Performance evaluation of DNA copy number segmentation methods

Morgane Pierre-Jean, Guillem Rigaill, Pierre Neuvial

Performance evaluation of DNA copy number segmentation methods

Morgane Pierre-Jean*[1], Guillem Rigaill†[2] and Pierre Neuvial‡[1]

[1]Laboratoire de Mathématiques et Modélisation d'Évry, Université d'Évry val d'Essonne, UMR CNRS 8071, USC INRA
[2]Unité de Recherche en Génomique Végétale (URGV), Université d'Évry val d'Essonne, UMR INRA 1165 - CNRS 8114

June 18, 2014

Biographical notes

Morgane Pierre-Jean is a PhD student in the Statistics and Genome team of the Laboratoire de Mathématiques et Modélisation d'Évry, University of Évry val d'Essonne / CNRS / INRA, France. She received her MS Degree in statistics from the University of Rennes in 2012.

Guillem Rigaill is an Assistant Professor at the University of Évry val d'Essonne, working in the Bioinformatics for Predictive Genomics team of the URGV / UMR INRA 1165 - CNRS 8114, France.

Pierre Neuvial is a CNRS researcher in the Statistics and Genome team of the Laboratoire de Mathématiques et Modélisation d'Évry, University of Évry val d'Essonne / UMR CNRS 8071 / USC INRA, France.

Key Points

- A number of methods are available for segmenting DNA copy number profiles in cancer studies.

*morgane.pierrejean@genopole.cnrs.fr
†rigaill@univ-evry.fr
‡pierre.neuvial@genopole.cnrs.fr

- A robust and reproducible comparison of such methods requires the definition of a framework for generating realistic copy number profiles, and a framework for assessing methods' performance.
- A data generation framework based on resampling from real data makes it possible to compare different methods across a large number of different realistic scenarios.
- The performance of segmentation methods is mainly driven by biological parameters such as the proportion of tumor cells in the sample and the proportion of heterozygous markers.
- Using the open source and cross-platform `R` package `jointseg`, the present comparison study may be reproduced either on the data sets provided or on other data sets.

Abstract

A number of bioinformatic or biostatistical methods are available for analyzing DNA copy number profiles measured from microarray or sequencing technologies. In the absence of rich enough gold standard data sets, the performance of these methods is generally assessed using unrealistic simulation studies, or based on small real data analyses.

In order to make an objective and reproducible performance assessment, we have designed and implemented a framework to generate realistic DNA copy number profiles of cancer samples with known truth. These profiles are generated by resampling publicly available SNP microarray data from genomic regions with known copy-number state. The original data have been extracted from dilutions series of tumor cell lines with matched blood samples at several concentrations. Therefore, the signal-to-noise ratio of the generated profiles can be controlled through the (known) percentage of tumor cells in the sample.

This paper describes this framework and its application to a comparison study between methods for segmenting DNA copy number profiles from SNP microarrays. This study indicates that no single method is uniformly better than all others. It also helps identifying pros and cons of the compared methods as a function of biologically informative parameters, such as the fraction of tumor cells in the sample and the proportion of heterozygous markers.

This comparison study may be reproduced using the open source and cross-platform `R` package `jointseg`, which implements the proposed data generation and evaluation framework: `http://r-forge.r-project.org/R/?group_id=1562`.

Keywords: DNA copy number, segmentation, realistic data generation, performance evaluation.

1 Background

Changes in DNA copy numbers are a hallmark of cancer cells [1]. Therefore, the accurate detection and interpretation of such changes are two important steps toward improved diagnosis and treatment. The analysis of copy number profiles measured from high-throughput technologies such as array-comparative genomic hybridization (array-CGH), Single Nucleotide Polymorphism array (SNP array) or high-throughput DNA sequencing data raises a number of statistical and bioinformatic challenges.

Various methods have been proposed in the past decade for analyzing such data. From a practitioner's point of view, it is quite difficult to find which method is best for a given scientific question. In fact, it is likely that the overall difficulty of the problem depends on the context (technology, type of cancer, percentage of tumor cells). It is also likely that certain methods are more appropriate for certain contexts. Therefore, it is important to take this context into account when evaluating a set of methods, in order to 1) get a sense of the overall difficulty of the problem when interpreting the results and 2) choose appropriate methods for this context. Typically, a practitioner chooses among available data analysis methods or calibrates their parameters using a trial and error approach. A limitation of such an approach is that it is subjective, hardly reproducible and non quantitative.

The present work tackles this problem by proposing a reproducible framework for evaluating the performance of existing segmentation methods for identifying change-points from DNA copy number profiles from cancer patients. As any performance evaluation strategy, addressing this question requires the definition of three objects:

1. data with known "truth";
2. methods to be compared;
3. criteria for performance assessment.

In this paper, we propose such a definition and illustrate how it may be used to compare segmentation methods. The main contributions of this work are

- a framework to generate realistic DNA copy-number profiles with known "truth". This framework is generic and may be applied to any copy number data set;
- a framework to address the question of which SNP array data segmentation method performs best, depending on biologically relevant parameters.

These frameworks are implemented in the `R` package `jointseg`. The rest of this paper is organized as follows. We start by giving some background on DNA copy number segmentation (Section 2) and describe our proposed data

generation framework (Section 3). Then, we describe the pipeline we use for evaluating segmentation methods (Section 4). Finally, the result of our comparison study on two data sets are reported in Section 5.

2 DNA copy number segmentation

2.1 DNA copy number data

Normal cells have two copies of DNA, inherited from each biological parent of the individual. In tumor cells, parts of a chromosome of various sizes (from kilobases to a chromosome arm) may be deleted, or copied several times. As a result, DNA copy numbers in tumor cells are piecewise constant along the genome. Copy numbers can be measured using microarray or sequencing experiments. For illustration, Figure 1 displays an example of copy number signals that may be obtained from SNP-array data. Red vertical lines represent change points. In this particular example, the first region [0-2200] is normal, the second one [2200-6100] is a region where one of the parental chromosomes has been duplicated, and the third one [6100-10000] is a region of uniparental disomy, that is, a region where one of the parental chromosomes has been duplicated and the other one deleted. The top panel represents estimates of the total copy number (denoted by c). The bottom panel represents estimates of allelic ratios (denoted by b). We refer to [2] for an explanation of how these estimates may be obtained. In the normal region [0-2200], the total copy number is centered around two copies and allelic ratios have three modes centered at 0, 1/2 and 1. These modes correspond to homozygous SNPs AA ($b = 0$) and BB ($b = 1$), and heterozygous SNPs AB ($b = 1/2$). We note that in the second region where the tumor has 3 copies of DNA, the average observed signal is substantially below the true copy number. This is due to the presence of normal cells in the "tumor sample", a phenomenon known as *normal contamination* which shrinks the observed signals toward two copies of DNA. The reader is referred to [2] for a more detailed explanation of this phenomenon and other sources of non-calibration in DNA copy number signals, such as the ploidy of the tumor. One important observation is that change points occur at the same position in both dimensions. This is explained by the fact that a change in only one of the parental copy numbers is reflected in both c and b. Therefore, it makes sense to analyze both dimensions of the signal jointly in order to identify change points.

In order to facilitate segmentation, allelic ratios (b) are generally transformed into unimodal signals, as originally proposed in [3]. This transformation is motivated by the fact that allelic ratios can be symmetrized ("folded") and that SNPs that are homozygous in the germline (these SNPs are plotted in gray in Figure 1) can be discarded as they do not carry any information about copy-number changes. Following [4], we define the "de-

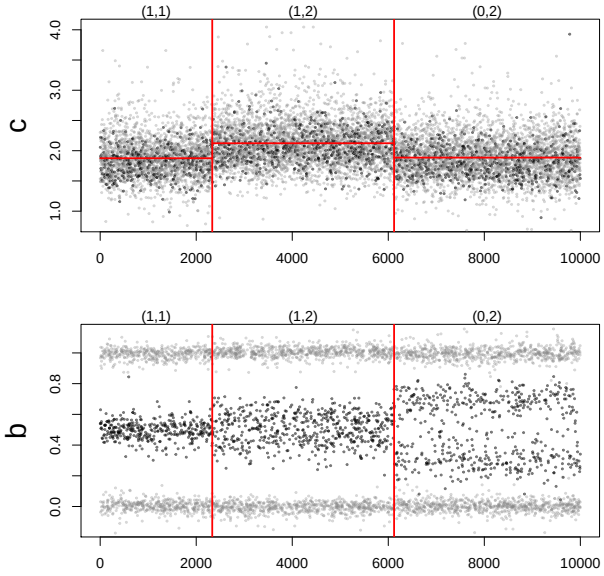

Figure 1: Example SNP array data. Total copy numbers (c), allelic ratios (b) along 10,000 genomic loci. Red vertical lines represent change points, and red horizontal lines represent mean signal levels between two change points. SNPs that are heterozygous in the germline are colored in black; all of of the other loci are colored in gray.

crease in heterozygosity" $d = 2|b - \frac{1}{2}|$ for SNPs that are heterozygous in the germline (referred to as "heterozygous SNPs" in the remainder of the paper for short), which is essentially a rescaled version of the "mirrored/folded BAF" defined by [3]. After this transformation, DNA copy numbers can be considered as a bivariate, piecewise-constant signal, as illustrated by Figure 2. It should be emphasized at this stage that because the proportion of heterozygous markers among SNPs is generally of the order of 1/3 for a given sample, the number of informative markers is several times larger for (c) than for (d). This feature of SNP array data has implications in terms of speed and performance of segmentation methods, which will be explained in detail later in the paper.

2.2 Typology of copy number segmentation methods

Many different methods have been proposed for the analysis of DNA copy number profiles. Most of them may be classified into four categories: methods based on Hidden Markov Models (HMM), multiple change-point methods, fused lasso-based methods and recursive segmentation methods.

1. HMM-based approaches rely on the idea that the recovered DNA copy number should be discrete and that these different levels can be mod-

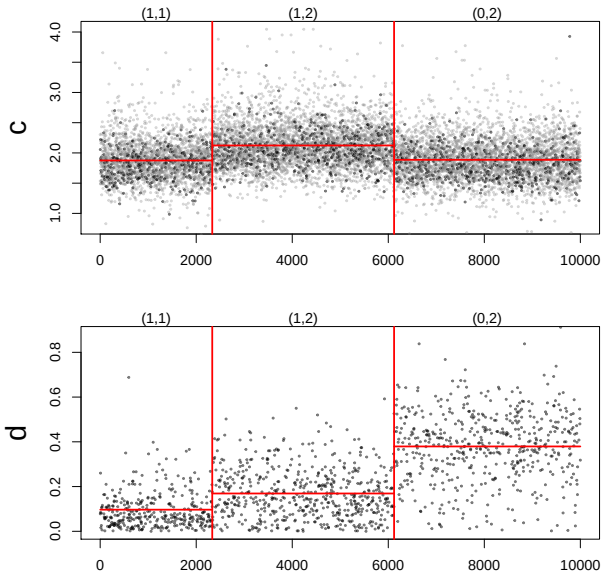

Figure 2: Example SNP array data along 10,000 genomic loci, after transformation of allelic ratios (b) into decrease in heterozygosity (d), following [4, 3]. Red vertical lines represent change points, and red horizontal lines represent mean signal levels between two change points. SNPs that are heterozygous in the germline are colored in black; all of of the other loci are colored in gray.

eled using a small number of HMM states. A typical example of such an HMM is the work of [5]. For the specific case of SNP array analysis in cancer samples, several dedicated HMM have been proposed [6, 7, 8].

2. Multiple change-point methods assume that the observed signal is affected by abrupt changes and that between these breaks the signal should be homogenous [9].

3. Methods based on a fused lasso penalty rely on the idea that, in most cases, two successive measurements should have the same estimate. This is encoded by a L_1 penalty on successive differences. The recovered signal is guaranteed to be piecewise constant. A typical example of such a fused model is the work of [10]. This class of methods can be viewed as solving a convex relaxation of the multiple change point problem.

4. Recursive segmentation approaches rely on the intuitive idea that a segmentation can be recovered by recursively cutting the signal into two or more pieces. A typical example of such an recursive approach is the work of [11].

We refer to [2, 12] for a more mathematical introduction to these methods. Here, we only note that all of these methods assume that the signals are Gaussian. The above classification is by no means exhaustive (see for example [13, 14]).

3 Generating data with known "truth"

3.1 Review of existing approaches

A number of data generation mechanisms have been proposed in the context of performance evaluation of DNA copy number analysis in cancer samples, either in comparison studies [15, 16, 17, 18], or in papers describing new analysis tools. The generation of data with known "truth" can be done using either simulated or real data, both of which have assets and drawbacks.

At first glance, simulated data are more appealing than real data because (i) "truth" is known with no ambiguity, (ii) the level of difficulty of the problem can be tuned as desired, and (iii) a large number of simulated data sets can be generated. As most DNA copy number segmentation methods rely on a Gaussian model (see Section 2), their performance is usually assessed using Gaussian simulations (see, for example, [9, 19]). While we do not question the usefulness of model assumptions for building statistical methods and for testing implementations, we believe that performance evaluation should as much as possible avoid relying on on a particular model. A recent study which compared several approaches for segmenting univariate

DNA copy number profiles using the multiple change point approach showed that the best performing methods on Gaussian simulations performed quite poorly on real data [20, Table 3]. In the remainder of this section, we briefly review some existing approaches that have tried to take the best of both the "simulated data" and the "real data" worlds:

An automatically annotated data set [15]. The authors analyzed real data using one particular segmentation method to generate "truth". They then used resampling to generate realistic copy-number profiles, where (Gaussian) noise was added in order to control the signal-to-noise ratio of the data set. Two drawbacks of this approach are that the notion of "truth" depends on the chosen segmentation method, and that the problem difficulty is not driven by biological considerations.

A dilution series [3]. In order to address the latter point, [3] have produced a dilution data set, where DNA from a lung cancer cell line is mixed with matched blood DNA from the same patient with varying (and known) mixture proportion (see description in Appendix A.1). Therefore, the fraction of tumor cells in the mixture controls the difficulty of the problem. The "truth" is a panel of regions whose DNA copy number status in the cell line (normal, gain, hemizygous deletion, copy-neutral LOH ...) is known. This evaluation method has been accepted as a *de facto* standard and has been used in several subsequent papers, including [8, 21, 22].

An important drawback of this evaluation framework is that it focuses on a very limited number of regions (ten), which results in very little discrimination between most methods in realistic settings. For example, four of the six methods compared in [21] reach maximum sensitivity in all 10 regions for tumor cell fractions greater than 25%. In practice, samples with less than 50% are rarely analyzed, in particular because the performance of most methods typically decreases severely when the fraction of tumor cells is less than 75%. We also note that sensitivity and specificity are evaluated separately in [3], and this weakness has been perpetuated in all subsequent papers based on the same evaluation framework.

A manually annotated data set [18]. The authors analyzed hundreds of neuroblastoma array-CGH profiles in order to define regions containing breakpoints (true signals), and regions not containing breakpoints (false signals). This data set is freely distributed on CRAN[1]. Based on this large data set with known truth, the authors have performed a comprehensive comparison of segmentation methods for array-CGH data based on ROC curves. A drawback of this evaluation framework is that once a particular data set is chosen, it is not possible to tune the signal-to-noise ratio of the

[1] `http://cran.r-project.org/web/packages/neuroblastoma/`

problem. Moreover, annotating a new data set is a challenging task, because it has to be large enough to contain a set of change-points that discriminate between competing segmentation methods.

A simulation model [16]. The authors designed a complex simulation model to generate "realistic" copy-number profiles. This model is implemented in the R package `CnaGen`, which is available from the authors' web page[2]. The simulation model depends on 24 parameters[3]. Some of them are directly driven by biological considerations, such as the percentage of tumor cells in the sample or intra-tumor heterogeneity. We empirically found it difficult to find a combination of parameters that yield realistic copy-number profiles. This may be due to the fact that the underlying data generation model is Gaussian. Table 1 summarizes the features of approaches reviewed above.

Reference	[15]	[3]	[18]	[16]	This paper
Based on real biological data?	√	√	√	-	√
Noise level based on a biological parameter?	-	√	-	√	√
Data generation possible?	√	-	-	√	√
Available as an R package?	√	-	√	√	√

Table 1: Features of existing frameworks for real copy number data with known "truth".

3.2 Proposed data generation mechanism

Based on these considerations, we propose an original data generation framework which aims at combining the advantages of all of the above-mentioned existing approaches. Two necessary and sufficient ingredients for generating a copy-number profile of length n are:

- truth, in the form of K breakpoint positions (out of $n-1$ intervals between two successive loci) and $K+1$ copy-number state labels for all $K+1$ regions between two consecutive breakpoints;
- signal, in the form of locus-level data. For SNP arrays, this is generally a $n \times 3$ matrix of total copy numbers (c), allelic ratios (b), and germline genotypes.

Our proposed approach is described below.

[2] `http://web.bioinformatics.cicbiogune.es/cnagen/`
[3] `CnaGen` version 2.1.

3.2.1 Generation of "truth"

When breakpoints and region labels are not user-supplied, we propose the following approach for generating them:

breakpoints: given a signal length n, draw K breakpoint positions uniformly out of the $n-1$ possible intervals between successive data points (vertical red lines in Figure 3);

region labels: draw $K+1$ region labels from a pre-defined set of copy-number state labels, such as normal, gain of one copy, hemizygous deletion, homozygous deletion, copy-neutral LOH (labels on top of each plot in Figure 3). By default, all region labels are equiprobable, but the user may provide a vector of probabilities for each desired region label. By default, successive regions are constrained in such a way that only one of the two parental copy numbers changes at the breakpoint. Not adding such a constraint would be equivalent to allowing two distinct biological events to occur at the same genomic position, which is possible in theory but rarely observed in practice.

3.2.2 Generation of locus-level data

Given breakpoint positions and region labels, we generate a copy-number profile as follows: for each region of size n_R between two breakpoints, we sample n_R data points from a real copy-number data corresponding to this type of region.

The data generation mechanism therefore relies on real data where the underlying region label is (assumed to be) known. We have made available two such "real data sets with known truth" in the package: each of them corresponds to a different SNP array platform (Affymetrix or Illumina), and both of them are taken from dilution series, consisting of mixtures of DNA from a tumor cell line and from blood cells originating from the same patient, with varying mixture proportions. For both data sets, we have selected several genomic regions which are representative of the diversity of copy-number states that are typically observed in tumor samples. Contrary to [15], these labels do not rely on any automatic segmentation or calling method. Both data sets are described in Appendix A.

3.3 Features of the proposed data generation mechanism

Our proposed data generation mechanism enjoys the following features:

- simplicity: small number of required parameters, all of which have a clear biological interpretation. In particular, for a given data set, the noise level is governed by the fraction of tumor cells. This is illustrated by Figure 3;

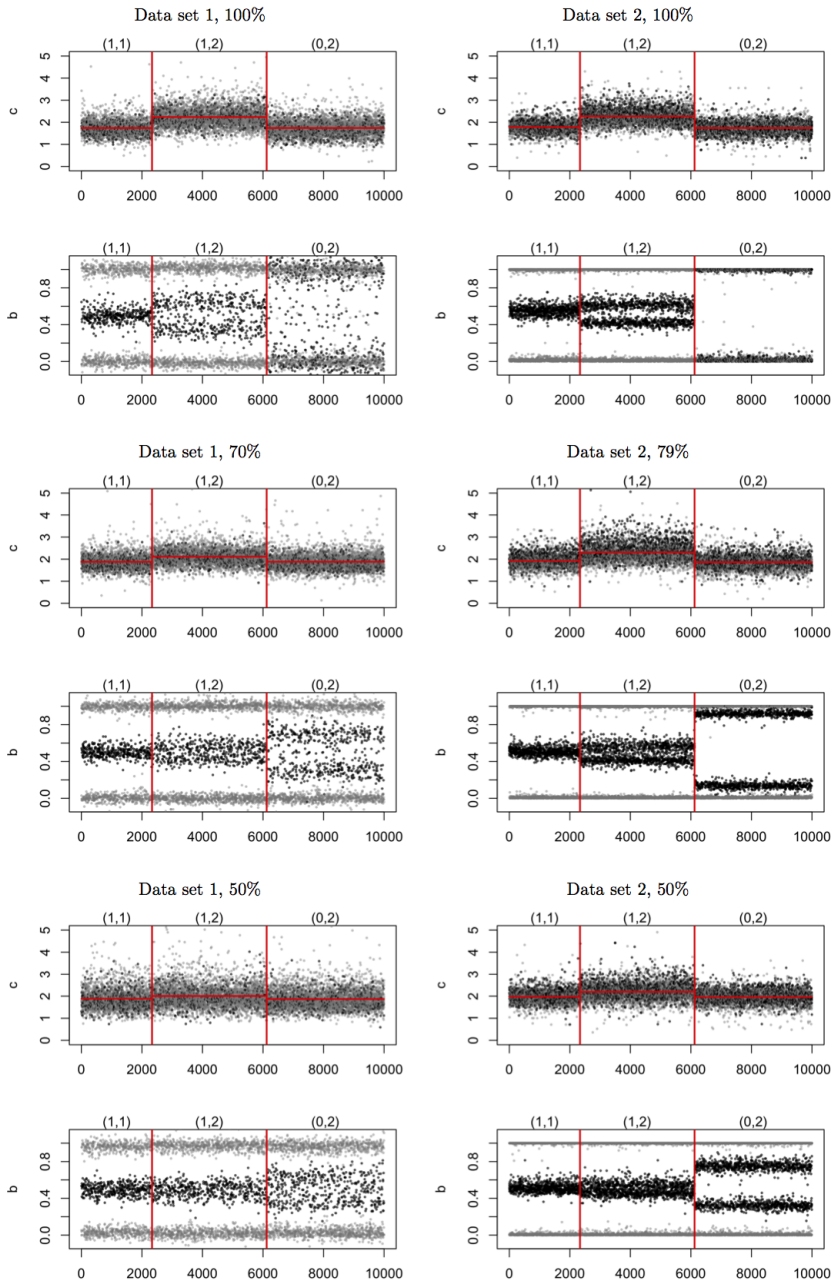

Figure 3: Illustration of the variety of copy-number profiles that can be generated from the same "truth" as in Figure 1. Each block of two plots corresponds to total copy numbers (c) and allelic ratios (b) for one particular combination of fraction of tumor cells (in rows) and data set (in columns). Red vertical lines represent change points. SNPs that are heterozygous in the germline are colored in black; all of of the other loci are colored in gray.

- flexibility: the user may specify breakpoint positions and region labels directly, if desired. Therefore, it is also possible to generate profiles with the same underlying "truth", but with different SNR, as illustrated by Figure 3;

- reliability: copy-number regions were identified using the profiles with 100% tumor cells. In these profiles, the region labels may be defined manually unambiguously. Because the same tumor cell line is used for the dilutions series from a given platform, the regions identified on the profiles with 100% tumor cells can also be considered as ground truth for the profiles with less tumor cells, where direct manual identification would have been more problematic;

- versatility: the design choice of separating "truth" generation from locus-level data generation implies that it is relatively easy to:

  - annotate a new data set. Although dilution series are not publicly available for all possible platforms, it is also possible to annotate representative profiles from a given data set. Moreover, annotating a new data set is not time-expensive, as one only needs to identify a few copy-number regions.
  - extend the framework to other data types (for example array-CGH or high-throughput exome capture or whole genome sequencing) is straightforward: only a set of annotated data is required.

4 Evaluation pipeline

Now that we have a framework to generate data, we describe how to evaluate the performance of segmentation methods.

4.1 Benchmark

Synthetic copy-number profiles were generated as described in Section 3:

region-level "truth" : Each profile contains $n = 200,000$ loci in copy number signal and $K = 20$ breakpoints. We chose to impose the constraint that on average, 90% of segments are either normal (1,1), copy-neutral LOH (0,2), single copy-gain (1,2) or hemizygous deletion (0,1). The remaining 10% of regions are given less common copy-number states, such as homozygous deletion, or balanced duplication. These parameters were inspired by our experience with SNP array data from The Cancer Genome Altas (TCGA), especially on ovarian cancers, where normal regions and regions of copy-neutral LOH, single

copy-gain, and hemizygous deletion are fairly common, while other types of alterations are much more rare [23].

locus-level data: for each of $B = 50$ such "truth" profiles, corresponding locus-level data are then generated for 100%, 70% and 50% of tumor cells for data set 1, and 100%. 79% and 50% of tumor cells for data set 2. These percentages are among those available from the dilution series from which real data was extracted, see Appendix A. Pure tumor samples (100%) are typically observed in studies about tumor cell lines, while percentages as low as 50% are typically observed in primary tumors.

4.2 Preprocessing

We log-transformed total copy numbers to stabilize their variance and smoothed outliers using `smooth.CNA` [11] as it improved segmentation results for all methods. Allelic ratios were converted to (unimodal) decrease in heterozygosity (d) as described in Section 2.1.

4.3 Compared segmentation methods

We evaluated different types of methods belonging to the different classes described in Section 2.2: multiple change-point, recursive, fused, and HMM-based methods. These methods are described in Table 2, where we mention which of them are able to process both signal dimensions (c and d) or only one of them. Not all of these methods were implemented in `R`. We ported from `Matlab GFLseg`[4] to `R` the implementation of multi-dimensional dynamic programming and the group-fused LARS [29], and we implemented recursive binary segmentation [27] in `R`. In practice, as recommended by [27, 28, 29], both group-fused LARS and recursive binary segmentation are used to quickly identify a list of *candidate* change points, which is then pruned using dynamic programming.

All of the compared methods are reasonably fast and memory-efficient, except those based on two-dimensional dynamic programming (DP): `cnaStruct` and our implementation of DP in `R` . Indeed, two-dimensional DP is quadratic in time and memory and thus cannot handle profiles of size $n = 10^5$. It may be surprising that the two-dimensional version of GFLars is faster than its one-dimensional counterpart. This is a consequence of the fact that the number of informative markers is several times larger for (c) than for (d) (as explained in Section 2.1). As the implementation of GFLars does not handle missing values, the 2d version of GFLars was applied to non-missing entries in (c, d), while the 1d version was applied to a much longer signal (all

[4] Available at `http://cbio.ensmp.fr/~jvert/svn/GFLseg/html`.

Name	R package	function	dims	Time (s) n=10^4	Time (s) n=10^5	Ref
		Multiple change-point				
DP	cghseg	segmeanCO	1d	0.24	2.37	[24]
CST	cnaStruct	segment	2d	120	fail	[25]
DP	jointseg	doDynamicProgramming	2d	140	fail	
		Recursive				
CBS	DNAcopy	segment	1d	0.34	1.69	[26]
PSCBS	PSCBS	segmentByPairedPSCBS	2d	1.04	4.00	[21]
RBS	jointseg	doRBS	2d	0.15	1.15	[27]
		Fused				
GFLars	jointseg	doGFLars	1d	0.29	3.70	[28]
GFLars	jointseg	doGFLars	2d	0.08	0.60	[29]
		HMM				
PSCN	PSCN	segmentation	2d	7.25	73	[8]

Table 2: List of DNA copy number segmentation methods evaluated.

(c) entries). This phenomenon does not happen for other two-dimensional segmentation methods as their implementation does handle missing values.

4.4 Criteria for performance evaluation

Comparison studies typically assess the performance of DNA copy number analysis methods either in terms of their ability to accurately identify breakpoint locations [17, 18], copy-number states [3, 16], or both [15]. This paper focuses on the former only, because we are interested in comparing segmentation methods. The problem of evaluating strategies for calling copy-number states is left for future work.

As our proposed data generation framework provides copy number profiles with known "truth", a natural way to evaluate the performance of a given method is to cast the problem of breakpoint detection as a binary classification problem. Specifically, for each generated copy number profile, we know where the true breakpoints are located. The number of true positives TP is the number of true breakpoints for which at least one breakpoint is detected closer than a given tolerance parameter. The number of false positives FP is defined as FP=P-TP, where P is the number of "positives", that is, the total number of detected breakpoints. With this definition, whenever a method identifies two or more breakpoints within the tolerance area of a true breakpoint, one of these breakpoints counts as a true positive, while all others count as false positives. This definition of true and false positives is

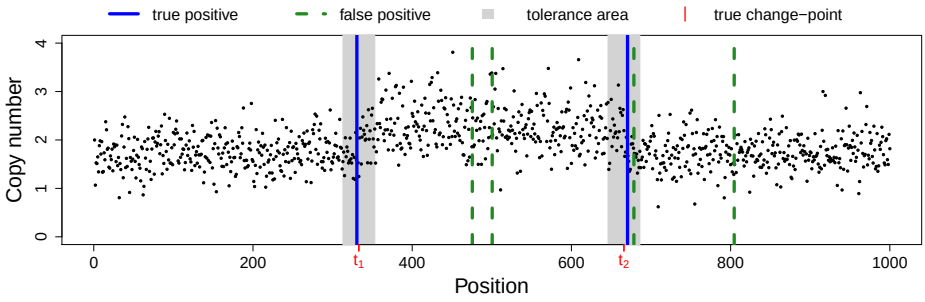

Figure 4: Definition of false positive and true positive to build performance evaluation.

illustrated by Figure 4, where gray areas highlight tolerance areas around the true change-points, whose positions are identified as t_1 and t_2 on the x axis. In this example, breakpoints were detected in both shaded areas, therefore the number of true positives (solid blue lines) is two. There are four false positives (dashed green lines): one in a gray area where there is already one true positive, and three which are not within the tolerance area of any true breakpoint. Alternative definitions of true and false positives may be considered. Some of these alternatives are implemented in the `jointseg` package, including one in which a second breakpoint found within a tolerance area is not counted as a false positive. We chose to stick with the above-described evaluation (where such breakpoints are called false positives) in order not to favor methods such as the (group) fused lasso that tend to systematically find multiple breakpoints very close to each other, which is generally inconsistent with the biology of cancers.

Related works. A similar definition of true and false positives is used in [15], although the authors do not mention how the above case of multiple breakpoints within the tolerance area is handled. Another related approach has been proposed in [18]. There, copy-number profiles are real, array-CGH profiles for which regions containing a breakpoint and regions containing no breakpoints have been delineated by experts. The main difference is that only a subset of the "true" and "false" breakpoints are annotated, and that the tolerance parameter cannot be tuned without the expert re-annotating the data set. Finally, a similar type of evaluation has been used by [17], at the the locus level instead of the breakpoint level. This locus-level based evaluation method tends to favor segmentation methods that accurately identify large altered regions, even if they fail to detect breakpoints delineating smaller altered regions.

4.5 ROC-based evaluation

Usually, each method provides a segmentation and its associated set of breakpoints. This can be translated into a measure of sensitivity and specificity using the above definition of true and false positives. However, the methods have to be compared at the same specificity or sensitivity level in order for this comparison to be fair. Ideally, we would like to compute a Receiver Operator Characteristic (ROC) curve for each method. In order to do this, one needs to explore a large set of possible segmentations with varying sensitivity and specificity, obtained by exploring the set of tuning parameters of each method. Such an exhaustive exploration is tedious and time consuming as soon as the number of parameters is larger than 2 or 3, and may lead to over-optimistic results. To overcome this problem, we adopted the following strategy: for any given method m, we recovered a segmentation in k_m change points using default parameters, and we retrieved for each $k \in \{1 \dots k_m\}$ the best k subset of these k_m using dynamic programming. Another possible strategy would be to sort the k_m change points according to a measure of confidence.

One could be worried that the range of explored sensitivity/specificity is highly variable across methods. In practice, our experience is that the default parameters of a method generally tend to over-segment the data and that typically, most of the true change points are found, at the cost of a more or less large number of false positives. This is in agreement with [18].

5 Results

5.1 Quantifying problem difficulty for known change points

Segmentation methods rely on a statistic to quantify the biological difference between any two regions. Based on this statistic, they aim at locating a good set of regions or equivalently, of change points. This location problem is combinatorial in nature. In this section, we try to quantify this biological difference independently of this combinatorial problem. In order to do this, we assume that change point positions are given *a priori* and we compare the power to call a change using total copy numbers (c) or allelic signals (d) for different types of change points. In order to perform this power study, we need to formally define the notion of power, or signal-to-noise ratio (SNR), between copy number regions. We chose a definition of SNR which is consistent with our proposed data-generation mechanism, in which DNA copy number data from a given region are sampled from a population which represents the corresponding copy-number state (see Section 3.3). Let us consider two regions and label by "0" and "1" the copy number state of two regions. For univariate signals (c or d), a natural definition of SNR is the (squared) Z statistic of the comparison between the sample means of

region "0" and region "1":

$$\mathrm{SNR}(c) = \frac{(\bar{c}_0 - \bar{c}_1)^2}{\sigma_{c,0}^2/n_0 + \sigma_{c,1}^2/n_1} \tag{1}$$

$$\mathrm{SNR}(d) = \frac{\left(\bar{d}_0 - \bar{d}_1\right)^2}{\sigma_{d,0}^2/n_0^\star + \sigma_{d,1}^2/n_1^\star}\,, \tag{2}$$

where n_i is the total number of loci in region i, $\bar{c}_i$ and $\sigma_{c,i}$ are the sample mean and population standard deviation of total copy numbers in state i and $\bar{d}_i, \sigma_{d,i}$ are the sample mean and population standard deviation of the decrease in heterozygosity in state i. Note that the decrease in heterozygosity is only defined for SNPs that are heterozygous in the germline, whereas the total copy number is defined for all loci. Therefore, $\bar{d}_i$ is calculated based on $n_i^\star$ heterozygous SNPs, while $\bar{c}_i$ is calculated based on all n_i loci. For a given DNA sample, the fraction of heterozygous SNPs among those present on the microarray is typically close to 1/3; moreover, data set 1 contains not only SNP probes but also non-polymorphic loci, with a 1:1 ratio. As a result, the fraction $n_i^\star/n_i$ is approximately 1/6 for data set 1 and 1/3 for data set 2. A natural extension of this definition of SNR to the two-dimensional case of the statistic (c, d) is

$$\mathrm{SNR}(c,d) = \left(\bar{c}_0 - \bar{c}_1, \bar{d}_0 - \bar{d}_1\right) \left(S_0 + S_1\right)^{-1} \left(\bar{c}_0 - \bar{c}_1, \bar{d}_0 - \bar{d}_1\right)' \,, \tag{3}$$

where S_i is the population covariance matrix of the bivariate vector (c, d), that is $S_i = \begin{pmatrix} \sigma_{c,i}^2/n_i & \tau_{cd,i}/n_i^\star \\ \tau_{cd,i}/n_i^\star & \sigma_{d,i}^2/n_i^\star \end{pmatrix}$ with $\tau_{cd,i}$ the covariance between c and d in state i. In practice, the population parameters for copy-number state i (that is, $\sigma_{d,i}$, $\tau_{cd,i}$, and $\sigma_{d,i}$) are calculated from the annotated data. The sample parameters ($\bar{c}_i$ and $\bar{d}_i$) are calculated from samples of n_i and $n_i^\star$ loci, respectively. Note that SNR(c) and SNR(d) are comparable with each other since they follow (non-centered) χ^2 distributions with 1 degree of freedom under the null hypothesis of no breakpoint between state 0 and state 1.

By definition, SNR is an increasing function of the length of each flanking segment. For $i \in \{0, 1\}$, we chose $n_i = 500$. $n_i^\star$ depends on the proportion of heterozygous SNPs in the sample; as explained above, it is very close to $n_0/6$ for data set 1 and $n_0/3$ for data set 2. Therefore, the length of the flanking regions essentially acts as a constant scaling factor across all transitions and settings. Therefore, SNR only reflects differences between the underlying copy number states. Figure 5 shows the average (and standard error) of log(SNR) across 100 samplings for three levels of tumor purity level, for three common types of copy number transitions for data set 1 (top panel) and data set 2 (bottom panels). Several conclusions may be drawn:

- **Difficulty generally increases with normal contamination**: SNR generally increases with the percentage of tumor cells. This is true for

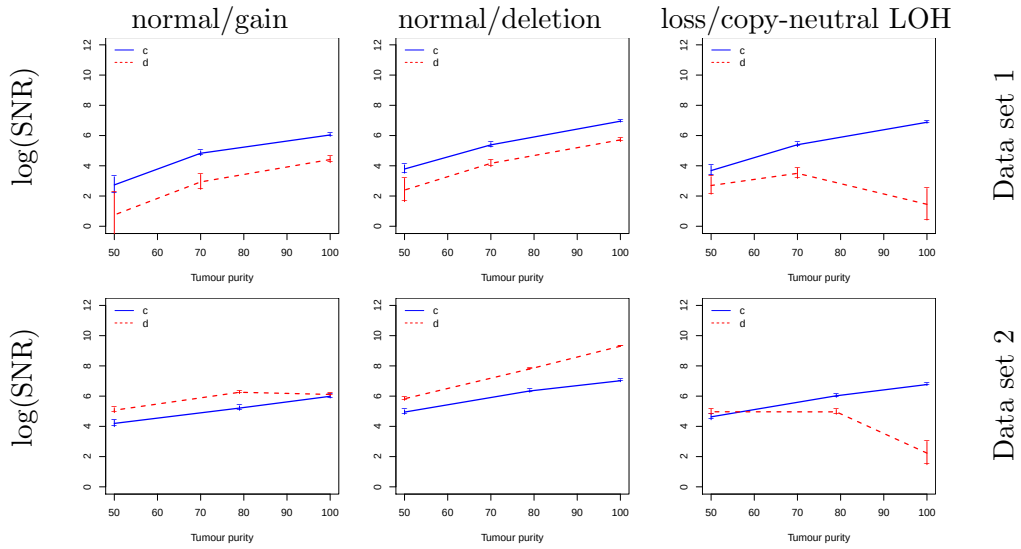

Figure 5: Average log(SNR) and corresponding standard errors across 100 samples as a function of the percentage of tumor cells for total copy numbers (c, solid blue lines) and allelic ratios (d, dashed red lines). Each column corresponds to a type of copy number transition. Each row corresponds to a given data set.

all types of transitions for c. For d, the only situation in which SNR is not an increasing function of tumor purity is the case of transitions between loss and copy-neutral LOH (Figure 5, rightmost column). This is expected theoretically because both of these states correspond to LOH in the tumor cells of the sample, implying that the true d in these cells is 1. In presence of normal cells, d estimates in both states are shrunk d toward 0, but in a state-specific way (see [4, Figure 4] for a detailed explanation of this phenomenon);

- **SNR levels depend on the type of copy number transition** for a given data set (that is, for a given row in Figure 5). This holds for both statistics (c or d). Note that in the case of c, this is unexpected, as all plotted transitions correspond to a one-copy gain.

- **Possibly low power.** Note that in some cases (e.g. data set 1, (a) and (c)), the computed SNR is lower than 2. Under the null hypothesis of no difference in mean levels, SNR follows a centered $\chi^2(1)$ distribution, so that this range of observed SNR correspond to p-values of the order of 1%, which is not low considering the large number of data points ($n_i = 500$).

- **Neither c or d is always the best statistic**. For a given type

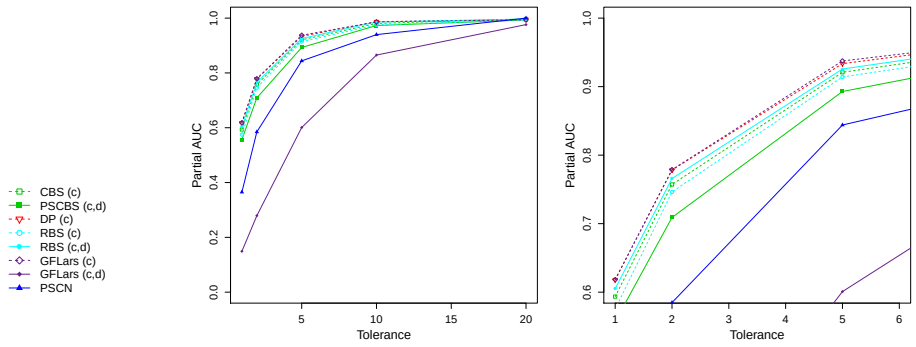

Figure 6: Method performance increase with the tolerance parameter for both data sets. Partial AUC for FP $\leq$ 10 for data set 1 and 100% tumor cells.

of transition (that is, for a given column in Figure 5) and a given statistic, the trend in SNR is comparable across data sets. However, the relative power of c with respect to d is much higher for data set 1 than for data set 2. This is directly related to the above-mentioned difference between ratios $n_i^\star/n_i$ of the number of informative loci for each statistic.

In this subsection, we assessed the intrinsic difficulty of calling a change point if the positions to test are known *a priori*. This study suggests that c and d are complementary sources of information, implying that change point detection methods should ideally take both of them into account. This study also sheds light on the fact that low percentages of tumor cells severely impacts SNR. In the remaining subsections, we assess the ability of segmentation methods to recover the true location of change points.

5.2 Robustness of the evaluation to the tolerance parameter

Our first goal was to check the influence of the tolerance parameter on the methods' performance. Our simulations were run using data generation as described in section 4.1. We computed partial areas under the ROC curves (pAUC) with a number of false positives between 0 and 10. Mean and 95% confidence intervals of pAUCs across simulation runs were calculated for each method for 5 values of the tolerance parameter (1, 2, 5, 10 and 20). For example, a tolerance of 5 means that a breakpoint is considered correct if it lies within 5 data points of the true breakpoints (see section 4.4 for more details). These results are reported in Figure 6 in the scenario without normal contamination. Similar results were observed for other scenarios.

Increasing tolerance clearly increases pAUC for all methods. This is the case even in the arguably "simple" scenario where no normal cells are

present. However, in most cases the ranking of all methods is not affected by tolerance. Based on these results, we decided to report only pAUC for one particular value of tolerance: 5 loci on each side of the breakpoints.

5.3 Joint segmentation generally increases performance

This section aims at comparing the quality of segmentations obtained using total copy numbers only (c), allelic ratios only (d), and both of them (c, d) and how the quality of the segmentation is affected by the purity of the sample. As explained in section 5.1, it is typically expected that localization of the breakpoints is easier using both dimensions of the signal. In order to do so, we compared 6 scenarios corresponding to two data sets and three levels of purity (high, intermediate and low). Table 3 reports the pAUC of the best (c), (d) and (c, d) methods for data set 1 and 2, respectively. Detailed results for all methods are presented in Table 4.

For both data sets it is quite clear that performance in terms of pAUC severely deteriorates when the level of contamination increases. (c) methods perform better than (d) methods for high level of purity. For example in the case of data set 2 the minimum difference in pAUC between (c) and (d) is 19% for high level (Table 4). For an intermediate level of purity, for data set 1 (c) outperforms (d) with a minimum pAUC difference of 41% and for data set 2 (c) is similar to (d). For a low level of purity, the pAUCs are low or very low for both data sets; for data set 1, (c) outperforms (d) with a minimum pAUC difference of 6%; for data set 2, (d) outperforms (c) with a minimum pAUC difference of 15%. These observations are in agreement with the results of Section 5.1. The difference between data sets 1 and 2 can be explained by the fact that the proportion of informative markers is different, namely around 1/6 and 1/3, respectively. This low proportion of informative markers also explains the poor performance of GFLars (c, d) (which could also be seen in Figure 6), as the current implementations of 2d GFLars do not handle missing values in one of the dimensions.

Not all (c, d) methods outperform (c)-only and (d)-only methods. For example, for data set 1 and 100%, although PSCBS has good performance, it is outperformed by 2 to 5 % by all (c) methods. However, as can be seen in Table 3, there are always several (c, d) approaches among top performers.

5.4 Choosing the appropriate method for a given context

In practice, when analyzing SNP array data, biostatisticians and bioinformaticians will choose one particular method to perform data segmentation. This choice is often *ad hoc* and based on personal experience. Our purpose here is not to make a comparison of all existing segmentation methods, but to compare relevant candidates in different classes of approaches. In the settings that we have considered it seems that RBS (c, d) performs very

well. However, the point of our framework is not to select once and for all a best segmentation tool, but rather to justify the use of one method for one particular type of scenario (cancer type, cellularity, data set). In particular, we make no claim about the performance of RBS for other data sets.

	Data set 1			Data set 2		
Statistic	100%	70%	50%	100%	79%	50%
(c, d)	0.93	0.63	0.22	0.97	0.95	0.75
(c)	0.94	0.64	0.18	0.96	0.89	0.49
(d)	0.35	0.18	0.10	0.71	0.84	0.67

Table 3: Best pAUC across methods for each combination of statistic, data set and percentage of tumor cells.

		Data set 1			Data set 2		
Statistic	Method	100%	70%	50%	100%	79%	50%
(c, d)	PSCBS	0.89	0.60	0.16	0.97	0.88	0.51
	GFLars	0.60	0.42	0.14	0.97	0.91	0.60
	RBS	0.93	0.63	0.22	0.97	0.95	0.75
(c)	CBS	0.92	0.59	0.16	0.91	0.84	0.45
	GFLars	0.94	0.64	0.18	0.96	0.89	0.49
	RBS	0.91	0.62	0.17	0.90	0.84	0.48
	cghseg	0.93	0.61	0.18	0.95	0.88	0.49
(d)	CBS	0.35	0.17	0.10	0.71	0.83	0.64
	GFLars	0.35	0.18	0.10	0.71	0.84	0.66
	RBS	0.34	0.17	0.09	0.69	0.83	0.65
	cghseg	0.35	0.18	0.10	0.70	0.84	0.67

Table 4: pAUC by for each combination of method, statistic, data set and percentage of tumor cells.

5.5 Heterogeneity of breakpoint detection difficulty

An important question when using a biostatistical or bioinformatic tool is to assess its ability to recover events and to know which events they are likely to find and which of them are harder to detect. In Table 3 it can be seen that the pAUC is never at 100%. This is not necessarily surprising as the signal is quite noisy and in fact considering noise level the pAUC is quite high. Figure 7 demonstrates that (as could be expected) missed change-points are those for which we have a low signal to noise ratio (the right panel is darker than the left panel). However, the signal to noise ratio substantially depends on the type of change-point. Typically, in Figure 7 the column corresponding to the (0,2)-(1,2) transition is much darker than that

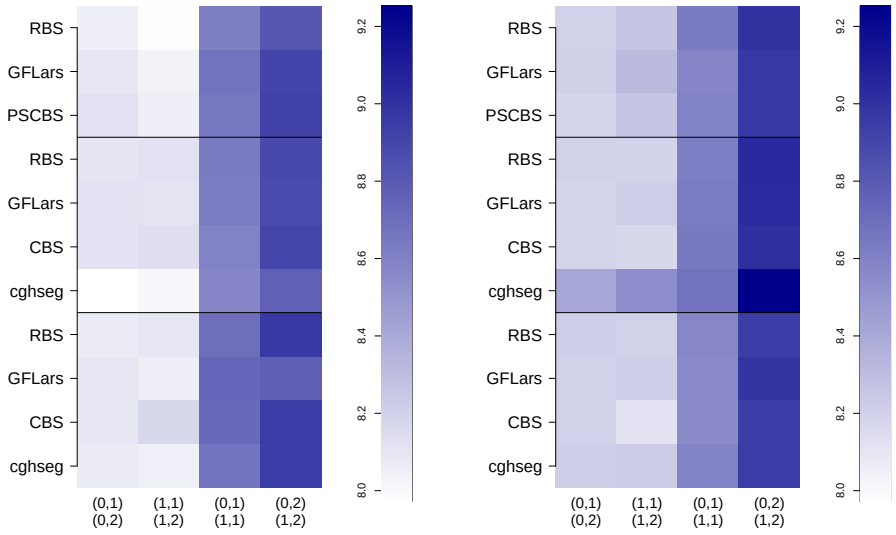

Figure 7: log(SNR) for missed (left) and caught (right) breakpoints for four types of breakpoints on data set 2 with 50% normal cell contamination.

of the (1,1)-(1,2) transition. This is confirmed by Table 5, which indicates that for a high level of normal contamination in data set 2, the proportion of missed (1,1)-(1,2) change-points is greater than 1/2.

6 Summary and discussion

We have developed a framework to assess the performance of various DNA copy number segmentation methods. A critical aspect of this framework is that it generates realistic copy-number profiles by resampling real SNP array data. This allows us to study a large number of scenarios without relying on a particular statistical model. It is our opinion that this framework is simple to use as it depends on few parameters, all of which have a straightforward biological interpretation. An `R` package is available and we believe that our proposed data generation scheme can be used readily as well as applied to other data sets and technologies. It is also possible to extend the set of segmentation methods compared, as explained in the package documentation. In this paper, we illustrated the usage of this framework on two SNP array data sets from Affymetrix and Illumina.

We were able to identify which technological and biological parameters drive the performance of segmentation methods. First, it appears that the percentage of tumor cells in the sample plays a critical role: for a percentage lower than 70%, it is probably hopeless to recover the whole set of breakpoints with a high accuracy. We emphasize the relevance of the considered

Statistic	Method	(0,1)-(0,2)	(1,1)-(1,2)	(0,1)-(1,1)	(0,2)-(1,2)
(c, d)	RBS	**0.40**	**0.47**	**0.32**	0.31
	GFLars	0.51	0.66	0.44	0.34
	PSCBS	0.55	0.63	0.51	0.47
(c)	RBS	0.57	0.69	0.52	0.63
	GFLars	0.54	0.70	0.45	0.58
	CBS	0.59	0.71	0.52	0.62
	cghseg	0.66	0.79	0.55	0.69
(d)	RBS	0.49	0.54	0.39	0.24
	GFLars	0.49	0.51	0.34	**0.20**
	CBS	0.51	0.49	0.41	0.23
	cghseg	0.51	0.51	0.38	0.23

Table 5: Proportion of missed breakpoints by method, statistic and type of copy-number transition (data set 2, 50% of tumor cells).

range of cellularity for applications: we expect that tumor cell lines should be well represented by the 100% setting, while the 50% is not unusual for clinical practice. Second, it seems that different microarray technologies might lead to different performances. Specifically, the ratio between the number of informative allelic probes (heterozygous SNPs) to the total number of probes is a crucial aspect, particularly for a high level of normal contamination. Finally, not all methods achieve similar performance across the scenarios that we have considered. Interestingly, we show that methods that take advantage of both signal dimensions are generally but not always better than those using only one of them. This variability between segmentation methods may be attributed to some extent to the biological and technological contexts, in the sense that some methods might be more adapted to certain scenarios.

Our framework provides a way to critically evaluate the performance of segmentation methods, and therefore to rationally select one or several of them for a particular data set. Such a quantitative assessment is also useful for interpretation. For example, we showed that even in favorable scenarios, performances are not perfect. Furthermore, perhaps unexpectedly, we showed that copy number transitions involving the gain or loss of a single DNA copy are not equally easy to recover, meaning that the proportion of different types of copy number transitions recovered by a particular segmentation method may not be directly interpretable.

7 Acknowledgements

The authors would like to thank Henrik Bengtsson and Cyril Dalmasso for very instructive discussions and feedback. The authors are also grateful

to the three referees whose constructive comments helped to improve the clarity of the paper.

References

[1] Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[2] Neuvial, P., Bengtsson, H., and Speed, T. P. Statistical analysis of single nucleotide polymorphism microarrays in cancer studies. In *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics. Springer, 1st edition, March 2011.

[3] Staaf, J., Lindgren, D., Vallon-Christersson, J., et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol*, 9(9):R136, October 2008.

[4] Bengtsson, H., Neuvial, P., and Speed, T. P. TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC bioinformatics*, 11(1):245, May 2010.

[5] Fridlyand, J. Hidden Markov models approach to the analysis of array-CGH data. *Journal of Multivariate Analysis*, 90(1):132–153, July 2004. ISSN 0047259X. doi:10.1016/j.jmva.2004.02.008. URL `http://linkinghub.elsevier.com/retrieve/pii/S0047259X04000260`.

[6] Sun, W., Wright, F. A., Tang, Z., et al. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucl. Acids Res.*, 37(16):5365–5377, September 2009.

[7] Greenman, C. D., Bignell, G., Butler, A., et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, 11(1):164–175, 2010.

[8] Chen, H., Xing, H., and Zhang, N. R. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Computational Biology*, 7(1):e1001060, 2011.

[9] Picard, F., Robin, S., Lavielle, M., et al. A statistical approach for array-CGH data analysis. *BMC bioinformatics*, 6:27, January 2005. ISSN 1471-2105. doi:10.1186/1471-2105-6-27. URL `http://www.ncbi.nlm.nih.gov/pubmed/15705208`.

[10] Tibshirani, R., Saunders, M., Rosset, S., et al. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 91–108, 2005.

[11] Olshen, A. B., Venkatraman, E. S., Lucito, R., et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[12] Zhang, N. DNA copy number profiling in normal and tumor genomes. In Feng, J., Fu, W., and Sun, F., editors, *Frontiers in Computational and Systems Biology*, pages 259–281. Springer-Verlag, 2010.

[13] Hupé, P., Stransky, N., Thiery, J.-P., et al. Analysis of array-CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, 2004.

[14] Ben-Yaacov, E. and Eldar, Y. C. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):i139—-i145, August 2008.

[15] Willenbrock, H. and Fridlyand, J. A comparison study: applying segmentation to array-CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–91, Nov 2005. doi:10.1093/bioinformatics/bti677.

[16] Mosén-Ansorena, D., Aransay, A., and Rodríguez-Ezpeleta, N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC bioinformatics*, 13(1):192, 2012.

[17] Lai, W. R., Johnson, M. D., Kucherlapati, R., et al. Comparative analysis of algorithms for identifying amplifications and deletions in array-CGH data. *Bioinformatics (Oxford, England)*, 21(19):3763–70, October 2005. ISSN 1367-4803. doi:10.1093/bioinformatics/bti611. URL `http://www.ncbi.nlm.nih.gov/pubmed/16081473`.

[18] Hocking, T., Schleiermacher, G., Janoueix-Lerosey, I., et al. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1):164, 2013.

[19] Zhang, N. R. and Siegmund, D. O. A modified Bayes Information Criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32, 2007.

[20] Rigaill, G., Hocking, T. D., Bach, F., et al. Learning sparse penalties for change-point detection using max margin interval regression. *Proceedings of the 30th International Conference on Machine Learning,* JMLR W&CP, 28(3):172–180, 2013.

[21] Olshen, A. B., Bengtsson, H., Neuvial, P., et al. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, 27(15):2038–2046, Aug 2011.

[22] Rancoita, P. M., Hutter, M., Bertoni, F., et al. An integrated Bayesian analysis of LOH and copy number data. *BMC bioinformatics*, 11(1):321, 2010.

[23] Research Network, T. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, jun 2011.

[24] Rigaill, G. Pruned dynamic programming for optimal multiple change-point detection. Technical report, http://arXiv.org/abs/1004.0887, 2010.

[25] Mosen-Ansorena, D. and Aransay, A. M. Bivariate segmentation of SNP-array data for allele-specific copy number analysis in tumour samples. *BMC Bioinformatics*, 14(1):84, 2013. ISSN 1471-2105. doi:10.1186/1471-2105-14-84. URL `http://www.biomedcentral.com/1471-2105/14/84`.

[26] Venkatraman, E. S. and Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array-CGH data. *Bioinformatics*, 23(6):657–663, March 2007. ISSN 1460-2059 (Electronic). doi: 10.1093/bioinformatics/btl646.

[27] Gey, S. and Lebarbier, E. Using CART to detect multiple change points in the mean for large sample. Technical report, Statistics for Systems Biology research group, 2008.

[28] Harchaoui, Z. and Lévy-Leduc, C. Catching change-points with lasso. *Advances in Neural Information Processing Systems*, 20:161–168, 2008.

[29] Bleakley, K. and Vert, J.-P. The group fused Lasso for multiple change-point detection. Technical report, http://hal.archives-ouvertes.fr/hal-00602121/en, June 2011.

[30] Rasmussen, M., Sundström, M., Göransson Kultima, H., et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol*, 12(10):R108, October 2011.

[31] Edgar, R., Domrachev, M., and Lash, A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[32] Bengtsson, H., Wirapati, P., and Speed, T. P. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, 27(17):2149–2156, 2009.

[33] Illumina, inc. Illumina's Genotyping Data Normalization Methods. White paper, 2006.

[34] Illumina, inc. Beadstudio genotyping module v3.2. User Guide.

[35] Peiffer, D. A., Le, J. M., Steemers, F. J., et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, 16(9):1136–1148, September 2006. doi: 10.1101/gr.5402306. URL `http://dx.doi.org/10.1101/gr.5402306`.

Appendix

A SNP array data sets

A.1 Data set 1

We have worked with a lung cancer data [30], for which raw data is accessible at NCBI GEO database [31], accession GSE29172. DNA from patient-matched lung cancer and blood cell lines `NCI-H1395` and `NCI-BL1395` were mixed to simulate tumor tissue with 30, 50, 70, 100% cancer cells. DNA was analyzed on Affymetrix SNP6.0 microarray. Data were normalized using ASCRMAv2 [32] followed by TumorBoost [4]. For the sake of reproducibility, the `R` scripts that were written to normalize this data set are distributed in the `jointseg` package, together with the normalized data itself.

A.2 Data set 2

We have also worked with a breast cancer data [3], for which raw data is accessible at NCBI GEO database [31], accession GSE11976. DNA from patient-match breast cancer cell line (`HCC1395`) and its match normal `HCC1395BL` were mixed to simulate tumor tissue with 14, 34, 50, 79, 100% cancer cells. DNA was analyzed on Illumina HumanCNV370-Duov1 microarrays. We obtained the BAF-normalized and summarized data as calculated by the Illumina BeadStudio software [33, 34, 35]

A.3 Description of annotated copy-number regions

The list below describes the different copy number states available for data generation. They are labeled as a pair (c_1, c_2), where c_1 corresponds to the minor copy number (the smallest of the two parental copy numbers), and c_2 corresponds to the major copy number (the largest of the two) [2].

(1,1): normal (one copy from each parent)

(0,1): hemizygous deletion (loss of one parental copy)

(0,0): homozygous deletion (loss of both parental copies)

(0,2): copy-neutral LOH (loss of one parental copy and gain of the other)

(0,3): loss of one parental copy and gain of two copies from the other parent)

(1,2): single copy gain

(1,3): unbalanced two-copy gain (gain of two copies from the same parent)

(2,2): balanced two-copy gain (gain of one copy from each parent)

(2,3): three-copy gain (gain of one copy from each parent, and two copies from the other parent)

CN state	(0,1)	(0,2)	(0,3)	(1,1)	(1,2)	(1,3)	(2,2)	(2,3)	(0,0)
Data set 1	22615	24135	25405	21539	19048	20903	27924	31098	0
Data set 2	2492	5484	6545	3196	2746	0	3044	0	838

Table 6: Size of annotated copy-number regions for each of the 2 data sets.

B Reproducing the figures and tables of this paper

```
library(jointseg)
path <- system.file("figures", package="jointseg")
filenames <- list.files(path, pattern="*.R$")
for (filename in filenames) {
  print(filename)
  pathname <- file.path(path, filename)
  source(pathname, local=TRUE)
}
```

The scripts used for the performance evaluation reported in this paper are available in the subdirectory `"eval"` of the `jointseg` package:

```
path <- system.file("eval", package="jointseg")
```

C Session information

```
> sessionInfo()
R version 3.0.2 (2013-09-25)
Platform: x86_64-apple-darwin10.8.0 (64-bi)

locale:
[1] fr_FR.UTF-8/fr_FR.UTF-8/fr_FR.UTF-8/C/fr_FR.UTF-8/fr_FR.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base
```

```
other attached packages:
 [1] PSCBS_0.39.1       DNAcopy_1.36.0     PSCN_1.0.1         MASS_7.3-29
 [5] changepoint_1.1    zoo_1.7-10         cghseg_1.0.1       jointseg_0.5.1
 [9] acnr_0.1.4         R.utils_1.27.5     R.oo_1.15.8        R.methodsS3_1.5.2
[13] matrixStats_0.8.12

loaded via a namespace (and not attached):
[1] grid_3.0.2      lattice_0.20-24 R.cache_0.9.0
```