We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 5

Enhancing Estimates of Breakpoints in Genome Copy Number Alteration using Confidence Masks

Jorge Muñoz-Minjares, Yuriy Shmaliy and
Oscar Ibarra-Manzano

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63913

Abstract

Chromosomal structural changes in human body known as copy number alteration (CNA) are often associated with diseases, such as various forms of cancer. Therefore, accurate estimation of breakpoints of the CNAs is important to understand the genetic basis of many diseases. The high-resolution comparative genomic hybridization (HR-CGH) and single-nucleotide polymorphism (SNP) technologies enable cost-efficient and high-throughput CNA detection. However, probing provided using these profiles gives data highly contaminated by intensive Gaussian noise having white properties. We observe the probabilistic properties of CNA in HR-CGH and SNP measurements and show that jitter in the breakpoints can statistically be described with either the discrete skew Laplace distribution when the segmental signal-to-noise ratio (SNR) exceeds unity or modified Bessel function-based approximation when SNR is <1. Based upon these approaches, the confidence masks can be developed and used to enhance the estimates of the CNAs for the given confidence probability by removing some unlikely existing breakpoints.

Keywords: copy number alterations, HR-CGH, SNP, breakpoints, confidence masks

1. Introduction

It is well known that the deoxyribonucleic acid (DNA) of a genome essential for human life often demonstrates structural changes [1–3] called genome copy number alterations (CNAs) [4–6], which are associated with disease such as cancer [7]. Analysis of the breakpoint locations in the CAN structure is still an important issue because it helps detecting structural alterations, load

of alterations in the tumor genome, and absolute segment copy numbers. Thus, efficient estimators are required to extract information about the breakpoints with accuracy acceptable for medical needs. To produce CNA profile, several technologies have been developed such as comparative genomic hybridization (CGH) [8], high-resolution CGH (HR-CGH) [9], whole genome sequencing [10], and most recently single-nucleotide polymorphism (SNP) [11]. The HR-CGH technology is still used widely in spite of its low resolution [12]. It has been reported in [13] that the HR-CGH arrays are accurate to detect structural variations (SVs) at the resolution of 200 bp (*base pairs*). Most recently, the single-nucleotide polymorphism technology was developed in the study of Wang et al. [11] to provide high-resolution measurements of the CNAs. In spite of their high resolution, the modern methods still demonstrate the inability in obtaining good estimates of the breakpoint locations because of the following factors: (1) the nature of biological material (tumor is contaminated by normal tissue, relative values, and unknown baseline for copy number estimation), (2) technological biases (quality of material and hybridization/sequencing), and (3) intensive random noise. The HR-CGH and SNP profiles have demonstrated deficiency in detecting the CNAs, but noise in the detected changes still remains at a high level [14] and accurate estimators are required to extract information about structural changes.

In the HR-CGH microarray technique, the CNAs are often normalized and plotted as $\log_2 R/G = \log_2$ ratio, where R and G are the fluorescent Red and Green intensities, respectively [12]. The CNA measurements using SNP technologies are represented by the Log-R ratios (LRRs), which are the log-transformed ratios of experimental and normal reference SNP intensities centered at zero for each sample [14]. From the standpoint of signal processing, the following properties of the CNA function are of importance [15]:

- It is piecewise constant (PWC) and sparse with a small number of alterations on a long base-pair length.
- Constant values are integer, although this property is not survived in the log-R ratio.
- The measurement noise in the log-R ratio is highly intensive and can be modeled as additive white Gaussian.

The CNA estimation problem is thus to predict the breakpoint locations and the segmental levels with a maximum possible accuracy and precision acceptable for medical applications. In this work, we developed our methods to two types of cancer: B-cell chronic lymphocytic leukemia (B-CLL) and BLC primary breast carcinoma. Nevertheless, the methods were designed to any samples of cancer with the characteristics described above.

2. Methods

2.1. CNA model and problem statement

Consider a chromosome section observed with some resolution Δ, bp at M discrete breakpoints, $n \in [1, M]$. An example of the CNA probes with a single breakpoint and two segments is shown in **Figure 1**. Suppose that the copy numbers change at K breakpoints,

http://dx.doi.org/10.5772/63913

$1<i_1<\ldots<i_K<M$, united in a vector $\boldsymbol{I}=[i_1 i_2 \ldots i_K]^{\mathrm{T}}$. The measurement can thus be represented with a vector $\boldsymbol{y}\in\mathcal{R}^M$ as

$$y=\left[y_1 y_2 \ldots y_{i_1} y_{i_1}+1 \ldots y_{i_2} \ldots y_{i_K} \ldots y_M\right]^T. \tag{1}$$

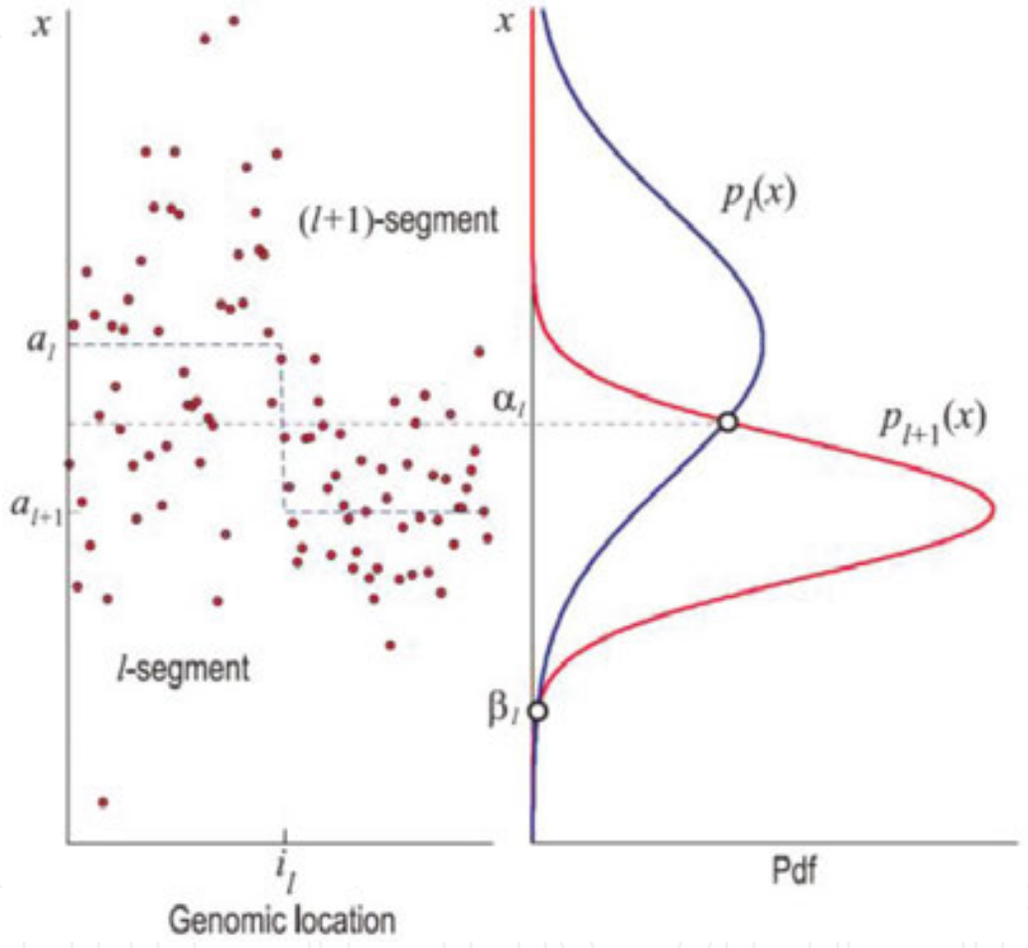

Figure 1. Typical CNA measurements with white Gaussian noise with a single breakpoint, between two segments l and $l+1$ having different segmental variances. The pdf for neighboring segments are depicted as $p_l(x)$ and $p_{l+1}(x)$.

Introduce a vector $\boldsymbol{a}\in\mathcal{R}^{K+1}$ of segmental levels, $\boldsymbol{a}=[a_1 a_2 \ldots a_{K+1}]^{\mathrm{T}}$, where a_1 corresponds to the interval $[1,\ i_1]$, a_{K+1} to $[i_K,\ M]$, and a_K, $k\geq 2$, to $[i_{k-1},\ i_k]$. In such a formulation, y obeys the linear regression model

$$\boldsymbol{y}=\boldsymbol{A}(\boldsymbol{I})\boldsymbol{a}+\ \mathbf{v} \tag{2}$$

where the regression matrix $\boldsymbol{A} \in \mathcal{R}^{M \times (K+1)}$ is sparse,

$$\boldsymbol{A} = \left[\boldsymbol{A}_1^T \boldsymbol{A}_2^T \ldots \boldsymbol{A}_{K+1}^T \right]^T, \tag{3}$$

having a component

$$\boldsymbol{A}_k = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}, \tag{4}$$

in which the kth column is filled with unity and all others are zeros. The number of the columns in $\boldsymbol{A}_k$ is exactly $K+1$. However, the number or the row depends on the interval $i_k - i_{k-1}$. Thus, the row-variant matrix (4) is $\boldsymbol{A}_k \in \mathcal{R}^{(i_k - i_{k-1}) \times (K+1)}$. Additive noise $\boldsymbol{v}$ in Eq. (2) is zero mean, $E\{\mathbf{v}\}=0$, and white Gaussian with the covariance $\boldsymbol{R}=\sigma_v^2 \boldsymbol{I}$, where $\boldsymbol{I} \in \mathcal{R}^{M \times M}$ is an identity matrix and σ_v^2 is a know variance.

The CNA estimation problem is thus to predict the breakpoint locations and evaluate the segmental changes $\boldsymbol{x}=\boldsymbol{A}(I)\boldsymbol{a}$ with a maximum possible accuracy and precision acceptable for medical applications. The problem is complicated by short number of the probes in each neighboring segment and indistinct edges. Therefore, an analysis of the estimation errors caused by the segmental noise and jitter in the breakpoints is required.

2.2. Jitter probability in the breakpoints

Consider a typical genomic measurement of two neighboring CNA segments in white Gaussian noise with different segmental variances as shown in **Figure 1**. A constant signal changes from level a_l to level a_{l+1} around the *breakpoint* i_l. In the presence of noise, the location of i_l is not clear owing to commonly large segmental variances σ_l^2 and σ_{l+1}^2. As an example, the Gaussian noise probability density functions (pdfs) $p_l(x)$ and $p_{l+1}(x)$ are shown in **Figure 1** for $\sigma_l^2 > \sigma_{l+1}^2$. Let us notice that $p_l(x)$ and $p_{l+1}(x)$ cross each other in two points, α_l and β_l, provided that $\sigma_l^2 \neq \sigma_{l+1}^2$.

Now considerer N probes in each segment neighboring to i_l with an average resolution. We thus may assign an event A_{lj} meaning that measurement at point $i_l - N \leq j < i_l$ belongs to the lth segment. Another event B_{lj} means that measurement at $i_l \leq j < i_l + N - 1$ belongs to the $(l+1)$th segment. In our approach, we think that a measured value belongs to one segment if the probability is larger than if it belongs to another segment. For example, any measurement point in the interval between α_l and β_l (**Figure 1**) is supposed to belong to the $(l+1)$th segment.

http://dx.doi.org/10.5772/63913

Following **Figure 1** and assuming different noise variances σ_l^2 and σ_{l+1}^2, the events A_{lj} and B_{lj} can be specified as follows [16]:

$$\underset{i_l - N_l \le j < i_l}{A_{lj}} \text{ is } \begin{cases} (\alpha_l < x_j) \vee \left(x_j < \beta_l\right), & \sigma_l^2 > \sigma_{l+1}^2, \\ x_j > \alpha_l, & \sigma_l^2 = \sigma_{l+1}^2, \\ \alpha_l < x_j < \beta_l, & \sigma_l^2 < \sigma_{l+1}^2, \end{cases} \tag{5}$$

$$\underset{i_l \le j < i_l + N_l - 1}{B_{lj}} \text{ is } \begin{cases} \beta_l < x_j < \alpha_l & \sigma_l^2 < \sigma_{l+1}^2, \\ x_j < \alpha_l, & \sigma_l^2 = \sigma_{l+1}^2, \\ (x_j < \alpha_l) \vee \left(x_j > \beta_l\right), & \sigma_l^2 > \sigma_{l+1}^2. \end{cases} \tag{6}$$

Because each point can belong only to one segment, the inverse events are $\bar{A} = 1 - A_{lj}$ and $\bar{B} = 1 - B_{lj}$.

Events A_{lj} and B_{lj} can be united into two blocks: $\boldsymbol{A}_l = \{A_{l(i_l - N)} A_{l(i_l - N + 1)} \dots A_{l(i_l - 1)}\}$ and $\boldsymbol{B}_l = \{B_{l(i_l)} B_{l(i_l + 1)} \dots B_{l(i_l + N - 1)}\}$.

If $\boldsymbol{A}_l$ and $\boldsymbol{B}_l$ occur simultaneously with unit probability each, then jitter at i_l will never occur. However, some other events may be found, which do not obligatorily lead to jitter. We ignore such events and define approximately the probability $P(\boldsymbol{A}_l, \boldsymbol{B}_l)$ of the jitter-free breakpoint as

$$P\left(\boldsymbol{A}_l, \boldsymbol{B}_l\right) = P\left(\mathrm{A}_{\mathrm{i}_l - N} \dots \mathrm{A}_{\mathrm{i}_l - 1} \mathrm{B}_{\mathrm{i}_l} \dots \mathrm{B}_{\mathrm{i}_l + N - 1}\right). \tag{7}$$

The inverse event $\bar{P}(\boldsymbol{A}_l, \boldsymbol{B}_l) = 1 - P(\boldsymbol{A}_l, \boldsymbol{B}_l)$ can be called the *approximate jitter probability* [17].

2.3. Jitter distribution in the breakpoints

To determine the confidence limits for CNAs using high-resolution genomic arrays, jitter in the breakpoints must be specified statistically for the segmental Gaussian distribution. This can be done approximately if to employ either the discrete skew Laplace distribution or, more accurately, the modified Bessel function of the second kind and zeroth order.

2.3.1. Approximation with discrete skew Laplace distribution

Following the definition of the *jitter probability* given in Section 2.1 and taking into consideration that all the events are independent in white Gaussian noise, Eq. (7) can be rewritten as: $P(\boldsymbol{A}_l, \boldsymbol{B}_l) = P^N(A_l) P^N(B_l)$, where, following Eqs. (5) and (6), the probabilities $P(A_l)$ and $P(B_l)$ can be specified as, respectively,

$$P(A_l)=\begin{cases}1-\int_{\beta_l}^{\alpha_l} p_l(x)dx, & \sigma_l^2>\sigma_{l+1}^2,\\ \int_{\alpha_l}^{\infty} p_l(x)dx, & \sigma_l^2=\sigma_{l+1}^2,\\ \int_{\alpha_l}^{\beta_l} p_l(x)dx, & \sigma_l^2<\sigma_{l+1}^2,\end{cases} \tag{8}$$

$$P(B_l)=\begin{cases}\int_{\beta_l}^{\alpha_l} p_{l+1}(x)dx, & \sigma_l^2>\sigma_{l+1}^2,\\ \int_{-\infty}^{\alpha_l} p_{l+1}(x)dx, & \sigma_l^2=\sigma_{l+1}^2,\\ 1-\int_{\alpha_l}^{\beta_l} p_{l+1}(x)dx, & \sigma_l^2<\sigma_{l+1}^2,\end{cases} \tag{9}$$

where $p_l(x)=1\Big/\sqrt{2\pi\sigma_l^2}e^{-((x-a_l)^2)/\sigma_l^2}$ is the Gaussian density.

Suppose that jitter occurs at some point $i_l \pm k$, $0\le k\le N$, as shown, for example, in **Figure 1**, and assign two additional blocks of events $\boldsymbol{A}_{lk}=\{A_{i_l-N}\dots A_{i_l-1-k}\}$ and $\boldsymbol{B}_{lk}=\{B_{i_l+k}\dots B_{i_l+N-1}\}$. The probability $P_k^-\triangleq P_k^-\left(\boldsymbol{A}_{lk}\overline{A}_{l(i_l-k)}\dots\overline{A}_{i_l-1}\boldsymbol{B}_l\right)$ that jitter occurs at the kth point to the left from i_l (left jitter) and the probability $P_k^+\triangleq P_k^+\left(\boldsymbol{A}_l\overline{B}_{l(i_l+1)}\dots\overline{B}_{l(i_l+k-1)}\boldsymbol{B}_k\right)$ that jitter occurs at the kth point to the right from i_l (right jitter) can thus be written as, respectively,

$$P_k^-=P^{N-k}(A_l)\left[1-P(A_l)\right]^k P^N(B_l), \tag{10}$$

$$P_k^+=P^N(A_l)\left[1-P(B_l)\right]^k P^{N-k}(B_l), \tag{11}$$

By normalizing Eqs. (11) and (12) with Eq. (8), one can arrive at a function that turns out to be independent on N:

$$f_l(k)=\begin{cases}[P^{-1}(A_l)-1]^{|k|}, & k<0,\ (left)\\ 1 & k=0\\ [P^{-1}(B_l-1)]^k, & k<0.\ (right)\end{cases}. \tag{12}$$

Further normalization of $f_l(k)$ to have a unit area leads to the pdf $p_l(k)=\frac{1}{\varphi_l}f_l(k)$, where φ_l is the sum of $f_l(k)$ for all k,

$$\phi_l = 1 + \sum_{\infty}^{k=1} [\varphi_l^A(k) + \varphi_l^B(k)], \tag{13}$$

where $\varphi_l^A(k) = [P^{-1}(A_l) - 1]^k$ and $\varphi_l^B(k) = [P^{-1}(B_l) - 1]^k$.

It follows from the approximation admitted that $f_l(k)$ converges with k only if $0.5 < \widetilde{P} = \{P(A),\ P(B)\} < 1$. Otherwise, if $\widetilde{P} < 0.5$, the sum φ_l is infinite and $f_l(k)$ cannot be transformed to $p_l(k)$. It has been shown in [18] that such a situation is practically rare. It can be observed with extremely small and different segmental SNRs when the probabilities are comparable that the measurement point belongs to one of another segment.

Accepting $0.5 < \widetilde{P} = \{P(A),\ P(B)\} < 1$, one concludes that $\widetilde{P} < 0$, $\ln(1 - \widetilde{P}) < 0$, and $\ln(1 - \widetilde{P}) < \ln(\widetilde{P})$. Next, using a standard relation $\sum_{k=1}^{\infty} x^k = 1 \Big/ (x^{-1} - 1)$, where $x < 1$, and after little transformations, Eq. (14) can be brought to

$$\phi_l = \frac{P(A_l) + P(B_l) - 1}{[1 - 2P(A_l)][1 - 2P(B_l)]}. \tag{14}$$

The *jitter pdf* $p_l(k)$ associated with the *lth* breakpoint can finally be found to be

$$p_l(k) = \frac{1}{\phi_l} \begin{cases} [P^{-1}(A_l) - 1]^{|k|}, & k < 0, \\ 1, & k = 0, \\ [P^{-1}(B_l) - 1]^k & k > 0, \end{cases} \tag{15}$$

where ϕ_l is specified by Eq. (15) and $0.5 < P(A_l),\ P(B_l) < 1$.

If now to substitute $q_l = P^{-1}(A_l) - 1$ and $d_l = P^{-1}(B_l) - 1$, find $P(A_l) = 1/(1 + q_l)$ and $P(B_l) = 1/(1 + d_l)$, and provide the transformations, then one may arrive at a conclusion that Eq. (16) is the discrete skew Laplace pdf [19].

$$p(k|d_l q_l) = \frac{(1 - d_l)(1 - q_l)}{1 - d_l q_l} \begin{cases} p_l^k, & k \geq 0 \\ q_l^{|k|} & k \leq 0 \end{cases}, \tag{16}$$

where $d_l = e^{-(\kappa_l/\nu_l)} \in [0,\ 1]$ and $q_l = e^{-(1/\kappa_l \nu_l)} \in [0,\ 1]$ and in which κ_l and $\nu_l > 0$ still need to be connected to Eq. (16). With this aim, consider Eqs. (16) and (17) at $k = -1$, $k = 0$, and $k = 1$. By equating Eqs. (16) and (17), first obtain $((1 - d_l)(1 - q_l)d_l)/(1 - d_l q_l) = 1/\phi_l (1 - P(B_l))/P(B_l)$ for $k = 1$ and $((1 - d_l)(1 - q_l)q_l)/(1 - d_l q_l) = 1/\phi_l (1 - P(A_l))/P(A_l)$ for $k = -1$ that yields

$$\nu_l = \frac{1-\kappa_l^2}{\kappa_l \ln \mu_l}, \tag{17}$$

where $\mu_l = (P(A_l)[1-P(B_l)])/(P(B_l)[1-P(A_l)])$. For $k=0$ we have $((1-d_l)(1-q_l))/(1-d_l q_l) = 1/\phi_l$ and transform it to the equation $x_l^2 - (\phi_l(1+\mu_l))/(1+\phi_l)x - (1-\phi_l)/(1+\phi_l)\mu_l = 0$, where a proper solution is

$$x = \frac{\phi_l(1+\mu_l)}{2(1+\phi_l)}\left(1 - \sqrt{1 - \frac{4\mu_l(1-\phi_l^2)}{\phi_l^2(1+\mu_l)^2}}\right) \tag{18}$$

and which $x_l = \mu_l^{-(\kappa_l^2)/(1-\kappa_l^2)}$ gives us

$$\kappa_l = \sqrt{\ln \frac{(x_l)}{\ln(x_l/\mu_l)}}. \tag{19}$$

By combining Eq. (18) with Eq. (20), one may also get a simpler form for ν_l, namely $\nu_l = -\kappa_l / \ln x_l$.

Now, introduce the segmental signal-to-noise ratios (SNRs): $\gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}$, and $\gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2}$, where $\Delta_l = a_{l+1} - a_l$, substitute the Gaussian pdf to Eqs. (9) and (10), provide the transformations, and rewrite Eqs. (9) and (10) as

$$P(A_l) = \begin{cases} 1 + \frac{1}{2}\left[\operatorname{erf}\left(g_l^\beta\right) - \operatorname{erf}\left(g_l^\alpha\right)\right], & \gamma_l^- < \gamma_l^+ \\ \frac{1}{2}\operatorname{erfc}\left(g_l^\alpha\right), & \gamma_l^- = \gamma_l^+ \\ \frac{1}{2}\left[\operatorname{erf}\left(g_l^\beta\right) - \operatorname{erf}\left(g_l^\alpha\right)\right], & \gamma_l^- > \gamma_l^+ \end{cases}, \tag{20}$$

$$P(B_l) = \begin{cases} \frac{1}{2}\left[\operatorname{erf}\left(h_l^\alpha\right) - \operatorname{erf}\left(h_l^\beta\right)\right], & \gamma_l^- < \gamma_l^+ \\ 1 - \frac{1}{2}\operatorname{erfc}\left(h_l^\alpha\right), & \gamma_l^- = \gamma_l^+ \\ 1 + \frac{1}{2}\left[\operatorname{erf}\left(h_l^\alpha\right) - \operatorname{erf}\left(h_l^\beta\right)\right], & \gamma_l^- > \gamma_l^+ \end{cases}, \tag{21}$$

http://dx.doi.org/10.5772/63913

where $g_l^{\beta}=(\beta_l-\Delta_l)\big/\,|\Delta_l|\sqrt{\gamma_l^-/2}$, $g_l^{\alpha}=(\alpha_l-\Delta_l)\big/\,|\Delta_l|\sqrt{\gamma_l^-/2}$, $h_l^{\beta}=\beta_l\big/\,|\Delta_l|\sqrt{\gamma_l^+/2}$, $h_l^{\alpha}=\alpha_l\big/\,|\Delta_l|\sqrt{\gamma_l^+/2}$, $\mathrm{erf}(x)$ is the error function, $\mathrm{erfc}(x)$ is the complementary error function, and

$$\alpha_l,\beta_l=\frac{a_l\gamma_l^- - a_l\gamma_l^+}{\gamma_l^- - \gamma_l^+}\mp\frac{1}{\gamma_l^- - \gamma_l^+}\times\sqrt{(a_l-a_{l+1})\gamma_l^-\gamma_l^+ + 2\Delta_l^2\left(\gamma_l^- - \gamma_l^+\right)\ln\sqrt{\frac{\gamma_l^-}{\gamma_l^+}}}\ . \tag{22}$$

2.3.2. Approximation of jitter distribution using the modified Bessel functions

An analysis shows that the discrete skew Laplace pdf (17) gives good results only if SNR is >1. Otherwise, real measurements do not fit well, and a more accurate function is required. Below, we show that better approach to real jitter distribution can be provided using the modified Bessel functions.

2.3.2.1. Modified Bessel function

Figure 2 demonstrates the jitter pdf measured experimentally (dotted) for different SNRs. The breakpoint corresponds here to the peak density and the probability of the breakpoint location diminishes to the left and to the right of this point. Note that the discrete skew Laplace pdf (17) behaves linearly in such scales. Therefore, Eq. (17) cannot be applied when SNR is <1 and a more accurate function is required.

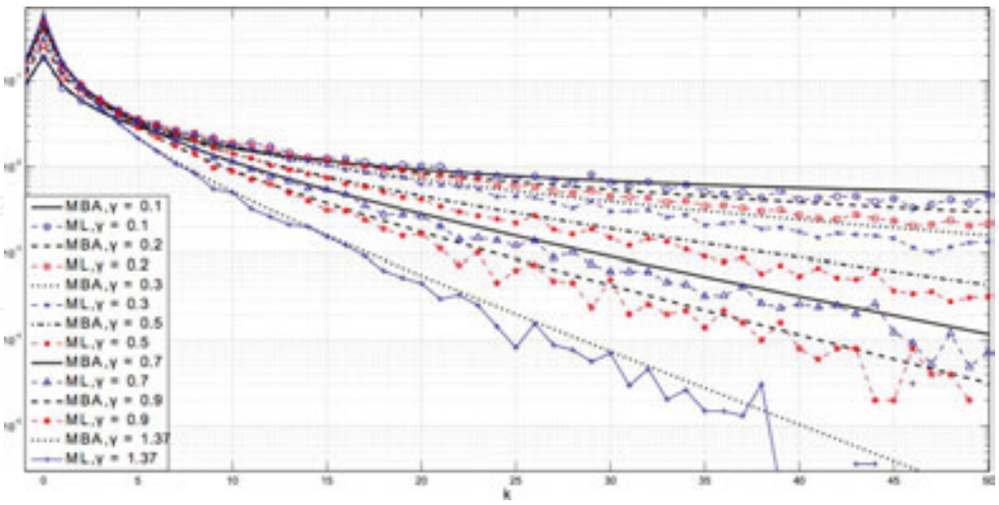

Figure 2. Experimentally defined one-sided jitter probability densities (dotted) of the breakpoint location for equal segmental SNR γ in the range of $M=400$ points with a true breakpoint at $n=200$. The experimental density functions were found using the *Maximum Likelihood* (ML) estimator. The histogram was plotted over 50×10^3 runs repeated nine times and average. Approximations (continuous) are provided using the proposed Bessel-based approximation depicted as MBA.

Among available functions demonstrating the pdf properties, the modified Bessel function of the second kind $K_0(x)$ and zeroth order is a most good candidate to fit the experimentally measured densities (**Figure 2**). The following form of $K_0(x)$ can be used:

$$K_0\left[x(k)\right] = \int_0^\infty \cos\left[x(k)\sinh t\right]dt = \int_0^\infty \frac{\cos\left[x(k)t\right]}{\sqrt{t^2+1}}dt > 0, x(k) > 0\ , \tag{23}$$

in which a variable $x(k)$ depends on index k, which represents a discrete departure from the assumed breakpoint location. Because $K_0[x(k)]$ is a positive-valued for $x(k)>0$ smooth function decreasing with x to zero, it can be used to approximate the probability density.

2.3.2.2. Approximation

In order to use Eq. (24) as an approximating function

$$\boldsymbol{B}(k|\gamma) = K_0\left[x(k)\right] \tag{24}$$

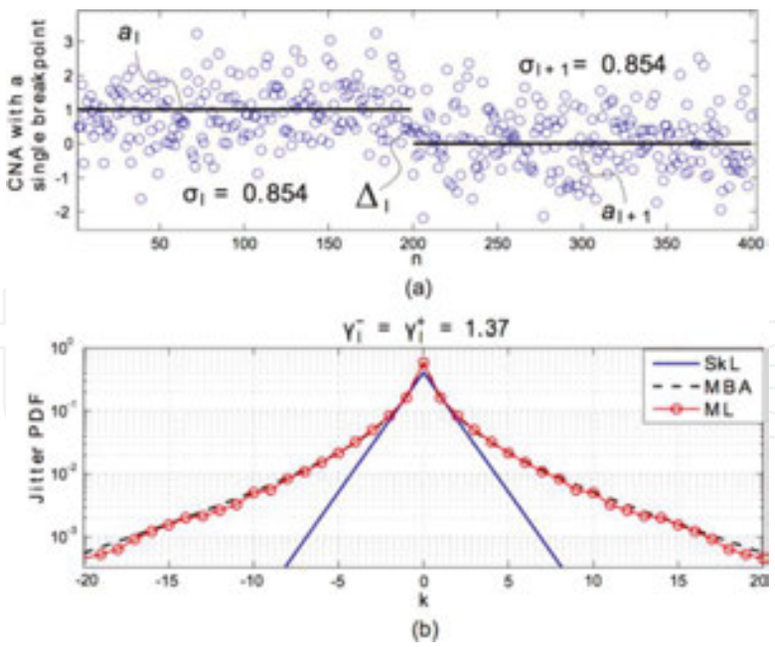

Figure 3. Simulated CNA with a single breakpoint at n = 200 and segmental standard deviations σ_l and σ_{l+1} corresponding to SNRs $\gamma_l^-=\gamma_l^+=1.37$: (a) measurement and (b) jitter distribution. Here, ML (circled) is the jitter pdf obtained experimentally using an ML estimator through a histogram over 50×10^3 runs, SkL (solid) is the Laplace distribution, and MBA (dashed) is the Bessel-based approximation.

conditioned on γ for the one-sided jitter probability densities shown in **Figure 2**, we represent a variable x via k as $x(k, \gamma)=\ln[\Phi(k, \gamma)]$ in a way such that small $k \geq 0$ correspond to large values x of and vice versa. Among several candidates, it has been found empirically that the following function $\Phi(k, \gamma)$ fits the histograms with highest accuracy:

$$\Phi(k,\gamma) = \left(|k|+1\right)^{\rho+\tau|k|}\left[\frac{1+\sqrt{\gamma}}{\gamma} - \epsilon\right], \tag{25}$$

if to set $\gamma=\gamma_l^-$ for $k<0$, $\gamma=\frac{\gamma_l^- + \gamma_l^+}{2}$ for $k=0$, and $\gamma=\gamma_l^+$ for $k>0$, and represent the coefficients and as $\tau(\gamma)$, $\rho(\gamma)$, and (γ) as

$$\tau(\gamma) = a_0\gamma + a_1 \tag{26}$$

$$\rho(\gamma) = \gamma\left(b_0\gamma^{b_1} + a_0\right) + b_2 \tag{27}$$

$$\epsilon(\gamma) = c_0\gamma^{c_1} + c_2 \tag{28}$$

where $a_0=0.02737$, $a_1=-4.5\times10^{-3}$, $b_0=0.3674$, $b_1=-0.3137$, $b_2=0.8066$, $c0=0.8865$, $c_1=-1.033$, and $c_2=-1.233$ were found in the mean square error (MSE) sense. These values were found in several iterations until the MSE reached a minimum.

In summary, **Figure 3** gives a typical example of a simulated CNA, where the modified Bessel function-based approximation (depicted as MBA) demonstrates better accuracy than the approximation obtained using the skew Laplace distribution (depicted as SkL).

2.4. Probabilistic masks

It follows from **Figure 3** that, in view of large noise, estimates of the CNAs may have low confidence, especially with small SNR $\gamma \leq 1$. Thus, each estimate requires confidence boundaries within which it may exist with a given probability [20, 21].

Given an estimate $\hat{a}_l$ of the lth segmental level in white Gaussian noise, the probabilistic upper boundary (UB) and lower boundary (LB) can be specified for the given confidence probability $P(\vartheta)$ in the ϑ-sigma sense as [20]

$$\hat{a}_l^{UB} \cong \hat{a}_l + \varepsilon = \hat{a}_l + \vartheta\sqrt{\frac{\sigma_j^2}{N_l}} = \hat{a}_l + \vartheta\hat{\sigma}_l \tag{29}$$

$$\hat{a}_l^{LB} \cong \hat{a}_l - \varepsilon = \hat{a}_l + \vartheta\sqrt{\frac{\sigma_j^2}{N_l}} = \hat{a}_l + \vartheta\hat{\sigma}_l \tag{30}$$

where ϑ indicates the boundary wideness in terms of the segmental noise variance $\hat{\sigma}_l$ on an interval N_l points, from $\hat{n}_{l-1}$ to $\hat{n}_l - 1$.

Likewise, detected the lth breakpoint location $\hat{n}_l$, the jitter probabilistic left boundary J_l^L and right boundary J_l^R can be defined, following [20], as

$$J_l^L \cong \hat{n}_l - k_l^R, \tag{31}$$

$$J_l^R \cong \hat{n}_l + k_l^L, \tag{32}$$

where $k_l^R(\vartheta)$ and $k_l^L(\vartheta)$ are specified by the jitter distribution in the ϑ-sigma sense.

By combining Eqs. (30) and (31) with Eqs. (32) and (33), the probabilistic masks can be formed as shown in [20] to bound the CNA estimates in the ϑ-sigma sense for the given confidence probability $P(\vartheta)$. An important property of these masks is that they can be used not only to bound the estimates and show their possible locations on a probabilistic field [20, 21] but also to remove supposedly wrong breakpoints. Such situations occur each time when the masks reveal double UB and LB uniformities in a gap of three neighboring detected breakpoints. If so, then the unlikely existing intermediate breakpoint ought to be removed.

Noticing that the segmental boundaries (30) and (31) remain the same irrespective of the jitter in the breakpoints, below we specify the masks for the jitter represented with the Laplace distribution (17) and Bessel-based approximation (25).

2.4.1. Masks for Laplace distribution

For the Laplace distribution (17), the jitter left boundary J_l^L (32) and right boundary J_l^R(33) can be defined in the ϑ–-sigma sense if to specify $k_l^R(\vartheta)$ and $k_l^L(\vartheta)$ as shown in [18],

$$k_l^R = \left[\frac{\nu}{\kappa}\ln\frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)}\right], \tag{33}$$

$$k_l^L = \left[\nu\kappa\ln\frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)}\right], \tag{34}$$

where $[x]$ means a maximum integer lower than or equal to x. Note that functions (34) and (35) were obtained in [18] by equating (17) to $\xi(N_l)=\text{erfc}(\vartheta/\sqrt{2})$ and solving for k_l.

The probabilistic UB mask $\mathcal{L}_l^{UB}$ and LB mask $\mathcal{L}_l^{LB}$ for the Laplace distribution were formed in [17,20,21] by the segmental upper boundary $\hat{a}_l^{\text{UB}}$ and lower boundary $\hat{a}_l^{\text{LB}}$ and by the jitter left boundary $J_l{}^{\text{L}}$ and jitter right boundary $J_l{}^{\text{R}}$. The algorithm for computing $\mathcal{L}_l^{UB}$ and $\mathcal{L}_l^{LB}$ masks has been developed and applied to the CNA probes in [22].

2.4.2. Masks for Bessel-based approximation

The UB mask $\boldsymbol{\mathcal{B}}_l^{UB}$ and LB mask $\boldsymbol{\mathcal{B}}_l^{LB}$ for the Bessel-based approximation can be formed using the same equations as for the Laplace distribution. Suppose that the Laplace pdf (17) is equal to the approximating function $\mathbf{B}_l(k)$ at $k=0$,

$$p\left(k=0|d_l,q_l\right)=\mathbf{B}_l\left(k=0\right), \tag{35}$$

that yields $\mathbf{B}_l(k=0)=\frac{1}{\phi_l}$. Then, define the probabilities $P^{\boldsymbol{B}}(A_l)$ at $k=-1$ and $P^{\boldsymbol{B}}(B_l)$ at $k=1$ as

$$P^{\mathbf{B}}\left(A_l\right)=\frac{\boldsymbol{B}_l\left(k=0\right)}{\mathbf{B}_l\left(k=-1\right)+\mathbf{B}_l\left(k=0\right)}, \tag{36}$$

$$P^{\mathbf{B}}\left(B_l\right)=\frac{\boldsymbol{B}_l\left(k=0\right)}{\mathbf{B}_l\left(k=1\right)+\mathbf{B}_l\left(k=0\right)}. \tag{37}$$

Next, substitute Eqs. (37) and (38) into Eqs. (19) and (20) to obtain $\kappa_l{}^{\boldsymbol{B}}$ and $\nu_l{}^{\boldsymbol{B}}$. The right-hand jitter $k_l{}^{\boldsymbol{B}\text{R}}$ and left-hand jitter $k_l{}^{\boldsymbol{B}\text{L}}$ can now be specified by, respectively,

$$k_l^{\boldsymbol{B}\text{R}}=\left[\frac{\nu_l^{\boldsymbol{B}}}{\kappa_l^{\boldsymbol{B}}}\ln\frac{1}{\xi\boldsymbol{B}_l(k=0)}\right], \tag{38}$$

$$k_l^{\boldsymbol{B}\text{L}}=\left[\nu_l^{\boldsymbol{B}}\kappa_l^{\boldsymbol{B}}\ln\frac{1}{\xi\boldsymbol{B}_l(k=0)}\right]. \tag{39}$$

Finally, define the jitter left boundary $J_l{}^{\boldsymbol{B}\text{L}}$ and right boundary $J_l{}^{\boldsymbol{B}\text{R}}$ as, respectively,

$$J_l^{\boldsymbol{B}\text{L}}\cong\hat{n}_l-k_l^{\boldsymbol{B}\text{R}}, \tag{40}$$

$$J_l^{BR} \cong \hat{n}_l - k_l^{BL}, \tag{41}$$

and use the algorithm previously designed in the study of Munoz-Minjares and Shmaliy [22] for the confidence masks based on the Laplace distribution.

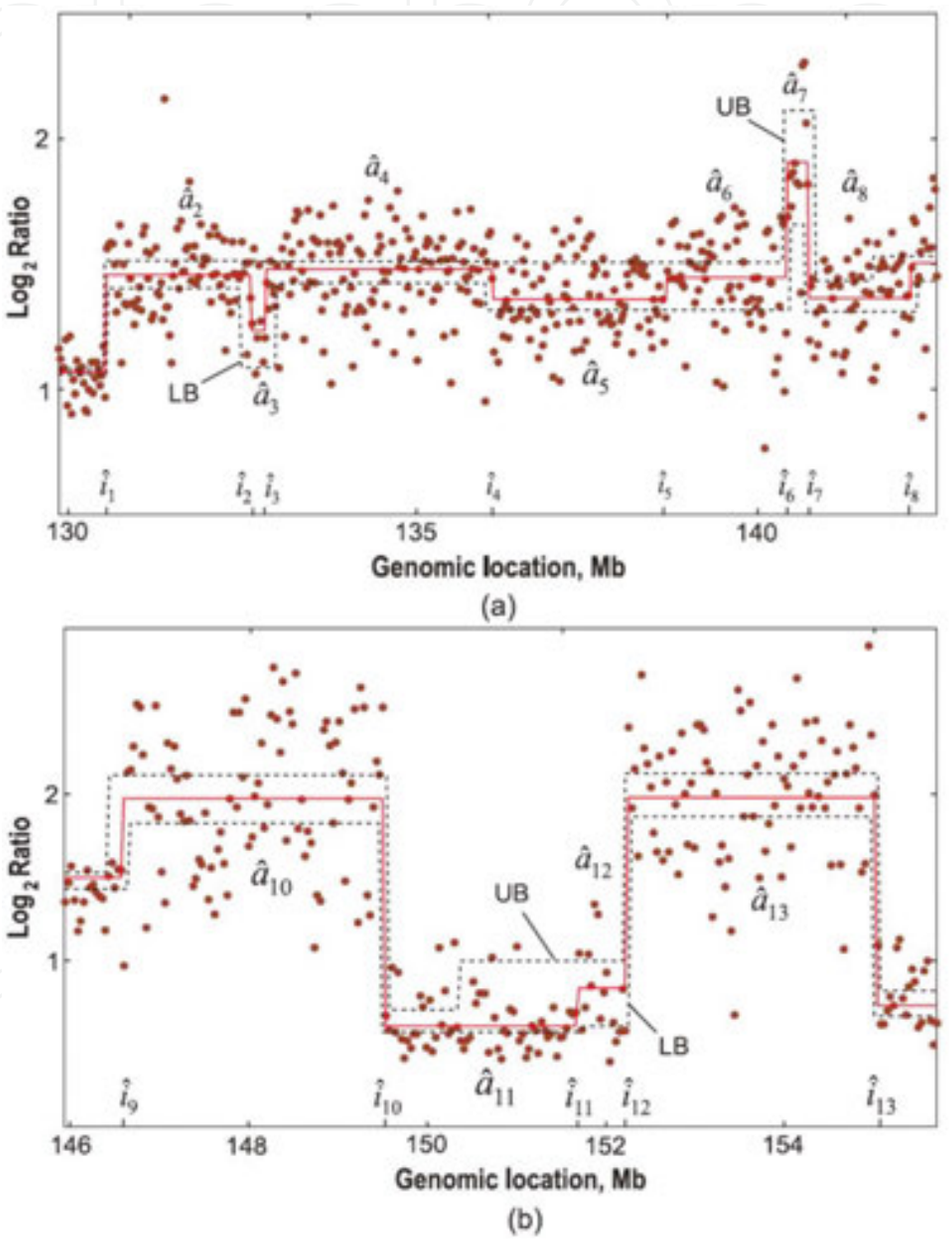

Figure 4. $\mathcal{L}_l^{UB}$ and $\mathcal{L}_l^{LB}$ masks for the seventh chromosome taken from "159A–vs–159D–cut" of ROMA: (a) genomic location from 130 to 146 Mb and (b) genomic location from 146 to 156 Mb. Breakpoints $\hat{\imath}_1$, $\hat{\imath}_6$, $\hat{\imath}_7$, $\hat{\imath}_9$, $\hat{\imath}_{10}$, $\hat{\imath}_{12}$, and $\hat{\imath}_{13}$ are well detectable because jitter is moderate. Owing to large jitter the breakpoints $\hat{\imath}_2$, $\hat{\imath}_3$, $\hat{\imath}_4$, $\hat{\imath}_5$, $\hat{\imath}_8$, $\hat{\imath}_9$, and $\hat{\imath}_{11}$ cannot be estimated correctly. There is a probability that the breakpoints $\hat{\imath}_2$, $\hat{\imath}_3$, $\hat{\imath}_4$, $\hat{\imath}_5$, and $\hat{\imath}_{11}$ do not exist. There is a high probability that breakpoint $\hat{\imath}_5$ does not exist.

http://dx.doi.org/10.5772/63913

3. Results

In this section, we test some CNA measurements and estimates by the algorithm developed in [22] based on the Laplace and Bessel approximations. In order to demonstrate the efficiency of the probabilistic masks and getting practically useful results, we exploit probes obtained by different technologies. First, we employ the results obtained with the HR-CGH profile and test them by the probabilistic masks using Laplace distribution. We next demonstrate the efficiency of the Bessel-based probabilistic masks versus the Laplace-based masks for the probes obtained with the SNP profile.

3.1. HR-CGH-based probing

The first test is conducted in the three-sigma sense suggesting that the CNAs exist between the UB and LB masks with high probability of $P=99.73\%$. The tested HR-CGH array data are available from the representational oligonucleotide microarray analysis (ROMA) [23]. The breakpoint locations are also given in [23]. Voluntarily, we select data associated with potentially large jitter and large segmental errors. For clarity, we first compute some characteristics of the detected CNAs and notice that the segmental estimates found by averaging [18] are in a good correspondence with [23]. The database processed is a part of the seventh chromosome in archive "159A–vs–159D–cut" of ROMA a sample of B-cell chronic lymphocytic leukemia (B-CLL). It is shown to have 14 segments and 13 breakpoints (**Figure 4a** and **b**). Below, we shall show that, owing to large detection noise, there is a high probability that some breakpoints do not exist.

It follows from **Figure 4a** that the only breakpoint which location can be estimated with high accuracy is i_1. Jitter in $\hat{i}_6$ and $\hat{i}_7$ is moderate. All other breakpoints have large jitter. It is seen that the UB mask covering second to sixth segments is almost uniform. Thus, there is a probability that the second to fifth breakpoints do not exist. If to follow the LB mask, the locations of the second to fourth breakpoints can be predicted even with large errors. At least they can be supposed to exist. However, nothing definitive can be said about the fifth breakpoint and one may suppose that it does not exist. It is also hard to distinguish a true location of the eighth breakpoint. In **Figure 4b**, i_{10}, i_{12}, and i_{13} are well detectable owing to large segmental SNRs. The breakpoint i_9 has a moderate jitter. In turn, the location of i_{11} is unclear. Moreover, there is a probability that i_{11} does not exist.

3.2. SNP-based probing

Our purpose now is to apply the probabilistic mask with SNP profile that represents the CNA with low levels of SNR. Specifically, we employ the probes of the first chromosome available from "BLC_B1_T45.txt" a sample of primary breast carcinoma.

Inherently, the more accurate Bessel-based approximation extends the jitter probabilistic boundaries with respect to the Laplace-based ones, especially for low SNRs. We illustrate it in **Figure 5**, where the estimates of the first chromosome were tested by $\mathcal{B}_l^{UB}$, $\mathcal{B}_l^{LB}$, $\mathcal{L}_l^{UB}$, and $\mathcal{L}_l^{UB}$ for $\vartheta=3$ (confidence probability $P=99.73\%$).

In **Figure 6**, the masks $\mathcal{B}_l^{UB}$ and $\mathcal{B}_l^{LB}$ are placed in the vicinity of segment $\hat{a}_{18}$ for several confidence probabilities: $\vartheta = 0.6745 (P = 50\%)$, $\vartheta = 1 (P = 68.27\%)$, $\vartheta = 2 (P = 95.45\%)$, and $\vartheta = 3 (P = 99.73\%)$. What the masks suggest here is that the CNA evidently exists with high probability, but the segmental levels and the breakpoint locations cannot be estimated with high accuracy, owing to low SNRs.

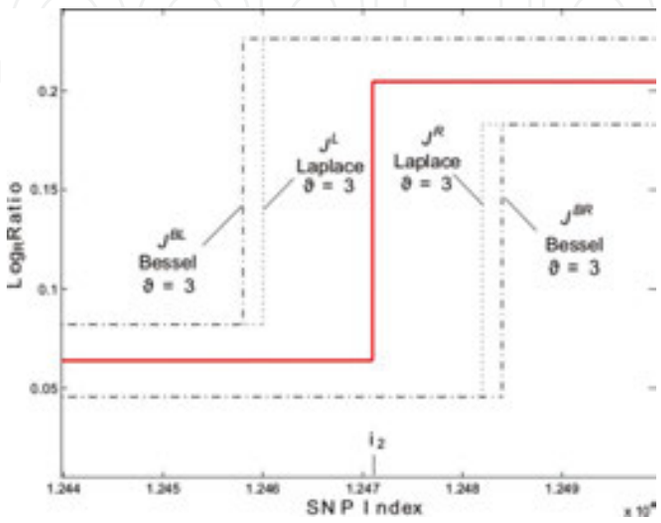

Figure 5. Jitter left boundaries $\mathcal{B}_l^{\mathcal{B}L}$, J_l^{L} and right boundaries $J_l^{\mathcal{B}R}$, J_l^{R} for the breakpoint $\hat{i}_2$ of first chromosome from sample BLC_B1_T45.txt (primary breast carcinoma). The probabilistic masks detect a breakpoint with a confidence probability $\vartheta = 3 (P = 99.73\%)$.

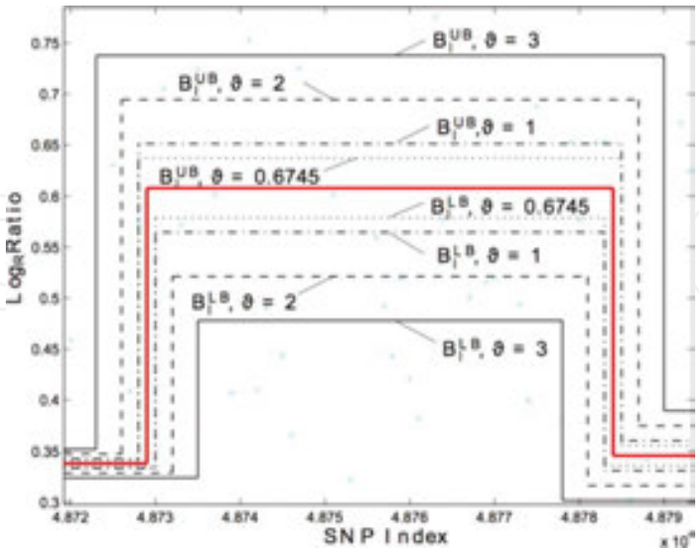

Figure 6. The $\mathcal{B}_l^{UB}$ and $\mathcal{B}_l^{LB}$ masks placed around the segmental level a_{18} for several confidence probabilities [20]. Here, the CNA exists with high probability, but the segmental levels and the breakpoint locations cannot estimate with high accuracy.

http://dx.doi.org/10.5772/63913

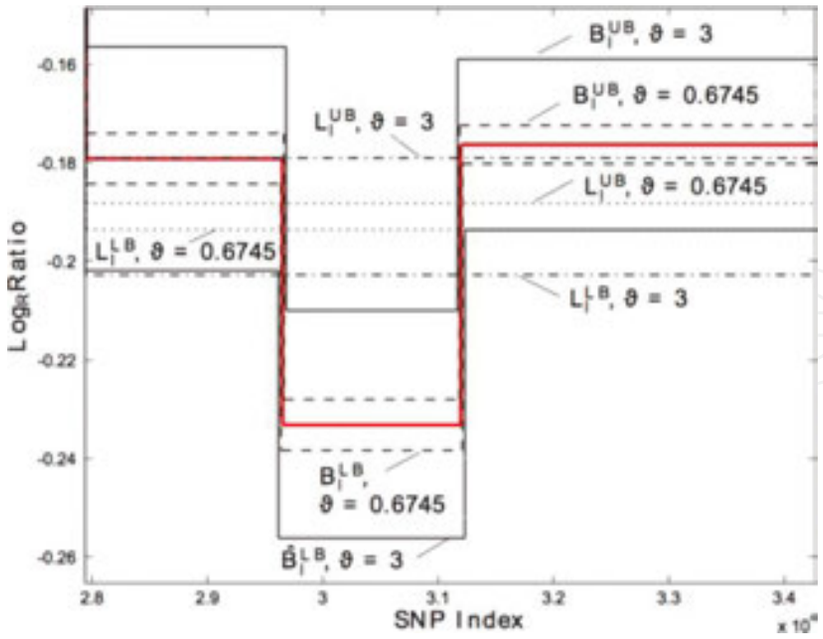

Figure 7. The confidence masks placed around a_{10} for ϑ=0.6745(P=50%) and ϑ=3(P=99.73%). Masks $\mathcal{L}_l^{UB}$ and $\mathcal{L}_l^{LB}$ do not confirm an existence of segmental changes while $\boldsymbol{\mathcal{B}}_l^{UB}$ and $\boldsymbol{\mathcal{B}}_l^{LB}$ indicate a small change.

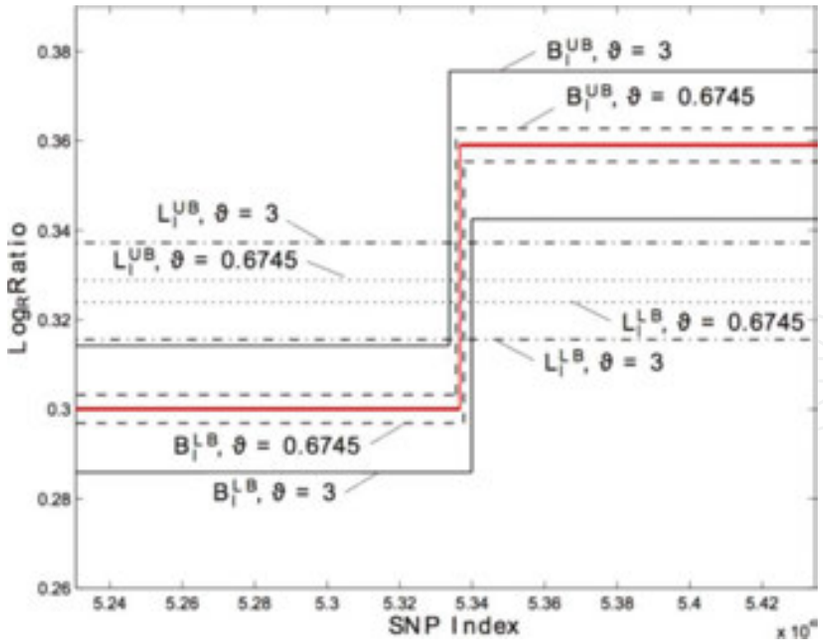

Figure 8. The confidence masks $\mathcal{L}_l^{UB}$, $\mathcal{L}_l^{\mathrm{LB}}$, $\mathcal{B}_l^{\mathrm{UB}}$ and $\boldsymbol{\mathcal{B}}_l^{LB}$. placed around the breakpoint $\hat{i}_{20}$ for confidence probabilities ϑ=0.6745 and ϑ=3 of first chromosome from sample BLC_B1_T45.txt. The confidence masks based on Laplace distribution cannot detect the breakpoint $\hat{i}_{20}$.

A special case can also be noticed when the masks $\mathcal{L}_l^{UB}$ and $\mathcal{L}_l^{LB}$ are not able to confirm or deny an existence of segmental changes with high probability, owing to the inability of computing the Laplace-based masks for extremely low SNRs. **Figures 7** and **8** illustrate such situations. Just on the contrary, masks $\boldsymbol{\mathcal{B}}_l^{UB}$ and $\boldsymbol{\mathcal{B}}_l^{LB}$ can be computed for any reasonable SNR.

A conclusion that can be made based on the results illustrated in **Figures 5–8** is that the Bessel-based probabilistic masks can be used to improve estimates of the chromosomal changes for the required probability.

We finally notice that the computation time required by the masks to process the first chromosome from sample "BLC B1 T45.txt" with a length of n = 905215 was 2.634599 s using MATLAB software on a personal computer with a processor Intel Core i5, 2.5 GHz.

4. Discussion

We evaluate the breakpoints obtained by the projects representational oligonucleotide microarray analysis [23] and GAP [14] with the confidence masks. As has been shown before, not all of the detected chromosomal changes have the same confidence to mean that there is a probability that some breakpoints do not exist. In order to improve the CNA estimates for the required confidence, the following process can be used:

1. Obtain estimates of the CNA using the standard CBS algorithm [24, 25] or any other algorithm.

2. Compute masks $\boldsymbol{\mathcal{B}}_l^{UB}$ and $\boldsymbol{\mathcal{B}}_l^{LB}$ for the given confidence probability P, % and bound the estimates.

3. If the masks reveal double uniformities, in UB and LB, in a GAP of any three neighboring breakpoints, then remove the intermediate breakpoint and estimate the segmental level between the survived breakpoints by simple averaging. The CNAs estimated in such a way will be valid for the given confidence P, %.

Application of this methodology to the CNA structure detected in frames of the Project GAP is shown in **Figure 9**. Its special feature is a number of hardly recognized small chromosomal changes (**Figure 9a**). We test them by the proposed masks $\mathcal{B}_l^{\mathrm{UB}}$ and $\mathcal{B}_l^{\mathrm{LB}}$. To this end, we first start with equal confidence probabilities of P =50% for each estimate to exist or not exist and find out that three breakpoints demonstrate no detectability. We remove these breakpoints and depict their locations with "×". Reasoning similarly, we remove four breakpoints to retain only probable changes, by P =75%, nine breakpoints to show a picture combined with almost certain changes, by P =93%, and 10 breakpoints in the three-sigma sense, P =99.73%. Observing the results, we infer that the masks are able to correct only the estimates obtained under the low SNRs. The relevant chromosomal sections S1–S7 are circled in **Figure 9**. It is not surprising because changes existing with high SNRs are seen visually. An estimator thus can easily detect them with high confidence.

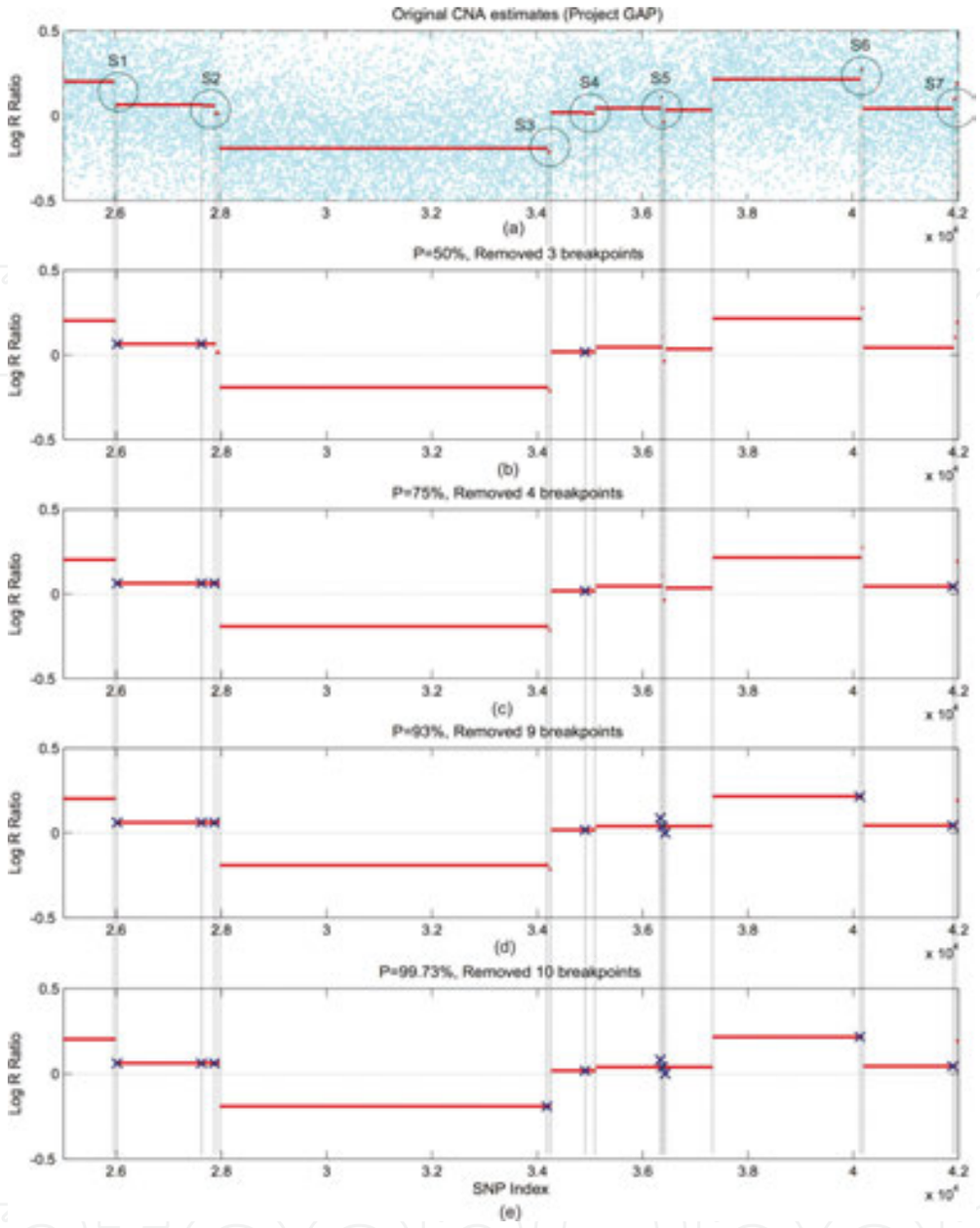

Figure 9. Improving estimates of the CNAs obtained in Project GAP [25] by removing some unlikely existing breakpoints: (a) original estimates, (b) even changes, $P=50\%$, (c) probable changes, $P=75\%$, (d) almost certain changes, $P=93\%$, and (e) three-sigma sense, $P=99.73\%$.

5. Conclusions

Modern technologies developed to produce the CNA profiles with high resolution still admit intensive white Gaussian noise. Accordingly, not one estimator even ideal is able to provide jitter-free estimation of segmental changes. Thus, in order to avoid wrong decisions, the estimates must be bounded for the confidence probability. Jitter exists in the CNA's break-

points fundamentally. When SNR is >1, it can statistically be described using the discrete skew Laplace distribution. Otherwise, if SNR is <1, the Bessel-based approximation produces more accuracy. By the jitter distribution, it is easy to find a region within which the breakpoint exists for the required probability. Of practical importance are the confidence UB and LB masks, which can be created based on the segmental and jitter distributions for the given confidence probability. The masks can serve as an auxiliary tool for medical experts to make decisions about the CNA structures. Applications to probes obtained using the HR-CGH and SNP technologies confirm efficiency of the confidence masks.

Author details

Jorge Muñoz-Minjares*, Yuriy Shmaliy and Oscar Ibarra-Manzano

*Address all correspondence to: ju.munozminjares@ugto.mx

Department of Electronics Engineering, University of Guanajuato, Salamanca, Mexico.

References

[1] Reymond, A., Henrichsen, C. N., Harewood, L., and Merla, G. Side effects of genome structural changes. Curr Opin Genet Dev. 2007;17(5):381–386.

[2] Alkan, C., Coe, B. P., and Eichler, E. E. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363–376.

[3] Feuk, L., Carson, A. R., and Scherer, S. W. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85–97.

[4] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., *et al*. Global variation in copy number in the human genome. Nature. 2006;444(7118):444–454.

[5] Hastings, P. J., Lupski, J. R., Rosenberg, S. M. and Ira, G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009; 10 (8): 551–564.

[6] Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O. *et al*. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289): 704–712.

[7] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011): 931–945.

[8] Forozan, F., Karhu, R., Kononen, J., Kallioniemi, A., and Kallioniemi, O. P. Genome screening by comparative genomic hybridization. Trends in Genetics. 1997; 13 (10): 405– 409.

[9] Speicher, M. R. and Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. Nat Rev Genet.2015;6(10): 782–792.

[10] Ng, P. C. and Kirkness, E. F. Whole genome sequencing. Methods MolBiol. 2010;628: 215–226.

[11] Wang, D. G., Fan, J. B., Siao, C. J., Bermo, A., Young, P., *et al*. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998;280 (5366): 1077–1082.

[12] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30(4): 1–10.

[13] Urban, A. E., Korbel, J. O., Selzer, R., Richmond, T., Hacker, A., Popescu, G. V., Cubells, J. F., Green, R., Emanuel, B. S., Gerstein, M.B., Weissman, S. M., Snyder, M. High-resolution mapping of DNA copy alterations inhuman chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci U S A. 2006;103(12): 4534–4539.

[14] Popova, T., Boeva, V., Manie, E., Rozenholc, Y., Barillot, E., and Stern, M.H. Analysis of Somatic Alterations in Cancer Genome: From SNP Arrays to Next Generation Sequencing. Genomics I Humans, Animals and Plants, 2013. <hal-01108425> Observation to publisher: Reference from https://hal.archives-ouvertes.fr/hal-01108425

[15] Pique-Regi, R., Ortega, A., Tewfik, A., and Asgharzadeh, S. Detection changes in the DNA copy number. IEEE Signal Process Mag. 2012;29(1): 98–107.

[16] Munoz-Minjares, J., Cabal-Aragón, J., and Shmaliy, Y.S. Jitter probability in the breakpoints of discrete sparse piecewise-constant signals. Proc. 21st Eur. Signal Process. Conf. (EUSIPCO-2013), Marrakech, Marocco. 2013; pp. 1–5.

[17] Munoz-Minjares, J., Shmaliy, Y. S., and Cabal, A. J. Noise studies in measurements and estimates of stepwise changes in genome DNA chromosomal structures. Adv. App Pure Math. 2014.

[18] Munoz-Minjares, J., Cabal-Aragón, J., and Shmaliy, Y. S. Probabilistic bounds for estimates of genome DNA copy number variations using HR-CGH microarray. Proc. 21st European Signal Process. Conf. (EUSIPCO-2013). 2013; pp. 1–5.

[19] Kozubowski, T. J. and Inusah, S. A skew Laplace distribution on integers. Ann Inst Stat Math 2006;58(3):555–571.

[20] Munoz-Minjares, J., Cabal-Aragon, J., and Shmaliy Y. S. Confidence masks for genome DNA copy number variations in applications to HR–CGH array measurements. Biomed Signal Process Contr. 2014;13:337–344.

[21] Munoz-Minjares, J., Shmaliy, Y. S., and Cabal-Aragon, J. Confidence limits for genome DNA copy number variations in HR–CGH array measurements. Biomed Signal Process Contr. 2014;10:166–173.

[22] Munoz-Minjares, J. and Shmaliy, Y. S.: Bounding error in estimates of genome copy number variations using SNP array. International Journal of Biomedical Engineering. 2015;9:127–132.

[23] Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C. Q., Powers, S., Ye, K. Q., Olshen, A., Venkatraman, E., Norton, L., Wigler, M. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res. 2003;13(10): 2291–2305.

[24] Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. Circular binary segmentation for the analysis of arraybased DNA copy number data. Biostatistics. 2004;5(4): 557–572.

[25] Venkatraman E. S. and Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23(6):657–663.