BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 1 2012, pages 40–47
doi:10.1093/bioinformatics/btr593

Sequence analysis
Advance Access publication October 28, 2011

Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data

Arief Gusnanto[1,*], Henry M. Wood[2], Yudi Pawitan[3], Pamela Rabbitts[2] and Stefano Berri[2]
[1]Department of Statistics, University of Leeds, Leeds LS2 9JT and [2]Leeds Institute of Molecular Medicine, University of Leeds, Leeds LS9 7TF, UK and [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Comparison of read depths from next-generation sequencing between cancer and normal cells makes the estimation of copy number alteration (CNA) possible, even at very low coverage. However, estimating CNA from patients' tumour samples poses considerable challenges due to infiltration with normal cells and aneuploid cancer genomes. Here we provide a method that corrects contamination with normal cells and adjusts for genomes of different sizes so that the actual copy number of each region can be estimated.

Results: The procedure consists of several steps. First, we identify the multi-modality of the distribution of smoothed ratios. Then we use the estimates of the mean (modes) to identify underlying ploidy and the contamination level, and finally we perform the correction. The results indicate that the method works properly to estimate genomic regions with gains and losses in a range of simulated data as well as in two datasets from lung cancer patients. It also proves a powerful tool when analysing publicly available data from two cell lines (HCC1143 and COLO829).

Availability: An `R` package, called `CNAnorm`, is available at http://www.precancer.leeds.ac.uk/cnanorm or from Bioconductor.

Contact: a.gusnanto@leeds.ac.uk

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

Received on May 18, 2011; revised on October 19, 2011; accepted on October 21, 2011

1 INTRODUCTION

Cancer cells often exhibit severe karyotypic alterations: whole chromosome gain or loss and structural rearrangements such as amplifications, deletions and translocations result in widespread aneuploidy (Hartwell and Kastan, 1994). The ability to detect copy number alterations (CNAs) of cancer cells is a crucial step to access the severity of chromosomal rearrangements and to find chromosomal regions where breakpoints are located. Furthermore, comparison of CNAs across tumours from different patients makes it possible to find regions commonly duplicated or lost to highlight the locations of cancer-related genes. Several methodologies are available to detect CNAs. Comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1992), array CGH (aCGH) (Pinkel *et al.*, 1998), single nucleotide polymorphism (SNP) array (Bignell *et al.*, 2004) and, more recently, a new generation of sequencing machines enabled massively parallel sequencing (Roche 454, Illumina GAII, HiSeq, MiSeq, ABI SOLiD, Ion Torrent PGM), making it possible to sequence full genomes at affordable cost.

We previously showed (Wood *et al.*, 2010) how it is possible to multiplex several samples in one Illumina GAII lane making copy number analysis by sequencing affordable and competitive with aCGH or SNP arrays. Between one and eight million aligning reads (compared to approximately half a billion reads for full coverage) are enough to provide genome-wide CNA at 50 kb resolution. As we expect sequencing technologies to become more widespread, affordable and accurate, copy number analysis by low coverage sequencing will become even more convenient and informative. Furthermore, sequencing is possible even with low amounts of DNA extracted from formalin-fixed paraffin-embedded specimens (Wood *et al.*, 2010).

Finally, one advantage of sequencing compared with array technology is that the signal scales linearly to the input DNA and does not show saturation nor background noise typical of hybridization techniques.

One of the first steps to take when analysing these data is normalization. The total intensity of signal from array technology or the total number of reads from sequencing does not reflect the total DNA content of the cells of interest, but it is largely determined by various technical aspects, tuned to achieve the maximum intensity range (array technology) or highest number of reads (sequencing). The normalization is a crucial, non-trivial and often underestimated step which can have enormous repercussions on downstream analysis and conclusions. In this article, we present a computational tool, called CNAnorm, to correct, normalize and scale the data from low coverage sequencing experiments.

1.1 A problem with multiple solutions

From a theoretical point of view, it is often impossible with cancer cells to determine the actual ploidy using array or sequencing technology. As an example, we can consider a fully tetraploid genome and a normal one. Since input DNA is adjusted to a given amount (usually dictated by technical requirements), signal from these two samples will be comparable and it would not be possible to distinguish the tetraploid from the diploid genome. However if, for instance, a chromosome loss (gain) occurred in the tetraploid genome, we could detect a ratio of 3/4 (5/4) between the signal

[*] To whom correspondence should be addressed.

from that chromosome and the rest of the genome. A similar loss (gain) in a diploid genome would result in a ratio of 1/2 (3/2). A further complication arises when we deal with DNA from patient tumour samples as these are infiltrated with normal cells resulting in inevitable contamination with the patients' normal DNA. This means that, if we observe a ratio of 3/4 for a given region, we cannot determine if it is due to a loss from a tetraploid genome of 100% tumour sample (as described above) or a loss from a diploid genome contaminated with 50% normal cells.

This scenario is quite common in tumours from patients because genomes are frequently aneuploid and contamination is largely inevitable. These aspects are often overlooked by algorithms that analyse CNA data. They are usually designed to analyse cell lines with the assumption that the underlying genome is not only pure, but also largely diploid or the average or median ploidy is to be considered 'normal'. There are some algorithms that perform normalization without assuming that the size of the cancer genome is comparable with the normal one (Castle *et al.*, 2010; Chen *et al.*, 2008; Staaf *et al.*, 2007; van Houte *et al.*, 2009) and, most likely, the best results can be achieved when using SNP array (Greenman *et al.*, 2010; Yau *et al.*, 2010) or high coverage sequencing, because information on the frequency of the underlying variant can inform on the absolute copy number.

Similarly, tools that analyse high-throughput sequencing data to obtain CNA are already available. Some are focused on germ line CNV (Xie and Tammi, 2009; Yoon *et al.*, 2009) where aneuploidy and contamination are not an issue, whereas others are designed to detect cancer CNA (Chiang *et al.*, 2009; Ivakhno *et al.*, 2010; Kim *et al.*, 2010). These algorithms, however, do not consider tumour contamination and mostly assume implicitly that the overall size of the two genomes are comparable. To our knowledge, the only exception is FREEC (Boeva *et al.*, 2011) which optionally performs correction for contaminating normal tissue if the ploidy of the most abundant copy number is provided. Unfortunately, this information is not usually available when dealing with patients' tumour samples. The main goal of those tools, however, is the segmentation, i.e. detecting where a change in copy number occurs.

In this study, we do not make any assumption on the overall size of the tumour genome, nor its purity. We designed CNAnorm to take advantage of the linearity of signal to provide information on underlying ploidy and tumour content, with the only assumption that the tumour is largely monoclonal or, if polyclonal, most clones share most of the alterations. When multiple solutions are possible, we select the most conservative among the compatible ones, where the most abundant copy number has the smallest number of deviations from diploidy. Furthermore, it allows the user to correct the most conservative solution provided by the software if other independent analyses (e.g. FISH, flow cytometry, known tumour content) can guide the experimenter's choice. Finally, CNAnorm uses a third-party segmentation tool, DNAcopy (Olshen *et al.*, 2004), to output data not only corrected for contamination and aneuploidy, but also segmented.

2 METHODS

2.1 Samples

Using an Illumina GAII, we produced 1836450 and 1653081 reads from DNA isolated from a fresh frozen lung tumour resection specimen and paired blood, respectively, from patient LS041. Similarly, we produced 3089173 test and 2545305 control reads from patient LS010. Details on sample preparation, DNA extraction and library preparation are described by Wood *et al.* (2010). We also considered publicly available datasets and we used 44762968 test and 34293547 control reads from cell line HCC1143 (Chiang *et al.*, 2009) as well as 18546568 test and 22269150 control reads from cell line COLO-829 (Pleasance *et al.*, 2010).

2.2 Sequence alignments, filtering and GC content

Sequences were aligned using the bwa suite (Li and Durbin, 2009) against assembly hg19 of the human genome. Only sequences that could be uniquely aligned and with mapping quality ≥ 37 were used. For each window, we calculated the average genomic GC content. A Perl script (bam2window.pl) that reads sam/bam files and optionally calculated GC content can be freely downloaded from the CNAnorm website. It produces the table required as input to CNAnorm.

2.3 Read counts

To identify the copy number, we count the number of reads per non-overlapping fixed-width genomic regions (window). Throughout the analysis, unless specified differently, we set the window size so that the median number of reads for each window in the sample with least reads is 30. For samples with more reads (HCC1143 and COLO-829), we set the window size to 50 kb wide.

The window size is a tuning parameter that can be optimized for the data available. However, this is beyond the scope of this article. From our experience in several different samples, selecting window size in which there are 30–180 read counts per window on average strikes a reasonable balance between error variability and bias of CNA. Using a much smaller window size, e.g. on average 1–5 reads per window, will result in many genomic regions with zero read count and make the overall analysis non-informative. At the other extreme, using a much bigger window size will 'smooth out' some pattern of alteration (i.e. increasing bias).

2.4 CNA

To proceed, let x_{jk} be the observed number of reads from a tumour sample or genome in chromosome $j=1,\dots,h$, and window $k=1,\dots,n_j$, where n_j is the number of windows in chromosome j so that the total number of reads in the genome is $n=\sum_{j=1}^{h} n_j$. Let y_{jk} be the observed number of reads in the normal sample. The normalization is performed on a sample-by-sample basis. To identify CNA ρ_{jk}, either as gains and losses in the tumour genome, we intuitively estimate it as an observed ratio between the tumour and normal genomes in each genomic window

$$\hat{\rho}_{jk}=r_{jk}=\frac{x_{jk}}{y_{jk}} \qquad (1)$$

for each chromosome j, and window k. However, this is a bias estimate as we describe next.

A normal genome has two copies of (autosomal) chromosomes, while a tumour genome may have zero, one, two and further multiple duplications. So, the ratios ideally takes any value in $G\equiv\{g_u\}=\{0,0.5,1,1.5,2,\dots\}$ corresponding to tumour copy numbers $P\equiv\{p_u\}=\{0,1,2,3,\dots\}$. In reality, due to (i) error, (ii) different number of reads recorded (sequencing coverage), (iii) different size of tumour and normal genomes and (iv) contamination by normal sample in the tumour sample, the estimates $\hat{\rho}_{jk}$ will not necessarily take a value in G (see Section 2.8 on contamination). Moreover, the observed CNA corresponding to normal genomic regions may not be centered to ratio one. We deal with the first problem by delineating the error variability in a linear model as described briefly in Section 2.6. We deal with the other problems by shifting and scaling the ratio to estimate CNA as described below.

Taking the above problems into considerations, the simplest ratio r_{jk} is a bias estimator, because it has not taken into account the different number of reads in each genome. Each sample may acquire a different level of coverage from the experiment, so that the ratio in Equation (1) may not be properly aligned. A common approach to solve this problem is by normalizing for the total number of reads. However, this assumes implicitly that both genomes involved are of equal size. This is a reasonable assumption when searching for germline CNV, but cannot be applied when one genome could be aneuploid.

We estimate the CNA in several steps below. These steps will be elaborated further in the subsequent sections.

(1) Calculate the ratios $r_{jk} = \frac{x_{jk}}{y_{jk}}$, and correct them for GC content (Section 2.5).

(2) Smooth the signal from r_{jk} to obtain $\tilde{r}_{jk}$. This step reduces noise and highlights the information about genomic alterations. In this article, we use the smooth segmentation approach (Huang *et al.*, 2007) as described in Section 2.6.

(3) Perform normalization on the distribution of $\tilde{r}_{jk}$ so that the most common genomic regions are centered to ratio one. This can be written as

$$\hat{\rho}^{a}_{jk} = \tilde{r}_{jk}\hat{\delta} \tag{2}$$

where δ is a genome-wide alignment coefficient. The coefficient takes into account the different size of tumour and normal genomes. This step is elaborated further in Section 2.7.

(4) The above estimates $\hat{\rho}^{a}_{jk}$ have not taken into account the tumour sample contamination. In this step, we estimate the level of contamination, $\hat{\psi}$, and correct the distribution of $\hat{\rho}^{a}_{jk}$ to obtain the estimates of CNA $\hat{\rho}_{jk}$. This is discussed further in Section 2.8. The observed ratio for each window r_{jk} (not only $\tilde{r}_{jk}$) can be corrected accordingly once the estimates $\hat{\delta}$ and $\hat{\psi}$ are obtained.

(5) At this point, the original data can be segmented using any segmentation tool and results are corrected accordingly. CNAnorm uses DNAcopy (Olshen *et al.*, 2004).

2.5 GC correction

It is known that the ratio r_{jk} can be influenced by the GC content in the window (Boeva *et al.*, 2011; Ivakhno *et al.*, 2010). We acknowledge this dependency by performing a loess correction to rectify the distribution of the ratio and hence removing the dependencies on GC content. We use the loess transformation with a multiplicative correction. Specifically

$$r^{\text{norm}}_{jk} = \frac{\kappa}{A_{jk}} r_{jk}, \tag{3}$$

where, across all j and k, κ is the median of r_{jk}, and $A_{jk} \equiv A(r_{jk})$ is the estimated loess point-wise mean of r_{jk}. In the subsequent steps, we assume that GC correction has been performed in advance so that we can drop the superscript *norm* in r^{norm}_{jk}. Further details on the GC correction and some results are available in the Supplementary Figures S2–S5.

2.6 Smooth segmentation

With thousands of windows to consider, we need to use the spatial information in the data to identify patterns, through smoothing. This step is necessary in the context of low-coverage data, because random error variability can severely affect the normalization and correction of ratio distribution. This, in turn, would result in bias estimates of proportion which would lead to wrong assignment of the diploid group. In our context, this step is critical to guide the normalization and scaling process. However, in case of large excess of reads, typically >500 per window, this step could be skipped.

It is important to note that we do not propose a new segmentation method in CNAnorm. In this study, we implement the smoothing approach as proposed by Huang *et al.* (2007) that employs a linear model where we assume that the second-order difference of the random-effect parameter follow a Cauchy distribution. The estimates of the random effects are the segmented ratio $\tilde{r}_{jk}$. After normalization and scaling, CNAnorm optionally performs a segmentation as implemented by Olshen *et al.* (2004), although other segmentation methods can generally be used.

2.7 Genome-wide normalization

Genome-wide normalization is a step to correct the location of the distribution of the copy number ratio by estimating δ from the segmented ratio data $\tilde{r}_{jk}$. Owing to systematic gains and losses, the ratio r_{jk} shows a multi-modal distribution. The segmented ratio $\tilde{r}_{jk}$ exhibits the multi-modality of the distribution more clearly after removing unwanted random errors in the smoothing step. The modes of the distribution of $\tilde{r}_{jk}$ indicate the (biased) position of CNA in G, corresponding to different copy numbers in P. At this stage, the modes of the distribution is not centered to the expected CNA in G, and the estimation of δ requires us to characterize the distribution of $\tilde{r}_{jk}$.

Reflecting on the multi-modality, we fit a mixture normal distribution to the distribution of smoothed ratio of tumour over normal samples $\tilde{r}_{jk}$

$$p(\tilde{r}_{jk}) = \sum_{m=1}^{M} \pi_m N(\tilde{r}_{jk}; \mu_m, \sigma^2_m), \tag{4}$$

where π_m are the mixing proportions, $\sum_{m=1}^{M} \pi_m = 1, 0 \le \pi_m \le 1$, for $m = 1,\dots,M$, μ_m and σ^2_m are the mean and variance of normal distribution. In this formulation, each of μ_m's corresponds to a value in G that reflects the ratio of tumour to normal copy numbers, and a tumour copy number in P. At this stage, the estimates of μ_m are still biased estimates for CNA in G. We will use the estimates of μ_m to guide us through all the steps in our normalization and scaling methods as described next.

Our experience suggested that the normal distribution as a component is adequate to model the distribution of $\tilde{r}_{jk}$, as the distribution of $\tilde{r}_{jk}$ does not show some heavy tails. Some genomic regions may exhibit extreme values of r_{jk} (but *not* $\tilde{r}_{jk}$) because of low counts in x_{jk} and y_{jk}. This is rare and mainly due to problems of mapping reads to some regions of the genome. We find that the smooth segmentation approach is relatively robust, where extreme values in the r_{jk} do not affect our fitting of the mixture distribution on the smoothed ratio $\tilde{r}_{jk}$.

We estimate the mixture components in model (4) using a standard expectation–maximization (EM) algorithm (e.g. McLachlan and Krishnan, 1997). In the algorithm, we put some constraints to impose identifiability (see also the Supplementary Material). The number of components in the model, M, is chosen using Akaike's information criterion (AIC) across different plausible values.

Once the estimates $\hat{\boldsymbol{\mu}} \equiv \hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_M$ are obtained, it is important for the subsequent steps to describe the relationship between $\hat{\boldsymbol{\mu}}$ and their corresponding tumour copy number (ploidy) in P. For example, we generally have a set of copy numbers $P^* \equiv \{p^*_m\} = \{0,1,2,3,\dots,M-1\}$ from M components identified in the mixture model (4). However, we allow a 'leap' in the ploidy so that the M copy number in P^* may contain non-unit increment. To deal with this, we acknowledge that the ideal CNA values in G increase linearly with copy number or ploidy P. So, for our purpose of aligning $\tilde{r}_{jk}$, we model $y_{\boldsymbol{\mu}} \equiv \hat{\boldsymbol{\mu}}$ in a simple linear regression as

$$y_{\boldsymbol{\mu}} = \alpha^* + \beta^* P^* + \varepsilon \tag{5}$$

where α^* is a fixed intercept, β^* is the regression slope for P^*, an M-vector of ploidy or copy number and ε is the error term. If the model fit (in terms of R^2) falls below a certain threshold, we 'juggle' the different copy numbers in P^* to include values greater than $M-1$, and select a collection of M values that give the best R^2. The threshold is set at 0.95 by default, although it is adjustable by user intervention.

Our next task is to identify a component in the mixture model (4), denoted v, $v \in \{1,\dots,M\}$, which corresponds to the normal ploidy (ratio one in G). This is determined as the most common component, i.e.

$$v = \arg\max_{m} \hat{\pi}_m.$$

Since multiple solutions are often possible, we choose the most conservative solution. This means that if $v \le 3$ we assign the most common component to have two copy numbers (diploid). Otherwise, we assign v-th component to have $v-1$ copy number (the first component corresponds to the total loss, zero copy number).

The genome-wide normalization coefficient δ is then estimated as

$$\hat{\delta} = \frac{1}{\hat{\mu}_v}. \tag{6}$$

In practice, we replace $\hat{\mu}_v$ with $E(\hat{\mu}_v)$, the fitted value of $\hat{\mu}_v$ from the above model (5). The estimation of δ indicates that the process of genome-wide normalization involves identifying the mixture component which corresponds to the normal ratio and shift the whole distribution of $\tilde{r}_{jk}$ multiplicatively so that the normal ratio is centered to one.

Once the estimate $\hat{\delta}$ is obtained, $\hat{\rho}^a_{jk} = \tilde{r}_{jk}\hat{\delta}$ is the estimate of 'crude' CNA where contamination is still present. When we have different pair of samples from different individuals, we are dealing with different degrees of contamination between the pairs. So, the 'crude' CNA between individuals are not comparable for the purpose of, e.g. statistical testing of the genomic regions. To make the estimate of CNA that are comparable between samples, we need to characterize the contamination, and make a proper correction to the distribution of $\hat{\rho}^a_{jk}$ as described in the next section.

2.8 Contamination correction

In clinical situations, pure tumour cells are difficult to obtain when the material comes from tissues of patients' tumour. Contamination with normal cells are inevitable. If there were no contamination, the smoothed ratio $\tilde{r}_{jk}$ is expected to take any one values in $G \equiv \{g_u\} = \{0, 0.5, 1, 1.5, 2, \ldots\}$. When contamination by a normal genome is present, the smoothed ratio $\tilde{r}_{jk}$ will be shrunk towards ratio one (Supplementary Fig. S1).

To estimate the contamination, we first assume that the contamination shrinks linearly the CNA towards ratio one, e.g. $\rho_{jk} = 2$ will be shrunk to $1 < \hat{\rho}_{jk} < 2$ and $\rho_{jk} = 0.5$ will be shrunk to $0.5 < \hat{\rho}_{jk} < 1$. Given the previous step to centre the normal copy number to ratio one, the estimate of 'crude' CNA $\hat{\rho}^a_{jk}$ can be assumed to have come up from a shrinkage on the non-contaminated $\hat{\rho}_{jk}$ around ratio one

$$\hat{\rho}^a_{jk} = 1 + (\hat{\rho}_{jk} - 1) \times (1 - \hat{\psi}) \tag{7}$$

where $\hat{\psi}$ is the estimate of contamination proportion $(0 \leq \hat{\psi} < 1)$.

We estimate ψ by investigating how the estimates in $\hat{\boldsymbol{\mu}}$ have been shrunk towards $\hat{\mu}_v$ that corresponds to the copy number two. We first normalize the estimates $\hat{\boldsymbol{\mu}} \equiv (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M)$ into $\hat{\boldsymbol{\mu}}^c \equiv \{\hat{\mu}^c_m\} = \{\hat{\mu}_m \hat{\delta}\}$, for $m = 1, \ldots, M$. The estimate of $\hat{\psi}$ is given by

$$\hat{\psi} = \frac{1}{M-1} \sum_m \left\{ 1 - \frac{|\hat{\mu}^c_m - \hat{\mu}^c_v|}{\hat{\mu}^c_v} \frac{1}{0.5 \times |p^*_m - p^*_v|} \right\}, \tag{8}$$

where the summation is performed on the set $m = 1, \ldots, v-1, v+1, \ldots, M$, and p^*_m is the copy numbers in P^*, excluding p^*_v.

Now, following the relationship in Equation (7), the estimates of CNA can be expressed as

$$\hat{\rho}_{jk} = 1 + (\hat{\rho}^a_{jk} - 1) \times \frac{1}{(1 - \hat{\psi})}. \tag{9}$$

The estimates of CNA $\hat{\rho}_{jk}$ in the above equation have taken into account the different read depths, genome sizes and contamination. This makes the estimates comparable between different pairs of samples, which is important when statistical tests are performed to infer the pattern of genome-wide CNA.

2.9 Simulation study

We performed a simulation study to test our working model where some complications were included in the simulated data. We produced a female highly aneuploid genome (largely tetraploid) with a series of large and small deletions and duplications using the human reference genome assembly hg19. From the aneuploid (test) genome t and the normal genome c of size N_t and N_c, respectively, we simulated 100 datasets each with 3000000 reads. For each dataset i, we simulated S_{ti} and S_{ci} reads so that

$$\begin{cases} \frac{S_{ti}}{N_t} : \frac{S_{ci}}{N_c} = (1 - \psi_i) : \psi_i \\ \\ S_{ti} + S_{ci} = 3000000 \end{cases}$$

with $N_t = 11515563746$ and $N_c = 6175502642$ nt (note: the normal genome is diploid) and $0.15 < \psi_i < 0.80$ is the contamination level. We simulated a mixture of tumour and normal cells in ratio $(1 - \psi_i) : \psi_i$ and then produced three million reads coming from the genomes of such a mixture of cells. Clearly, since the tumour genome is larger than the normal one, a tumour content of $(1 - \psi_i)$ will produce more than $(1 - \psi_i)$ of the reads. To simulate the reads, we used wgsim 2.6 (default parameters), a tool part of the bwa suite (Li and Durbin, 2009). Similarly, we produced three million reads for the control genome.

We performed analysis on 100 simulated datasets using CNAnorm and FREEC v3.93 (Boeva *et al.*, 2011) using a window size of 50 kb, and no GC correction (being simulated data). We use the default parameters for both programs, with an exception in FREEC where we included the option to correct for contamination. We then compared the results from both methods. FREEC can adjust for contamination with normal cells as long as the most frequent ploidy is provided. To compare the two tools in equivalent situations, we set CNAnorm to normalize the simulated data so that the most abundant copy number was set to four. However, we also report the estimates CNAnorm would produce when the most common ploidy is estimated. Since FREEC forces the copy number to be an integer, we rounded the segmented value of CNAnorm output in this comparison. To assess the two programs, for each window j, we calculate a score

$$\Delta = \frac{1}{n} \sum_j |\hat{\rho}_j - \rho_j| \tag{10}$$

where n is the number of windows in the simulated genome, $\hat{\rho}_j$ is the estimated copy number for window j and ρ_j is the expected copy number for the same window. In Equation (10), a high score Δ suggests that the estimates of CNA $\hat{\rho}_j$ is far from the true values, and vice versa.

3 RESULTS

3.1 Simulation study

First, we present the results of one of the 100 simulated datasets (a single realization) with 30% contamination ($\psi = 0.30$). The results on all simulated datasets are presented subsequently below. Figure 1 shows the histogram of the ratio and smoothed ratio (see Supplementary Fig. S6 for more details on ratios across

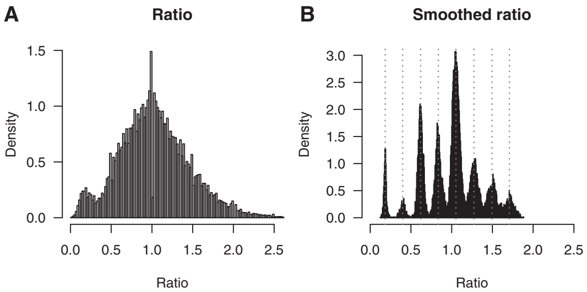

Fig. 1. Histogram of ratio (**A**) and smoothed ratio (**B**) across the genome in a simulated data (single realization) with $\psi = 0.30$ (30% contamination). In (B), the dotted vertical lines mark the estimates of means when we fit a mixture distribution on the smoothed ratios. More details on the ratios across genome are presented in the Supplementary Figure S6.

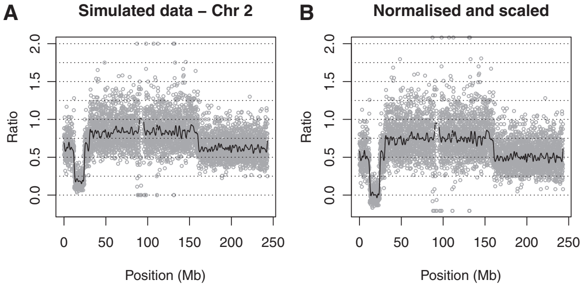

Fig. 2. The ratio in Chromosome 2 of simulated data with $\psi = 0.3$ (30% contamination) before (**A**) and after (**B**) normalization and scaling. The solid line is the estimate of CNA $\hat{\rho}_{jk}$ as a smoothed signal. The horizontal dashed lines are the different expected ratios if there is no contamination in the samples. Note that the 'normal' copy number is four, so that the expected ratios are in $G^* = \{0/4, 1/4, 2/4, \ldots\}$. The patterns of genomic CNA are shown in Supplementary Figures S6 and S7.

the genome). From the figure, we now can see the underlying distribution of CNAs in the simulated data. The vertical dotted lines in the Figure 1B are the estimates of mean when we fit the mixture model (4) onto the smoothed ratios. These estimates are $\{\hat{\mu}_m\} = \{0.18, 0.40, 0.62, 0.84, 1.05, 1.27, 1.50, 1.71\}$, corresponding to copy numbers $P^* = \{0, 1, \ldots, 7\}$.

It is estimated by AIC that we have eight components in the mixture model, and this can be clearly seen from the figure. The most common mixture component in the simulated data is the fifth component which corresponds to four copy number in the tumour sample. The estimates of the proportions $\hat{\pi}_m$ are (in percentages) $\{4.4, 2.4, 15.9, 15.1, 34.2, 13.6, 9.0, 5.3\}$. From this result, we align the fifth component to the ratio one with $\hat{\delta} = 0.943$. The contamination is estimated at $\hat{\psi} = 29.54\%$, which is close to the true value of 30%.

To see the results of the normalization in a chromosome, Figure 2 presents the estimates of CNA in Chromosome 2 before and after the normalization and scaling for contamination. Since the most common component is tetraploid, we find that the expected ratio is in the increment of 0.25 as shown in the figure as horizontal dashed lines. The figure indicates the estimates of CNA are now properly centered.

Next, we compare the results of the 100 simulations obtained using CNAnorm with those obtained using FREEC (Boeva *et al.*, 2011) as presented in Figure 3. The figure indicates that the performance is getting worse as the contamination level increases using all methods. When the information on ploidy is provided, CNAnorm (solid black line) performs better than FREEC (dash-dotted line) 95% of the time. When ploidy is not provided, CNAnorm still performs better than FREEC 72% of the time. Understandably, all three methods perform better at low rather than high contamination levels. Remarkably, CNAnorm has very good and consistent performance (Δ is very close to zero) for contamination up to 40% even when the information on ploidy is not provided.

A closer inspection on the above results indicates that, while the performance of FREEC steadily deteriorates with increased contamination, CNAnorm performs consistently well unless it fails to correctly detect some mixture components. When this happens, there are mainly three consequences. First, the mixture components are correctly detected but the copy number is 'shifted'. The estimates of CNA are off by a discrete number of ploidies. This is easily rectified if the most abundant ploidy is provided. Experiments with average distance of 1, but low Δ when ploidy is available (Fig. 3) fall in this category. Secondly, if a limited number of mixture components are missed or wrongly estimated, but the most abundant one is correctly identified, the ratio is correctly shifted but over- or under-scaled. Thirdly, several mixture components could be missed or wrongly assigned. In this case, the normalization is severely affected. Providing the ploidy of the most abundant mixture may or may not improve the normalization.

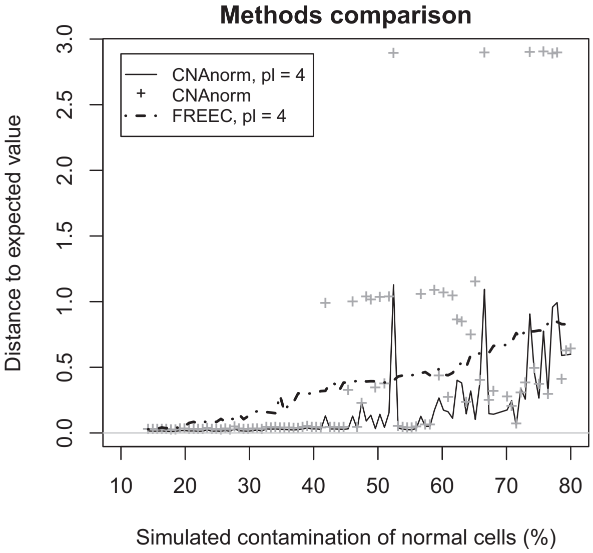

Fig. 3. The scores Δ from 100 simulated data as a function of normal genome contamination level (percent). A low value of Δ indicates that the estimate of CNA is close to the true value. The score for FREEC is shown in the dash-dotted line and for CNAnorm is shown in solid line. The plus (+) grey points are the scores for CNAnorm when the information on the most frequent ploidy is not provided.

3.2 Normalization of patients' sample data

We present the results of analysis on LS041 dataset here, while the results on LS010 dataset are presented mainly in the Supplementary Material. Figure 4 shows the effect of smoothing on the distribution of copy number ratio (see Supplementary Fig. S8 for more details on the ratios r_{jk} and smoothed ratios $\tilde{r}_{jk}$).

From Figure 4A, we are not able to see clearly the multi-modality in the distribution of the ratio r_{jk} in the genome. The smoothing process accentuates the multi-modality as shown Figure 4B . The fitting of the mixture model (4) is performed on the distribution of $\tilde{r}_{jk}$, which is shown in the figure. Based on AIC, for LS041, we came to a result that the optimal number of mixture components is $M = 7$. Given $M = 7$, the estimates of the means are $\hat{\mu}_m$ = (0.47, 0.65, 0.92, 1.15, 1.35, 1.53, 2.43), and they are marked as vertical dotted lines in Figure 4B.

The estimated proportion (in percentages) of the mixture components are $\hat{\pi}_m = (1.0, 26.7, 29.5, 28.7, 10.3, 1.0, 2.8)$, indicating that the third mixture component is the most common one ($\nu = 3$). This suggests that the tumour genome is largely diploid. Next, we

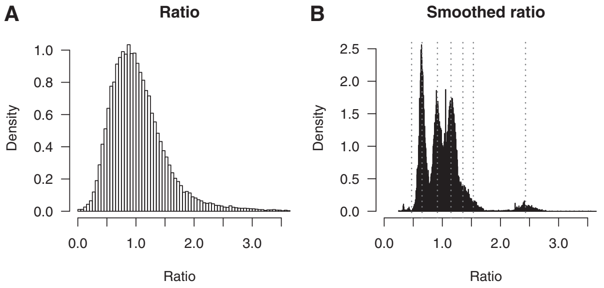

Fig. 4. Histogram of ratio (**A**) and smoothed ratio (**B**) across the genome. In (B), the dotted vertical lines mark the estimates of means when we fit a mixture distribution on the smoothed ratios. More details are shown in Supplementary Figure S8.

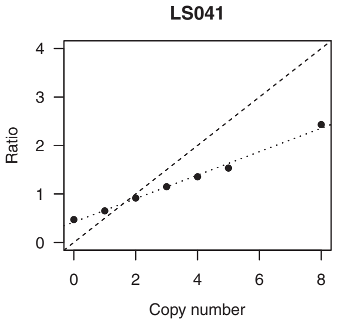

Fig. 5. Relationship between estimates of mean $\hat{\mu}_m$ from fitting mixture models and copy number. The dotted line is the fitted linear regression line, and the dashed line is the expected line with slope 0.5 if there is no contamination in the tumour sample.

plot the estimates $\hat{\mu}_i$ against the copy number as shown in Figure 5. The figure shows a linear relationship between the estimates $\hat{\mu}_{im}$ with the copy number. For the last mixture component, we allow a 'leap' in the copy number so that we have a better fit of the linear regression (dotted line). The fitted line has a lower slope estimate than the expected line due to contamination.

In this step, we multiplicatively align the whole distribution of r_{jk} and $\tilde{r}_{jk}$ so that the mixture component which corresponds to the normal copy number is centered to ratio one. We obtained the estimates $\hat{\mu}_3 = 0.92$ and $\hat{\delta} = 0.904$ as the multiplicative factor.

3.3 Contamination correction on patients' sample data

The estimate of slope in Figure 5 is lower than the expected (0.5) due to the contamination of tumour sample. We assume in this study that the effect is proportional to the distance of $\tilde{r}_{jk}$ to the ratio one (Supplementary Material). After scaling for the diploid component to have ratio one, the scaled estimates $\hat{\mu}^c_m$ are $\{0.52, 0.72, 1.01, 1.27, 1.50, 1.70, 2.69\}$. The value for the diploid component is not exactly one, due to the use of fitted value in Equation (6). This gives the estimate of contamination $\hat{\psi} = 0.491$, or 49.1%. We correct the whole distribution of $\tilde{r}_{jk}$ (and applicable to r_{jk}) using a multiplication, centered on ratio one, so that the mean estimates in Figure 5 are aligned to be close to the expected distribution as presented in Figure 6A.

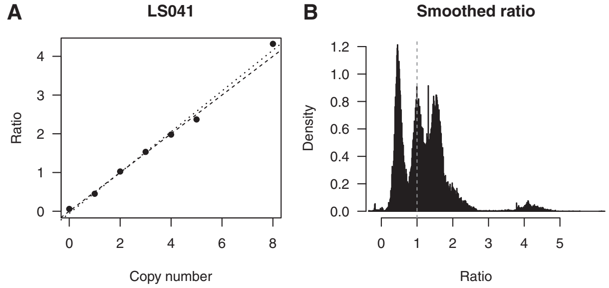

Fig. 6. (**A**) Relationship between estimates of mean $\hat{\mu}^c_m$ (after correction for contamination) and copy number. The dotted line is the fitted linear regression line, and the dashed line is the expected line with slope 0.5 when there is no contamination in the tumour sample. (**B**) Histogram of the segmented ratio across the genome after correction for contamination (estimated at 49.1%). The vertical dashed line marks the mean for normal copy number after alignment.

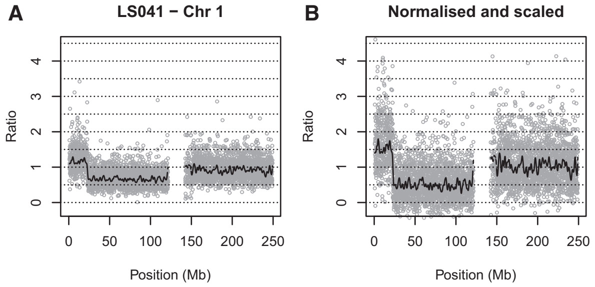

Fig. 7. Chromosome 1, before (**A**) and after (**B**) the normalization. The solid line is the estimate of CNA.

In Figure 6B, we can see that the distribution of $\tilde{r}_{jk}$ is expanded from reference point one compared to Figure 4. The result of this contamination correction across the genome is presented in Supplementary Figures S8 and S9. The figures indicate that the estimated CNA are now close to the expected copy number ratios in G. To see the estimates in more detail in a chromosome, Figure 7 shows the result of our proposed method on chromosome 1 for LS041 data. After the normalization step and correction for contamination that we have performed, the estimates of CNA are now correctly aligned to their expected values in G.

3.4 Analysis of cell line data: HCC1143 and COLO829

Although CNAnorm was developed to deal with low coverage data from clinical samples, we tested the package by analysing two publicly available datasets of higher coverage sequencing of cells lines: HCC1143, obtained from a human breast cancer genome (Chiang *et al.*, 2009) and COLO829, obtained from a human malignant melanoma (Pleasance *et al.*, 2010). The results of analysis on the COLO829 data are presented in the Supplementary Material.

For cell line HCC1143, we performed the default analysis and CNAnorm found that the third component was the most abundant and conservatively assigned it to $P^* = 2$. In doing so, it also estimated a tumour content of 47.5%. Although, from a technical point of view, this is a plausible solution, our knowledge about the starting material informs us that a tumour content of 47.5% is suspiciously low for a cell line. We then shifted the estimates of P^* and obtained P^*+1 and P^*+2 that would then predict a more likely 87% tumour content or an unrealistic 153.5%, respectively. We then accepted P^*+1 and normalized the data. The results for whole genome and a subset of chromosomes are shown in Supplementary Figures S15 and S16. Here, we used external clues to perform the most sensible normalization.

HCC1143 is a very well-characterized cell line and SKY karyotypes are available from http://www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC1143.html. We were able to compare the estimation of CNAnorm against independently observed evidence and confirm that the normalization is appropriate. In particular, CNAnorm estimates that Chromosome 5 is largely tetraploid, Chromosome 12 largely triploid and the acrocentric Chromosome 13 largely diploid (Supplementary Fig. S16). These estimations are easily confirmed by the karyotype. As HCC1143 is extensively rearranged, other abnormalities are harder to validate, but given that CNAnorm performs a genome-wide normalization, there is no reason why the copy number estimation on other chromosomes should not be correct, in particular within the 2–4 ploidy. Furthermore, CNAnorm estimates the median ploidy at 3.2 copies, that is compatible with the description of 'near-tetraploid' or 'hypo-tetraploid' as reported on the above Cancer Genomics Program web page.

4 DISCUSSION

We have investigated a method to estimate CNA from patient tumour samples using next-generation sequencing. We showed how the method performs with a range of simulated data, on low coverage data from patients' samples and on higher coverage data from cell lines. Compared with several segmentation tools available to analyse high-throughput sequencing data, CNAnorm focuses on correcting the data for contamination, different read depth and different genomic size. We acknowledge that the problem could lead to several equally valid solutions, but provide an easy way for the user to correct the estimation from CNAnorm when independent clues (such as tumour content, or strong and independent evidence about ploidy of certain regions) are available. We believe that the normalization step is often underestimated or, due to its intrinsic difficulty and plurality of solution, left to a simplistic approach that assumes that the overall size of a cancer genome is comparable with that of a normal cell.

With the data from the two cell lines (HCC1143 and COLO829), we have shown how information on tumour content, ploidy of some chromosomes and overall ploidy could guide the experimenter to perform a more meaningful normalization. At the same time, CNAnorm could provide further insight on the cancer material analysed. When no external information is available, e.g. in the case of patients' tumours LS041 and LS010, CNAnorm performs the most conservative normalization. In this regard, if regions of homozygous deletions could be identified and confirmed, they would be a valuable guide during the normalization process.

Despite being a powerful tool, CNAnorm is not a 'silver bullet' for CNA analysis. In particular, we would like to point out how polyclonal tumours could produce misleading results. If several tumour clones, each with its gains and loss, constitute the tumour sample, it will not be possible to detect the underlying ploidy and the mixture model approach would over-fit distributions within a single ploidy range. CNAnorm is meant to be robust and clonal variability in a few chromosomal regions would not be problematic. However, in these cases, CNAnorm would tend to underestimate tumour content. This is, in our opinion, what happens with cell lines HCC1143 and COLO829. Although the cell line should be 100% tumour, CNAnorm estimates only 87 and 90%, respectively. We think this is due to some variability within cells. This variability can be observed, for instance, in HCC1143 where the copy number of the long arm of Chromosome 2, or Chromosome 6 are, unlike most of the rest of the genome, not close to any integer copy number. Since we are aware of possible polyclonal variability, we chose an approach and a segmentation tool, DNAcopy, that does not force every region of the genome to fit into an integer copy number. Clues about polyclonal variation are, *per se*, potentially informative.

5 CONCLUSION

Next-generation sequencing data from clinical samples obtained directly from patients presents a serious challenge. Analysis of CNAs in tumour samples is not straightforward. This is because the observed raw copy number ratios do not necessarily take the expected values due to random error, different sequencing coverage and contamination with normal cells. We deal with the random error using smoothing methods. The other challenges are dealt with by acknowledging the multi-modality in the distribution of the segmented copy number ratios. This allows us to model the locations of the distribution of ratios corresponding to the different copy numbers and make the necessary correction. The simulation study shows that the method works properly to estimate genomic regions with gains and losses and that it works well in a range of real datasets from patients' samples and cell lines.

ACKNOWLEDGEMENTS

A.G. would like to thank Terry Speed and Yuval Benjamini for a helpful discussion. We thank the surgical team at St James's University Hospital (Leeds, UK).

Funding: Yorkshire Cancer Research (L341PA).

Conflict of Interest:none declared.

REFERENCES

Bignell,G. *et al.* (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.*, **14**, 287–295.

Boeva,V. *et al.* (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.

Castle,J. *et al.* (2010) DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics*, **11**, 244.

Chen,H-I. *et al.* (2008) A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics*, **24**, 1749–1756.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Greenman,C.D. *et al.* (2010) Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Hartwell,L.H. and Kastan,M.B. (1994) Cell cycle control and cancer. *Science*, **266**, 1821–1828

Huang,J. *et al.* (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, **23**, 2463–2469

Ivakhno,S. *et al.* (2010) CNAseg - a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, **26**, 3051–3058.

Kallioniemi,A. *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Kim,T.M. *et al.* (2010) rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics*, **11**, 432.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

McLachlan,G. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572 .

Pawitan,Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.

Pinkel,D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pleasance,E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Staaf,J. *et al.* (2007) Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics*, **8**, 382.

van Houte,B. *et al.* (2009) CGHnormaliter: an iterative strategy to enhance normalization of array CGH data with imbalanced aberrations. *BMC Genomic*, **10**, 401.

Wood,H. *et al.* (2010) Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.*, **38**, e151.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing, *BMC Bioinformatics*, **10**, 80.

Yau,C. *et al.* (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.*, **11**, R92.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.