Finding Recurrent Copy Number Alteration Regions: A Review of Methods

Oscar M. Rueda*,[1,2] and Ramon Diaz-Uriarte*,[1]

[1]*Structural and Computational Biology Programme, Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain;* [2]*Breast Cancer Functional Genomics, Cancer Research UK, Cambridge, UK*

Abstract: Copy number alterations (CNA) in genomic DNA are linked to a variety of human diseases. Although many methods have been developed to analyze data from a single subject, disease-critical genes are more likely to be found in regions that are common or recurrent among diseased subjects. Unfortunately, finding recurrent CNA regions remains a challenge. We review existing methods for the identification of recurrent CNA regions. Methods differ in their working definition of "recurrent region", the type of input data, the statistical and computational methods used to identify recurrence, and the biological considerations they incorporate (which play a role in the identification of "interesting" regions and in the details of null models used to assess statistical significance). Very few approaches use and/or return probabilities, and code is not easily available for several methods. We emphasize that, when analyzing data from complex diseases with significant among-subject heterogeneity, methods should be able to identify CNAs that affect only a subset of subjects. We suggest that finding recurrent CNAs would benefit from clearly specifying the types of pattern to be detected and the intended usage of the regions found (CNA association with disease, CNA effects on gene expression, clustering of subjects). We finish with suggestions for further methodological research.

Keywords: aCGH, copy number alterations, recurrent, common regions.

1. INTRODUCTION

Copy number alterations (CNAs) are changes in the number of copies of DNA in specific regions of the genome, and can vary in size from 1 kb to a complete chromosome arm [1-4]. CNAs have been linked to many different types of disease, such as cancer, HIV acquisition and progression, autoimmune diseases, and Alzheimer and Parkinson's disease [5-10]. Identification of CNAs uses mainly chip- or array-based technologies, such as aCGH arrays (including Agilent, NimbleGen, BAC, and cDNA arrays [11, 12]), and SNP-based arrays [13, 14], as well as sequencing-based approaches [15-18]. Many methods exist for analyzing a single array of CGH (e.g., see references in [19-23]) but location of CNAs in individual samples, however, is only the initial step in the search for disease-critical genes: the regions more likely to harbor disease-critical genes are those that are recurrent or common among diseased individuals or samples (e.g., [12, 24-26]). Recurrent CNAs regions are likely to contain "driver" alterations (functionally important changes in terms of disease initiation or progression), whereas CNAs that are subject-specific would represent "passenger" alterations (random somatic events without pathological relevance) [3, 27]. Finding common or recurrent CNA regions, however, remains a challenge [2], both computationally and conceptually. In this review we discuss the available methods (many developed in the last few years), and some of the reasons why this task is a challenge.

*Address correspondence to these authors at the Breast Cancer Functional Genomics, Cancer Research UK, Cambridge, UK and The Spanish National Cancer Research Centre, Spain; E-mail: **rueda.om@gmail.com**; rdiaz02@gmail.com

On Section 2 we start with a simple definition of what a recurrent CNA region is. We will then elaborate on this basic definition and will examine possible departures from it, as well as different objectives in terms of the exact patterns that researchers might be interested in detecting. In Section 3, we provide a brief overview of each of the existing methods. Then, we highlight common issues relevant to more than one method, and conclude with suggestions for further research.

2. RECURRENT REGIONS: SCENARIOS

Intuitively, the idea of a "recurrent CNA region" seems straightforward. For instance, Rouveirol *et al.* [28] provide the following definition (p. 849): "We define a recurrent region as a sequence of altered probes common to a set of CGH profiles". More generally, we can define a recurrent or common CNA region as a set of contiguous probes (a region) that, as a group, shows a high enough probability (or evidence) of being altered (e.g., gained) in at least some samples or arrays. Scenario I in Fig. (**1**) represents a simple case that fits the above definition: there is a recurrent CNA region that covers probes 1 and 2: probes 1 and 2 are altered (gained, in this case) in all five samples. Scenario I, however, is a very simplified scenario. We will discuss next several additional scenarios, as well as departures from the above definition.

In Scenario II each of the regions affects only a fraction of the subjects (blue region: 40%; red region: 60%). The two regions of this scenario might be detected by many methods if tunable parameters are modified. For instance, some methods (e.g., MAR, CMAR; see Section 3.4) incorporate a frequency parameter, making it straightforward to detect some of these cases. Other methods (e.g., KC-SMART [29]) incorporate a largest FDR or largest p-value parameter so increasing this threshold would allow us to detect more regions

1574-8936/10 $55.00+.00

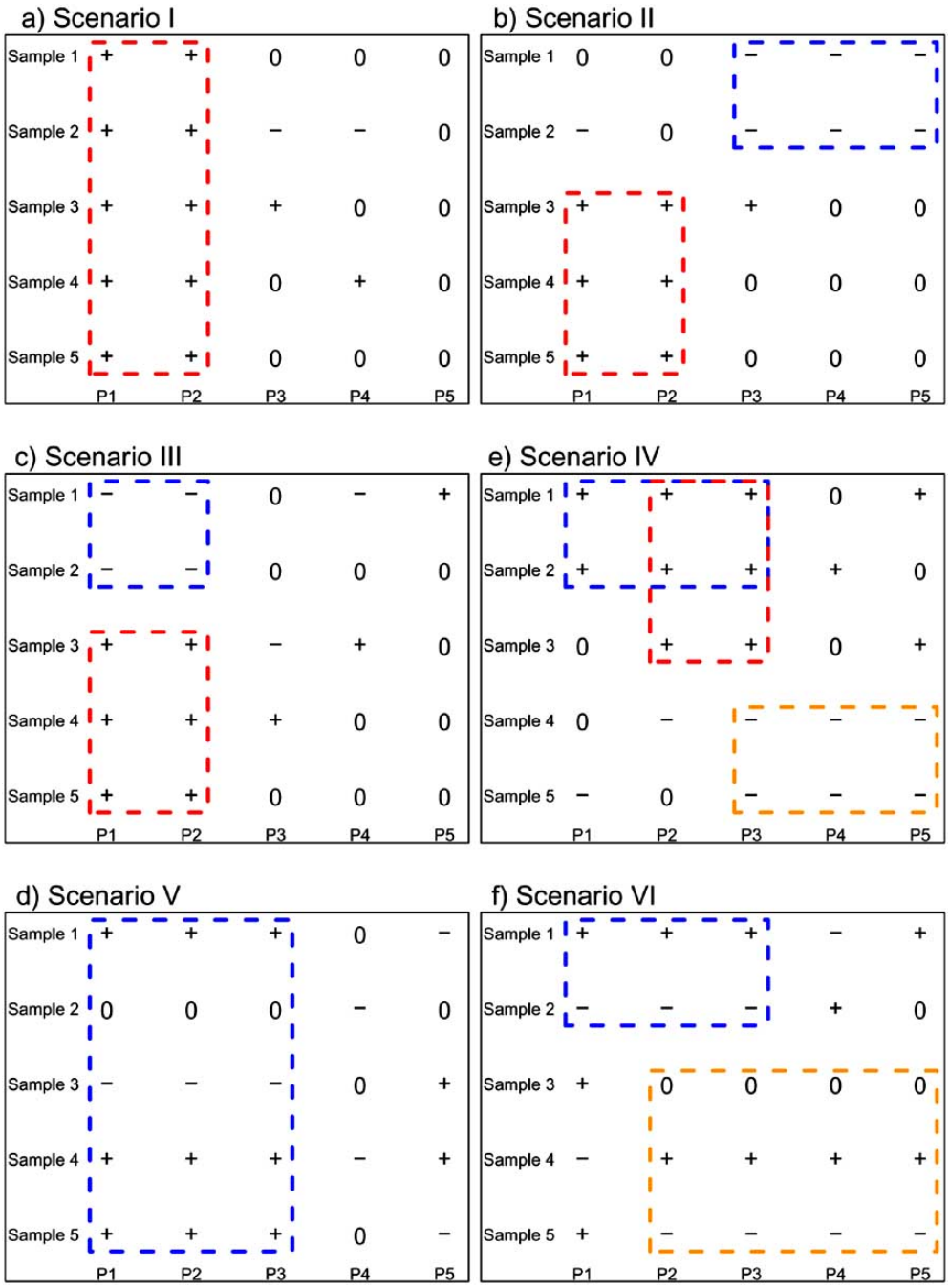

Fig. (1). Six possible scenarios of patterns of recurrent CNA regions. The dotted colored lines enclose a region. Data are shown as segmented data, where a "0" denotes no alteration, a "+" denotes gain, and a "-" denotes loss.

(but might also increase the false positive rate beyond acceptable levels). Regardless of how this pattern is detected, it does represent a case of heterogeneity among subjects.

Scenario III is a hybrid of Scenarios I and II. Only 40% of the samples show a loss, but the regions of gain and loss share the same boundaries. This might not be a most plausible biological scenario, but it is necessary to recognize it as a distinct case. For instance, a method that averages over all probes might not detect any region here as gains and losses could cancel out.

Scenario IV represents a case that only a few methods can detect (but one that might be easily found by biclustering approaches —see Section 4.8): there are three regions, each of which affects only a subset of the individuals, and two of the regions overlap. Properly identifying the red and blue regions requires that we work with recurrent regions and not just sets of recurrent probes (see Section 4.1). pREC-S (see Section 3.13) specifically deals with these type of patterns. The three regions are also regions according to the methods in Rouveirol *et al.* (Section 3.4), but the blue region would not be reported as a minimal region (see Section 3.4).

In scenario V we want to detect a single region. Within the blue region, the pattern of alteration remains constant within sample over contiguous probes, even if some of the

Table 1. Methods Available. log2 Ratios: Either log2 Ratios, as from Two Colour Arrays, or Equivalent Measures (such as log signal intensities and similar values returned from SNP arrays). Smoothed log2 Ratios: the Smoothed, Predicted or fitted log2 Ratio Returned by Some Segmentation Methods. Gains/Losses: Data Reduced to the Values 0, 1, -1, or Equivalent, Denoting no Alteration, Gain, Loss, or Genomic DNA. We make no mention of multiple testing control issues: All methods incorporate some form of control, usually via FDR or bonferroni.

Name	Input	Output (Significance)	Null Model (for Significance)
CGHregions	Gains/Losses	None	None
Master HMMs	log2 ratios	Probabilities of alteration for each probe	Homogeneous Hidden Markov Model
cghMCR	Smoothed log2 ratios	None	None
MAR / CMAR	Gains/Losses	None	None
GEAR	log2 ratios	p-values	Permutation of the alterations over the entire genome
KC-SMART	log2 ratios	p-values	Permutation of the log-ratios over the entire genome
STAC	Gains/Losses	Confidence for regions	Permutation of the regions within chromosomes
MSA	log2 ratios	p-values	Permutation of the regions within chromosomes
GISTIC	Gains/Losses	p-values	Permutation of the probes over the entire genome
RAE	log2 ratios	p-values	Permutation of copy number values using hotspots information
Interval Scores	log2 ratios	Scores for each interval	Large deviation bound for iid Gaussian data
CoCoA	Gains/Losses	Scores for each interval	Binomial distribution on probes and intervals
BSA	log2 ratios	Bayes Factors	Bayesian hierarchical model
pREC-A	log2 ratios	Probabilities of alteration for each region	Non-Homogeneous Hidden Markov Model
pREC-S	log2 ratios	Probabilities of alteration for each region	Non-Homogeneous Hidden Markov Model

samples show gains (Samples 1, 4, 5), some losses (Sample 3), and some no alterations (Sample 2). The key to understanding the difference between Scenario V and Scenario III is to focus on the detection, in V, of one single region that includes all samples. Therefore, the blue region in Scenario V does not fulfill the definition of recurrent region above.

Only one method, CGHregions (see Section 3.1), has been specifically developed for Scenario V. This method uses another definition of "common"that refers to a contiguous set of probes that, within sample, remains (almost) constant (see also Section 3.1). To further understand Scenario V and what CGHregions attempts to capture (see also Section 3.1), instead of focusing on the actual gains/losses, we can focus on the difference in the state of two successive probes. For each of the five samples, the differences are zero between P1 and P2 and between P2 and P3, but the differences are not zero between P3 and P4. Thus, the blue rectangle delimits a region of homogeneous behavior between probes P1 and P3.

After seeing Scenario V, Scenario III can be considered another instance of the pattern shown in Scenario V, if we are not interested in differentiating between samples with amplifications and samples with deletions. In fact, for Scenario III CGHregions will report a common region in probes P1 and P2, since there is no change, within-sample, in the state of the probes in those two locations. Most other methods, in contrast, will do two passes over the data: one for gains and one for losses, as they focus on the actual type of alteration. Thus, with most other methods we will obtain "recurrent CNA region with loss of DNA" (or "recurrent deletion") for samples 1 and 2 and "recurrent CNA region with gain of DNA" (or "recurrent amplification") for samples 3 to 5. This is why Scenario III might represent a problem for some methods: only 40% of the samples have a deletion.

Scenario VI is an extension of Scenario V where we allow for the existence of subsets ofj subjects with different boundaries and regions. No existing method is designed to capture the patterns of Scenario VI.

3. OVERVIEW AND DETAILS OF EXISTING METHODS

Some of the main features of existing methods are summarized in Table **1** and **2** and in Fig. (**2**). In this section we review each method in turn, providing further details and pointing out potential problems and limitations. For practical reasons, we focus mainly on methods with available code. Issues common to several methods are discussed in Section 4.

3.1. CGHregions [30]

As mentioned before (Section 2), this method is designed to capture patterns such as those in Scenario V of Fig. (**1**), and has therefore been considered a dimension reduction approach. The authors clearly state (p. 56, [30]) "Note that

Table 2. Software Available

Name	Availability	Operating System and other Dependencies	License
CGHregions	R/BioConductor package http://www.bioconductor.org	R dependent	GPL 2
Master HMMs	MATLAB toolbox http://www.cs.ubc.ca/~sshah/acgh/CNA-HMMer-v0.1.zip	MATLAB dependent	GNU General Public License
cghMCR	R/BioConductor package http://www.bioconductor.org	R dependent	GPL 2
MAR, CMAR	From the authors upon request Also part of VAMP and CAPweb programs (see text)		
GEAR	Standalone application http://www.systemsbiology.co.kr/GEAR/	Windows	Copyright stated in the setup program
KC-SMART	R/BioConductor package and standalone application based on Matlab Component Runtime http://www.bioconductor.org http://bioinformatics.nki.nl/~klijn/	R or MATLAB dependent	GPL 2 in the case of R package
STAC	Standalone Java application http://cbil.upenn.edu/STAC/	Multiplatform	Unknown
MSA	Standalone Java application or as part of GenePattern http://www.cbil.upenn.edu/MSA/	Multiplatform	Unknown
GISTIC	Standalone based on Matlab (Component Runtime version 7.7 needed) http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=162	Linux 64-bit or as part of Gene Pattern	Unknown
RAE	R script with a standalone wrapper http://cbio.mskcc.org/downloads/rae/	Linux for the wrapper.	GPL 2
Interval Scores	Stepgram, CNVDetector http://bioinfo.cs.technion.ac.il/stepgram http://www.csie.ntu.edu.tw/~kmchao/tools/CNVDetector/	Windows	Unknown
CoCoA	None		
BSA	R scripts http://www.mshri.on.ca/mitacs/software/SOFTWARE.HTML	R dependent	Shareware
pREC-A	R package RJaCGH http://cran.r-project.org	R dependent	GPL 2
pREC-S	R package RJaCGH http://cran.r-project.org	R dependent	GPL 2

we do not require the clones in a region to be constant *across* samples." (italics in original). Each "region" identified by this method is a collection of rows (clones) in the matrix of segmented data organized as clones by subjects. Thus a region can be used to summarize the data, as it captures a pattern that remains (almost) constant over several (many) contiguous clones. But, for any probe, some of the samples might present a gain, some others a loss, and some others might show no alteration. Thus, the "regions" identified do not represent recurrent or common patterns of copy number alteration over subjects (i.e., the copy number or copy number state need not be common to all, or even most, of the subjects).

3.2. Master HMMs [31]

In [31] a single-subject Hidden Markov Model (HMM) is extended to simultaneously model several subjects: a "master" sequence captures the common or recurrent pattern over subjects. Specific individual deviations from the master sequence are modeled in several different possible ways, introducing private and undefined state sequences. The HMMs, however, are all restricted to three hidden states (plus an "unidentified" state in one type of model); using only three hidden states, to represent just the states "loss", "neutral", "gain", is a questionable decision [21, 32]. This approach has also been criticized because it "contains a biologically irrelevant tuning parameter" (p. 1670 in [32]).

A recurrent alteration identified by this approach is "(...) a CNA found at the same location in multiple samples" (see p. i450; also p. 348 in [3]); thus, the authors identify recurrent probes, but do not address the identification of recurrent regions (see also Section 4.1). The authors also state that their approach cannot identify subgroups, although their method has been extended to investigate this problem (see [33]).

3.3. cghMCR [25]

Using segmented (i.e., smoothed data), this algorithm [25] first identifies altered segments within subject (those above the 97th or below the 3rd perdentile of the data) and next joins adjacent segments separated by less than 500 kb. Then, the algorithm identifies Minimal Common Regions, defined as "contiguous spans having at least 75% of the peak recurrence as calculated by counting the occurrence of highly altered segments. If two MCRs are separated by a gap of only one probe position they are joined." (p. 9068). When measuring recurrence, a sample will count as having the alteration in the altered segment if its smoothed ratio is larger (smaller) than 0.13 (-0.13). To provide further biological information, the authors prioritize the MCRs based on the recurrence of high-amplitude alterations (p. 9069). This paper was one of the first to attempt to identify recurrent regions of alteration. It addresses the problems inherent in the structural complexity of many copy number alterations by considering how to define boundaries and joining contiguous segments, as well as emphasizing the potential relevance of high-amplitude alterations. The results of this approach, however, seem to depend strongly on parameters such as the gap to join segments (500 kb by default); moreover, it is common for this method to identify common regions that do not correspond to any regions of gain/loss found by individual-sample segmentation methods (personal observation).

3.4. MAR, CMAR [28]

Rouveirol *et al*. "(...) define a recurrent region as a sequence of altered probes common to a set of CGH profiles and a minimal recurrent region as a recurrent region that contains no smaller recurrent regions." (p. 849 in [28]). The authors then formalize these definitions and develop two algorithms, MAR and CMAR, for finding the minimal common regions using segmented data, not the original ratio data (see also Section 4.2 —an open question is whether some of the ideas formalized in this paper could be extended to smoothed data or to probabilities). This approach can detect regions that affect only a small fraction of the subjects (see, e.g., p. 854 in [28]). The description of recurrent regions given in the paper covers Scenarios I to IV (in Scenario V, a region of gain would be detected as it affects three samples). However, in Scenario IV, only the red and orange regions would be reported as minimal regions. In [28] terminology, the red region is a "closed subsequence" of the blue region, and thus the blue region cannot be a minimal region (see p. 851 and examples 3 and 4 in [28]).

This is a rigorous attempt to define and detect common regions, but the the paper is hard to follow. One reason is the usage of an unjustifiably complex formalization and terminology. A second reason is that the explanation of what is being searched for (the type of regions and why) is tangled with the algorithmic solutions (how to find the regions in a computationally efficient way). However, the methods are rich and flexible and can capture a variety of patterns, incorporating several additional user-specified constraints (number of samples that share a region, size of region, whether the region contains —or not— a specified observation, and how different a region is from surrounding probes —the "well bounded" criterion). In fact, these restrictions are likely to capture the biologically motivated ideas found in more recent methods such as GISTIC (see Section 3.8) and RAE (Section 3.9).

Code is not easily available from public repositories. It is part of the VAMP [34] program and has also been incorporated in CAPweb [35]. The VAMP implementation, in Java, can be requested from http: //bioinfo-out.curie.fr/projects/vamp/. There might soon be another version available from http: //eric.voirin.-free.fr/regions/.

3.5. GEAR [36]

GEAR [36] implements several methods. The individual clone-based method uses as working definition of recurrent that a given alteration be shared by more than a pre-specified proportion of samples (frequency cuttoff) or be more frequent than expected by chance (p-value cutoff) under a null model where observed alteration frequencies are position independent and constant over the genome. This approach is not suited to detect regions over unknown subsets of samples.

Alternatively, GEAR allows us to use a modified version of the SW-ARRAY method [37]: instead of analyzing the ratios of an array, GEAR applies SW-ARRAY to the mean (or the scaled mean) of the ratios over all samples. The possible advantage of this approach is that SW-ARRAY is designed to detect contiguous regions, but see Section 4.1.

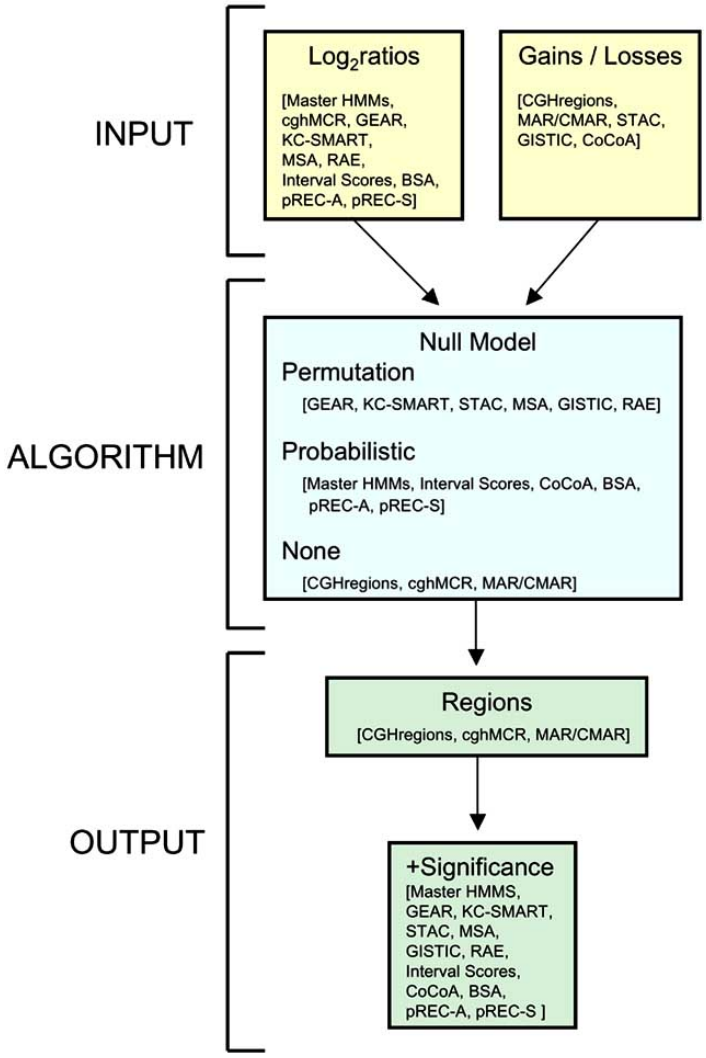

Fig. (2). Relationships between methods and flow-chart of main procedures.

Moreover, dealing with means precludes detecting aberrations common only to a small subset of samples.

GEAR has a nice and user-friendly interface but, unfortunately, it is only available for Microsoft Windows operating systems.

3.6. KC-SMART [29]

This is another method that uses a form of weighted average of amplitude of alteration by frequency over subjects to call a gain (or a loss) recurrent across an entire tumor set. The basic approach is straightforward: the positive and negative ratios are summed (separately) across tumors for each clone, and a kernel estimate of the density of this summation is determined. The kernel function used (flat top Gaussian) is based on the assumption that nearby probes provide more information than distant ones, and accounts for unequal distances between probes. To identify "relevant" peaks in that density, a permutation test (with Bonferroni correction for multiple testing) is used: first, ratios are randomly shuffled within tumor; next, for each permutation, positive and negative ratios are summed over tumors for each location, and the kernel density determined again; finally, the peaks from the observed data are compared to those from the kernel density estimates of the randomly shuffled data. By construction, this method is not suited to identify recurrent regions that affect only a small subset of subjects.

The user needs to specify a significance level, and it is necessary to use several kernel widths to detect both high-amplitude alterations over a small region and low-amplitude alterations that span a large region. According to the authors, the usage of several kernel widths facilitates the analysis of complex aberrations (p. 13 in [29]).

3.7. STAC [24] and MSA [38]

STAC [24] and MSA [38] are two closely related methods. STAC was developed first, and MSA can be considered an improvement over STAC. STAC used as input segmented data, and considered both the frequency of an aberration (or the frequency of a stretched of altered probes) and its "footprint" (the number of locations c such that c is contained in some interval of a set of intervals over samples; see p. 3 in [24]; or the length of the projection of a set of intervals onto the genome, see p. 1466 in [38]). The intuitive notion behind footprints is that smaller footprints are less likely to arise by chance, and thus such a tight alignment of aberrations might indicate the presence of critical genes. MSA [38] builds upon

Table 3. Scenarios (as depicted in Fig. 1) detected by each method. For scenarios II and III, many of the methods listed could detect them, provided the appropriate thresholds are modified. The entries in the table represent what seems to be the canonical or standard procedure of a method. See discussion in section 2. entries with a "? " are discussed further in the description of each method

	Scenarios					
Method	I	II	III	IV	V	VI
CGHregions	x		x		x	
Master-HMM	x	x	x			
cghMCR	x					
MAR / CMAR	x	x	x	?		
GEAR	x	x	x			
KC-SMART	x					
STAC	x	x	x			
MSA	x	x	x	?		
GISTIC	x					
RAE	x					
Interval Scores	x	x	x	?		
CoCoA	x	x	x	?		
BSA	x					
pREC-A	x	x	x			
pREC-S	x	x	x	x		

the notions of frequency and footprint but extends the method. First, MSA uses the original ratio data, not previously segmented data, by searching over a set of possible cutoff values. Second, several algorithmic and heuristic enhancements increase considerably the execution speed of MSA. In what follows, we focus on MSA.

In the canonical implementation, MSA (and STAC) use permutations of the entire regions within chromosomes (instead of over the complete genome) to assess significance in patterns. This permutation scheme might preclude detecting large aberrations (see also Section 4.5). Although MSA uses the original ratios (not the segmented data, as STAC), for each probe it uses a common threshold over all arrays and thus ignores possible differences in variability between arrays.

The actual size and type of region found by MSA is not clear. Although it is not explicit in the paper (but see documentation, http: //www.cbil.upenn.edu/MSA/doc/MSADoc.doc) the user of their program needs to specify the "binParam", defined as "number of positions per bin". In other words, each chromosome is divided in a set of consecutive bins of predetermined size. A bin is regarded as altered if a single probe within the bin is altered –personal observation. In the permutations tests, entire within-sample intervals (where each interval spans one or more bins) are randomly placed in another location, so within-individual intervals are not broken up; (see Fig. **1**) in [38]. However, the patterns of recurrence are reported per bin, not per interval. In this sense, MSA is finding "common bins", not "common regions" (see Section 4.1). In terms of scenarios, MSA should detect scenarios I, II, III, with the caveat that we might find different regions if we alter the "binParam". As Scenario IV requires common regions, not just common probes, MSA cannot really detect this type of patterns (see Section 4.1).

3.8. GISTIC [27, 39]

This method aggregates data over different tumors to differentiate between driver and passenger aberrations. Somewhat similar to RAE (see next), the method explicitly tries to identify "driver aberrations", aberrations that "rise above the background rate of random passenger aberrations" (see also Section 4.5). This method involves three main steps: first, data-preprocessing and identification of copy number alterations tumor by tumor; second, data aggregation over tumors (computation of *G*-score and permutation test); third, identification of "peak regions".

The authors use SNP arrays, and include several initial steps designed to minimize the effects of systematic and random errors in the accuracy with which aberrations are identified, but the key elements of their approach can be used with any platform. In the description of the paper, the data are first segmented to obtain smoothed means (the authors originally used GLAD [40]), and a common threshold applied so that smoothed values below the threshold are not regarded as altered). In the current software implementation (ftp: //ftp.broadinstitute.org/pub/genepattern/modules_public_server_doc/GISTIC.pdf) the user must input smoothed data and the common threshold). Next, very small segments (less than four probes) or datasets with high noise (lack of separate peaks) are discarded. The aggregation step uses a single statistic (*G*-score) that combines prevalence and amplitude: the authors explicitly assume that "(...) prevalence and average amplitude of these events independently indicate the likelihood with which a region is affected by such driver aberra-

tions" (Supplementary Information text in [27]). Their combined score is the prevalence of the copy-number change times the average amplitude. The significance of the observed *G*-scores is evaluated with a semi-exact approximation to a permutation test (see Section 4.5). Using the significant locations identified in the previous step, the authors finally try to find the most likely locations of the oncogenes and tumor suppressor genes, by incorporating several biological considerations: "peak regions" with maximal *G*-scores and minimal p-values are selected (thus focusing only on regions "most frequently aberrant to the highest degree" [27]); independent peaks (peaks which are independently aberrant) are recaptured via a "peel-off" algorithm; boundaries of peak regions are recomputed to eliminate shifts from random passenger mutations; focal aberrations are distinguished from broad ones (those that affect more than half a chromosome arm).

This method seems, initially, a rather complex one. However, biological considerations and assumptions enter mainly in steps first and third, with the second step being statistically very straightforward. The main limitations of the method are the computation of the *G*-score: it does not take into account inter-array variability (as it is simply the average amplitude of an aberration times its frequency), and equates amplitude with strength of evidence of alteration (see also section 4.2). In addition, in the original paper [27] segmentation is performed using GLAD [40]: GLAD has been shown to perform worse than several alternative segmentation approaches [19-21], and require tunning of several parameters of non-intuitive meaning (but GLAD is one of the few segmentation methods, together with RJaCGH [21] and ACE [41], that explicitly attempts to classify regions as gained, lost, or not-altered, although this feature of GLAD is not used in GISTIC —see Supplementary Information text to [27], under "Identification of Copy-Number Aberrations"). GISTIC is not designed to detect regions of alteration common only to a small subset of subjects.

3.9. RAE [42]

RAE [42] starts from an initial copy number assessment from a segmentation procedure (CBS [43, 44] in the canonical procedure) and tries to identify "genomic regions of interest". RAE uses individual tumor noise models instead of a single global threshold to deal with reliability in the detection of copy number alterations. (The authors emphasize "soft thresholding" for making more robust assessments of alterations in noisy systems; but it seems to us that this procedure just falls short of providing a probability assessment, which also avoids making a discretized, 0/1, call —see [21]–, with the advantage that the probability assessment does not need to regard as equivalent amplitude and strength of evidence of alteration; see Section 4.3).

For RAE [42], the resolution of genomic regions of interest is targeted towards identifying "(...) manageable and interpretable events, perhaps involving a single gene." (p. 6, [42]); this objective strongly affects the rest of the procedure. Assessment of common regions is done initially through an average across samples that leads to a summary score. The significance of the summary score is then evaluated via a complex permutation test (see Section 4.5). Finally, boundaries for regions of interest are located, incorporating notions of spatial and measurement imprecission; the end result should be the location of biologically relevant recurrent regions of alteration common to all subjects in the study (the "manageable and interpretable events, perhaps involving a single gene", mentioned above).

We find that, in contrast to many of the other methods, the biological assumptions and the statistical and computational approaches are too closely intertwined, which results in a complex method (see also Section 4.5) that can be hard to understand. This is further complicated because the method introduces several terms (e.g., unified breakpoint, genomic regions of interest, peak threshold) that seem crucial in the development but are rarely succintly defined. Moreover, it is unclear how changes in the assumptions or in the research questions (e.g., trying to detect recurrent copy numbers that affect more than a single gene; encoding gains with more components than "single-copy gain" and "amplification"; changing the null model for the permutations) could be incorporated in this method. However, it might be precisely the tightly integrated biology + statistics that could make this method attractive, if the biological assumptions make sense to the researcher.

3.10. Interval scores [45] and CoCoA [46]

These two approaches are closely related, and developed by the same research group. Both methods assume that the observed ratios are independently (and identically) distributed across the chromosome [31], a biologically unrealistic assumption. In addition, the procedure in [45] has been criticized because "it relies in unntested parametric assumptions and does not make multiple testing considerations" (p. 8 in [24]). Both methods are capable of detecting patterns of aberration over subsets of subjects, when neither the region nor the subsets are pre-specified (e.g., equations 6 and 7 and section 5.2 in [45]). But it is unclear whether patterns such as those in Scenario IV can be directly distinguished: "An interesting aspect of the problem, which we did not attempt to address here, is the separation and visualization of different located aberrations, many of which contain significant intersections" (Section 5.2 in [45]). CoCoA [46] provides, as output, probabilistic scores (see pp. 127 and 128 in [46]) and carefully deals with the preservation of within-sample integrity of patterns (see also Section 4.1). No code or program seems available for the methods in [46]. The method in [45] has been implemented in "Stepgram" (http: //bioinfo.cs.technion.ac.il/-stepgram/). It is also available, with fewer assumptions on noise distribution (but, apparently, without the option for "class discovery" —location of regions over subsets of samples), in [47]. However, these two programs are only distributed as Windows executables, and source code in not available.

3.11. BSA [32]

This method carries out segmentation and assignment of copy number status simultaneously using a hierarchical Bayesian model. Its goal is to detect a set of signal regions and differentiate if from the background region (the collective region with no copy number changes). The algorithm performs a sequential segmentation based on a combination of marginal likelihood and Bayes factors to evaluate each candidate segment.

The authors show through simulations that this method "outperforms other segmentation methods in terms of accuracy and power for both breakpoint detection and segmentation for recurrent CNAs using multiple samples" (p. 1674), although they do not compare it with methods specifically designed for common regions, as the ones in this review.

This method should detect patterns such as those of Scenario I, although it should be possible to detect some individual deviations from them, by using the Bayes factors: "this Bayesian approach can assign each individual a posterior probability to have a CNV at a given intensity level" (p. 1678). In general, however, to deal with more complex scenarios ("(...) complex architecture including smaller CNVs contained within larger ones, CNVs with interindividual breakpoints variation or CNVs with juxtaposed duplications and deletions within the same individuals.") the authors advise against a joint analysis of all samples (p. 1678): "Since these CNVs with complex architectures are of great individuality, joint analysis using all samples will not be effective, because the individuality will be lost when information from multiple samples is aggregated. Applications of single sample methods such as CBS or our Bayesian approach by setting the sample size equal to one may be helpful."

3.12. pREC-A [48]

This method is also based on a Hidden Markov Model (HMM). Given a set of samples, it locates all regions with an average (average over all arrays) probability of alteration larger than a user selected threshold. The probabilities refer to the joint probability of alterations of the probes within a region. Therefore, this algorithm is designed for Scenario I and, by choosing a smaller threshold, Scenarios II and III. This is one of the few algorithms that uses probabilities as input; moreover, this algorithm has just a single parameter (the threshold probability) of immediate interpretation.

3.13. pREC-S [48]

This algorithm is also based on a HMM. It locates all regions shared by at least *freq.array* arrays or samples given that each region in each array has a probability of alteration of at least . Note that there are two parameters here, but again both have immediate interpretation. pREC-S is specially suited to detect patterns such as those in Scenario IV (and will also detect Scenarios I, II, III). As pREC-A, this method deals with common regions, not probes, since the probability used is the joint probability of alteration of the complete region.

3.14. Related approaches

As explained in the Introduction, this review focuses on methods that try to locate recurrent regions, *de novo*, from a set of arrays. There are other methods in the literature with different objectives that, however, present partial overlap with the location of recurrent CNA regions. We discuss them here briefly.

Liu and collaborators, in two papers [49, 50], focus on the problem of clustering subjects using aCGH data. In the process of clustering, markers that characterize subsets of samples are found. Note, however, that the methods do not identify regions of alteration, but rather markers that are defined by a (single) position (see p. 451 in [49], where it is stated "Each marker is represented by two numbers <p, q>, where p and q denote the position and the type of aberration, respectively"). Thus, whereas these markers might be relevant when the focus is only clustering subjects, these markers do not satisfy the idea of a "recurrent region", or "recurrent set of contiguous probes".

Some authors [51, 52] have tried to use pre-existing information about regions that show copy number polymorphism to improve the search for rare copy number variants. Location and definition of regions, however, is not a result of the usage of these methods. The delimitation of copy number variation (CNV) regions from normal samples has been carried out carefully by Komura *et al.* [53], who describe the method used to produce the first global map of copy number variation in the human population using the HapMap data [5]. A modification of SW-ARRAY [37] is used on all possible pairs of samples to extract candidate CNV regions from each sample; SNP information and signal ratios are integrated to better define boundaries and copy number; finally, diploid samples are defined, (using a maximum clique algorithm) for all regions and precise boundaries and copy numbers estimated. The details of this method are highly specific for one platform (the 500K EA SNP array; but should be applicable to other 500K Affymetrix arrays). They incorporate SNP and intensity information (again, making them specific to SNP-based date) and have been developed for normal samples. It is unclear if some of the approaches used here could be adapted to other platforms to detect complex patterns (such as those in Scenario IV) of CNA regions.

There has also been some work, most notably that of [54-56] on the joint estimation of copy number alterations, so that instead of analyzing one array at a time, we use all arrays simultaneously to improve the detection of alterations by using information across samples. In GADA [54], normalization and estimation of copy number are carried out simultaneously, and complex patterns of polymorphic copy number variation can be dealt with (p. 1230), but finding common CNA regions is not an objective of the method, *per se*. Similarly, Engler *et al.* [56] analyze simultaneously several arrays, but estimated parameters seem to be per-array (e.g., equation 4.5 in [56]), no systematic procedure is given to identify breakpoints or regions (p. 407) and users are to use visual inspection of plots. LaFramboise *et al*. [55] use a multi-sample rank-based approach to improve detection of boundaries and provide a working definition of region based on changes in p-values (but note that very common aberrations might be difficult to detect; see p. 727 in [55]). The final results are always per array, not over the complete set of arrays.

In fact, the later is one defining characteristic of most of these methods: many samples are analyzed simultaneously yielding improved performance (in terms of segmentation, delimitation of boundaries or breakpoints, etc) compared to sample-by-sample analysis. However, the results (and, consequently, output) is still per-array. In other words, the input data of dimensions number of arrays by number of probes is mapped into another matrix (or two matrices, one for gains, one for losses) of the same dimensions where entries are probabilities, p-values, or calls. But there is no decrease in the dimension "number of arrays", so there is no notion of

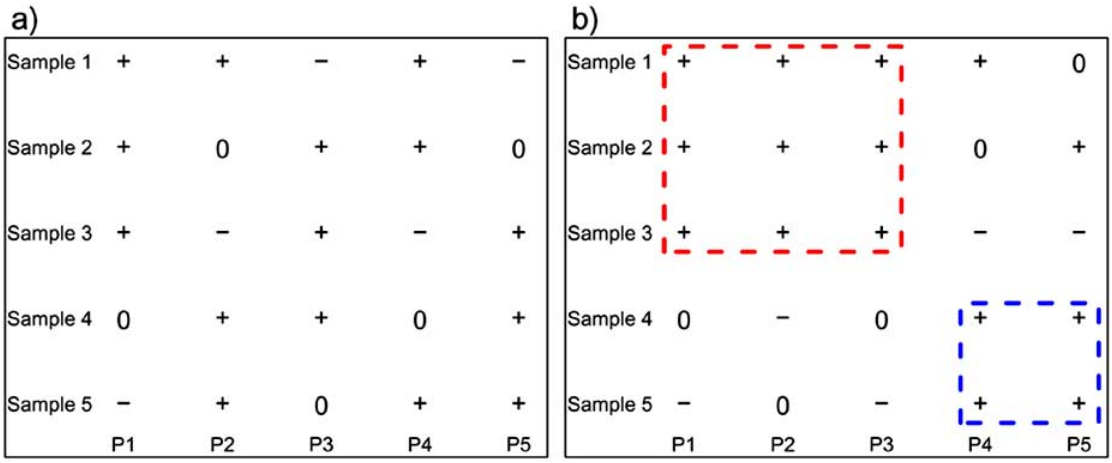

Fig. (3). Comparison of "probe-by-probe" vs. "region" approach (see Section 4.1). Meaning of symbols as in Fig. (**1**).

"common": for each probe there are still "number of arrays" entries. With these methods, therefore, if the user analyzes, say, 100 arrays, it is still completely up to the user to then define or find whatever is common over those 100 arrays (for instance, by post-processing the output with other techniques, maybe some of the ones discussed previously). In summary, these are segmentation methods that use inter-array information to improve the intra-array segmentation, without returning any common patterns. (Note that several of the methods we have discussed above, such as M-HMM [Section 3.2], MSA [Section 3.7], KC-SMART [Section 3.6], BSA [Section 3.11]) also use information across arrays in what could be considered segmentation steps, but these methods do return some kind of statistic over all arrays that is designed to measure the degree of recurrence).

QuantiSNP [57] is another HMM method that can also borrow information across samples to improve the resolution of boundaries of copy number change. The scenarios covered would include those in I, II, III. The Bayes Factors reported (see equation 12 in [57]) are specific for a region; however, it is unclear whether the multi-sample inferences refer to shared regions or to shared probes (see Fig. **3** in [57]). This method is limited to SNP-based arrays. The recent approach by Wang *et al*. [58] also uses a HMM to locate probes that are lost or duplicated with high frequency compared to the rest of the genome, or that present a rate of loss/duplication that differs between pre-specified subsets of samples (see p. 11 in [58]). As with M-HMM [31], there is no notion of region, and this paper is highly specific for the Illumina platform. These two methods, thus, present the information per probe, not per probe times per array (e.g., see equations 7 and 8 in [58]) and, therefore, are very similar to the method in M-HMM [31] (Section 3.2).

4. COMMON ISSUES

4.1. Recurrent Regions or Recurrent Probes?

Not all approaches try to locate regions, but some methods instead try only to locate common probes, without any notion of region. There are several reasons to try to locate common regions, and not just common probes: locating regions facilitates summarization of information [30], can improve the power of tests of association between CNAs and disease [59-61] as well as the integration of gene expression data [62, 63] (see also Section 4.9), and seems biologically reasonable, since most genes are interrogated by several successive probes (so identifying a single isolated recurrent probe might not be biologically relevant).

Sometimes a recurrent region is taken to be a set of consecutive recurrent probes. But a set of recurrent probes does not necessarily represent a recurrent region. Fig. (**3**) illustrates this point. On both Fig. (**3a** and **3b**) each of the five probes is altered (gained) on 60% of the samples. However, there is no single common region on Fig. (**3a**), whereas there are two common regions on Fig. (**3b**); the red region is common to 60% of the samples, whereas the blue one is common to 40% of the samples. The key difference between panels a) and b) is that panel b) shows patterns of "joint alteration", but the patterns of joint alteration cannot be recovered from the overall (marginal) frequency of alteration of each of the individual probes. The overall, marginal frequencies, are unlikely to preserve the within-sample integrity. Similar examples can be constructed with smoothed data or probabilities instead of calls. (In fact, when considering probabilities, the situation is even more complicated, since just for a single sample or array, the joint probability that a set of probes is altered is rarely the same as any simple function of the marginal probability of alteration of each probe (see discussion in [48]).

This also explains why detection of regions blue and red in Scenario IV requires common regions and not just common probes. Unless we preserve the within-sample integrity (see also [46]), all we can tell is that P1 has a frequency of amplification of 40% and P2 and P3 of 60%, but that is not enough to recognize the two partially overlapping regions.

Finding common probes is much simple than finding common regions. If we are using segmented data (i.e., data that have been classified as either altered or non-altered, or gained/lost/not-changed —see also Section 4.2—), locating a common probe could be as immediate as identifying any probe that is altered (gained or lost) in more than a pre-specified fraction of the data (and, for example, this is what one of the approaches in GEAR does). In Fig. (**3a**), for instance, each of the five probes is recurrent or common if our threshold for recurrent is set to 60% (or less).

Only a few of the methods we have reviewed return "regions" in the sense of preserving the within-sample integrity. CoCoA does, and its authors discuss this issue (see p. 133 in [46]). The methods in MAR and CMAR do keep track of the shapes of the rectangles and the changes in individual patterns (see, for instance, the definitions of "extension", "closure", and "well bounded" in [28]). Any method based on biclustering (see Section 4.8) should also preserve the within-sample patterns. pREC-S and pREC-A use joint probabilities of alteration of a set of successive probes within sample, and work with regions. CGHregions also uses regions as it emphasizes dimension reduction based on homogeneity within sample. BSA tries to locate common regions, called "signal regions", as can be seen from (Fig. **1**) and equation 1 in [32].

In contrast, Master-HMMs searches for recurrent probes: the authors state that "A recurrent CNA in a cohort of patients is a CNA found at the same location in multiple samples". Thus, there is no notion of region or sequence of contiguous probes. Methods that use weighted averages of intensity, such as KC-SMART are also likely looking just for probes and can suffer from the problems in Fig. (**3a**): the smoothed density estimate in KC-SMART uses the sum of values over all samples (i.e., within-sample integrity of a pattern cannot be preserved; see equation 2 and Fig. (**1**) in [29]). In other words, regions are obtained from averages over all samples. Similar comments apply to GEAR: the second method in GEAR uses SW-ARRAY on the average probe intensity over all arrays; the first method in GEAR directly thresholds each individual probe. GISTIC uses smoothed values, where the smoothing is based on a previous segmentation. The smoothed values, themselves, carry over the notion of segment. However, since the statistic G is computed probe by probe, problems similar to those of Fig. (**3a**) can occur.

MSA does not directly use probes, but rather "bins" of probes. The method returns, as output, confidences (1 - p-value) for each bin. The method also provides single sample calls, which are obtained by finding "the tightest multiple sample concordance" (p. 1471 in [38]), roughly equivalent to finding the samples that provide maximal evidence for the confidences of each of those bins. It is not clear if the problems in Fig. (**3a**) are present, because the permutation tests that lead to the confidences (and, thus, also to the single sample calls) are based on permuting the location of entire within-sample intervals, not of individual bins. However, the method basically yields, as output, a matrix of 0s and 1s of dimension number of samples by number of bins, not a list of regions and samples in each region (actually, the method returns the equivalent of two matrices, one for gains and one for losses). Now, it is up to the user to define and find common regions in these two matrices (maybe by applying some other technique, such as biclustering; see Section 4.8). Moreover, in addition to the need for further post-processing, it is undefined how the confidence of consecutive bins should be combined to give us the confidence of a region (made up of those consecutive bins). In this sense, MSA resembles some of the methods that borrow among-array information to make within-array calls (see discussion in Section 3.14).

The working definition of region used in RAE is based on their notion of "unified breakpoint profile" (p. 4 and (Fig. **3**) in [42]). A unified breakpoint is the set of all breakpoints from all samples (i.e., the union of the breakpoints obtained from the segmentation conducted on every sample or array). A region is then the sequence of probes between two consecutive unified breakpoints. Most of the statistical operations (e.g., computing averages or conducting permutation tests) are carried out within regions. Therefore, a region in RAE cannot join segments that are separated by a breakpoint in any individual sample. However, segments that are highly homogeneous within many samples can be broken apart by the unified break point procedure, so they can end up in different (even if consecutive) regions. Therefore, within-sample integrity of patterns is not preserved.

In summary, only a few of the methods try to locate recurrent regions. Does it matter whether we are searching for recurrent probes or recurrent regions? There is no definite answer. For some problems, locating common probes might be all that is needed. In some other cases, even if regions are ultimately the objective, locating common probes might be a good enough place to start the search. In other problems, however, it is arguable that we really are looking for recurrent regions in the sense of their being an underlying unit, with functional and biological meaning, and that has within-individual integrity; for instance, when using regions in subsequent studies of association with disease or for the integration of gene expression data (see Section 4.9).

4.2. Segmented Data vs. Original Log Ratio Data

Some methods (e.g., MAR and CMAR, STAC) use, as input, data reduced or discretized to the values "gain", "loss", "no change". In other words, instead of using the original intensity or their ratios, or the smoothed ratios (the "predicted" or "estimated true" values from a segmentation analysis), the original signal is mapped to three possible categories. These approaches have been critized because of the potentially large loss of information they entail [29], a problem that can be more severe in very noisy systems [42] and when the aCGH measurements come from heterogenous populations of tumor cells [29]. According to Klijn *et al*. [29], and when comparing KC-SMART with STAC, the discretization (neglecting the amplitude of the aberration) could lead to an increase in false positives. Note also that methods that use as input the segmented data implicitly assume that the classification of probes into states of gain/loss/no-change is done without error, and do not provide a way to propagate the uncertainty in these calls to the rest of the downstream analysis [48].

On the other hand, using the observed intensity data when trying to locate common regions, as done in KC-SMART and MSA is also open to criticisms. For instance, Choi *et al*. [64] say "(...) experimental copy number of a gene is not directly comparable across samples (...) every tumor biopsy results in a mixture of tumor and normal cells and the ratio of this mixture varies by sample. (...) Hence approaches that take the raw copy number data as measurements comparable across the samples (...) may be subject to unexpected errors (...)." This problem probably also affects approaches that use smoothed data, such as GISTIC: as the smoothing is done per array, the smoothed values might not

be comparable across arrays. Note that this criticism does not apply to HMM-based approaches [31, 48, 57], as the HMM models incorporate array-to-array parameters or estimate probabilities per-array. RAE also uses intensity data as input, but the soft-discriminator model is applied per array.

Based on these arguments, it is arguable that the least problematic approach is to use probabilities, since they avoid the loss of information from using segmented data, and also avoid the non-comparability inherent in the original intensity data or the per-array smoothed intensity data.

4.3. Amplitude and Strength of Evidence

Some methods (e.g., KC-SMART, MSA) use the original ratios for the computation of a statistic that should measure the evidence that a probe or region is altered. Thus, amplitude of change (ratio) is equated to strength of evidence: increase in amplitude should be reflected in monotonic increases in the likelihood that a region or probe is gained (and similarly for decreases below a ratio of 0 and evidence of loss).

However, this mapping is not always so straightforward (see also cite from Choi *et al*. [64] in previous Section), and the relation between amplitude and strength of evidence should be mediated by the variance in the ratios, both inter-array (e.g., the meaning of an observed is not the same in high-variance and low-variance arrays) and type of alteration and segment. This non-direct mapping is easily and implicitly incorporated in Hidden Markov Models [21, 31, 57, 64], but not with other approaches. The "soft thresholding" method in RAE tries to address this problem without explicitly returning probabilities of alteration. Using the smoothed (and possibly scaled between arrays) ratios, as in GISTIC or cghMCR, can also ameliorate this problem (since the scaled and smoothed ratio is more likely to have a monotonically increasing relation with likelihood of alteration).

4.4. Refinements of Common Regions

Some authors further refine their objectives, when dealing with the inherent complexities in patters of structural aberration, with the aim of identifying "driver mutations", oncogenes, etc. Rouveirol *et al.* [28] define (p. 849) a "minimal recurrent region as a recurrent region that contains no smaller recurrent region" because, they argue, "the accurate determination of minimal regions of chromosomal alterations is the first, crucial step towards the identification of new oncogenes and tumor suppressor genes". A similar objective is behind the procedures applied in other methods, such as MSA, GISTIC, and RAE: the common regions located are further examined and post-processed, often with methods much more complex than those used for locating common regions themselves. This post-processing often incorporates many more biological considerations and assumptions and, thus, is likely to yield more biologically-relevant results. For instance, notice that the straightforward scenarios in Fig. (**1**) make no mention of possible restrictions such as length of a region (e.g., a few hundreds of base pairs vs. whole chromosome arms), location in the chromosome relative to hotspots, additional biological annotation, etc.

The problem of refining the search lies in the potential ambiguity of subsequent steps. Rouveirol *et al.* [28], with MAR and CMAR, present a complex but carefully defined approach, and Klijn *et al*. [29] use a straightforward method where we vary a scale parameter in a kernel smoothing function. In contrast, the authors of RAE (p. 6 in [42]) acknowledge that "regions of interest are not rigorously defined, but are intuitive and motivated primarily by two issues. First (...) manageable and interpretable events, perhaps involving a single gene. Second (...) we see where peaks of alteration exist but are confounded by noisy data." This can easily lead to vague and complex method descriptions, as well as algorithmic implementations that are difficult to extend or modify (see also Sections 4.5 and 5).

4.5. Null Models

Most methods that return p-values for the regions found obtain those p-values via permutation tests. To find the p-value (how unlikely the statistic we have observed is in the absence of common regions), we need to generate the distribution of the statistic under the null model (i.e., a scenario of absence of common regions). Obviously, large differences in the null model can lead to large differences in results. The problem is that there are a variety of null models in use, without a careful and reasoned comparison among them.

The null models used in STAC, MSA, KC-SMART, and GISTIC are relatively straightforward: the observed log2 ratios (KC-SMART and GISTIC) or the observed intervals of aberration (STAC, MSA) are placed in a random location. (Strictly, GISTIC does not use random relocations, but a semi-exact approximation to the distribution of the statistic under a random permutation of the marker locations). However, the random relocations of regions in STAC and MSA are within chromosome, whereas the reshuffling of log2 ratios in KC-SMART is over the whole genome (although the analysis in MSA can also be conducted at the genome level to detect whole chromosome alterations: see p. 1484 of [38]). Klijn *et al.* [29] argue that relocation over the entire genome is to be preferred, because relocations within chromosome will prevent detecting recurrent losses or gains that affect whole-chromosome arms, a result that we have also observed. Moreover, relocating within chromosome is likely to penalize the detection of large aberrations: a very large aberration can only be randomly relocated in a small number of ways (i.e., the denominator of the permutation test is small), and most of those will have a large overlap. Therefore, it is unlikely that we will obtain a small p-value. However, relocating an interval of aberration (and intervals of aberration are the "natural units" to be relocated in STAC and MSA) might not be possible over the genome since, for instance, a very large aberration in chromosome 1 would just simply not fit inside chromosome 22.

The above methods are a direct application of the usual statistical approach in permutation tests [65]: the null distribution of the test statistic is computed conditional on random permutations of the observed data. In the methods above, under the null hypothesis of no common regions, any location of the log2 ratios or the intervals of aberration should be equally likely.

In contrast, RAE uses a much more complicated model that does not simply condition on random permutations of the observed values but, instead, uses information about hostspots. This approach is motivated by the attempt to differentiate between "tumor-associated breakpoints" and total breakpoints in the genome, the later being related to a "be-

nign genetic background". RAE's authors therefore develop a model that incorporates this genetic background using recombination hostpots.

The approach in RAE might be superior to the much more straightforward approaches of STAC, MSA, and KC-SMART for identifying "tumor-associated breakpoints". The later methods might detect common regions that belong to what [42] regard as simply "benign genetic background". However, the approach in RAE is not a straightforward, direct, permutation test, and its justification is completely contingent on their background model being an appropriate biological model. GISTIC might remind us of RAE, because of the incorporation of several biological considerations into the core of the procedure; however, the steps where those biological considerations are incorporated are clearly distinct from the permutation test step (see comments in Section 3.8). It is interesting to note, for instance, that whereas the procedures of STAC, MSA, KC-SMART, and GISTIC are invariant to the passage of time (i.e., the p-values obtained today ought to be the same as those we would obtain ten years from now), the results of the approach in RAE are completely contingent on the information available about recombination hotspots. This feature thus highlights this major difference: STAC, MSA, KC-SMART, and GISTIC conduct a typical permutation test, whereas RAE mixes the idea of a permutation test with the incorporation of additional background knowledge for the generation of the null distribution of the statistic.

Null models and their extensions are also used to evaluate the performance of methods. First, generating data under the null model and running a given method against the generated data will provide information on how often the method makes a wrong call (Type I error rate, false positive rate). Moreover, some papers examine the performance of methods (sensitivity, false negatives, power) by generating "true signal" relative to the null model. In other words, data are generated using especific deviations from the null model, and those data are analyzed by the method. The data thus generated are supposed to represent the type of data we would obtain when there really are common regions of alteration; therefore, the mechanism for data generation depends crucially on what the working definition of common region is, and what is regarded as a reasonable model for locating the common and the discordant regions of copy number variation. An interesting example is Fig. (**11B**) of the STAC paper [38]: it is arguable that there are many common regions for high values of Lambda (i.e., there are many aberration intervals that overlap considerably in different individuals) that are not included among the theoretical "true" concordant regions. And data "with signal" generated under a given null model might be of a very specific type, and very different from data "with signal" generated under a different null model.

4.6. Probabilities and p-Values

Most methods use p-values (with correction for multiple testing, usually via FDR or Bonferroni) to provide a measure of strength of evidence that the region or probe detected is a real alteration or is really common. It must be rememberd, however, that the mapping from a p-value to a "probability that this region is altered" (or "probability that this region is commonly altered over these set of samples") is not straightforward at all: a p-value measures the probability of obtaining a statistic as extreme as (or more extreme than) the observed one under a specific null hypothesis. Even when we are conducting simple, well understood, hypothesis tests, the mapping between a p-value and the probability of the null is complicated [66]. In the present case, the situation is much more complicated, both because the null hypotheses are often more complex (see Section 4.5) and because of the added layer introduced by multiple testing corrections. Moreover, using only p-values we cannot rank by relevance the non-significant regions. Of the available methods for recurrent CNA regions, posterior probabilities of alteration of probes or scores directly related to those probabilities are only provided by pREC-S, pREC-A, Master-HMMs, CoCoA, and BSA; see also [54, 56-58] for related methods that perform closely related tasks using HMMs (see Section 3.14).

4.7. Common Regions Over Subsets of Samples

Some complex diseases are quite heterogeneous and present molecular subgroups [3, 67, 68]. Thus, it is often crucial to differentiate between two different cases. In one case, we consider all the samples (subjects or arrays) in the study as a homogeneous set of individuals, and we want to focus on the major, salient, patterns in the data: we will try to locate regions of the genome that present a constant alteration over all (or most of) the samples. This is, for instance, what Scenario I in Fig. (**1**) shows. In a second case, we suspect that the subjects are a heterogeneous group such as shown in Scenario II or Scenario IV in Fig. (**1**). In this case, we want to identify clusters or subgroups of samples that share regions of the genome that present a constant alteration. In other words, we want to detect recurrent alterations in subtypes of samples when we do not know in advance which are these recurrent alterations nor the subtypes of samples. This second case is arguably much more common than the first one in many of the diseases where CNA studies are being conducted [3]. In this second case, using an algorithm appropriate for the first case (one that, by construction, tries to find alterations common to most arrays), or using settings (e.g., minimal frequency) that only allow finding the most common aberration, is clearly inappropriate: it does not answer the underlying biological question, risks missing relevant signals, and leads to conceptual confusion.

4.8. Biclustering

It is somewhat surprising that the connection between finding common regions and biclustering has not been made explicitly more often (but see p. 853 in [28]), especially when one is interested in locating alterations that might be common only to subsets of subjects. Biclustering has been widely used with genomic data with the objective of identifying "(...) groups of genes that show similar activity patterns under a specific subset of the experimental conditions" (p. 2 in [69]) or "(...) sets of genes sharing compatible expression patterns across subsets of samples" (p. 1122 in [70]). These objectives are very similar to those of locating common regions of copy number alteration. Exploiting these similarities might prove worhtwhile given that the biclustering problem has been extensively studied (see reviews in [69, 70]) and that there are fast and simple reference models [70] that could be applied directly to the segmented data. It is

likely that this might require carefully considering similarity measures and type of linkage; work in defining linkage and similarity that are specific for aCGH data has been conducted by Van Wieringen and colleagues [71].

4.9. Association with Disease and Integration of Expression Data

Recurrent regions of alteration (or variation) are sometimes used as input for further downstream analysis. In studies of association between disease status and copy number, increases in statistical power can be achieved if regions (instead of single probes) are used, as shown in [59-61]. Note, however, that those three methods take regions as given. They cannot be used to define them. SIRAC [72] attempts to identify regions that are useful for differentiating between sets of tumors, and uses an operational definition that is completely tailored to just that objective. Thus, this method is not a general method for detecting common regions of aberration. GEAR implements two approaches for detecting what it regards as "class-specific alterations"; one is the usual and straightforward comparison of the frequency of aberrations between two pre-specified classes using Fisher's test, and the second uses SW-ARRAY [37] on the mean difference between the two groups.

Since many studies of recurrent CNA regions are carried out with the ultimate purpose of relating CNA to disease, this area is open for much further work. First, in terms of Fig. (**1**), it needs to be clarified what type of pattern of recurrent CNA we want to detect when studying association with disease. Initially, we will probably want to focus on Scenarios II, III and V. Scenarios IV and VI are likely too complex (at least initially), and Scenario I is obviously unrelated to differential phenotype if all samples show the CNA. Second, the search for recurrent regions can be carried out before the association analysis or simultaneously. SIRAC takes the second approach, but the statistical method used is too simple for most case-control studies or with continuous dependent variables (i.e., the statistical method in SIRAC, SAM, does not allow the rich modelling available in both CNVassoc [59] and CNVtools [60]). The first method in GEAR is a common one in the literature, is not related to regions (it is a probe-by-probe approach), only allows very simple statistical models, and can be severely affected by uncertainties in the aberration calls [59]. The second approach in GEAR is conceptually not very different from the approach in SIRAC: differences between groups in the signal are computed first, and then a method is applied to search for "regions" in those differences (p-values in the case of SIRAC). As before, only very simple statistical models can be fitted, and it is unclear that the "regions" found are really regions since they are regions from averages over subjects (see Section 4.1).

Likewise, many studies have attempted to assess the effects of changes in genomic DNA copy number on gene expression. VanWier *et al*. [62] developed two statistics to carry out "regional analysis" because "We (...) expect analysis that takes regional effects into account to yield better results that might offset the negative effects of noise in the data or low penetrance." (p. 5 in [62]) and "(...) low penetrance (not all cells in the sample) and low prevalence (not all samples in the study) alterations might affect expression below the 2-fold mark and only in some of the samples, but in a significant manner when a genomic region is considered." (p. 2 in [62]). These authors do report improved performance from using regional analysis with pre-defined region as found by CGHregions [30]. Thus, regions are taken as given, and the association between copy number and gene expression is examined with weighted test statistics (with shrinkage) applied over regions. A similar effect is reported in Lipson *et al*. [63] using a simple average correlation over regions. In this paper, another method (GCSM) is developed: the authors search for submatrices of the data that contain amplified (deleted) genes and over-expressed (repressed) genes; thus, in this second method, regions are not taken as given, but rather searched for so as to maximize association with expression data.

4.10. Comparisons Among Methods

There is no comprehensive comparison of the different approaches, and very few of the published papers present any comparison with other methods (but see [29] for comparisons between KC-SMART and STAC, [38] for comparisons between MSA and STAC, and [42] for comparisons between RAE and GISTIC). Carrying out these comparisons is difficult because of some issues already mentioned:

- The meaning of common region is vague, and different methods have different objectives and types of regions they will detect. Thus, it is unclear how to define a metric to measure performance. For instance, many method comparisons are not meaningful if we are interested in detecting Scenario IV.
- Some methods depend strongly on specific null models. Since settling down which of the null models is the correct one is unlikely to happen soon, comparison ought to be done using several of the proposed null models.
- There are no real reference data sets that can be used as gold standards; any comparisons using real data will, thus, always be incomplete and inconclussive (are the detected patterns real? are the undetected patterns just not there?).

In spite of those difficulties, however, the field is ready for such a comprehensive, careful, comparison of the relative strengths of methods using a variety of simulated data sets. Only by using carefully planned simulation studies can we get an idea of which methods are likely to perform better with any given real data set. It is worth noting that similar comments have been recently made by Shah [3].

4.11. Code Availability and Code Licenses

Several of the methods do not have code available. We find this a most unfortunate situation, since a method without code is, basically, a method that will remain unused: given that there are many competing approaches, it is unlikely anybody will implement a method that someone else has developed. Claiming "software available upon request from the authors", or similar formulas is, often, a red flag that software is not really available, or is only available in a difficult to use form. We emphatically suggest to reviewers and editors to require that code be publicly available for any new method published, if that method is to have any chance of making a difference and being used by other researchers. Of course, when we say "code" we mean not only an executable

but, most importantly, the source code. Only the source code allows researchers to extend the method, detect and fix bugs, and know what an implementation is really doing [73-75]).

Some methods are only available for Matlab. Again, this is often unfortunate, since it makes the method inaccessible to researchers that do not have a Matlab license. While it is true that developers can distribute stand-alone Matlab applications, this precludes modifying, improving, and debugging the code, which are some of the key advantages of having the source code available, and a definite need in Bioinformatics [73, 75]. Similarly, some methods are only available as Windows executables, precluding their usage under other operating systems. This is particularly unfortunate since many clusters and high-performance workstations run GNU/Linux and Unix operating systems. In terms of R packages, authors and editors are strongly encouraged to have the R packages deposited in either CRAN or BioConductor. This ensures both that the package will remain available regardless of what happens to the authors' personal web pages, and provides additional quality control checks.

Finally, licenses are often times not specified. We do have a strong preference for free software licenses, for reasons articulated elsewhere by us and by others [73-75]. Regardless of the type of license, it must be clearly spelled out: lack of a license hinders using, modifying, and further developing a method, since it is unclear for any prospective developer whether changes to a code base can be further distributed, and what are the terms of usage of the output of the program.

5. FURTHER WORK

To summarize, we think there are several areas where further work is needed. First, we need a clear delineation between the statistical and computational steps and the biological assumptions and ultimate objectives. In terms of the Scenarios in Fig. (**1**), it is rarely explicit what is ultimate biological objective of many methods and, therefore, it is difficult to choose a method based on a specified objective. In addition, some current procedures are very difficult to modify and adapt, since the statistical approaches and biological assumptions are intertwined in a convoluted way (see also Section 4.5). Likewise, we need a clear delineation between the type of patterns we want to detect (what is a common region, why we regard that pattern as a biologically relevant one) and how those patterns are to be detected (the algorithmic solution). A biologist needs to be able to assess whether a given method detects a pattern she or he is interested in, without having to plod through an algorithm.

Second, and as discussed in Section 4.2, probabilities do not suffer from either discarding information or mixing non-comparable intensity ratios. Probabilities also do not confound amplitude and strength of evidence (Section 4.3), and output in terms of probabilities of alteration is directly interpretable (Sections 4.6). Therefore, methods that use probabilities, both as input and as output, would be conveniently suited for finding recurrent regions.

Third, choice of a method for finding recurrent CNA regions should depend on the intended usage of the detected recurrent CNA regions but, so far, there is very little work on carefully matching intended usage to method for finding recurrent CNA regions. If the detected regions are to be used to cluster individuals, investigation of biclustering approaches are likely to be fruitful (see Section 4.8). In contrast, if recurrent regions are to be used to study the association of disease with CNAs, or the impact of CNAs on gene expression data, then other types of methods and biological scenarios in terms of Fig. (**1**) might be needed (see Section 4.9).

Finally, comprehensive, through comparisons, of performance of different methods under different scenarios are missing (see Section 4.10), making it hard to base choice on the actual performance of different approaches. Of course, method comparison makes sense only after the intended usage and the scenarios are well defined.

6. ACKNOWLEDGEMENTS

Work partially funded by Fundacion de Investigacion Medica Mutua Madrileña. One anonymous reviewer for comments on the ms. Publication charges covered by projects CONSOLIDER: CSD2007-00050 of the Spanish Ministry of Science and Innovation and by RTIC COMBIOMED RD07/0067/0014 of the Spanish Health Ministry.

REFERENCES

[1] Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* **2007**; 39: S48-S54.

[2] Scherer SW, Lee C, Birney E, *et al.* Challenges and standards in integrating surveys of structural variation. *Nat Genet* **2007**; 39: S7-S15.

[3] Shah SP. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet Genome Res* **2008**; 123: 343-351.

[4] Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet* **2009**; 10: 551-564.

[5] Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* **2006**; 444: 444-454.

[6] Lupski JR. Genomic rearrangements and sporadic disease. *Nat Genet* **2007**; 39: S43-S47.

[7] McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet* **2007**; 39: S37-S42.

[8] Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* **2007**; 8: 639-646.

[9] Wain LV, Armour JAA, Tobin MD. Genomic copy number variation, human health, and disease. *Lancet* **2009**; 374: 340-350.

[10] Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* **2009**; 1: 62.

[11] Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* **2006**; 34: 445-450.

[12] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **2005**; 37 (Suppl): S11-S17.

[13] Huang J, Wei W, Chen J, *et al.* CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics* **2006**; 7: 83.

[14] Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **2007**; 39: S16-S21.

[15] Korbel JO, Urban AE, Affourtit JP, *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **2007**; 318: 420-426.

[16] Xie C, Tammi M. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **2009**; 10: 80.

[17] Lee S, Cheran E, Brudno M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **2008**; 24: i59-67.

[18] Kidd JM, Cooper GM, Donahue WF, *et al*. Mapping and sequencing of structural variation from eight human genomes. *Nature* **2008**; 453: 56-64.

[19] Lai WRR, Johnson MDD, Kucherlapati R, Park PJJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **2005**; 21: 3763-3770.

[20] Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **2005**; 21: 4084-4091.

[21] Rueda OM, Diaz-Uriarte R. Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput Biol* **2007**; 3: e122.

[22] Rueda OM, Diaz-Uriarte R. A response to Yu *et al.* 'A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array', BMC Bioinformatics 2007, 8: 145. *BMC Bioinformatics* **2007**; 8: 394.

[23] Pique-Regi R, Monso-Varona J, Ortega A, Seeger R, Triche T, Asgharzadeh S. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **2008**; 24: 309-318.

[24] Diskin S, Eck T, Greshock J, *et al*. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* **2006**; 16: 1149-1158.

[25] Aguirre AJ, Brennan C, Bailey G, *et al.* High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* **2004**; 101: 9067-9072.

[26] Misra A, Pellarin M, Nigro J, *et al.* Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res* **2005**; 11: 2907-2918.

[27] Beroukhim R, Getz G, Nghiemphu L, *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* **2007**; 104: 20007-20012.

[28] Rouveirol C, Stransky N, Hupé P, *et al*. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **2006**; 22: 2066-2073.

[29] Klijn C, Holstege H, de Ridder J, *et al.* Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res* **2008**; 36: e13.

[30] van de Wiel MA, van Wieringen W. CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics* **2007**; 2: 55-63.

[31] Shah S, Lam W, Ng R, Murphy K. Modeling recurrrent CNA copy number alterations in array CGH data. *Bioinformatics* **2007**; 23: i450-i458.

[32] Yang L, Chipman HAA, Bull SBB, Briollais L, Wang K. A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics* **2009**; 25: 1669-1679.

[33] Shah SP, Cheung KJ, Johnson NA, *et al*. Model-based clustering of array CGH data. *Bioinformatics* **2009**; 25: i30-38.

[34] Rosa PL, Viara E, Hupe P, *et al.* VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics* **2006**; 22: 2066-2073.

[35] Liva S, Hupé P, Neuvial P, *et al.* CAPweb: a bioinformatics CGH array Analysis Platform. *Nucleic Acids Res* **2006**; 34: W477-W481.

[36] Kim TM, Jung YC, Rhyu MG, Jung MH, Chung YJ. GEAR: genomic enrichment analysis of regional DNA copy number changes. *Bioinformatics* **2008**; 24: 420-421.

[37] Price TS, Regan R, Mott R, *et al*. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **2005**; 33: 3455-3464.

[38] Guttman M, Mies C, Dudycz-Sulicz K, *et al.* Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* **2007**; 3: e143+.

[39] Weir B, Woo M, Getz G, *et al*. Characterizing the cancer genome in lung adenocarcinoma. *Nature* **2007**; 450: 893-898.

[40] Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **2004**; 20: 3413-3422.

[41] Lingjaerde OC, Baumbusch LO, Liestol K, Glad IK, Borresen-Dale AL. CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **2005**; 21: 821-822.

[42] Taylor BSS, Barretina J, Socci NDD, *et al.* Functional copy-number alterations in cancer. *PLoS ONE* **2008**; 3: e3179.

[43] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **2004**; 5: 557-572.

[44] Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **2007**; 23: 657-663.

[45] Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhinim Z. Efficient calculation of interval scores for DNA copy number data analysis. *J Comput Biol* **2006**; 13: 215-228.

[46] Ben-Dor A, Lipson D, Tsalenko A, *et al.* Framework for identifying common aberrations in DNA copy number data. *Proc RECOMB '07* **2007**; 4453: 122-36.

[47] Chen PA, Liu HF, Chao KM. CNVDetector: locating copy number variations using array CGH data. *Bioinformatics* **2008**; 24: 2773-2775.

[48] Rueda OM, Diaz-Uriarte, R. Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics* **2009**; 10: 308.

[49] Liu J, Ranka S, Kahveci T. Markers improve clustering of CGH data. *Bioinformatics* **2007**; 23: 450-457.

[50] Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M. Distance-based clustering of CGH data. *Bioinformatics* **2006**; 22: 1971-1978.

[51] Korn JMM, Kuruvilla FGG, McCarroll SAA, *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **2008**; 40: 1253-1260.

[52] Wang K, Li M, Hadley D, *et al*. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **2007**; 17: 1665-1674.

[53] Komura D, Shen F, Ishikawa S, *et al.* Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res* **2006**; 16: 1575-84.

[54] Pique-Regi R, Ortega A, Asgharzadeh S. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics* **2009**; 25: 1223-1230.

[55] LaFramboise T, Winckler W, Thomas RK. A flexible rank-based framework for detecting copy number aberrations from array data. *Bioinformatics* **2009**; 25: 722-728.

[56] Engler D, Mohaptra G, Louis D, Betensky R. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **2006**; 7: 399-421.

[57] Colella S, Yau C, Taylor JMM, *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **2007**; 35: 2013-2025.

[58] Wang H, Veldink JHH, Blauw H, van den Berg LHH, Ophoff RAA, Sabatti C. Markov models for inferring copy number variations from genotype data on illumina platforms. *Human Hered* **2009**; 68: 1-22.

[59] Gonzalez J, Subirana I, Escaramis G, *et al.* Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformatics* **2009**; 10: 172+.

[60] Barnes C, Plagnol V, Fitzgerald T, *et al*. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* **2008**; 40: 1245-1252.

[61] Ionita-Laza I, Perry GH, Raby BA, *et al*. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epidemiol* **2008**; 32: 273-284.

[62] van Wieringen WNN, van de Wiel MAA. Nonparametric testing for DNA copy number induced differential mRNA Gene Expression. *Biometrics* **2009**; 65: 19-29.

[63] Lipson D, Ben-Dor A, Dehan E, Yakhini Z. Joint analysis of DNA copy numbers and gene expression levels. In Algorithms in Bioinformatics New York*;* Springer-Verlag **2004**:135-146.

[64] Choi H, Qin ZS, Ghosh D. A double-layered mixture model for joint analysis of copy number and gene expression data. *J Comput Biol* **2009**; in press.

[65] Edgington E, Onghena P. Randomization Tests, 4th ed. (Statistics: a Series of Textbooks and Monographs). London; Chapman & Hall/CRC **2007**.

[66] Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. *Am Stat* **2001**; 55: 62-71.

[67] Wood LDD, Parsons DWW, Jones S, *et al*. The genomic landscapes of human breast and colorectal cancers. *Science* **2007**; 318: 1108-1113.

[68] Sebat J. Major changes in our DNA lead to major changes in our thinking. *Nat Genet* **2007**; 39: S3-S5.

[69] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE TCBB* **2004**; 1: 24-45.

[70] Prelić A, Bleuler S, Zimmermann P, *et al*. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **2006**; 22: 1122-1129.

[71] Van Wieringen WNN, Van De Wiel MAA, Ylstra B. Weighted clustering of called array CGH data. *Biostatistics* **2008**; 9: 484-500.

[72] Lai C, Horlings HHM, van de Vijver MMJ, *et al*. SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets. *BMC Bioinformatics* **2007**; 8: 422.

[73] Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. *Biotechniques* **2003**; (Suppl): 45-51.

[74] Stallman RM, Gay J. Free Software, Free Society: Selected Essays of Richard M. Stallman. Free Software Foundation **2002**.

[75] Dìaz-Uriarte R. Supervised methods with genomic data: a review and cautionary view. In Azuaje F, Dopazo J, Eds. Data analysis and visualization in genomics and proteomics. New York: Wiley **2005**: pp. 193-214.

Received: January 01, 2009 Revised: August 12, 2009 Accepted: August 20, 2009