Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2008 *Paper* 42

Finding Recurrent Regions of Copy Number Variation: A Review

Oscar M. Rueda* Ramon Diaz-Uriarte†

*Spanish National Cancer Research Centre (CNIO)

†CNIO, rdiaz02@gmail.com

Finding Recurrent Regions of Copy Number Variation: A Review

Oscar M. Rueda and Ramon Diaz-Uriarte

Abstract

Copy number variation (CNV) in genomic DNA is linked to a variety of human diseases, and array-based CGH (aCGH) is currently the main technology to locate CNVs. Although many methods have been developed to analyze aCGH from a single array/subject, disease-critical genes are more likely to be found in regions that are common or recurrent among subjects. Unfortunately, finding recurrent CNV regions remains a challenge. We review existing methods for the identification of recurrent CNV regions. The working definition of "common" or "recurrent" region differs between methods, leading to approaches that use different types of input (discretized output from a previous CGH segmentation analysis or intensity ratios), or that incorporate to varied degrees biological considerations (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 suggest that finding recurrent CNVs could benefit from reframing the problem in a biclustering context. We also emphasize that, when analyzing data from complex diseases with significant among-subject heterogeneity, methods should be able to identify CNVs that affect only a subset of subjects. We make some recommendations about choice among existing methods, and we suggest further methodological research.

1 Introduction

Copy number variations (CNVs) are often defined as DNA segments longer than 1 kb for which copy number differences are observed when comparing two or more genomes [1, 2]. CNVs have turned out to be much more abundant than previously thought [3-5] and have been linked to many different types of disease, including cancer, HIV acquisition and progression, autoimmune diseases, and Alzheimer and Parkinson's disease [4-7]. Identification of CNVs in individual samples nowadays uses mainly array-based Comparative Genomic Hybridization (aCGH), encompassing ROMA, oaCGH (including Agilent, NimbleGen, and many non-commercial, in-house oligonucleotide arrays), BAC, and cDNA arrays [8, 9], and SNP-based arrays [10, 11]. Location of CNVs in individual samples, however, is only the initial step in the search for "interesting genes". The regions more likely to harbor disease-critical genes are those that are recurrent or common among samples (e.g., [9, 12-14]). Many methods exist for analyzing a single array of CGH (e.g., see references in [15, 16]), but integrating several samples and finding common regions of alteration, however, remains a challenge [2], both computationally and conceptually. In this review we discuss the available methods (many developed in the last two years), and some of the reasons for the difficulties.

1.1 What are recurrent common regions?

One of the first problems that a user faces when choosing a method to "find recurrent regions" is that "recurrent region" has different meanings for different authors.

Few authors have attempted a rigorous definition of recurrent or common region of copy number alteration, with the notable exception of Rouveirol et al. [17]. In addition to the complexity of Rouveirol et al.'s definition (although complexity of a definition might be unavoidable in this problem), their scheme is tied to using segmented data (the data reduced to three possible states, "loss", "gain", "no alteration"), not the original intensity data (see also section 3.1). Most other papers do not attempt a rigorous definition of what a recurrent region is; they seem to accept a definition of a CNV common 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 (thus the usage of terms such as "common" or "recurrent"). There are, however, several departures from this definition and the terms used:

- Not all approaches try to locate regions. Some methods deal with a much simpler objective: finding common

probes (e.g., [18-20]), which does not require addressing the problem of the location of boundaries of regions. Why dealing boundaries is a potentially difficult problem can be seen already in [13], one of the earliest papers that attempts to locate common regions.

If we ignore the problem of locating boundaries, and if we are using segmented data, 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.

- The definition above might suggest that homogeneity is considered relative to the status (gained or lost), but several methods use also information about the magnitude of the alteration. For instance, some authors differentiate between common regions of low amplitude gain vs. common regions of high amplitude gain.
- Most methods try to locate regions that are common to all the subjects in the study. A few methods, however, can also locate regions common only to a subset of the samples and, thus, can incorporate among-subject heterogeneity in common regions (see further discussion in section 3.5).
- Some papers do focus on locating common regions or common probes as the main (methodological) objective. In other papers, however, the location of common regions is only a step of a method that attempts to locate relevant oncogenes, tumor suppressor genes, etc [12, 21-23]. Any of the common regions located are further examined and post-processed, often with methods much more complex than those used for locating common regions themselves, and incorporating many more biological considerations and assumptions, with the aim of identifying "interesting genes".

Finally, a few other methods seem to search for something akin to common regions, but their objectives are actually quite different. SIRAC [24] identifies regions that are useful for differentiating between sets of tumors, and uses an operational definition that is completely tailored to just that objective. The methods of [19, 20] locate markers (not regions) with the only objective of clustering and, similarly to SIRAC, use an operational definition of marker tailored just to the clustering objective. Finally, in sharp contrast to most other methods, CGHregions [25] does not try to find probes that are constant across sets of samples; thus, it would not fit at all the definition above.

Our focus in this review is the location of common regions as informally defined above. However, since the variety of approaches and objectives can often lead to confusion, in this review we have chosen to include methods

that encompass all the objectives mentioned above. We first provide a brief overview of each of the existing methods, next highlight common issues relevant to more than one method, and conclude with suggestions for further research and choice among methods.

2 Overview and details of existing methods

Some of the main features of existing methods are summarized in Table 1 and 2. In this section we review each method in turn, provide further details on their working, 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 3.

Name	Input	Output (Significance)	Null model (for significance)	Detect subgroups of common regions ?
MSA	log2 ratios	p-values	Permutation of the regions within chromosomes	Yes
GISTIC	log2 ratios	p-values	Permutation of the probes over the entire genome	No
RAE	log2 ratios	p-values	Permutation of copy number values using hotspots information	No
MAR, CMAR	Segmented data	None	None	No
cghMCR	Smoothed log2 ratios	None	None	No
CGHregions	Segmented data	None	None	No
Master HMMs	log2 ratios	Probabilities of alteration for each probe	Based on a HMM	No
STAC	Segmented data	Confidence for regions	Permutation of the regions within chromosomes	Yes
Interval Scores	log2 ratios	Scores for each interval	Large deviation bound for iid Gaussian data	No
CoCoA	Segmented data	Scores for each interval	Binomial distribution on probes and intervals	Yes
KC-SMART	log2 ratios	p-values	Permutation of the log-ratios over the entire genome	No
SIRAC	log2 ratios	p-values	Hypergeometric distribution	No
GEAR	log2 ratios	p-values	Permutation of the alterations over the entire genome	No
Markers	Segmented data	None	None	Yes

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). This is to contrast the ``original signal'' with the segmented values (below). segmented data: data reduced to the values 0, 1, -1, or equivalent, denoting no alteration, gain, loss, or genomic DNA. smoothed log2 ratios: the smoothed, predicted or fitted log2 ratio returned by some segmentation methods. A simple example is using the median of a set of probes to estimate its smoothed value. We make no mention of multiple testing control issues: all methods incorporate some form of control, usually via FDR or Bonferroni.

Name	Software available	Software OS	Software license
MSA	Standalone Java application or as part of GenePattern[1]	Multiplatform	Unknown
GISTIC	Standalone based on Matlab Component Runtime version 7.7 (needed [2]	Linux 64-bit or as part of Gene Pattern	Unknown
RAE	R script with a standalone wrapper[3]	Linux for the wrapper	GPL 2
MAR, CMAR	From the authors upon request		
cghMCR	R package [4]	R dependent	GPL 2
CGHregions	R package [5]	R dependent	GPL 2
Master HMMs	MATLAB toolbox [6]	MATLAB dependent	GPL 2
STAC	Standalone Java application [7]	Multiplatform	Unknown
Interval Scores	None		
CoCoA	None		
KC-SMART	R package[8]and standalone application based on Matlab Component Runtime[9]	R or MATLAB dependent	GPL 2 in the case of R package
SIRAC	Matlab function[10]	Matlab dependent	Unknown
GEAR	Standalone application[11]	Windows	Copyright stated in the setup program
Markers	From the authors upon request		

Table 2: Software available

1 http://www.cbil.upenn.edu/MSA/
2 http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=162
3 http://cbio.mskcc.org/downloads/rae/
4 http://www.bioconductor.org (cghMCR)
5 http://www.bioconductor.org (CGHregions)
6 http://www.cs.ubc.ca/~sshah/acgh/CNA-HMMer-v0.1.zip
7 http://cbil.upenn.edu/STAC/
8 http://www.bioconductor.org (KCsmart)
9 http://bioinformatics.nki.nl/~klijn/
10 http://bioinformatics.nki.nl/software.php
11 http://www.systemsbiology.co.kr/GEAR/

2.1 CGHregions [25]

This method is really a dimension reduction approach, not a method to locate common alterations. In fact, the authors clearly state (p. 56, [25]) “Note that 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).

2.2 Markers [19, 20]

Liu and collaborators, in two papers [19, 20], 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 [19], 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”.

2.3 SIRAC [24]

This method attempts to identify regions that can be used to differentiate between classes of subjects. Relevant probes (those that differ between groups) are located using SAM [26], and their “significance” evaluated via sliding windows and a hypergeometric test (comparing observed vs. expected relevant probes). A region is defined using a consensus over the different window sizes. The researcher needs to decide in advance the number of relevant probes and the range of window sizes. This method is not a general method for detecting common regions of aberration, but only for

detecting regions that are useful to differentiate between pre-specified types of tumors. The method assumes ratios are independently distributed along the chromosome (i.e., it ignores correlations among probes), as the intially relevant probes are identified using SAM.

2.4 Master HMMs [18]

In [18] a single-subject 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. Their 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 [15].

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

2.5 cghMCR [13]

Using segmented (i.e., smoothed data), this algorithm [13] 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 adrresses 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).

2.6 MAR, CMAR [17]

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 [17]). The authors then formalize these definitions and develop two algorithms, MAR and CMAR, for finding the minimal common regions. This is one of the most rigorous attempts to define (and detect) common regions, but the formalization is complex and is carried out assuming segmented data, not the original ratio data (see also 3.1). This approach cannot detect regions over subsets of subjects. Code is not easily available.

2.7 GEAR [27]

GEAR [27] actually 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. With either of these approaches, thus, we are trying to locate recurrent clones, not necessarily recurrent regions. 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 [28]: 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. The problem, however, is that SW-ARRAY was designed to work with ratios, not their mean over several samples. 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.

2.8 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 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.

2.9 STAC [12] and MSA [21]

STAC [12] and MSA [21] 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 [12]; or the length of the projection of a set of intervals onto the genome, see p. 1466 in [21]). 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 [21] builds upon 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.

Both STAC and MSA can detect recurrent regions over subsets of subjects. In our experience, MSA is capable of detecting complex patterns of regions over subsets of subjects, except in extreme cases of very small sample sizes.

In the canonical implementation, both STAC and MSA use permutations of the entire regions within chromosomes (instead of over the complete genome) to assess significance in patterns; this permutation scheme might not be the most appropriate, and could preclude detecting large aberrations (see also section 3.3). Although MSA uses the original ratios (not the segmented data, as STAC), MSA uses a common threshold for all arrays and chromosomes and thus ignores possible differences in variability between chromosomes and arrays.

2.10 GISTIC [22, 30]

In a nutshell, this method aggregates data over different tumors to try 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 3.3). (After identification of driver aberrations, tumors are classified according to whether or not they have them). This method involves three main steps: first, data-preprocessing and copy number alteration 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. Then, GLAD is used to locate copy-number changes; 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 aberrations" (Supplementary Information text in [22]). 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 permutation test (see section 3.3). 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” [22]); 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 3.1). In addition, segmentation is performed using GLAD [31]: GLAD has been shown to perform worse than several alternative segmentation approaches [15, 32, 33], and require tunning of several parameters of non-intuitive meaning (but GLAD is one of the few segmentation methods, together with RJaCGH [15] and ACE [34], 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 [22], under “Identification of Copy-Number Aberrations”). GISTIC is not designed to detect regions of alteration common only to a subset of subjects.

2.11 RAE [23]

RAE [23] starts from an initial copy number assessment from a segmentation procedure (CBS [35, 36] 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 their usage of “soft thresholding”, with a sigmoid function, 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 [15]–, with the advantage that the probability assessment does not need to regard as equivalent amplitude and strength of evidence of alteration; see section 3.2).

For RAE [23], the resolution of genomic regions of interest is targeted towards identifying “(...) manageable and

interpretable events, perhaps involving a single gene." (p. 6, [23]); 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 3.3). 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 3.3) 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.

2.12 Interval scores [37] and CoCoA [38]

These two approaches are closely related, and developed by the same research group. Unfortunately, no software seems available for any of the procedures. Both methods assume that the observed ratios are independently (and identically) distributed across the chromosome [18], a biologically unrealistic assumption.

3 Common issues

3.1 Segmented data vs. original log ratio data

Some methods (e.g., MAR and CMAR [17], STAC [12]) use, as input, data reduced or discretized to the values "gain", "loss", "no change". In other words, instead of using the original 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 [23] and when the aCGH measurements come from heterogenous populations of tumor cells [29]. 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 analyses.

3.2 Amplitude and strength of evidence

Some methods (e.g., KC-SMART [29], MSA [21]) 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 straightforwad, 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 [15, 18], but not with other approaches. The "soft thresholding" method in RAE [23] tries to address this problem without explicitly returning probabilities of alteration. Using the smoothed (and possibly scaled between arrays) ratios, as in GISTIC [22] or cghMCR [13], can also ameliorate this problem (since the scaled and smoothed ratio is more likely to have a monotonically increasing relation with likelihood of alteration).

3.3 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 find or 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 [12], MSA [21], KC-SMART [29], and GISTIC [22] 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 actually 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 [21]). [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 the methods of [12, 21]) 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 [39]: 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 [23] uses a much more complicated model that does not simply condition on random permutations of the observed values and, 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 "benign genetic background". RAE's authors [23] therefore develop a model that incorporates this genetic background using recombination hostpots.

The approach in RAE [23] might be superior to the much more straightforward approaches of STAC [12], MSA [21] and KC-SMART [29] for identifying "tumor-associated breakpoints". The later methods might detect common regions that belong to what [23] 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 [22] 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 2.10). 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 runing 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 Figure 11 B of [21]: 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, as before, 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.

3.4 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 uderstood, hypothesis tests, the mapping between a p-value and the probability of the null is complicated [40]. In the present case, the situation is much more complicated, both because the null hypothesis are often more complex (see section 3.3) 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 published methods, only the one of Shah and collaborators, using Hidden Markov Models [18], provides posterior probabilities of alteration of probes.

3.5 Common regions over subsets of samples

Most existing methods try to find regions that are common to all the arrays in the sample and, thus, presuppose that a disease is homogeneous with respect to the pattern of CNVs. It is known, however, that many complex diseases, such as cancer [41] or autism "(...) consist largely of a constellation of rare, highly penetrant mutations" (p. S4 in [3]): we can observe a similar phenotype but we could arrive at this phenotype from several alternative DNA copy number alterations. Thus, it is often crucial to differentiate between two different scenarios. In one scenario, we consider all the samples (subjects or arrays) in the study as a homogeneous set of individuals, so we want to focus on the major, salient, patterns in the data and thus we will try to locate regions of the genome that present a constant alteration over all (or most of) the samples. This is what existing methods for the study of common regions try to do. In a second

scenario, we suspect that the subjects are really a heterogeneous group. What we really want here is 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 scenario is arguably much more common than the first one in many of the diseases where CNV studies are being conducted. In this second scenario, using an algorithm appropriate for the first scenario (one that, by construction, tries to find alterations common to most arrays) is clearly inappropriate: it does not answer the underlying biological question, risks missing relevant signals, and leads to conceptual confusion.

3.6 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. 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. Thus, it is unclear how to define a metric to measure performance.
- 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.

3.7 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. Note that 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.

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 [42, 43].

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 [42-44]. 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.

3.8 Biclustering

It is somewhat surprising that the connection between finding common regions and biclustering has not been made explicitly, especially when one is interested in locating alterations that might be common only to subsets of subjects. Biclustering has been used extensively 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 [45]) or "(...) sets of genes sharing compatible expression patterns across subsets of samples" (p. 1122 in [46]). 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 [45, 46]) and that there are fast and simple reference models [46] that could be applied directly to the segmented data.

4 Further work

We think there are four major areas where further work is needed:

- A clear delineation between the statistical and computational steps and the biological assumptions and ultimate objectives, so that complex procedures can be analyzed, and modified if needed, by examining or changing their different constituent components.
- The incorporation of probabilities, so that end users obtain not just p-values but, much better, probabilities that regions are altered.
- Comprehensive, through comparisons, of performance of different methods under different scenarios.
- Evaluation of the utility of biclustering approaches for the detection of common regions of copy number changes.

5 User recommendations

Method recommendation is difficult given that there are no comprehensive comparisons among methods. With the available information, however, we can make the following summary suggestions. For end users, methods without available code are of no interest. Among the remaining methods, which one to use depends strongly, of course, on our objectives. There are a few extreme cases where the choices are clear cut: if we are only interested in clustering (e.g., [19, 20]), or we only want dimensional reduction or feature selection for classification [24, 25].

Interest in common regions (not common probes) excludes the methods based on master HMMs [18]. cghMCR [13] and STAC [12] have arguably been superseded by more recent methods. GEAR [27] and KC-SMART [29] have the advantage of being relatively straightforward methods. GISTIC [22] and RAE [23] and, to a certain extent, MSA [21], are much more complex methods that attempt to incorporate additional biological considerations to identify "interesting genes" (and, therefore, choice between these methods could be dictated by how reasonable these

biological assumptions are). Only MSA [21] would be appropriate if one is interested in detecting subsets of subjects with respect to regions of alteration.

6 Acknowledgements

Work partially funded by Fundacion de Investigacion Medica Mutua Madrilena.

References

[1] Lee C, Iafrate AJ, Brothman AR, Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genetics* **2007**, 39:S48–S54.

[2] Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L, Challenges and standards in integrating surveys of structural variation. *Nat Genet* **2007**, 39(7 Suppl):S7–S15.

[3] Sebat J, Major changes in our DNA lead to major changes in our thinking. *Nature Genetics* **2007**, 39:S3–S5.

[4] 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(8):639–646.

[5] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews DT, F%egler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, Macdonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MrJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang Ja, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani C Hiroyuki an d Lee, Jones KW, Scherer SW, Hurles ME, Global variation in copy number in the human genome. *Nature* **2006**, 444(7118):444–454.

[6] Lupski JR, Genomic rearrangements and sporadic disease. *Nat Genet.* **2007**, 39(7 Suppl):S43–S47.

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

[8] 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.

[9] Pinkel D, Albertson DG, Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **2005**, 37 Suppl:S11–S17.

[10] Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, Shapero MH, CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide

arrays. *BMC Bioinformatics* **2006**, 7.

[11] Carter NP, Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **2007**, 39(7 Suppl):S16–S21.

[12] Diskin S, Eck T, Greshock J, Mosse Y, Naylor T, Stoeckert CJ, Weber B, Maris J, Grant G, STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.* **2006**, 16(9):1149–1158.

[13] Aguirre AJ, Brennan C, Bailey G, Sinha R, Feng B, Leo C, Zhang Y, Zhang J, Gans JD, Bardeesy N, Cauwels C, Cordon-Cardo C, Redston MS, Depinho RA, Chin L, High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* **2004**, 101:9067–9072.

[14] Misra A, Pellarin M, Nigro J, Smirnov I, Moore D, Lamborn KR, Pinkel D, Albertson DG, Feuerstein BG, Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res* **2005**, 11:2907–2918.

[15] Rueda OM, Diaz-Uriarte R, Flexible and Accurate Detection of Genomic Copy-Number Changes from aCGH. *PLoS Comput. Biol.* **2007**, 3(6):e122.

[16] 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+.

[17] Rouveirol C, Stransky N, Hupé P, La Rosa P, Viara E, Barillot E, Radvanyi F, Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **2006**, 22:2066–2073.

[18] Shah S, Lam W, Ng R, Murphy K, Modeling recurrrent CNA copy number alterations in array CGH data. *Bioinformatics* **2007**, 23(13):i450–i458.

[19] Liu J, Ranka S, Kahveci T, Markers improve clustering of CGH data. *Bioinformatics* **2007**, 23(4):450–457.

[20] Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M, Distance-based clustering of CGH data. *Bioinformatics* **2006**, 22(16):1971–1978.

[21] Guttman M, Mies C, Dudycz-Sulicz K, Diskin SJ, Baldwin DA, Stoeckert CJ, Grant GR, Assessing the

Significance of Conserved Genomic Aberrations Using High Resolution Genomic Microarrays. *PLoS Genetics* **2007**, 3(8):e143+.

[22] Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR, Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences* **2007**, :0710052104+.

[23] Taylor BSS, Barretina J, Socci NDD, Decarolis P, Ladanyi M, Meyerson M, Singer S, Sander C, Functional Copy-Number Alterations in Cancer. *PLoS ONE* **2008**, 3(9).

[24] Lai C, Horlings HHM, van de Vijver MMJ, van Beers EH, Nederlof PM, Wessels LFA, Reinders MJT, SIRAC: Supervised Identification of Regions of Aberration in aCGH datasets. *BMC Bioinformatics* **2007**, 8:422+.

[25] van de Wiel MA, van Wieringen W, CGHregions: Dimension reduction for array CGh data with minimal information loss. *Cancer Informatics* **2007**, 2:55–63.

[26] Tusher VG, Tibshirani R, Chu G, Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **2001**, 98(9):5116–5121.

[27] Kim TM, Jung YC, Rhyu MG, Jung MH, Chung YJ, GEAR: genomic enrichment analysis of regional DNA copy number changes. *Bioinformatics* **2008**, 24(3):420–421.

[28] Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, Smith L, Greenfield A, Tiganescu A, Buckle V, Ventress N, Ayyub H, Salhan A, Pedraza-Diaz S, Broxholme J, Ragoussis J, Higgs DR, Flint J, Knight SJ, 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.

[29] Klijn C, Holstege H, de Ridder J, Liu X, Reinders M, Jonkers J, Wessels L, Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic acids research* **2008**, 36(2).

[30] Weir B, Woo M, Getz G, Perner S, Ding L, Beroukhim R, Lin W, Province M, Kraja A, Johnson L, Shah K,

Sato M, Thomas R, Barletta J, Borecki I, Broderick S, Chang A, Chiang D, Chirieac L, Cho J, Fujii Y, Gazdar A, Giordano T, Greulich H, Hanna M, Johnson B, Kris M, Lash A, Lin L, Lindeman N, Mardis E, Mcpherson J, Minna J, Morgan M, Nadel M, Orringer M, Osborne J, Ozenberger B, Ramos A, Robinson J, Roth J, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz M, Tsao MS, Twomey D, Verhaak R, Weinstock G, Wheeler D, Winckler W, Yoshizawa A, Yu S, Zakowski M, Zhang Q, Beer D, Wistuba I, Watson M, Garraway L, Ladanyi M, Travis W, Pao W, Rubin M, Gabriel S, Gibbs R, Varmus H, Wilson R, Lander E, Meyerson M, Characterizing the cancer genome in lung adenocarcinoma. *Nature* **2007**, 450(7171):893–898.

[31] 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.

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

[33] 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.

[34] Lingjaerde OC, Baumbusch LO, LiestÃ¸l K, Glad IK, Borresen-Dale AL, CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **2005**, 21:821–822.

[35] 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.

[36] Venkatraman ES, Olshen AB, A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **2007**, 23(6):657–663.

[37] 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(2):215–228.

[38] Ben-Dor A, Lipson D, Tsalenko A, Reimers M, Baumbusch L, Barrett M, Weinstein J, Borresen-Dale A, Yakhini Z, Framework for Identifying Common Aberrations in DNA Copy Number Data. *Proceedings of RECOMB '07* **2007**, 4453:122–136.

[39] Edgington E, Onghena P: *Randomization Tests, Fourth Edition (Statistics: a Series of Textbooks and*

Monographs). Chapman & Hall/CRC **2007**.

[40] Sellke T, Bayarri MJ, Berger JO, Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician* **2001**, 55:62–71.

[41] Wood LDD, Parsons DWW, Jones S, Lin J, Sjöblom T, Leary RJJ, Shen D, Boca SMM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PAA, Kaminker JSS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKVK, Sukumar S, Polyak K, Park BHH, Pethiyagoda CLL, Pant PVKV, Ballinger DGG, Sparks ABB, Hartigan J, Smith DRR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SDD, Parmigiani G, Kinzler KWW, Velculescu VEE, Vogelstein B, The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* **2007**, 318:1108–1113.

[42] Dudoit S, Gentleman RC, Quackenbush J, Open source software for the analysis of microarray data. *Biotechniques* **2003**, Suppl:45–51.

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

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

[45] Madeira SC, Oliveira AL, Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2004**, 1:24–45.

[46] Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E, A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **2006**, 22(9):1122–1129.