, respectively. However, this modest improvement in recall came at a dramatic decrease in precision. The corresponding average decreases in precision are 39, 33 and 32%, respectively. In terms of recall, propseg-sqrt was found to be significantly better than cbs and cbs-sqrt. We detected no difference with other methods. However, propseg-sqrt was found to be better or equivalent to cbs-out, cbs-sqrt-out, exomecnv-0, exomecnv-15 in, respectively, three, four, six and seven out of the nine cell lines.

Based on this cell line dataset, we conclude that propseg-sqrt is in better agreement with the SNP 6.0 data than exomecnv-0 or exomecnv-15 in terms of Euclidean distance and Pearson correlation (Fig. 5 and Supplementary Materials). It also achieves a 12% improvement in terms of precision without loss in terms of recall.

4 DISCUSSION

4.1 Statistical testing approach for the identification of gains and losses

The main focus of our work was to predict the copy number ratio for a given region. Yet, very often, one is also interested in the simpler problem of detecting (calling) gained and lost regions. Our model reduces the complexity of the data as all gains (respectively, all losses) are treated similarly. Nonetheless, our model can be used to call gains and losses: for a given segmentation, our model is a linear model and testing whether the proportionality coefficient of each segment is different from one is straightforward (in fact, we implemented this simple approach in our code).

This is a slightly different approach than what is often used for DNA copy number analysis of CGH or SNP arrays. Indeed, for CGH, a classification approach incorporating the existence of different classes of gains and losses in the model is often used (see, for example Fridlyand, 2004; Picard *et al*., 2007). However, the number of classes is typically unknown *a priori* and needs to be inferred from the data. Furthermore, very often there is a class imbalance problem as the normal regions are much more abundant than both gains and losses. In comparison, our testing approach is simpler. However, this test needs to be further refined as it does not account for the fact that the segmentation is estimated. Arguably, from a biological perspective, one is typically interested in detecting regions that show clear evidence of being different from normal (two copies) which perfectly fits with the statistical testing paradigm.

To the best of our knowledge, all methods proposed for capture sequencing data do not take the discrete nature of the data into account. Sathirapongsasuti *et al*. (2011); Deng (2011) and Lonigro *et al*. (2011) all used CBS, which was developed for continuous CGH data, and in our model we also assumed the error to be normal. For low coverage regions, this is a problem. From a modeling perspective, this could be taken into account using a Poisson or binomial negative distribution; however, this leads to an increase in the complexity of the model selection part.

We applied our new methodology to estimate DNA copy number from capture sequencing data. Our model explicitly takes into account the fact that we have two channels, one being the test sample and the other one the mean of a set of reference samples. This situation is not specific to capture sequencing data and is also encoutered in CGHarray, SNParray and full genome sequencing copy number analysis. We believe that our approach is particularly relevant for sequencing technologies because it gives low-weights to low-intensity regions which are problematic. Nonetheless, our approach could also be applied to array data.

In this article, we proposed a statistical method for the analysis of DNA copy number variation and illustrate its performances using capture-sequencing data. Rather than calculating a ratio, or log-ratio, we used a proportional model to compare normal coverage and test coverage. We have shown that the ratio approach suffers from outliers while our approach overcomes this problem. Furthermore, our approach has a higher similarity (measured with both the Euclidean distance and correlation) with results obtain with a different technology (Affymetrix SNP 6.0) than the recently proposed ExomeCNV (Sathirapongsasuti *et al*., 2011). Modeling the ratio or the log-ratio of the coverage is the most convenient choice as most methods for SNP or CGH arrays input this type of profile. However, by combining the two measurements (test and control) in a single value, there is loss of information and low values in the control sample result in outliers. We therefore advocate the approach where both measurements are modeled.

We used a quadratic time algorithm to recover the best segmentation of the data. For capture sequencing technology, it is not a problem as the number of baits per chromosome are relatively limited. However, with the improvement of the technology and for higher density technologies, the quadratic complexity could become prohibitive. In that case, the pruned dynamic programming algorithm proposed by Rigaill (2010) (available on arxiv) should be able to recover the best segmentations much faster.

There is only a single parameter to calibrate with our approach, namely the maximum number of breakpoints ($K_{\max}$) the dynamic programming algorithm will be looking for. This parameter is relatively easy to calibrate as one can quite easily choose $K_{\max}$ large enough to ensure that the optimal K is most probably smaller.

DNA capture datasets focus sequencing efforts on a genomic region of interest and can be effectively used to detect single-nucleotide variants, small insertions or deletions and CNAs. As these types of datasets will be increasing dramatically in the near future, methods for detecting CNAs in sequencing data are urgently needed. Here, we present a sound and highly competitive statistical modeling approach for detecting CNAs for capture data which overcomes basic limitations in existing approaches.

Funding: The Cancer Systems Biology Center grant by NWO (to G.J.R) and NHMRC research (575581 to I.J.M.).

Conflict of Interest: none declared.

REFERENCES

Albert,T.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Meth.*, **4**, 903–9.

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

Birgé,L. and Massart,P. (2006) Minimal penalties for gaussian model selection. *Probab Theory Relat Fields*, **138**, 33–73.

Deng,X. (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome- sequencing data. *BMC Bioinformatics*, **12**, 267.

Forbes,S.A. *et al.* (2010) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.

Fridlyand,J. (2004) Hidden markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, **90**, 132–153.

Gnirke,A. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

Guédon,Y. (2008) Exploring the segmentation space for the assessment of multiple change-point models. *Technical Report*, 6619, INRIA (http://hal.inria.fr/inria-00311634).

Harbour,J.W. *et al.* (2010) Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science*, **330**, 1410–1413.

Hodges,E. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–1527.

Jones,S. *et al.* (2010) Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science*, **330**, 228–231.

Lebarbier,E. and Mary-Huard,T. (2007) Une introduction au critre bic: fondements thoriques et interprtation. *J. de la SFDS*, **147**, 39–58.

Lonigro,R.J. *et al.* (2011) Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. *Neoplasia*, **13**, 1019–1025, PMID: 22131877.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Ng,S.B. *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome. *Nat. Genet.*, **42**, 790–793.

Okou,D.T. *et al.* (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Meth.*, **4**, 907–909.

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.

Picard,F. *et al.* (2007) A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, **63**, 758–766.

Picard,F. *et al.* (2011) Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, **12**, 413–428.

Rigaill,G. (2010) Pruned dynamic programming for optimal multiple change-point detection. arXiv:1004.0887.

Rigaill,G. *et al.* (2011) Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Stat. Comput*, **22**, 917–929.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, (ArticleType: research-article/Full publication date: Mar., 1978/Copyright 1978 Institute of Mathematical Statistics).

Varela,I. *et al.* (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, **469**, 539–542.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Zhang,N.R. and Siegmund,D.O. (2007) A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.