BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1823–1829
doi:10.1093/bioinformatics/btu123

Sequence analysis

Advance Access publication March 5, 2014

# Estimating optimal window size for analysis of low-coverage next-generation sequence data

Arief Gusnanto[1,*], Charles C. Taylor[1], Ibrahim Nafisah[1,2], Henry M. Wood[3], Pamela Rabbitts[3] and Stefano Berri[3,4]

[1]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom, [2]Department of Statistics, Faculty of Science, King Saud University, Riyadh, Saudi Arabia, [3]Leeds Institute Cancer and Pathology, University of Leeds, Leeds LS9 7TF, UK and [4]Illumina UK Ltd., Chesterford Research Park, Saffron Walden, CB10 1XL, UK

Associate Editor: Inanc Birol

## ABSTRACT

**Motivation:** Current high-throughput sequencing has greatly transformed genome sequence analysis. In the context of very low-coverage sequencing (≤0.1×), performing 'binning' or 'windowing' on mapped short sequences ('reads') is critical to extract genomic information of interest for further evaluation, such as copy-number alteration analysis. If the window size is too small, many windows will exhibit zero counts and almost no pattern can be observed. In contrast, if the window size is too wide, the patterns or genomic features will be 'smoothed out'. Our objective is to identify an optimal window size in between the two extremes.

**Results:** We assume the reads density to be a step function. Given this model, we propose a data-based estimation of optimal window size based on Akaike's information criterion (AIC) and cross-validation (CV) log-likelihood. By plotting the AIC and CV log-likelihood curve as a function of window size, we are able to estimate the optimal window size that minimizes AIC or maximizes CV log-likelihood. The proposed methods are of general purpose and we illustrate their application using low-coverage next-generation sequence datasets from real tumour samples and simulated datasets.

**Availability and implementation:** An *R* package to estimate optimal window size is available at http://www1.maths.leeds.ac.uk/~arief/R/win/.

**Contact:** a.gusnanto@leeds.ac.uk

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

Received on February 4, 2013; revised on December 31, 2013; accepted on January 27, 2014

## 1 INTRODUCTION

The recent advent of high-throughput sequencing (or NGS, 'next-generation sequencing') has revolutionized the quantity and quality of data produced. The ability to sequence a large number of DNA or cDNA fragments at reasonable cost is proving to be flexible and powerful. One of the applications is to use NGS to assess copy-number alterations (CNA) in tumour cells. Although information about copy number is often obtained by analysing high-coverage data (>20×) (Pleasance *et al.*, 2010), we have previously shown (Gusnanto *et al.*, 2012; Wood *et al.*, 2010) that it can also be reliably obtained by more affordable low-coverage data (<0.05×) from small amounts of fragmented DNA obtained from formalin-fixed paraffin-embedded samples. We expect low-coverage data will still be a valuable choice because, regardless of falling sequencing costs, data storage, analysis time and infrastructure costs associated with large datasets will not decrease as quickly. Wider use of NGS means it will be used more and more in diagnostic settings, where low costs and rapid analysis time are critical. Finally, the recently launched sequencing machines (Illumina MiSeq, Life Technologies PGM) have allowed costs to be within reach of individual laboratory budgets, although this means that they produce fewer reads. For these reasons, NGS low-coverage data could become common and partially replace hybridization-based technologies such as aCGH and SNP arrays.

Owing to the sparse nature of the data, however, it is important to extract the maximum information. In particular, the size of the genomic window used for binning reads is a critical 'tuning' parameter: if it is too wide, the analysis will miss some genomic regions that exhibit important features, and if it is too narrow, the noise level will be dominant and many windows will contain zero reads.

We expect that in high-coverage cases (>20×), the choice of window size is less crucial because by the time reads windows are so small, they are of the same order of magnitude as the reads themselves, and information about chromosomal junctions can be precisely obtained by reads spanning across chromosomal regions that are disjointed in the reference genome.

The number of reads per window, should, in theory, follow a Poisson distribution. It was, however, evident from the early experiments that there is overdispersion (Bentley *et al.*, 2008). This is due to a number of reasons including GC content bias (Benjamini and Speed, 2012; Bentley *et al.*, 2008; Gusnanto *et al.*, 2012), mappability (Lee and Schatz, 2012), underlying changes in copy number (both somatic and germ line) and possibly other biological and technical factors still to be described.

In this study, our focus is on estimating an optimal window size for analysis of NGS data to best track the depth of coverage signal in a very low-coverage setting, motivated by our study of CNA. The principle that we use to answer this problem is by considering that the 'binning' or 'windowing' is basically the

*To whom correspondence should be addressed.

same process as a histogram construction. In the construction, we calculate either the count of the data falling in each window or, equivalently, the density in each window. By assuming a statistical model for the underlying density in each window, we are able to quantify the statistical distance of the model to the 'truth' across different window sizes. From here, we can identify the optimal window size as the one that minimizes the distance. Once the optimal window size is estimated, further analysis, e.g. GC correction, comparison with matched normal, CNA analysis, can be performed based on the optimal window size.

Before we propose our method, we consider in the next section some studies that have discussed the notion of an optimal window size in the context of next-generation sequence data.

### 1.1 Previous consideration of window size in the context of NGS

Some studies have discussed or considered the problem of estimating the optimal window size in the context of next-generation sequence data. They all agree in principle that an optimal window size is important, if not critical, in inferring any pattern from the data [Xi *et al*. (2011)]. To illustrate this point, we show in Figure 1 the pattern of read counts in our tumour sample (LS041) at different window sizes.

The figure shows the distribution of read counts in chromosome 8 of LS041 tumour sample using 10 kb, 50 kb, 200 kb and 2 Mb window (~6, 31, 124 and 1227 reads per window, respectively). The figure indicates that the variability of the read counts is higher in the smaller window size and lower in the wider window size. Moreover, the pattern of read counts at window size 2 Mb indicates that a potential change around the position of 40 Mb is smoothed away, whereas it is visible in window sizes of 50 and 200 kb. If we decrease the window size further to 10 kb, the change is obscured under high variability of the read counts.

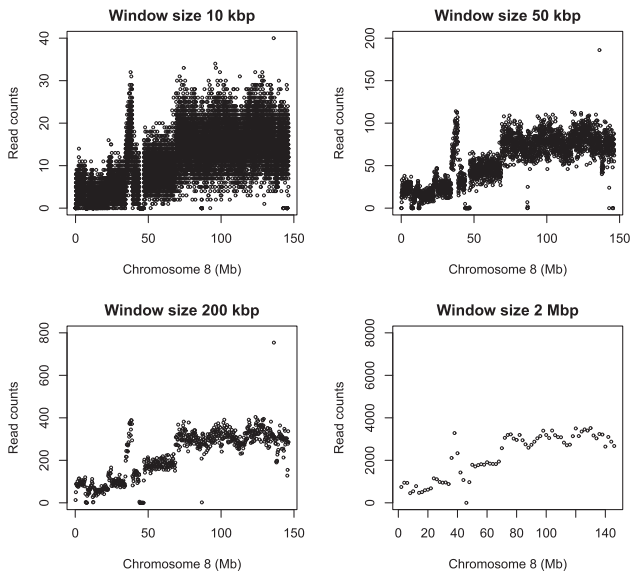


**Fig. 1.** The distribution of read counts in the chromosome 8 of LS041 cancer sample, at window sizes of 10 kb, 50 kb, 200 kb and 2 Mb

This example clearly shows why estimating an optimal window size is important.

Previously, in the context of CNA analysis, and given a fixed coverage, Xie and Tammi (2009) described the dependencies of optimal window size on the *P*-values and magnitude of copy-number ratio to be detected ($r'$) from ratio one. In practice, the optimal window sizes are estimated from some *preset P*-values and $r'$. In their proposed method, Xie and Tammi (2009) assumed normality in the distribution of observed copy-number ratio (Supplementary Material).

Castle *et al*. (2010) considered the 'optimal' window size based on detecting copy-number ratio 1.5 with a *P*-value of 0.001, assuming a Poisson error model. Using this approach, with ~3.3 million reads in their data, they came to a window size of 164 kb or ~110 reads per window. Yoon *et al*. (2009) considered an 'optimal' window size as the (minimum) window size that makes the distribution of read counts to approximately follow a normal distribution. They noted that, in the context of high coverage (30×), the optimal window size is found to be ~100 bp (0.1 kb). In other studies, the optimal window size is even chosen arbitrarily. For example, Cheng *et al*. (2011), in the context of RNA-seq analysis, defined subjectively a window size of 100 bp in their analysis.

In the above approaches, we have an immediate problem. When we *preset* the *P*-value of detection in advance, it relies heavily on the asymptotic distribution of the (observed) copy-number ratios being the normal distribution. As Yoon *et al*. (2009) have illustrated, this approximation is justifiable when the window size is 'large enough', but it falls short when the window size is small. Another related problem is that the estimation of optimal window size is done on the ratio data, which, in the case of data from cancer patients, implicitly assume that the ratio has been properly normalized to deal with the sample contamination (Gusnanto *et al*., 2012).

To deal with the above problem, we propose the following method. We acknowledge that the process of counting reads that fall in a window is basically a process of histogram construction. In the construction, the height of the histogram in a window is calculated as either the count or the count that has been normalized (divided) by the total number of reads in the genome. We refer to the latter as the 'density' because the histogram densities in the data sum up to one. In this study, the use of density is preferable because it takes into account different number of reads in different datasets. In our proposed method (Section 2.3), a window size is considered optimal if, given a statistical model, the observed density of the read counts is 'close enough' to the true underlying density of reads.

The use of a histogram to accentuate genomic features is not new [Johnson *et al*. (2007) and Robertson *et al*. (2007)]. On a related method, Boyle *et al*. (2008) discussed the estimation of biologically relevant features in the genome by detecting the presence of enrichment of mapped sequence reads using the kernel density. In the context of density estimation, both histogram and kernel density are trying to estimate the true underlying density of reads in the genome. The main difference being that the kernel density method estimates the underlying density as a smooth function, whereas the histogram method estimates it as a step function. The estimates of density in both methods also rely on the window size. Unfortunately, in their proposed method, Boyle

*et al*. (2008) did not describe how to estimate the optimal window size, although the kernel density method relies on the window size for its estimates to be meaningfully interpretable. They only set the default window size of 600 bp (0.6 kb) or any other window size that makes their estimates 'smooth enough' or until the kernel density estimates approximately follow a normal distribution.

To estimate the optimal window size in the context of histogram construction, previous studies in the statistical literature have published some methods (Section 2.2). Their objectives were the same: to estimate the window size that would make the distance between the observed and the true underlying density of the reads minimal. However, those approaches fall short because of the statistical nature of the next-generation sequence data that exhibit relatively non-regular density (as compared with, for example, probability density).

To deal with this problem, we use data-based approaches to estimate an optimal window. Specifically, given a statistical model, a window size is considered optimal if it minimizes Akaike's information criterion (AIC) or maximizes cross-validation (CV) log-likelihood as described in Section 2.3, across a wide range of window sizes. In Section 2.2, we first discuss some previously published methods of estimating an optimal window size in the general context of probability density function. As above, the term 'density' that we use from this point onwards is basically read count that is standardized with the number of reads in the data so that the reads densities sum up to one.

## 2 METHODS

### 2.1 Samples and sequence alignments

Using an Illumina GAII, we produced 1 836 450 and 1 653 081 reads from DNA isolated from a fresh frozen lung tumour resection specimen and paired blood, respectively, from patient LS041. Similarly, we produced 3 089 173 test and 2 545 305 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 dataset and used 44 762 968 test and 34 293 547 control reads from cell line HCC1143 (Chiang *et al.*, 2009) in the simulation study. Sequences were aligned using the bwa suite version 0.5.9-r16 (Li and Durbin, 2009) against assembly hg19 of the human genome. Only sequences that could be uniquely aligned and with mapping quality $\geq 37$ were used.

### 2.2 Optimal window size in histogram construction

We discuss here some previously published methods to estimate an optimal window size in the context of histogram construction. The problem of identifying an optimal window size is well known and has received much attention in the past few decades. In the construction of a histogram, defining the window size, hence, the location of the breaks, is critical so that the histogram can reflect the true underlying density of the data. In this article, we highlight this principle to define the window size for next-generation sequence data. Let $x_1, x_2, \ldots, x_n$ be the observed position of reads in the genome, assumed to be random sample from a density $f(x)$. The density $f(x)$ represents the underlying true density of reads. In this article, we consider the problem of 'binning' the observed reads into equally spaced mesh or windows.

Let $I_i(x)$ be the $i$-th genomic window, $i = 1, 2, \ldots, m$, and let $t_i(x)$ denote the left hand point of $I_i(x)$. We denote $h = t_{i+1} - t_i$ to be the width of the window. Obviously, the width depends on the number of reads in the data, i.e. $h_n$, but without loss of generality, we drop the subscript in the notation for simplicity.

Let $v_i(x)$ be the number of reads falling in the genomic window $I_i(x)$. This quantity, $v_i(x)$, has a binomial distribution Binom$(n, p_i(x))$ where $p_i(x)$ is the probability of reads in the window $I_i(x)$. Scott (1979) assumed that $f(x)$ is a continuous function and regular, or more specifically, a probability density function. In this setting, the histogram estimate is given by

$$\hat{f}(x) = \frac{v(x)}{nh} \tag{1}$$

which is the estimate of binomial parameter $p(x)$. Scott (1979) is interested to minimize the integrated mean squared error

$$IMSE = \int E\left\{\hat{f}(x) - f(x)\right\}^2 dx \tag{2}$$

Through some approximations, (Scott, 1979) arrived at

$$IMSE = \frac{1}{nh} + \frac{1}{12}h^2 \int_{-\infty}^{\infty} f'(x)^2 dx + O(1/n + h^3) \tag{3}$$

where "'" notation represents a differentiation with respect to $x$. Minimizing the first two terms in the above Equation (3), we have an approximate optimal window size

$$h^* = \left\{\frac{6}{n \int f'(x)^2 dx}\right\}^{\frac{1}{3}}. \tag{4}$$

The approaches taken by Scott (1979) and other authors [e.g. Freedman and Diaconis (1981); Hall and Marron (1987); Jones and Sheather (1991); Wand (1997)] to estimate optimal window size are mainly to approximate a regular density function or probability density function using histogram. In low-coverage next-generation sequence data, the data generally exhibit density irregularities and noticeable random error variability. So, linear approximations that are mainly used in the above studies are inadequate.

Moreover, many of the above approaches require an evaluation of the integral of $f'(x)^2$ or its similar form. The evaluation of those quantities will produce higher correction factor in the denominator of expression for $h$ compared with regular density function. As a result, we experience considerable underestimation of optimal window size $h$.

To deal with this problem, we propose to empirically estimate the optimal window size using CV and AIC, given a model on reads density. An important advantage of this approach is that we do not need to make an approximation of the underlying density of reads $f(x)$, which can be poorly estimated in sequence data. The only minor downside of these approaches is that they require more computational power. However, this can be overcome easily even with a moderate personal computer.

We describe those two methods in the following section.

### 2.3 AIC and CV

The main concern in the existing methods to estimate an optimal window size is the evaluation of an integral of $f'(x)^2$ or its similar form. This evaluation makes the optimal window size heavily underestimated or, equivalently, the number of windows $m$ in the genome grossly overestimated. To strike a balance between model complexity (in terms of number of windows) and model accuracy (histogram density estimation), we resort to AIC and CV to estimate the optimal window size.

AIC and CV have been used to compare models and, in our setting, they are used to compare models with different window sizes. When the minimum AIC is attained, i.e. at the optimal window size, the expected distance of the model to the underlying probability structure that generates the data is minimal.

To proceed, instead of assuming $f(x)$ to be a smooth function, Taylor (1987) defined $f(x)$ as a step function

$$f(x) = c_i, \; x \in I_i(x) \tag{5}$$

over a given window $I_i(x)$. The likelihood function is given by

$$L(c) = \prod_{j=1}^{n} \hat{f}(X_j) = \prod_{i=1}^{m} c_i^{v_i(x)} \tag{6}$$

Under the constraints $c \geq 0$ and $\sum_i c_i = 1/h$, we have the familiar maximum likelihood estimate as histogram

$$c_i = \frac{v_i(x)}{nh} \tag{7}$$

In the above Equation (7), the maximum likelihood estimate of histogram is just the standardized reads count in $i$-th window.

An optimal number of windows $m$ (hence, optimal window size) is estimated as the one that minimizes AIC

$$AIC = m - \sum_{i=1}^{m} v_i(x) \log\left\{\frac{v_i(x)}{nh}\right\} \tag{8}$$

across different values of $m$. In Equation (8) above, the penalty to the log-likelihood takes into effect in terms of the number of windows $m$ in the histogram construction. To have an idea on the magnitude of $m$, the number of windows we consider in our LS041 data example ranges from ~309 000 (in 10 kb window) to 1500 (in 2000 kb window).

In addition to AIC, we also consider CV of the likelihood in (6) in estimating the optimal window. The main purpose is to support the results that we obtain when estimating optimal window size based on AIC, as the connection between the two methods is well known (see below). We work with the log of the likelihood in (6), which can be written as

$$\begin{aligned} \log L(c) &= \log\left\{\prod_{i=1}^{m} c_i^{v_i(x)}\right\} \\ &= \sum_{i=1}^{m}\sum_{s=1}^{v_i(x)} \log c_i \\ &= \sum_{i=1}^{m}\sum_{s=1}^{v_i(x)} \log\left(\tfrac{v_i(x)}{nh}\right) \end{aligned} \tag{9}$$

where $v_i(x)$ is the number of reads in window $i$ as previously defined.

For a given window size, we then calculate a leave-one-out cross-validated log-likelihood

$$\log L^{CV}(c) = \sum_{i=1}^{m}\sum_{s=1}^{v_i(x)} \log\left(\frac{v_i^{(-s)}(x)}{(n-1)h}\right) \tag{10}$$

where $(-s)$ denotes that the $s$-th read in window $i$ is not taken into account in the calculation. Our aim is to identify the optimal window size that maximizes the CV log-likelihood (10). The calculation of CV likelihood (10) can be done relatively fast, as we only evaluate the quantities in each window, and not each read.

In a general context, the AIC and CV methods are theoretically equivalent in estimating the optimal window size. The equivalence of both methods can be explained by considering that the estimation of optimal window size is essentially a model selection problem, where the window size $h$ (or, equivalently, the number of windows in the genome $m$) is the model parameter. Technically, we compare some competing models and try to select the best model. In this context, the papers by Stone (1974, 1977) as well as Pawitan (2001, pp. 381–382) proved the equivalence of AIC and CV in model selection and showed that the best model obtained by AIC will also be the one obtained by CV.

In practice, however, there are some possibilities that the estimated optimal window size by the two methods is different. In some datasets, the AIC and CV log-likelihood curves can almost be flat around the minimum. When this happens, the differences in the estimated optimal window size between the two methods can be attributed to random chance. If this happens, it means that those window sizes around the flat area are almost equally optimal, and we can safely select whichever optimal window size identified by either method that is more practical in our application.

## 2.4 Simulation study

Our interest in the above estimation method is whether, at the optimal window we estimated, the reads density we estimated in the low-coverage data $\hat{f}(x)$ is the closest to the underlying true reads density $f(x)$. From a practical point of view, the true underlying reads density can be considered as the read density when we have very high coverage in our data. We simulated a high-coverage data, which are considered to contain the 'true' density of reads $f(x)$. From this high-coverage data, we take random samples to create low-coverage (simulated) data. From the low-coverage data, we estimate the true density with $\hat{f}(x)$ and estimate the optimal window size using our proposed model. The objective is to confirm that the optimal window size we estimated is the one that brings $\hat{f}(x)$ to be the closest to $f(x)$.

We produced a female highly aneuploid genome (tumour sample only) with 19 large and small deletions and duplications using the human reference genome assembly hg19. In these data, we have 81 624 495 reads (~27×), and this is considered as the high-coverage data in our simulation. We mapped these reads to the genome in the same way we would do for a real sample to introduce the same mapping biases. To simulate low-coverage data from this simulated data, we sample with 1% probability 100 times to create 100 low-coverage simulated data with an average of 816 245 reads. One thing to note here is that we simulate the high-coverage data once, and the low-coverage datasets were created by sampling from that high-coverage data. Further details of the simulated data are presented in the Supplementary Material.

We also consider the HCC1143 cell line dataset (Chiang *et al.*, 2009) in our simulation study with 44 762 968 tumour and 34 293 547 control reads (~15× and 11×). The datasets are considered as the high-coverage data, and to simulate a low-coverage dataset from these data, we sample with 10% probabilities 100 times. This step creates 100 low-coverage simulated data for each tumour and control sample. The results of simulation using this dataset are presented in the Supplementary Material.

As mentioned above, we consider the reads density estimates in the high-coverage data $\hat{f}_H(x; h = h_0)$ as the 'true' population density $f(x)$, at a fixed window size $h_0$. Being the 'true' density, $\hat{f}_H(x; h = h_0)$ is assumed to be fixed and known. The estimated reads density from each simulated low-coverage data $\hat{f}_L(x; h)$ will then be compared with $\hat{f}_H(x; h = h_0)$ at different window sizes. The dependency of the densities on the window size in $\hat{f}_L(x; h)$ is made explicit in the notation here. The choice for $h_0$ is not so critical here as they are for high-coverage data; however, we consider 5 and 10 kb for $h_0$ in the simulation study.

Specifically, we first calculate the squared distance between the densities of the $j$-th low-coverage simulated data and the high-coverage data

$$R_j = \int \left\{\hat{f}_H(x; h = h_0) - \hat{f}_L(x; h)\right\}^2 dx \tag{11}$$

where the integration means that the squared differences are summed up across the whole genome between the two data for each window size. We then take the mean across the $n_{sim} = 100$ low-coverage simulated data to come up with a single number summary for each window size

$$R = \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} R_j \tag{12}$$

At the same window size, the density estimates $\hat{f}_H(x; h)$ and $\hat{f}_L(x; h)$ are comparable because the estimates are adjusted for the different number of reads as in Equation (1). Our interest is to identify the window size that gives the lowest $R$, which indicates the closest 'distance' between the observed density and the simulated 'true' density. This is to confirm

whether the optimal window size we estimated using our proposed method in Section 2.3 gives the best estimation of the 'true' underlying density of reads.

## 3 RESULTS

For the LS041 data, the results on estimation of optimal window size are presented in Figure 2, based on the AIC and CV log-likelihood. The minimum AIC is achieved at the window size 100 kb, equivalent to an average of 60 reads per window in both tumour and normal samples. Similarly, the maximum CV log-likelihood is located at the window size 100 kb in both tumour and normal samples. Because the horizontal axis of the figures is in a log scale, the range of window sizes around the optimal one that can be considered near-optimal is actually wider than it seems to be. For example, although the window size 100 kb is optimal, window sizes between 50 and 250 kb can be considered near-optimal in Figure 2 (Supplementary Fig. S1 where the horizontal axis is in a linear scale). This can be regarded as an advantage as we will revisit in the discussion section.

### 3.1 Optimal window size for LS010 data

For the LS010 sample, the results on the estimation of optimal window size are presented in Figure 3. In the tumour sample, the optimal window size is estimated at 80 kb (61 reads per window) using the AIC and 70 kb (54 reads per window) using the CV log-likelihood. The difference is not substantive and it is not a contradiction because the AIC and CV log-likelihood curves happen to be nearly flat around 70–80 kb window sizes (with a difference of ~0.09%). In the normal sample, the optimal window size is estimated at 100 kb (63 reads per window), using both the AIC and CV log-likelihood.

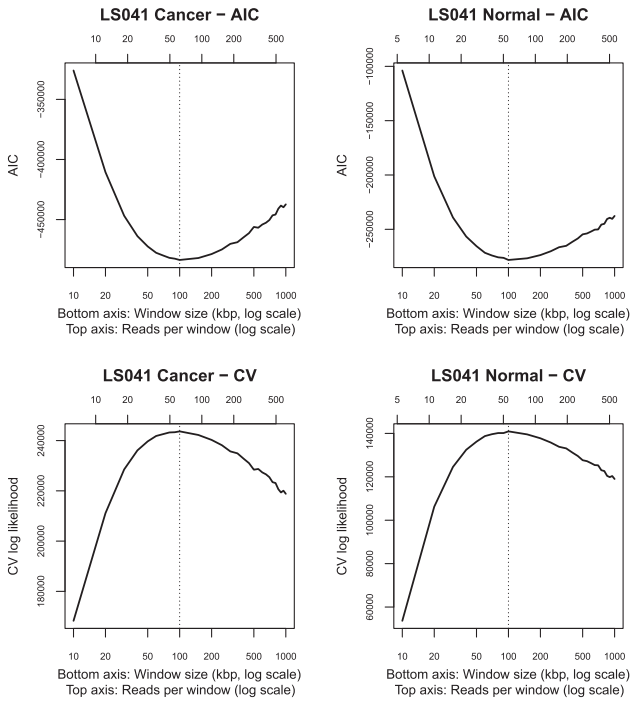


**Fig. 2.** AIC (top row) and CV log-likelihood (bottom row) as a function of different window sizes (bottom axis in each figure) or the corresponding number of reads per window (top axis in each figure) in LS041 cancer and normal sample. The horizontal axes are in log scale. The vertical dotted lines mark the optimal window size (or, equivalently, optimal number of reads per window). See Supplementary Figure S1 for figures with horizontal axes in linear scale

### 3.2 Simulated data

Figure 4 shows the mean of sum of squared distance $R$ in the simulated data as a function of window size. The distance $R$ is

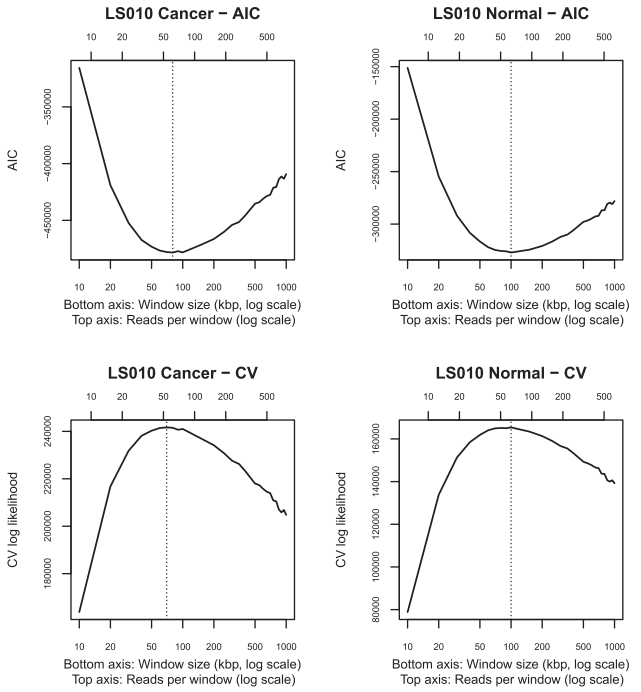


**Fig. 3.** AIC (top row) and CV log-likelihood (bottom row) as a function of different window sizes (bottom axis in each figure) or the corresponding number of reads per window (top axis in each figure) in LS010 cancer and normal sample. The horizontal axes are in log scale. The vertical dotted lines mark the optimal window size (or, equivalently, optimal number of reads per window). See Supplementary Figure S2 for figures with horizontal axes in linear scale

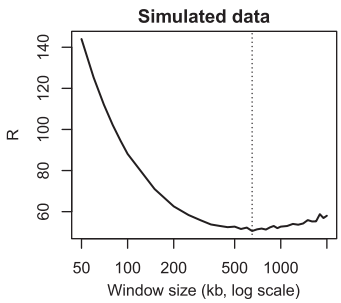


**Fig. 4.** Mean of sum of squared distance, $R$, in the simulated data as a function of window sizes. The mean distance $R$ is calculated between the read density in the (high-coverage) simulated data and its 100 low-coverage samples (at 1% sampling). The horizontal axis is in log scale

calculated between the read density in the high-coverage simulated data and its 100 low-coverage samples. The low-coverage samples were obtained from 1% random sampling from the high-coverage simulated data. The figure indicates that $R$ reaches the minimum at window size 600 kb. This result indicates that at this optimal window size, the distance between densities in the high and low coverage is at its closest. Our interest is to check whether we can estimate this optimal window size from the (sampled) low-coverage samples, using our proposed method. The results of this estimation are presented in Figure 5.

Figure 5 shows the AIC and CV log-likelihood from our model (5) in the low-coverage simulated data. The figure is the mean from 100 low-coverage simulated data, sampled from the high-coverage simulated data at 1%. The figure indicates that the optimal window size is estimated at 600 kb (140 reads per window), at which the AIC reaches its minimum and the CV log-likelihood also reaches its maximum. This confirms the result in Figure 4 that at this optimal window size the distance of density between the high-density and low-density simulated data is minimal.

The results of sensitivity analysis on the simulated data to detect gains/losses in CNA are presented in Figure 6. We use the CNAnorm package (Gusnanto *et al.*, 2012) to analyse the data, where we use both the circular binary segmentation (CBS) (Olshen *et al.*, 2004) and smooth segmentation (Huang *et al.*, 2007) to estimate CNA. For the CBS segmentation, the optimal window size at 600 kb shows better combination of sensitivity and specificity, compared with the window sizes 200 kb and 2 Mb. At low 1-specificity, the window size of 2 Mb shows slightly better sensitivity compared with the optimal window size 600 kb. This is a signature of CBS that prefers longer genomic continuous regions, which exhibit gains/losses in our data. When the data are segmented using smooth segmentation, the optimal window size 600 kb gives better sensitivity and specificity than the data based on window sizes 200 kb and 2 Mb.

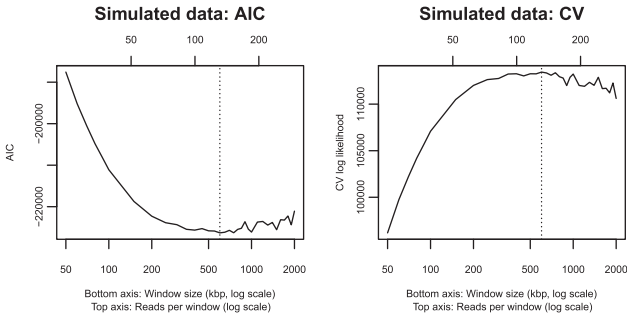


**Fig. 5.** The AIC (left) and CV log-likelihood (right) as a function of window size or number of reads per window, from the low-coverage samples. The horizontal axis is in log scale

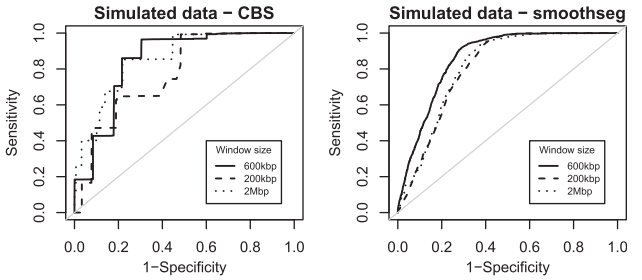


**Fig. 6.** Sensitivity analysis of the simulated data at different window sizes in detecting gains or losses, when the CNA are segmented using CBS and smooth segmentation. The solid grey diagonal line is the identity line

## 4 DISCUSSION

The estimation of optimal window size for analysis of low-coverage next-generation sequence data can be considered to be a histogram construction where the 'breaks' for histogram are the window limits. As such, each experiment will result in one particular optimal window size, depending on the underlying features and biases, both technical and biological.

Previous approaches to this problem are problematic. Some methods tend to underestimate the optimal window size in NGS data because of a poor approximation to the underlying reads density in the genome. For example, the method proposed by Xie and Tammi (2009) resulted in an estimated optimal window size of 40 kb (24 reads per window) to 10 kb (6 reads per window) in our datasets, for detecting a copy-number ratio of 1.2 to 1.4, respectively, with $P$-value of 0.0001 (Supplementary Material), although CNV-Seq is only meaningful for window size >10 reads per window. The latter requirement is to prevent the random variability to dominate the analysis. Some other approaches, which mainly rely on a subjective consideration, also tend to overestimate the optimal window size. This is because having a wide window size gives a smooth pattern of genomic features that can easily be interpreted. However, this mainly subjective approach ignores the potential discoveries of genomic features that are present in short regions, as we discussed previously. In this study, we proposed to use AIC and CV log-likelihood, given in the model in Section 2.3, because the methods do not depend on an approximation of the reads density, have a simple interpretation and are relatively easy to implement.

Results from our analysis with the LS041 and LS010 data suggest that some window sizes around the optimal one can be considered as near-optimal (Figs 1 and 2 with a linear-scale horizontal-axis in the Supplementary Material). This is an advantage when we want to analyse NGS data across different samples or patients. As each sample may exhibit a different optimal window size, there may be a window size that is in the overlapping regions of near-optimal window sizes across the different samples. Our tool will help the experimenter to make an informed decision when estimating an optimal window size to analyse the data. Our previous experience suggested that we lose little information when we use slightly suboptimal window size compared with the optimal one in the analysis across different samples.

One important message that we can take here is that the calculation of AIC and CV log-likelihood usually fails or becomes uninterpretable when the window size we evaluate has <5 reads per window on average. Although the optimal window size differs from one dataset to the next or from one experiment to the next, we found in our LS041 and LS010 datasets that the optimal window size is ~60 reads per window. We also found in our study that window size of 30–180 reads per window can be considered near-optimal. In the LS041 data, this corresponds to a

range of 50–300 kb window size, and in the LS010 data, 40–250 kb window size.

Our proposed methods assume a simple model for the underlying density of reads. Therefore, our estimation of the optimal window size depends on all technical and biological factors that contribute to the observed signal. The window size needs to be optimized so that it is capable of 'tracking' all the concurring factors that contribute to the final signal. As an example, to be able to correct for GC content, the signal needs to follow the GC content bias, and a window size unnecessarily too large or too small might prevent the best normalization. However, we have observed that reads mapped with low confidence sometimes cluster in particular regions of the genome and introduce extreme peaks in the signal that might negatively affect the optimal size of the window, as they dominate the calculation of AIC and CV log-likelihood. For this reason, we suggest that an estimation of optimal window size is made on the data with good mapping quality.

Our simulation study (Supplementary Material) indicates that the proposed methods are able to estimate an optimal window size that minimizes the distance between the observed reads density in the low-coverage data and the 'true' underlying density. This holds in many typical cases where, for example, structural rearrangements are reasonable. In a rare case where we have extreme structural chromosome rearrangements, our proposed methods can still estimate a near-optimal window size, which is close to the optimal one. In this regard, the proposed methods can still be useful to identify a good estimate of window size that can be used in the analysis.

Last but not the least, our proposed methods can also be used in a higher coverage context as well, or, possibly, in transcriptome or ChIP-seq experiments or whenever data need to be binned in windows of predefined size.

## 5 CONCLUSION

In the context of the analysis of very low-coverage next-generation sequence data, the estimation of optimal window size is critical. If the window is too narrow or too wide, we will potentially miss genomic features of interest. To estimate the optimal window size, we first assume the reads density to be a step function. Given this, the optimal window size is estimated as the one that minimizes AIC or maximizes CV log-likelihood across different window sizes that we evaluate. Our analysis on LS041 and LS010 data indicates that the optimal window size is approximately equivalent to 60 reads per window. Our simulation study confirms that the optimal window size we estimated produces the closest distance between the reads density in the low-coverage data and the 'true' underlying density.

## ACKNOWLEDGEMENTS

The authors would like to thank Helene Thygesen for useful feedback on the manuscript.

*Funding*: Yorkshire Cancer Research (L341PG); and the Leeds Cancer Research UK Centre (C37059/A11941).

*Conflict of Interest*: The last author (S.B.) is currently employed by Illumina UK Ltd.

## REFERENCES

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Boyle,A.P. *et al.* (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.

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

Cheng,C. *et al.* (2011) A statistical framework for modelling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.*, **12**, R15.

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

Freedman,D. and Diaconis,P. (1981) On the histogram as a density estimator: L2 theory. *Z. Wahrsheinllchkeffstheorie Verwandte Gebeite*, **57**, 453–476.

Gusnanto,A. *et al.* (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**, 40–47.

Hall,P. and Marron,J.S. (1987) Estimation of integrated squared density derivatives. *Stat. Probab. Lett.*, **6**, 109–115.

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

Johnson,D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.

Jones,M. and Sheather,S.J. (1991) Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Stat. Probab. Lett.*, **11**, 511–514.

Lee,H. and Schatz,M.C. (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, **28**, 2097–2105.

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

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.

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

Robertson,G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Scott,D.W. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

Stone,M. (1974) Cross-validatory choice and assessment of statistical prediction. *J. R. Stat. Soc. B*, **36**, 111–147.

Stone,M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. B*, **39**, 44–47.

Taylor,C.C. (1987) Akaike's information criterion and the histogram. *Biometrika*, **74**, 636–639.

Wand,M.P. (1997) Data-based choice of histogram bin width. *Am. Stat.*, **51**, 59–64.

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.

Xi,R. *et al.* (2011) Detecting structural variations in the human genome using next generation sequencing. *Brief. Funct. Genomics*, **9**, 405–415.

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.

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