BMC Bioinformatics

Methodology article

Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously

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

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

Email: Oscar M Rueda* - rueda.om@gmail.com; Ramon Diaz-Uriarte* - rdiaz02@gmail.com

* Corresponding authors

Published: 23 September 2009

BMC Bioinformatics 2009, **10**:308 doi:10.1186/1471-2105-10-308

Received: 1 April 2009
Accepted: 23 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/308

Abstract

Background: Alterations in the number of copies of genomic DNA that are common or recurrent among diseased individuals are likely to contain disease-critical genes. Unfortunately, defining common or recurrent copy number alteration (CNA) regions remains a challenge. Moreover, the heterogeneous nature of many diseases requires that we search for common or recurrent CNA regions that affect only some subsets of the samples (without knowledge of the regions and subsets affected), but this is neglected by most methods.

Results: We have developed two methods to define recurrent CNA regions from aCGH data. Our methods are unique and qualitatively different from existing approaches: they detect regions over both the complete set of arrays and alterations that are common only to some subsets of the samples (i.e., alterations that might characterize previously unknown groups); they use probabilities of alteration as input and return probabilities of being a common region, thus allowing researchers to modify thresholds as needed; the two parameters of the methods have an immediate, straightforward, biological interpretation. Using data from previous studies, we show that we can detect patterns that other methods miss and that researchers can modify, as needed, thresholds of immediate interpretability and develop custom statistics to answer specific research questions.

Conclusion: These methods represent a qualitative advance in the location of recurrent CNA regions, highlight the relevance of population heterogeneity for definitions of recurrence, and can facilitate the clustering of samples with respect to patterns of CNA. Ultimately, the methods developed can become important tools in the search for genomic regions harboring disease-critical genes.

Background

Genomic DNA copy number is often variable. Some of this variability, commonly referred as copy number variations or CNVs, is naturally present in the germ line and thus heritable [1-3], whereas somatic, large-scale alterations that often characterize tumor cells are called copy number alterations or copy number aberrations (CNAs) [3-6]. These CNAs are often longer than CNVs and have been linked to other diseases in addition to cancer, such as HIV acquisition and progression, autoimmune diseases, and Alzheimer and Parkinson's disease [7-10]. The most popular current approaches for the identification of DNA copy number differences are chip- or array-based. These include SNP arrays [11-13] and array-based Com-

parative Genomic Hybridization (aCGH). aCGH is a broad term that encompasses oligonucleotide aCGH (Agilent, NimbleGen, and occasionally in-house oligonucleotide arrays), BAC and, less frequently nowadays, ROMA and cDNA arrays [14,15]. In addition to the array-based technologies, sequencing-based approaches [2,16-18] are also used to study CNAs. (See [3] for differences on the identification of CNVs and CNAs, and the specific challenges associated to the reliable detection of CNAs, that are due to tissue heterogeneity and contamination and uncertainty about baseline ploidy). Location of CNAs 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 show alterations that are recurrent among diseased individuals [15,19-21]. In this context, we can define a recurrent CNA region as a set of contiguous genes (a region) that shows a high enough probability (or evidence) of being altered (e.g., gained) in at least some samples or arrays. Unfortunately, although many methods exist for analyzing a single array (e.g., see comparisons and references in [22-25]), few papers deal with the problem of integrating several samples and finding CNA regions that are common over sets of samples. Thus, merging data from several samples to find recurrent CNA regions remains a challenge [6], both methodologically and conceptually.

Two recent reviews [4,26] highlight the main features and difficulties of existing methods. Most methods [19,20,27-30] try to find recurrent CNA regions using, as starting point, the discrete output from an aCGH segmentation algorithm in the form of the classification of every probe into gained, normal or lost. Because these methods use discretized output, they discard any available estimate of the uncertainty of these estimates; as a consequence, a gain for which there is strong evidence will have the same weight in subsequent calculations as another gain for which there is less certainty. Moreover, the majority of these methods ignore within- and among-array variability in aCGH ratios as they use a common threshold for all probes and arrays. A few other methods perform the segmentation and search for recurrent CNA regions in the same step [31-33]. The method in [33], which does not use nor returns probabilities, employs elaborate and heuristic approaches to search over possible thresholds and adjustments for multiple testing. Another two methods, [34,35], intertwine, in a complex way, biological assumptions and statistical procedures, leading to convoluted, heuristically based methods, with critical assumptions and parameters of difficult interpretation and assessment (see also [4] for a critique of the attempts to differentiate between "driver" and "passenger" mutations). In [31] copy numbers of contiguous probes as treated as independent, which is clearly biologically unrealistic. Hidden Markov Models are used by [32], but this method seems to locate recurrent probes, not recurrent regions, and the number of states is restricted to four; therefore, all the gains are grouped into a single state with a common mean, which is biologically unreasonable, and makes it impossible to differentiate between samples with moderate amplitude changes and large-amplitude changes.

In addition to the above difficulties, one the most serious problems of existing methods is the inability to find common regions over subset of samples. The majority of approaches [27,28,31,32,34-37] try to find regions that are common to all the arrays in the sample. Thus, these methods presuppose that a disease is homogeneous with respect to the pattern of CNAs. It is known, however, that for many complex diseases, such as cancer or autism [38-40], molecular subphenotypes are common. It follows that heterogeneity should be appropriately addressed [4] in studies of recurrent CNA regions. Two methods [19,33] (see also reviews in [4,26]) try to find recurrent regions defined over a subset of the samples but, in addition to not using probabilities, they depend on a resolution (or number of bins) parameter that controls the number of probes considered within region, so that, given this parameter, the method, by construction, will regard either all or none of the probes as jointly altered. But the point of searching for regions is, precisely, to identify regions for which we do not know in advance location, number of subjects, or length. Moreover, there are concerns [36] about the permutation strategy used by the above two methods to assess the statistical significance of the patterns found, as it precludes locating large aberrations. Therefore, there are currently no satisfactory approaches for addressing among-sample heterogeneity.

To further clarify and understand this problem, we can 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 most existing methods for the study of recurrent CNA regions try to do. In a second scenario, we suspect that the subjects are 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 CNA 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 inappropri-

ate: it does not answer the underlying biological question, risks missing relevant signals, and leads to conceptual confusion.

Existing methods, therefore, have serious limitations and it is necessary to develop new approaches that fulfill the following three major requirements. First, we want to explicitly differentiate between the two scenarios in the last paragraph. As a consequence, we want to be able to locate either regions common to most of the arrays or regions common to only a subset of the arrays. Second, we want to preserve the uncertainty in the state of a probe (probability of alteration), and we want to return probabilities, as a probability is the single most direct answer to the question "is this region altered over this set of arrays?" (a p-value does not directly answer this question, but rather provides support against a specific null hypothesis). Third, we want that the biological meaning of the regions found be immediate, which we can try to achieve by using methods that depend on few parameters of straightforward interpretation. We have developed two approaches that fulfill these criteria.

Results

Two different approaches for finding recurrent CNA regions

Here we provide an intuitive understanding of our two different approaches. Further details are provided below.

Our first method, **pREC-A** (probabilistic recurrent copy number regions, common threshold over all arrays), finds those regions that, over the complete set of arrays, show an average (over arrays) probability of being altered that is above a predefined threshold. When using **pREC-A** we only need to provide one threshold, p_a, the minimal probability of alteration of a region over a set of arrays. p_a is chosen by the researcher, but generally cannot be too stringent (e.g., will rarely be larger than 0.80) because even with a large number of arrays, only a few arrays without that alteration will prevent finding the region (as we are averaging over arrays).

Our second method, **pREC-S** (probabilistic recurrent copy number regions, subsets of arrays), identifies all common regions over subsets of arrays; alternatively, we can think of this algorithm as identifying subsets of arrays that share regions of alteration. The regions of alteration found might not be common to most arrays, but within each array in the identified subset, the regions of alteration will have a probability of being altered above a threshold (p_w). When using **pREC-S**, therefore, the user needs to provide two thresholds, p_w, the minimal probability of alteration of a region in every array in the selected subset, and *freq.array*, the smallest number of arrays (i.e., the smallest size of the subset of arrays) that share a common region. Here we will often use more stringent thresholds for probability (e.g., p_w = 0.90), because those high probabilities might be attained over a highly homogeneous and small subset of arrays. We can use the output of **pREC-S** as the basis for clustering and to display patterns of groupings of arrays; an example is shown below (see "Simple numerical example: **pREC-S**").

For both methods, we will use probabilities of alteration as returned, for example, by RJaCGH [24]. RJaCGH is a Hidden Markov Model-based approach that returns probabilities of alteration of probes and segments; no hard thresholds are imposed, and thus the user decides what constituted sufficient evidence (in terms of probability of alteration) to call a probe gained (or lost). We have shown [24,25] that RJaCGH performs as well as, or better than, competing methods in terms of calling gains and losses, and the relative advantage of the method increases as the variability in distance between probes increases. It is essential to understand that the probabilities that we use are not the marginal probabilities of alteration but the joint probabilities of alteration of a region of probes (see details in "Computation of the joint probability of an arbitrary sequence of probes in an array"). Our approach incorporates both within-and among-array variability (as it is based on the hidden process of alterations and uses the probability of every probe in every array): we use the information on the certainty of each call of gain/loss (i.e., the probability) in all computations of recurrent CNA regions. Therefore, our approach is qualitatively different from using the same threshold over all probes and arrays. See further details below. Moreover, using probabilities of alteration (instead of magnitude of change), in addition to differentiating between evidence of alteration and estimated fold change, prevents inter-array differences in range of log_2 ratios and tissue mixture to get confounded with evidence of alteration. Finally, note that we use at most two parameters and that their biological meaning is immediate: probability of alteration, and number of samples that share an alteration (the later only needed for **pREC-S**).

Algorithms

Before we can develop algorithms for the two approaches, **pREC-A** and **pREC-S**, we will need to develop methodology that will allow us to: 1) compute the joint probability of alteration of an arbitrary sequence of probes; 2) combine that probability over arrays. The first two parts of this section detail this machinery before showing the details of the algorithms. For the rest of this section, please bear in mind that we are always referring to probabilities of alteration, and never to p-values. We are working on a Bayesian framework and are estimating posterior probabilities; we are not conducting hypothesis tests.

Computation of the joint probability of an arbitrary sequence of probes in an array

To find altered regions, that is, sets of contiguous probes, we have to compute the joint probability of alteration for a sequence of probes. In other words, we need to compute, for each array $i = 1,... r$, the probability that a subset of consecutive probes is, for example, gained (the problem for losses is equivalent). That is, if we denote as S_i the state of probe i and with 1 the state 'gain', we are interested in $P(S_j = 1,..., S_{j+p} = 1)$ for a subset of contiguous p probes. (Note that, strictly, we can find $P(S_j = 1,..., S_{j+p} = 1)$ also for the case of non-contiguous probes, but this scenario is unlikely to be of any interest in the search for recurrent CNA regions.)

Using RJaCGH (or other methods) we can compute the probability for every probe to belong to any of the states of gain and to any of the states of loss. The problem of these probabilities is that they are marginal probabilities: they are the probability of the event of an alteration of a probe without considering the alteration of other probes, in particular of neighboring probes. But the states of the probes are not independent [24], and thus the probability of alteration of a region (within an array) can not be computed simply as the product of the probability of the individual probes.

With HMM it is customary to obtain the most likely path of hidden states using the Viterbi algorithm which returns the maximum a posteriori sequence (MAP). The Viterbi algorithm, however, does not return any distributional statements about the states of the path [41]. It is straightforward, however, to compute the marginal probabilities of the state of a probe or the joint probabilities of an arbitrary sequence of probes, because the sequence of hidden states conditioned on the parameters of the HMM is a Markov Chain [41]. For instance, we could compute the probability that the first three probes are jointly gained: $P(S_1 = 1, S_2 = 1, S_3 = 1)$ using straightforward conditional probabilities as $P(S_1 = 1)P(S_2 = 1|S_1 = 1)P(S_3 = 1|S_2 = 1)$, and these conditional probabilities can be computed by backward-smoothing. The problem is that the classification of probes or regions into states given by these two approaches (Viterbi and backward-smoothing) does not always coincide, leading to inconsistencies. For example, we might obtain a sequence of hidden states with maximum marginal probabilities that is not the same as we obtain with Viterbi; that sequence might even contain two consecutive altered probes that can not be jointly altered [42]. This is a common problem that can arise when using maximum likelihood approaches to HMM.

To avoid these problems, we can use, as RJaCGH does, Markov Chain Monte Carlo (MCMC) instead of Maximum Likelihood (ML). With MCMC, however, we can not average the conditional probabilities obtained through the MCMC iterations, because that would break the Markovian property [43], as we are averaging over different runs with (potentially) different values for the model parameters (as new values for the parameters are drawn at each iteration of the MCMC). For instance, suppose we want to compute the probability that the first three probes are jointly gained: $P(S_1 = 1, S_2 = 1, S_3 = 1)$. We cannot compute $P(S_1 = 1)P(S_2 = 1|S_1 = 1)P(S_3 = 1|S_2 = 1)$, with those conditional probabilities obtained by averaging over the multiple MCMC runs. What we can do, instead, is compute the probability of an alteration for any arbitrary sequence as the frequency of that sequence being altered in the MAPs from each of the MCMC draws. For the previous example, we would count in how many MAPs (from Viterbi) we found $S_1 = S_2 = S_3 = 1$. We must note that, in this case, we are not obtaining the real distribution of the hidden states per se, but the distribution of the hidden states as members of the maximum a posteriori hidden sequence [44]. That is, we do not sample from the distribution of the hidden states, but from the distribution of the MAP. This is coherent with the classification method used with just one array, as every sequence is only accounted for if it has been part of the MAP sequence, and thus this is a stronger requirement as the regions obtained have always been part of the MAP.

Finally, the above scheme can be applied both to models that assign to hidden states probabilities of being altered of either 1 or 0, and to models that assign to hidden states probabilities of being altered between 0 and 1.

Combining regions over arrays

Once we have computed the probability that the above region is altered, for our first algorithm, **pREC-A**, we need to know how to average over the arrays to get a probability of alteration for that region over a set of arrays. Many HMM models (RJaCGH included) will model each array with a different HMM, to reflect the fact that they can have different characteristics, such as dispersion. Thus, for each array, we have a (potentially different) stochastic process for the log-ratios. Once the data are summarized as states (gain, loss, no-change), however, they are comparable across arrays as we are using the same approach to label probes as gained/lost/not-changed. In other words, a value of $S_j = 1$ has the same meaning regardless of the array. Thus, we can average directly all the probabilities for every array (the averages might be weighted if there are differences in the reliability or the precision of different arrays). Therefore, the probability that a given region of the genome is altered over a set of arrays is computed as:

$$P(S_i = 1,..., S_{i+p} = 1) = \sum_{j=1}^{r} P(S_i = 1,..., S_{i+p} = 1 \mid array_j) P(array_j) \quad (1)$$

where different $P(array_j)$ allow us to use different weights for different arrays (and, of course, the $P(array_j)$ are scaled, if needed, so that $\sum_j P(array_j) = 1$).

For notational convenience, when there is only one probe, we define

$$P(S_i = 1) = \sum_{j=1}^{r} P(S_i = 1 \mid array_j) P(array_j) \quad (2)$$

pREC-A: Finding regions with a probability of alteration of at least p_a

The following algorithm (Table 1) finds all the regions with an average (average over all arrays) probability of alteration of at least p_a. This algorithm is the one that is most similar to other existing approaches in objective. Notice, however, the simplicity of our algorithm, and the straightforward interpretation of its parameters. A detailed explanation of each line of the algorithm and its logic is provided in the Additional file 1.

pREC-S: Finding all the regions shared by at least freq.array *arrays where each region in each array has a probability of at least* p_w

In this algorithm (Table 2) we are imposing two thresholds: 1) p_w, the minimum joint probability, within array, for each region; 2) *freq.arrays*, the minimum number of arrays that share the alteration. Notice that p_w in this algorithm is different from p_a in the previous algorithm (where averaging over arrays is used). This algorithm has no equivalent in alternative methods. A detailed explanation of each line of the algorithm and its logic is provided in the Additional file 1.

Table 1: pREC-A algorithm

```
 1  Start ← 1
 2  while Start ≤ Total Number Of Probes do
 3    P1 ← P(S_Start = 1);
 4    if P1 ≥ p_a then
 5      End ← Start + 1;
 6      while End ≤ Total Number Of Probes do
 7        P2 ← P(S_Start,..., S_End = 1);
 8        If P2 <p_a then
 9          break out of the while loop;
10        else
11          P1 ← P2;
12          End ← End + 1;
13        end
14      end
15      UpdateRegionA(Start, End - 1, P1);
16      Start ← End;
17    else
18      Start ← Start + 1;
19    end
20  end
```

Table 2: pREC-S algorithm

```
 1  for Start ← 1 to Total Number Of Probes do
 2    SetArrays_A ← φ;
 3    for array ← 1 to Total Number Of Arrays do
 4      if P(S_Start = 1|array) ≥ p_w then
 5        SetArrays_A ← SetArrays_A ∪ array;
 6      end
 7    end
 8    if |SetArrays_A| ≥ freq.arrays then
 9      End ← Start + 1;
10      while End ≤ Total Number Of Probes do
11        SetArrays_B ← φ;
12      foreach candidate array in SetArrays_A do
13          if P(S_Start,..., S_End = 1|candidate_array) ≥ p_w then
14            SetArrays_B ← SetArrays_B ∪ candidate_array;
15          end
16        end
17        if |SetArrays_B| <freq.arrays then
18          break out of the while loop
19        else
20          if |SetArrays_B| < |SetArrays_A| then
21            UpdateRegionS(Start, End - 1, SetArrays_A);
22            SetArrays_A ← SetArrays_B;
23          end
24          End ← End + 1;
25        end
26      end
27      UpdateRegionS(Start, End - 1, SetArrays_A);
28    end
29  end
```

Simple numerical example: pREC-A

Suppose we have fit a model to six probes and four arrays and, after using RJaCGH's model averaging, we have obtained the marginal probabilities of gain shown in Table 3. We want to use **pREC-A** with $p_a = 0.6$. First, we average the probability for probe 1 for the four arrays:

$$P(S1 = Gain) = \frac{0.17+0.16+0.08+0.16}{4} = 0.14$$

As it does not reach the threshold of 0.6, S1 can not belong to a region. We do the same for S2, obtaining 0.35. For S3 the averaged probability is 0.97, so the first region will include this probe. To see if we can extend this region to the next probe, we compute for every array the joint probability of probes 3 to 4 to be gained. This probability

Table 3: Simulated data example.

	S1	S2	S3	S4	S5	S6
A1	0.17	0.17	0.97	0.97	0.97	0.17
A2	0.16	1.00	1.00	1.00	0.15	1.00
A3	0.08	0.07	0.93	0.07	0.06	0.92
A4	0.16	0.16	0.99	1.00	1.00	1.00

Marginal probabilities of being gained.

is not shown in the table above (which shows only marginal probabilities) but is obtained as explained above (see section "Computation of the joint probability of an arbitrary sequence of probes in an array"): the relative frequency of a sequence in the MAPs from all the MCMC samples.

$$P(S3 = Gain, S4 = Gain) = \frac{0.97+1+0.07+0.99}{4} = 0.76$$

As it is over the threshold, we join S4 to the region.

Now we check if S5 can be joined too. We compute the joint probability of gain for the probes 3 to 5 (again, the joint probability is computed from the relative frequency of this sequence in the MAPs from all the MCMC samples):

$$P(S3 = Gain, S4 = Gain, S5 = Gain) = \frac{0.97+0.15+0.06+0.99}{4} = 0.54$$

As it does not reach 0.6, S5 will not be part of the region, so we get:

Region 1: {(S3, S4)}.

Now we keep on searching from probe 5. S5 does not have a marginal probability higher than the threshold, so it will not form any region. But S6 will:

$$P(S6 = Gain) = \frac{0.17+1.00+0.92+1.00}{4} = 0.77$$

So it will form its own region. As there are no more probes, the regions found are {(S3, S4), (S6)}.

Boundaries of regions are forced to be common over all arrays: the algorithm finds the common regions. For instance, the left boundary of the first region of gain of sample A2 is located in probes S2, whereas the boundary for all the other three samples is located in S3. Thus, S2 is excluded from the first common region: a region that spanned {(S2, S3, S4)} would not reach, over all four arrays, the required $p_a = 0.6$.

Simple numerical example: pREC-S

We use the same data as above. We want to find all regions where at least two arrays have a joint probability of gain of at least 0.9 (note that we raise the probability threshold because we do not ask that, on average, all arrays reach it, but at least two of them do). In other words, we are using **pREC-S** with *freq.arrays* = 2 and $p_w = 0.90$. Line numbers below refer to the lines in the algorithm.

We start on S1, but there is no array that reaches the threshold of 0.9 for that probe (i.e., the condition in line 4 of Table 2 is not fulfilled for any array). We iterate (line 1 of Table 2) to the next probe, S2, but the threshold is reached only in Array 2, and we imposed that there should be at least 2 arrays. Thus, condition in line 8 is not met. We iterate to the next probe, S3. Here, when we iterate over all the arrays (line 3) we find all of the arrays reach the threshold, so in line 5 we end up with *SetArrays_A* = (*A*1, *A*2, *A*3, *A*4). As the condition in line 8 is fulfilled we try to increase the region by one probe: we set *End* to S4 (line 9) and enter the "while" loop (line 10) as we are not yet at the end of the total number of probes.

After looping over all four arrays (line 12) we find that line 13 is only fulfilled for Arrays 1, 2 and 4:

$$P(S3 = Gain, S4 = Gain \mid A1) = 0.97$$
$$P(S3 = Gain, S4 = Gain \mid A2) = 1.00$$
$$P(S3 = Gain, S4 = Gain \mid A4) = 0.99$$
$$P(S3 = Gain, S4 = Gain \mid A3) < 0.90$$

Note that the last expression is obvious since $P(S4 = Gain|A3) = 0.07$.

Therefore (from the iteration over line 14) we have *SetArrays_B* = (*A*1, *A*2, *A*4). We still fulfill the condition about *freq.arrays* in line 17, but the new set of arrays contains fewer than before (line 20) which means that in the step before a region was found. We call *UpdateRegionS* so that the region ((*S*3), (*A*1, *A*2, *A*3, *A*4)) is stored, and we set *SetArrays_A* = (*A*1, *A*2, *A*4) (line 22). We increase *End* to S5 (line 24), and consider it as the end of the new possible region. Iterating again (line 12) we find

$$P(S3 = Gain, S4 = Gain, S5 = Gain \mid A1) = 0.97$$
$$P(S3 = Gain, S4 = Gain, S5 = Gain \mid A4) = 0.99$$
$$P(S3 = Gain, S4 = Gain, S5 = Gain \mid A2) < 0.90$$

As above, this means that in the previous step we found a region (line 20 is true). Therefore, we call *UpdateRegionS* to store the region from the previous step: ((*S*3, *S*4), (*A*1, *A*2, *A*4)). We increase *End* to S6 and find

$$P(S3 = Gain, S4 = Gain, S5 = Gain, S6 = Gain \mid A1) < 0.90$$
$$P(S3 = Gain, S4 = Gain, S5 = Gain, S6 = Gain \mid A4) = 0.99 \quad (3)$$

Now, the condition in line 17 is true, because only one array satisfies being over p_w. We break out of the while loop (line 19) and we *UpdateRegionS* in line 27, so we store the region from the previous step:((*S*3, *S*4, *S*5), (*A*1, *A*4)).

We continue iterating over *Start* (line 1), so now *Start* = *S*4. Repeating the steps above we would find a first region ((*S*4), (*A*1, *A*3, *A*4)), and a second region ((*S*4, *S*5), (*A*1, *A*4)). However, when executing *UpdateRegionS*, we would find each of these regions is a subset of a previously found region (((*S*4), (*A*1, *A*3, *A*4)) of ((*S*3, *S*4), (*A*1, *A*3, *A*4)); ((*S*4, *S*5), (*A*1, *A*4)) of ((*S*3, *S*4, *S*5), (*A*1, *A*4))).

When we iterate over *Start* to *Start* = *S*5, we find only the region ((*S*5), (*A*1, *A*4)) which is again a subset of a previously found region.

Finally, we set *Start* = *S*6. We find (lines 3 and 4) that p_w is satisfied by arrays A2, A3, A4, so we end up with *SetArrays_A* = (*A*2, *A*3, *A*4). We fullfill the requirement about *freq.arrays*, but in line 10, however, we find we are at the end of the total number of probes, so we do not enter that loop (lines 11 to 24 are skipped). We therefore call *UpdateRegions*, and add the region ((*S*6), (*A*2, *A*3, *A*4)). (Note that the call to *UpdateRegions* in line 27 with *End* - 1 is correct, since we increased *End* one position over S6 in line 9). Therefore, we end up with the regions:

Regions = {((S3), (A1, A2, A3, A4)), ((S3, S4), (A1, A2, A4)), ((S3, S4, S5), (A1, A4)), ((S6), (A2, A3, A4))}

We can see the regions obtained in Figure 1. In contrast to **pREC-A**, boundaries need not be common over arrays; with **pREC-S** differences in boundaries will lead to different subsets and different regions (for instance, that is why the common region (S3, S4) includes only samples A1, A2, A4, but not A3).

We can also use the output of this algorithm as the basis for clustering and to display patterns of groupings of arrays. We can measure similarity between two arrays as the number of common probes in recurrent CNA regions between those two arrays or, alternatively, as the number of common regions (where the same probe might belong to more than one region) between two arrays. Once similarity is measured, we can immediately apply any clustering method of our choice. An example is show in Figure 2. At this stage, clustering is mainly a device for representing patterns of similarity, since the grouping of arrays with respect to recurrent CNVs is the very output of the **pREC-S** algorithm.

Implementation and testing

The algorithms above are part of the freely available and open-source RJaCGH R package (available from the R repositories), which uses R and C (the later, dynamically loaded from within R). For storage and efficiency reasons, we do not save directly all of the Viterbi paths (i.e., each Viterbi from each iteration of the MCMC sampler) but only the jumps in paths and the counts of different paths. This requires less storage, allows for faster access to the information and computation of the joint sequence, and of course permits reconstructing all of the sequences. The Viterbi paths are obtained as part of the regular execution of the C code for RJaCGH, saved in R as gzipped files, and read back by the C functions for **pREC-A** and **pREC-S** only once.

Execution time in all the examples of the paper is negligible: all the examples of pREC-A execute in less than 5 seconds. Execution time for pREC-S goes up to 160 seconds for the examples from [45] but less than 4 seconds for the remaining examples. (All these timings from a workstation with an AMD 280 processor running Debian GNU/Linux).

Testing was carried out by comparing the output from the algorithms with manually computed examples. Code for the examples and comparisons is included in the repository for the package http://launchpad.net/rjacgh.

Examples with real data and comparison to other approaches

All the examples below were analysed with RJaCGH, which provided the probabilities of alteration. Our examples use aCGH arrays because these are three "classic" sets of data that have been analyzed before with other approaches. Our methods, however, can also be applied to other platforms, including custom and commercial oligonucleotide arrays and SNP arrays (e.g., [25]) or any other platform for which we can obtain joint probabilities of alteration. The main objective of these examples is to illustrate the range of analysis that can be performed.

Colorectal cancer example (Nakao et al.): direct application of pREC-A

Nakao et al. [46] analyze 125 colorectal tumors. They apply a segmentation method based on a threshold and then find common regions of alteration studying the frequency of alterations. Rouveirol et al. [28] apply both of their algorithms for minimal common regions to the same data. As shown in the Additional file 1, using **pREC-A** with a threshold of 0.35, we find basically the same regions of alteration, and most of the reported differences come from regions with a probability (or frequency, in the case of [46]) in the limit of 35%. The only remarkable case is the gain in 11q which has a much lower probability in our analysis, probably because that alteration is based on a single BAC and the segmentation analysis used in [46] is based on a threshold and therefore is more likely to be affected by outliers.

Colorectal cancer example (Douglas et al.): comparing probability of alterations between groups using pREC-A

pREC-A can also be used to compare the probability of alteration between groups of samples. Douglas et al. [47] present data from 37 primary cancers. Seven show microsatellity instability (MSI) and 30 show chromosomal

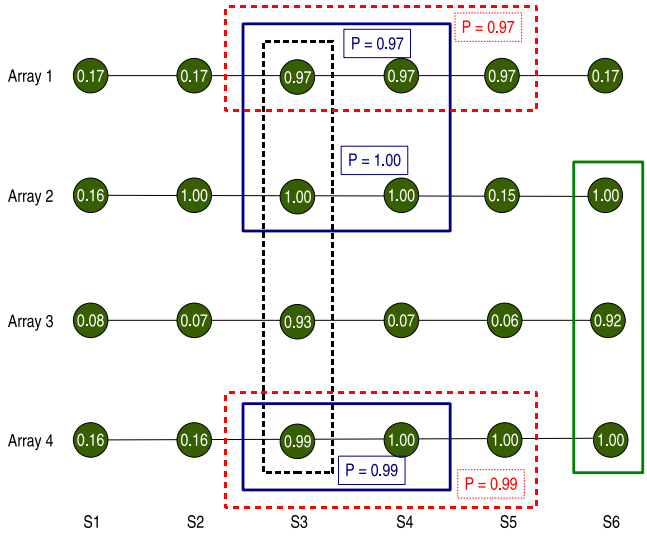

Figure 1
pREC-S, simple numerical example. Subsets of at least 2 arrays that share common regions of gain of at least 0.90 probability: *freq.arrays* = 2, p_w = 0.90. Boxes of the same color represent the same region. In circles, the marginal probabilities of gain. In boxes, the joint probabilities.

instabillity (CIN). (For a definition of genetic alterations, see [48]). They call alterations using a threshold-based method and compare their frequency between the two types using a chi-square statistic. van de Wiel and van Wieringen [49] analize the same data using a dimension reduction technique (CGHRegions) after segmenting the data with DNACopy [50]. They then use a Wilcoxon test with FDR correction for the difference between the two levels.

We have used a threshold of p_a = 0.50 to find the common regions of gain/loss and have then compared the probability of alteration in those regions for the two groups of samples. We have obtained a total of 21 regions of gain and 11 of loss, shown in Additional file 1 - Figure S2. Next, for every region found above we computed the joint probability of alteration for each of the 30 arrays of class CIN and the seven arrays of class MSI and, by region, we calculated the absolute value of the difference in mean probability between the MSI and CIN groups. To assess the significance of this statistic, we used a permutation test (randomly permuting the MSI and CIN labels and recomputing the absolute value of the difference in mean probability) to obtain a two-sided p-value. Then, we applied the FDR method [51] for multiple testing correction (to account for the multiple testing arising from comparing multiple regions). The regions found significantly different (at 0.05 level) between groups are listed in Additional file 1 - Table S3, where we also provide further details about the differences with the results in [47] and [49]. Our results are largely coincident with those in [47] and [49]. Some regions mentioned in [47] (a two-clone region in chromosome 8, a 29-clone region in chromosome 18, and the p arm of chromosome 20) are not detected by us as these are regions with probability of alteration just below 0.50. There are differences with the method of [49], CGH-regions, in the location of the breakpoints: CGHregions is a dimension reduction method that simplifies the complexity of the sample profiles, which probably leads to a larger imprecision in the location of region boundaries.

Breast cancer example (Pollack et al.): pREC-S and homogeneity index

Pollack et al. [45] analyze data from 44 breast tumors and 10 cancer cell lines. They search for common regions of alteration and then compare the frequency of aberrations in each arm of every chromosome as a function of other variables such as tumor grade, estrogen receptor (ER) and TP53 mutations. Rouveirol et al. [28] also analyze these data. We have applied our second method, pREC-S, to the 44 tumors to examine if there is any similarity in the alterations shared by the groups of arrays defined by those variables. We have computed common regions of at least 0.50 probability of alteration (Gains or Losses) shared by at least two arrays (i.e., *freq.array* = 2, p_w = 0.50).

To compare our approach with the results of [45], and to gain more insight on the patterns of recurrent CNA regions and their relationship to the other three variables (tumor grade, ER, TP53), we have defined a simple statistic to measure within-group homogeneity of recurrent CNA regions. Let Y_{ij} be the number of probes that array i and array j have altered in common, k a group of arrays (typically, with some common characteristic), n_k the number of different pairs of arrays in a given group k and n_k the number of different pairs formed by arrays in group k and arrays in a different group. Let us define

$$\bar{Y}_k = \sum_{i,j \in k} \frac{Y_{ij}}{n_k}$$

$$\bar{Y}_{-k} = \sum_{i \notin k, j \in k} \frac{Y_{ij}}{n_{-k}}$$

That is, $\bar{Y}_k$ is the average number of common altered probes between two arrays of group k, and $\bar{Y}_{-k}$ is the average number of common altered probes between one array of group k and other in a different group. We define the proportion of common alterations shared by the group k as $\bar{Y}_k / \bar{Y}_{-k}$. This index measures the homogeneity of the genomic alterations within a subset of arrays compared to the alterations shared with arrays of other group. If this

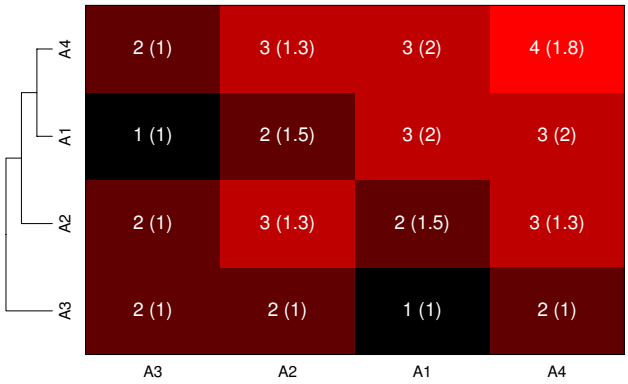

Figure 2
Clustering based upon pREC-S. Number of common regions shared by pairs of arrays. In parenthesis, the average length in probes of the regions. On the left, a dendrogram using hierarchical clustering (complete linkage) with number of common regions shared by pairs of arrays as similarity measure.

index is greater than 1, the arrays of this group share more alterations between themselves than arrays of different groups do. If this index is 0, no alterations are shared between any two arrays in the group. A value of 8 means that no alteration is shared between arrays of this group and others. We can compute this index for the groups defined by the three variables tumor grade, ER, and TP53 mutations; this is shown in the Additional file 1 - Tables S4 to S6). Those tables allow us to easily discern chromosomes that are very homogeneous with respect to shared alterations; for instance, gains in chromosomes 4 and 5 and losses in chromosome 8 are very homogeneous in the estrogen receptor negative samples (Additional file 1 - Table S4). We can display the patterns of similarity graphically, as is done in Figure 3, where we have ordered the arrays by tumor grade and show the number of common alterations for chromosome 8. Our results are not easy to compare with [45], because they define the regions and compare subgroups at chromosome arm resolution, while our method works at probe resolution.

Furthermore [45] consider every chromosome arm as altered or not without taking into account the number of altered probes in it.

To further understand the pattern of similarities, instead of comparing subgroups according to the number of alterations, we can analyze how homogeneous each group is over the whole genome (not chromosome by chromosome, as in previous tables). This is shown in Table 4. When we divide arrays according to tumor grade, Grade I and Grade III show high homogeneity within groups, meaning that the alterations are consistent in arrays within those grades. Arrays of grade II, however, show much more heterogeneity, sharing many aberrations with arrays of Grade I and/or Grade III. This is an indication that arrays of Grade II can be classified in one of the other two groups according to the pattern of alterations. Figure 2 provides a graphical illustration: four arrays of Grade II are very similar to the arrays of Grade III.

Discussion

We have developed two very different approaches for finding regions of recurrent, or common, copy number alteration. The lack of gold standards and the current non-existence of an unambiguous definition of what a region of recurrent CNA is [6], as well as the unique and qualitatively different nature of our methods from previous ones, make it difficult to compare performance, but at the same time highlight the relevance of our methods for current and future studies of CNA, their relation to phenotypic variation, and their usage for subject clustering.

The two methods we have developed share that they use as input probabilities of alteration and return probabilities. Regardless of whether the input probabilities are obtained from our RJaCGH method [24] or some other approach, it can be argued that probabilities are much better suited to the task at hand than p-values or discrete classifications into "gained", "lost", "not changed". By using probabilities as input, we incorporate uncertainty in the estimates of copy number states. By returning probabilites and using probabilities throughout all the analysis, the user can decide the appropriate thresholds (or, even, modify them depending on context) and define distances between arrays that incorporate the strength of evidence in favor of alteration. Precisely because of the conceptual simplicity of using probabilities, we can approach within a unified framework both questions related to "unsupervised problems" (e.g., identify subsets of regions that are common to subsets of arrays) and to "supervised problems" (e.g., measure how different two groups of arrays are with respect to recurrent regions of alteration). This unified approach is unique to our methods, and not shared by any others.

Our first method, **pREC-A**, searches for general, broad patterns of common gains (or losses) over all the samples in the study. This is the approach which is most similar to previous ones. This method is well suited to comparing pre-defined groups of samples. By its very nature (e.g., that an overall pattern is identified by a mean probability larger than a threshold) this method can only detect regions for which there is at least moderate evidence (medium probability) of alteration over almost all samples, or very strong evidence (high probability) of altera-

tion over an important fraction of the samples. Thus, it is easy to miss regions that are present with very high probability in a small subset of the samples. As well, mixing in the same sample very heterogenous groups will tend to smooth out the evidence of alteration, so that few common regions will be found. Alternatively, if there are very different sample sizes (different number of arrays) in the different heterogenous groups, the detected common regions will often be a subset of the common regions among the most abundant group. These features can be controlled to answer the specific study questions. First, as equation 1 shows, it is easy to weight different arrays differently, so as to increase the influence of some arrays in the final analysis. Moreover, if we know in advance that there are different subgroups of samples, we can use **pREC-A** independently in the different subgroups; for instance, when we have already subdivided the subjects in the study into homogeneous groups with respect to disease (e.g., [52]), and want to locate recurrent CNA regions common to most samples within a subgroup and possibly different from other subgroups. Finally, as our last example with the data of [47] shows, a user that understands these features of **pREC-A** can employ this algorithm to highlight the differences between subgroups and how these change as we modify the minimum required threshold for the probability of alteration. In particular, note the easy formulation of a permutation-based test for identifying the differences in the probabilities of alteration of regions between subgroups. This type of approach might be even more useful when two or more suspected subgroups are compared against a larger, reference group. The main advantages of this algorithm are that it is most similar to previous approaches, has a simple interpretation in terms of global patterns across most of the samples, and requires the specification of only one parameter. Thus, **pREC-A** will often be the method of choice if we are trying to relate major, global, recurrent patterns of CNA to variations in phenotype or to differentiate between subroups of samples. In contrast to **pREC-A**, the second method, **pREC-S**, can detect small subgroups of samples with respect to common alterations, without being adversely affected by averages over arrays or differences in number of samples in different subgroups. Moreover, different subgroups can be detected with respect to different alterations. **pREC-S**, therefore, addresses a common and distinct need that arises in any study of CNA with heterogeneous samples.

As seen in the results, this second algorithm allows us to elegantly approach some of the questions in the second example (breast cancer example, [45]). First, the derivation of a specially tailored statistic, $\overline{Y}_k / \overline{Y}_b$, to answer the relevant questions in this study is straightforward. More importantly, the second algorithm finds homogeneous subgroups, with respect to alterations, and these differences are associated with differences in three other markers (estrogen receptor status, TP53 mutation, tumor grade; see Additional file 1). In other words, **pREC-S** finds CNV that differentiate between groups. It must be emphasized that **pREC-S** has been applied to the complete set of data after specifying that the within-array probability of alteration be larger than 0.5 (i.e., $p_w = 0.50$) and that these regions be shared among, at least, two arrays (i.e., *freq.array* = 2), but the algorithm is blind to the "labels" of the arrays regarding the other markers (estrogen receptor, TP53, grade). Therefore, **pREC-S** allows us to find CNVs that differentiate between known groups (as in this case), but its systematic usage also opens the door to finding patterns of CNA that might differentiate between previously unknown groups. Moreover, there is no need for the association recurrent CNA regions-marker to be similar among different markers, specially since, as explained above, different subgroups of arrays can be detected with respect to different CNA recurrence patterns. These are features unique and characteristic of **pREC-S**, compared to all the alternative available methods.

We suggest that **pREC-S** is the method of choice when there is unknown heterogeneity among arrays in CNAs, and when we want to relate possibly non-identical subsets of samples, defined in terms of recurrent patterns of CNA, to phenotypic variation. Moreover, routine use of **pREC-S** even with apparently homogeneous groups of samples might help discover possible subtypes of diseases that might generate novel hypothesis or uncover previously unknown heterogeneities.

pREC-S is also a key method for clustering. Integrative studies that combine CNV data with other data (e.g., mRNA, SNP) often use clustering of subjects based upon the CNA data (e.g., [53,54]). The problem of most of these approaches is that, when clustering based upon the CNA data (either the gain/loss calls or the smoothed data), the measure of distance or similarity used ignores that some of the data show strong serial dependence (probes next to each other) whereas some of the data (e.g., probes in different chromosomes) are independent. Thus, in most cases the distance computed is likely to introduce serious distortions in the true distances among subjects (see also [29,55,56]). This problem is in addition to the aforementioned issues of not integrating variability and uncertainty in the gain/loss calls or smoothed means. In contrast, by using a biologically motivated and probabilistically based approach to CNA common regions, such as **pREC-S**, it will be possible to construct distance metrics and, therefore, clustering approaches, that make full usage of CNA

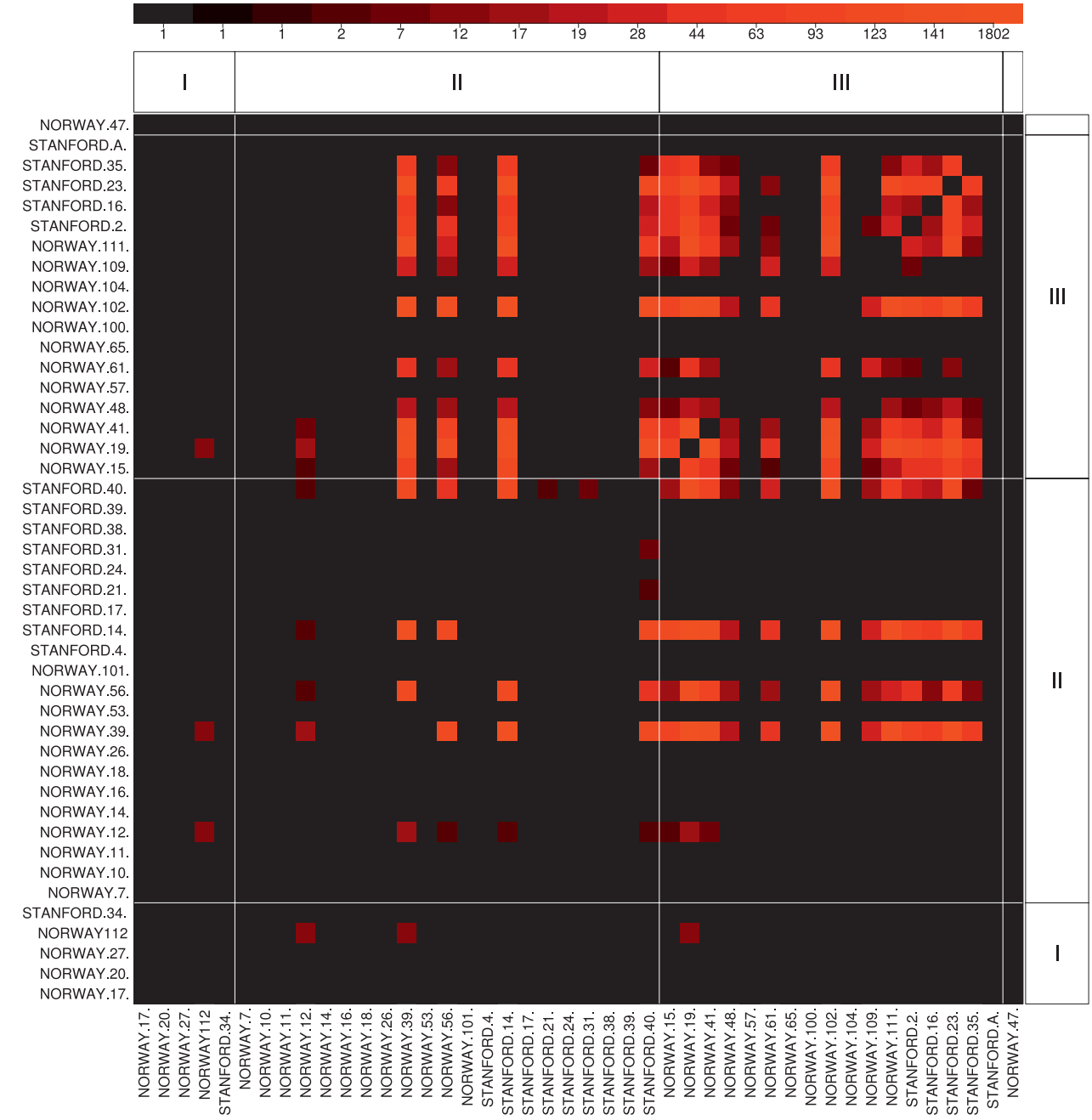

Figure 3

Chromosome 8 from the Pollack et al. example. Number of regions of gain with at least 0.50 probability shared by at least two arrays (i.e., pREC-S, *freq.arrays* = 2, p_w = 0.50). The arrays are ordered according to tumor grade. Arrays with grade III share many more alterations between them than the other arrays. Four arrays with grade II share the same gains in copy number with tumors of higher grade, so they are probably related. There is one array unidentified.

data when searching for groups of subjects. Fully developing a method for clustering based upon CNA data is outside the scope of this paper, but we have presented a simple example to motivate further work.

Moreover, an additional distinct feature of our methods is that both **pREC-S** and **pREC-A** have at most two parameters of straightforward biological interpretation (probability of alteration, number of samples that share the

Table 4: Alterations in Pollack et al. [45], genomewide.

		pREC-S (Homogeneity index)
ER	Positive	0.75
	Negative	1.12
p53	Wild Type	0.67
	Mutant	1.23
Grade	I	1.21
	II	0.56
	III	1.40

The homogeneity index, $\overline{Y}_k / \overline{Y}_b$, is computed over the whole genome, not chromosome by chromosome, as done in previous tables.

alteration). An added advantage of the type of input and output used by our methods is that probabilities allow researchers to modify thresholds as needed, and to easily (and intelligibly) examine the sensitivity of results to changes in thresholds.

Furthermore, as both methods are based on a Hidden Markov Model (HMM) with no restrictions on the number of states [24], we can use models involving an arbitrary number of states of gain and loss. The HMM (probabilistically) assigns probes to hidden states, but it is up to subsequent analysis to assign those states to specific or interesting "copy number states". This allows us to keep the two different concepts of "amplitude (or magnitude) of change" and "evidence of alteration" separate. Moreover, it is also immediate to restrict finding common regions to alterations above a certain threshold of amplitude or that belong only to a subset of states so that we can focus only on alterations of a certain type (e.g., only the largest hidden states of gain in a model with three hidden states for gain).

Finally, note that the problem we have been addressing is the location, *de novo*, of recurrent CNA regions. A different set of problems is using pre-existing information about regions that show copy number polymorphism to inform the search for rare copy number variants [57,58]. Likewise, another very different set of problems is the usage of previously identified variable regions in tests of association between copy number variation and disease [59-61]. These are, however, sufficiently related objectives, and methodological and conceptual advances in any one set of approaches could be highly beneficial for the other two sets of problems.

Conclusion

We have developed methods for finding regions of copy number alteration (CNA) common or recurrent over several arrays. Our methods have an immediate and intuitive biological interpretation, and incorporate both within- and among-array variability. Reanalysis of several data sets in the literature show that our methods can indeed recover patterns previously found but can also uncover additional patterns. Moreover, probabilities allow researchers to modify thresholds as needed, and to easily examine the sensitivity of results to changes in thresholds. In addition, the examples show how it is straightforward to derive tailored statistics and summary measures to answer specific research questions. The development of these two distinct algorithms highlights a key idea that has often been neglected: recurrent or common CNAs can refer to very distinct patterns in a group of samples, specially concerning heterogeneity among arrays and probability of alteration. We expect that these two algorithms will help advance efforts to standardize definitions of recurrent or common CNA regions, and ultimately the search for genomic regions harboring disease-critical genes.

Authors' contributions

OMR developed the statistical model, participated in the programming and conducted all of the analysis. RD-U conceived the original HMM model and participated in model development and programming. Both authors wrote, read, and approved the final manuscript.

Additional material

Additional file 1

Supplementary Material for "Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously". A PDF file with further details on the algorithms and tables and comments on the Examples.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-308-S1.PDF]

Acknowledgements

C. Lázaro-Perea, J. Poyatos, and four anonymous reviewers for comments on the ms. Funding provided by Fundación de Investigación Médica Mutua Madrileña. 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. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews DT, Fiegler 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 Hiroyuki an d Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444(7118):**444-454.
2. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM,

Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome.** *Science* 2007, **318(5849):**420-426.
3. Sun W, Wright FA, Tang Z, Nordgard SH, Loo PVV, Yu T, Kristensen VN, Perou CM: **Integrated study of copy number states and genotype calls using high-density SNP arrays.** *Nucleic acids research* 2009 in press.
4. Shah SP: **Computational methods for identification of recurrent copy number alteration patterns by array CGH.** *Cytogenetic and genome research* 2008, **123(1-4):**343-351.
5. Lee C, Iafrate AJ, Brothman AR: **Copy number variations and clinical cytogenetic diagnosis of constitutional disorders.** *Nature Genetics* 2007, **39:**S48-S54.
6. 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):**.
7. Lupski JR: **Genomic rearrangements and sporadic disease.** *Nature Genetics* 2007, **39:**S43-S47.
8. McCarroll SA, Altshuler DM: **Copy-number variation and association studies of human disease.** *Nat Genet* 2007, **39(7 Suppl):**S37-S42.
9. 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.
10. Wain LV, Armour JAA, Tobin MD: **Genomic copy number variation, human health, and disease.** *Lancet* 2009, **374:**340-350.
11. 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:**83.
12. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays.** *Nat Genet* 2007, **39(7 Suppl):**S16-S21.
13. Laframboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.** *Nucl Acids Res* 2009, **37(13):**4181-4193.
14. Ylstra B, Ijssel P van den, 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.
15. Pinkel D, Albertson D: **Array comparative genomic hybridization and its application in cancer.** *Nature Genetics* 2005, **37(Supplement):**S11-S17.
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(13):**i59-67.
18. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453(7191):**56-64.
19. 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.
20. Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L, Depinho RA: **High-resolution genomic profiles of human lung cancer.** *Proc Natl Acad Sci USA* 2005, **102:**9625-9630.
21. 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.
22. 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.
23. Willenbrock H, Fridlyand J: **A comparison study: applying segmentation to array CGH data for downstream analyses.** *Bioinformatics* 2005, **21:**4084-4091.
24. Rueda OM, Diaz-Uriarte R: **Flexible and accurate detection of genomic copy-number changes from aCGH.** *PLoS Comput Biol* 2007, **3(6):**1115-1122.
25. 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+.
26. Rueda OM, Diaz-Uriarte R: **Finding Recurrent Copy Number Alteration Regions: A Review of Methods.** *Current Bioinformatics* 2009 in press.
27. 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 USA* 2004, **101:**9067-9072.
28. 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.
29. Liu J, Ranka S, Kahveci T: **Markers improve clustering of CGH data.** *Bioinformatics* 2007, **23(4):**450-457.
30. 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.
31. 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.
32. Shah S, Lam W, Ng R, Murphy K: **Modeling recurrrent CNA copy number alterations in array CGH data.** *Bioinformatics* 2007, **23(13):**i450-i458.
33. 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+.
34. 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, **104:**20007-20012.
35. 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):**.
36. 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):**.
37. 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:**893-898.
38. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits.** *Nat Rev Genet* 2009, **10(4):**241-251.
39. 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.
40. Sebat J: **Major changes in our DNA lead to major changes in our thinking.** *Nature Genetics* 2007, **39:**S3-S5.
41. Cappé O, Moulines E, Ryden T: *Inference in Hidden Markov Models* New York: Springer; 2005.
42. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition.** *Proceedings of the IEEE* 1990, **77:**257-286.
43. Scott S: **Bayesian methods for hidden Markov models: Recursive computing in the 21st century.** *JASA* 2002, **97:**337-351.
44. Bilmes J: **What HMMs can do.** *IEICE Trans Inf & Syst* 2006, **E89-D(3):**869-891.
45. Pollack J, Sorlie T, Perou C, Rees C, Jeffrey S, Lonning P, Tibshirani R, Botstein D, Borresen-Dale A, Brown P: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci USA* 2002, **99(20):**12963. 12968
46. Nakao K, Mehta K, Fridlyand J, Moore D, Jain A, Lafuente A, Wiencke J, Terdiman J, Waldman F: **High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization.** *Carcinogenesis* 2004, **25(8):**1345-1357.
47. Douglas E, Fiegler H, Rowan A, Halford S, Bicknell D, Bodmer W, Tomlinson I, Carter N: **Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas.** *Cancer Res* 2004, **64(14):**4817-4825.
48. Lengauer C, Kinzler K, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396:**643-649.
49. Wiel MA van de, van Wieringen W: **CGHregions: Dimension reduction for array CGh data with minimal information loss.** *Cancer Informatics* 2007, **2:**55-63.
50. 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.
51. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc Ser B* 1995, **57:**289-300.
52. Kim JH, Dhanasekaran SM, Mehra R, Tomlins SA, Gu W, Yu J, Kumar-Sinha C, Cao X, Dash A, Wang L, Ghosh D, Shedden K, Montie JE, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative analysis of genomic aberrations associated with prostate cancer progression.** *Cancer Res* 2007, **67(17):**8229-8239.
53. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, Wiel MA van de, Green AR, Ellis IO, Porter PL, Tavare S, Brenton JD, Ylstra B, Caldas C: **High-resolution array-CGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biology* 2007, **8:**R215+.
54. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, Beroukhim R, Milner DA, Granter SR, Du J, Lee C, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson ML, Fisher DE, Sellers WR: **Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma.** *Nature* 2005, **436(7047):**117-122.
55. Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M: **Distance-based clustering of CGH data.** *Bioinformatics* 2006, **22(16):**1971-1978.
56. Van Wieringen WNN, Wiel MAA Van De, Ylstra B: **Weighted clustering of called array CGH data.** *Biostatistics* 2008, **9:**484-500.
57. Korn JMM, Kuruvilla FGG, McCarroll SAA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJJ, Darvishi K, Lee C, Nizzari MMM, Gabriel SBB, Purcell S, Daly MJJ, Altshuler D: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.** *Nature genetics* 2008, **40:**1253-1260.
58. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M: **PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.** *Genome Research* 2007, **17(11):**1665-1674.
59. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles MEE: **A robust statistical method for case-control association testing with copy number variation.** *Nature genetics* 2008, **40(10):**1245-1252.
60. Gonzalez J, Subirana I, Escaramis G, Peraza S, Caceres A, Estivill X, Armengol L: **Accounting for uncertainty when assessing association between copy number and disease: a latent class model.** *BMC Bioinformatics* 2009, **10:**172+.
61. Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, Laird NM, Weiss ST, Lange C: **On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test.** *Genetic epidemiology* 2008, **32(3):**273-284.