Noise Studies in Measurements and Estimates of Stepwise Changes in Genome DNA Chromosomal Structures

JORGE MUNOZ-MINJARES, YURIY S. SHMALIY, JESUS CABAL-ARAGON
Universidad de Guanajuato
Department of Electronics Engineering
Ctra. Salamanca-Valle, 3.5+1.8km, 36885, Salamanca
MEXICO
j.ulises_minjares@live.com, shmaliy@ugto.mx

Abstract: Measurements using the high resolution array-comparative genomic hybridization (HR-CGH) array are accompanied with large noise which strongly affects the estimates of the copy number variations (CNVs) and results in segmental errors as well as in jitter in the breakpoints. Based on the probabilistic analysis and algorithm designed, we show that jitter in the breakpoints can be well approximated with the discrete skew Laplace distribution if the local signal-to-noise ratios (SNRs) exceed unity. Using this distribution, we propose an algorithm for computing the estimate upper and lower bounds. Some measurements and estimates tested using these bounds show that the higher probe resolution is provided the more segmental accuracy can be achieved and that larger segmental SNRs cause smaller jitter in the breakpoints. Estimates of the CNVs combined with the bounds proposed may play a crucial role for medical experts to make decisions about true chromosomal changes and even their existence.

Key–Words: Genome copy number, estimate, jitter, breakpoint, error bound

1 Introduction

The deoxyribonucleic acid (DNA) of a genome essential for human life often demonstrates structural changes called copy-number variations (CNVs) associated with disease such as cancer [1]. The sell with the DNA typically has a number of copies of one or more sections of the DNA that results in the structural chromosomal rearrangements - deletions, duplications, inversions and translocations of certain parts [2]. Small such CNVs are present in many forms in the human genome, including single-nucleotide polymorphisms, small insertion-deletion polymorphisms, variable numbers of repetitive sequences, and genomic structural alterations [3]. If genomic aberrations involve large CNVs, the process was shown to be directly coupled with cancer and the relevant structural changes were called copy-number alterations (CNAs) [4]. A brief survey of types of chromosome alterations involving copy number changes is given in [5]. The copy number represents the number of DNA molecules in a cell and can be defined as the number of times a given segment of DNA is present in a cell. Because the DNA is usually double-stranded, the size of a gene or chromosome is often measured in base pairs. A commonly accepted unit of measurement in molecular biology is kilobase (kb) equal to 1000 base pairs of DNA [6]. The human genome with 23 chromosomes is estimated to be about 3.2 billion base pairs long and to contain $20000 - 25000$ distinct genes [1]. Each CNV may range from about one kb to several megabases (Mbs) in size [2].

One of the techniques employing chromosomal microarray analysis to detect the CNVs at a resolution level of 5–10 kbs is the array-comparative genomic hybridization (aCGH) [7]. It was reported in [8] that the high-resolution CGH (HR-CGH) arrays are accurate to detect structural variations (SV) at resolution of 200 bp. In microarray technique, the CNVs are often normalized and plotted as $\log_2 R/G = \log_2 \mathrm{Ratio}$, where R and G are the fluorescent Red and Green intensities, respectively [9]. An annoying feature of such measurements is that the Ratio is highly contaminated by noise which intensity does not always allow for correct visual identification of the breakpoints and copy numbers and makes most of the estimation techniques poor efficient if the number of segmental readings is small. It was shown in [10] that sufficient quality in the CNVs mapping can be achieved with tens of millions of paired reads of 29–36 bases at each. Deletions as small as 300 bp should also be detected in some cases. For instance, arrays with a 9-bp tiling path were used in [8] to map a 622-bp heterozygous deletion. So, further progress in the probe resolution

ISBN: 978-1-61804-240-8

of the CNVs measurements is desirable.

Typically, a chromosome section is observed with some average resolution $\bar{r}$, bp and M readings in the genomic location scale. The following distinct properties of the CNVs function were recognized [2, 5]:

1) It is piecewise constant (PWC) and sparse with a small number L of the *breakpoints* (edges) i_l, $l \in [1, L]$, on a long base-pair length. The breakpoints are places as $0 < i_1 < \cdots < i_L < \bar{r}M$ and can be united in a vector

$$\mathcal{I} = [i_1\, i_2 \ldots i_L]^T \in \mathcal{R}^L\,. \tag{1}$$

Sometimes, the genomic location scale is represented in the number of readings $n \in [1, M]$ with a unit step ignoring "bad" or empty measurements, where n represents the nth reading. In such a scale, the n_lth discrete point corresponds to the i_lth breakpoint in the genomic location scale and the points placed as $0 < n_1 < \cdots < n_L < M$ can be united in a vector

$$\mathcal{N} = [n_1\, n_2 \ldots n_L]^T \in \mathcal{R}^L\,. \tag{2}$$

An advantage of $\mathcal{N}$ against $\mathcal{I}$ is that it facilitates the algorithm design. However, the final estimates are commonly represented in the genomic location scale.

2) Its *segments* with constant copy numbers a_j, $j \in [1, L+1]$, are integer, although this property is not survived in the $\log_2 \mathrm{Ratio}$. The segmental constant changes can also be united in a vector

$$\mathbf{a} = [a_1\, a_2 \ldots a_{L+1}]^T \in \mathcal{R}^{L+1}\,, \tag{3}$$

in which a_j characterizes a segment between i_{j-1} and i_j on an interval $[i_{j-1}, i_j - 1]$.

3) The measurement noise in the $\log_2$ Ratio is highly intensive and can be modeled as additive white Gaussian.

The estimation theory offers several useful approaches for piecewise signals such as those generated by the chromosomal changes. One can employ the *wavelet*-based [11, 12] filters, *robust* estimators [12], adaptive *kernel smoothers* [13, 14], *maximum likelihood* (ML) based on Gauss's *ordinary least squares* (OLS), penalized *bridge* estimator [15] and *ridge* regression [16] (also known as Tikhonov regularization), fussed least-absolute shrinkage and selection operator (*Lasso*) [17], the *Schwarz information criterion*-based estimator [18, 19], and *forward-backward smoothers* [20–22].

We also find a number of solutions developed especially for needs of bioinformatics. Efficient algorithms for filtering, smoothing and detection were proposed in [11,12,19,23–28]. Methods for segmentation and modeling were developed in [10, 18, 24, 29–32].

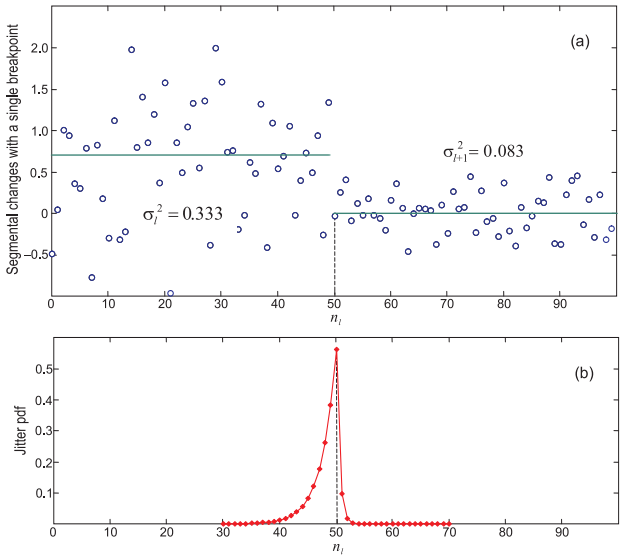

Figure 1: Simulated genome segmental changes with a single breakpoint at $n_l = 50$ and segmental variances $\sigma_l^2 = 0.333$ and $\sigma_{l+1}^2 = 0.083$ corresponding to segmental SNRs $\gamma_l = 1.47$ and $\gamma_{l+1} = 5.88$: (a) measurement and (b) jitter pdf. The jitter pdf was found by applying a ML estimator via a histogram over 10^4 runs.

Sparse representation based on penalized optimization and Bayesian learning were provided in [33–38]. These results show that a small number of readings N_j per a segment a_j in line with large measurement noise remain the main limiters of accuracy in the estimation of CNVs. Picard *et al.* have shown experimentally in [29] that each segmental estimate is accompanied with *errors* and each breakpoint has *jitter* which cannot be overcome by any estimator.

For clarity, we generalize an experiment conducted in [29] in Fig. 1. Here, a chromosomal part having two constant segments $a_l = 0.7$ and $a_{l+1} = 0$ and a breakpoint $n_l = 50$ is simulated in the presence of discrete white Gaussian noise having segmental variances $\sigma_l^2 = 0.333$ and $\sigma_{l+1}^2 = 0.083$ (Fig. 1a). For the local segmental signal-to-noise ratios (SNRs)

$$\gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}\,, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2}\,, \tag{4}$$

where $\Delta_l = a_{l+1} - a_l$ is a local segmental change, it corresponds to $\gamma_l^- = 1.47$ and $\gamma_l^+ = 5.88$.

The breakpoint location n_l was detected in Fig. 1 using a ML estimator [22] (one can employ any other estimator). Measurements and estimations were repeated 10^4 times with different realization of noise. Then the histogram was plotted for the detected breakpoint locations and normalized to have a unit area. The jitter probability density function (pdf) obtained in such a way is sketched in Fig. 1b. Even a quick

ISBN: 978-1-61804-240-8

look at this figure assures that jitter at a level of 0.01 (jitter probability of 1%) has 10 points to the left (left jitter) and 2 points to the right (right jitter). In other words, with the probability of 99%, the breakpoint n_l can be found at any point between $n = 40$ and $n = 52$ that may be too rough for medical conclusions, especially if $\bar{r}$ is large. Let us add that simple averaging which is optimal for the estimation of PWC changes between the breakpoints is able to reduce the noise variance by the factor of N_l. Noise reduction may thus also be insufficient for medical applications if N_l is small. So, effect of noise needs more investigations and the CNVs estimate bounds are required.

2 Jitter in the Breakpoints

In follows from the experiment conducted in [29] and supported by Fig. 1 that jitter in the breakpoints plays a critical role in the estimation of the CNVs. Large jitter may cause wrong conclusions about the breakpoint locations. On the other hand, it may cause extra errors in the determination of segmental changes especially if N_l and segmental SNRs occur to be small.

2.1 Laplace-Based Approximation

The results published in [29] and our own investigations provided in [39] and generalized in Fig. 1b show that jitter in the breakpoints has approximately the skew Laplace distribution. The discrete skew Laplace distribution was recently derived in [40],

$$p(k|d_l, q_l) = \frac{(1-d_l)(1-q_l)}{1-d_l q_l} \begin{cases} d_l^k\,, & k \geqslant 0\,, \\ q_l^{|k|}\,, & k \leqslant 0\,, \end{cases} \tag{5}$$

where $d_l = e^{-\frac{\kappa_l}{\nu_l}} \in (0,1)$ and $q_l = e^{-\frac{1}{\kappa_l \nu_l}} \in (0,1)$ and in which $\kappa_l > 0$ and $\nu_l > 0$ are coefficients defined by the process. Below, we shall show that (5) can serve as a reasonably good approximation for jitter in the breakpoints of PWC signals such as that shown in Fig. 1a if the segmental SNRs exceed unity.

Let us consider N neighboring to n_l readings in each segment. We may assign an event A_{lj} meaning that all measurements at points $n_l - N \leqslant j < n_l$ belong to lth segment. Another event B_{lj} means that all measurements at $n_l \leqslant j < n_l + N - 1$ belong to $(l+1)$th segment. We think that a measured value belongs to one segment if the probability is larger than if it belongs to another segment. Because noise is Gaussian and the segmental variances are different, the Gaussian pdfs cross each other in two points, α_l and β_l. The events A_{lj} and B_{lj} can thus be specified as follows:

$$A_{lj} \quad \text{is} \quad \begin{cases} (\alpha_l < x_j) \wedge (x_j < \beta_l)\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ x_j > \alpha_l\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ \alpha_l < x_j < \beta_l\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{6}$$

$$B_{lj} \quad \text{is} \quad \begin{cases} \beta_l < x_j < \alpha_l\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \\ x_j < \alpha_l\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ (x_j < \alpha_l) \wedge (x_j > \beta_l)\,, & \sigma_l^2 > \sigma_{l+1}^2\,. \end{cases} \tag{7}$$

The inverse events meaning that at least one of the points do not belong to the relevant interval are $\bar{A}_{lj} = 1 - A_{lj}$ and $\bar{B}_{lj} = 1 - B_{lj}$.

Both A_{lj} and B_{lj} can be united into two blocks

$$\begin{aligned} \mathbf{A}_l &= \{A_{l(i_l-N)} A_{l(i_l-N+1)} \dots A_{l(i_l-1)}\}\,, \\ \mathbf{B}_l &= \{B_{l(i_l)} B_{l(i_l+1)} \dots B_{l(i_l+N-1)}\}\,. \end{aligned}$$

We think that if $\mathbf{A}_l$ and $\mathbf{B}_l$ occur simultaneously then the breakpoint n_l will be jitter-free. However, there may be found some other events which do not obligatorily lead to jitter. We ignore such events and define approximately the probability $P(\mathbf{A}_l\mathbf{B}_l)$ of the jitter-free breakpoint as

$$P(\mathbf{A}_l\mathbf{B}_l) = P(A_{i_l-N} \dots A_{i_l-1} B_{i_l} \dots B_{i_l+N-1})\,. \tag{8}$$

The inverse event $\bar{P}(\mathbf{A}_l\mathbf{B}_l) = 1 - P(\mathbf{A}_l\mathbf{B}_l)$ meaning that at least one point belongs to another event can be called the *jitter probability*.

In white Gaussian noise, all the events are independent and (8) thus can be rewritten as

$$P(\mathbf{A}_l\mathbf{B}_l) = P^N(A_l) P^N(B_l)\,, \tag{9}$$

where, following (6) and (7), the probabilities $P(A_l)$ and $P(B_l)$ can be specified as, respectively,

$$P(A_l) \quad = \quad \begin{cases} 1 - \int\limits_{\beta_l}^{\alpha_l} p_l(x)dx\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ \int\limits_{\alpha_l}^{\infty} p_l(x)dx\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ \int\limits_{\alpha_l}^{\beta_l} p_l(x)dx\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{10}$$

$$P(B_l) \quad = \quad \begin{cases} \int\limits_{\beta_l}^{\alpha_l} p_{l+1}(x)dx\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ \int\limits_{-\infty}^{\alpha_l} p_{l+1}(x)dx\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ 1 - \int\limits_{\alpha_l}^{\beta_l} p_{l+1}(x)dx\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{11}$$

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

Let us now think that jitter occurs at some point $n_l \pm k$, $0 \leqslant k \leqslant N$, and assign two additional blocks of events

$$\begin{aligned} \mathbf{A}_{lk} &= \{A_{i_l-N} \dots A_{i_l-1-k}\}, \\ \mathbf{B}_{lk} &= \{B_{i_l+k} \dots B_{i_l+N-1}\}. \end{aligned}$$

The probability $P_k^- \triangleq P_k^-(\mathbf{A}_{lk}\bar{A}_{l(i_l-k)} \dots \bar{A}_{i_l-1}\mathbf{B}_l)$ that jitter occurs at kth point to the left from n_l (left jitter) and the probability $P_k^+ \triangleq P_k^+(\mathbf{A}_l\bar{B}_{l(i_l+1)} \dots \bar{B}_{l(i_l+k-1)}\mathbf{B}_{lk})$ that jitter occurs at kth point to the right from n_l (right jitter) can thus be written as, respectively,

$$P_k^- = P^{N-k}(A_l)[1-P(A_l)]^k P^N(B_l), \quad (12)$$

$$P_k^+ = P^N(A_l)[1-P(B_l)]^k P^{N-k}(B_l). \quad (13)$$

By normalizing (12) and (13) with (9), we arrive at a function that turns out to be independent on N:

$$f_l(k) = \begin{cases} [P^{-1}(A_l)-1]^{|k|} & , \quad k<0, \quad \text{(left)} \\ 1 & , \quad k=0, \\ [P^{-1}(B_l)-1]^k & , \quad k>0. \quad \text{(right)} \end{cases} \quad (14)$$

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

$$\phi_l = 1 + \sum_{k=1}^{\infty} [\varphi_l^A(k) + \varphi_l^B(k)], \quad (15)$$

where $\varphi_l^A(k) = [P^{-1}(A_l)-1]^k$ and $\varphi_l^B(k) = [P^{-1}(B_l)-1]^k$. Now observe that, in the approximation accepted, $f_l(k)$ converges with k only if $0.5 < \tilde{P} = \{P(A), P(B)\} < 1$. Otherwise, if $\tilde{P} < 0.5$, the sum ϕ_l is infinite, $f_l(k)$ cannot be transformed to $p_l(k)$, and the lth breakpoint cannot be detected. Considering the case of $0.5 < \tilde{P} = \{P(A), P(B)\} < 1$, we conclude that $\ln\tilde{P} < 0$, $\ln(1-\tilde{P}) < 0$, and $\ln(1-\tilde{P}) < \ln\tilde{P}$. Next, using a standard relation $\sum\limits_{k=1}^{\infty} x^k = \frac{1}{x^{-1}-1}$, where $x < 1$, and after little transformations we bring (15) to

$$\phi_l = \frac{P(A_l) + P(B_l) - 1}{[1-2P(A_l)][1-2P(B_l)]}. \quad (16)$$

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

$$p_l(k) = \frac{1}{\phi_l} \begin{cases} [P^{-1}(A_l)-1]^{|k|} & , \quad k<0, \\ 1 & , \quad k=0, \\ [P^{-1}(B_l)-1]^k & , \quad k>0, \end{cases} \quad (17)$$

where ϕ_l is specified by (16) and $0.5 < \{P(A_l), P(B_l)\} < 1$. By substituting $q_l = P^{-1}(A_l) - 1$ and $d_l = P^{-1}(B_l) - 1$, we find $P(A_l) = 1/(1+q_l)$ and $P(B_l) = 1/(1+d_l)$, provide the transformations, and finally go from (17) to the discrete skew Laplace distribution (5) in which κ_l and ν_l still need to be connected to (17). To find κ_l and ν_l, below we consider three points $k=-1$, $k=0$, and $k=1$. By equating (5) and (17), we first obtain $\frac{(1-d_l)(1-q_l)d_l}{1-d_lq_l} = \frac{1}{\phi_l}\frac{1-P(B_l)}{P(B_l)}$ for $k=1$ and $\frac{(1-d_l)(1-q_l)q_l}{1-d_lq_l} = \frac{1}{\phi_l}\frac{1-P(A_l)}{P(A_l)}$ for $k=-1$ that gives us $\nu_l = \frac{1-\kappa_l^2}{\kappa_l \ln\mu_l}$, where

$$\mu_l = \frac{P(A_l)[1-P(B_l)]}{P(B_l)[1-P(A_l)]}. \quad (18)$$

For $k=0$, we have $\frac{(1-d_l)(1-q_l)}{1-d_lq_l} = \frac{1}{\phi_l}$ and transform it to an equation $x_l^2 - \frac{\phi_l(1+\mu_l)}{1+\phi_l}x - \frac{1-\phi_l}{1+\phi_l}\mu_l = 0$, which proper solution is

$$x_l = \frac{\phi_l(1+\mu_l)}{2(1+\phi_l)}\left(1 - \sqrt{1 + \frac{4\mu_l(1-\phi_l^2)}{\phi_l^2(1+\mu_l)^2}}\right) \quad (19)$$

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

$$\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}. \quad (20)$$

By combining ν_l with (19), we also provide a simpler form for ν_l, namely

$$\nu_l = -\frac{\kappa_l}{\ln x_l}. \quad (21)$$

The discrete skew Laplace distribution (5) can thus be used to represent jitter in the breakpoints statistically.

Now substitute the Gaussian pdf to (10) and (11), provide the transformations, and find

$$P(A_l) = \begin{cases} 1 + \frac{1}{2}[\mathrm{erf}(g_l^\beta) - \mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- < \gamma_l^+, \\ \frac{1}{2}\mathrm{erfc}(g_1^\alpha) & , \quad \gamma_l^- = \gamma_l^+, \\ \frac{1}{2}[\mathrm{erf}(g_l^\beta) - \mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- > \gamma_l^+, \end{cases} \quad (22)$$

$$P(B_l) = \begin{cases} \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- < \gamma_l^+, \\ 1 - \frac{1}{2}\mathrm{erfc}(h_l^\alpha) & , \quad \gamma_l^- = \gamma_l^+, \\ 1 + \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- > \gamma_l^+, \end{cases} \quad (23)$$

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

$\gamma_l^- \neq \gamma_l^+$, the coefficients α_l and β_l are defined by

$$\alpha_l, \beta_l = \frac{a_l\gamma_l^- - a_{l+1}\gamma_l^+}{\Gamma_l} \mp \frac{|\Delta_l|}{\Gamma_l}\sqrt{\gamma_l^-\gamma_l^+ + 2\Gamma_l \ln\sqrt{\frac{\gamma_l^-}{\gamma_l^+}}} \tag{24}$$

where $\Gamma_l = \gamma_l^- - \gamma_l^+$. For $\gamma_l^- = \gamma_l^+$, set $\alpha_l = \Delta_l/2$ and $\beta_l = \pm\infty$. Using (22) and (23), below we investigate errors inherent to the Laplace-based approximation.

2.2 Errors in Laplace-Based Approximation

To realize how well the discrete skew Laplace distribution (5) fits real jitter distribution with different SNRs, we consider a measurement of length M with one breakpoint at $n = K$ and two neighboring segments with known changes a_l and a_{l-1}. The segmental variances σ_l^2 and σ_{l-1}^2 of white Gaussian noise are supposed to be known. In the ML estimator, the mean square error (MSE) is minimized between the measurement and the CNVs model in which the breakpoint location is handled around an actual value. Thereby, the breakpoint location is detected when the MSE reaches a minimum. In our experiments, measurements were conducted 10^4 times for different noise realizations and the histogram of the estimated breakpoint locations was plotted. Such a procedure was repeated several times and the estimates were averaged in order to avoid ripples. Normalized to have a unit area, the histogram was accepted as the jitter pdf. The relevant algorithm can easily be designed to have as inputs a_l, a_{l-1}, segmental SNRs γ_l^- and γ_l^+, M, K, and the number of point K_1 around K covering possible breakpoint locations. The algorithm output is the jitter histogram "Jitter". An analysis was provided for typical SNR values peculiar to the CNVs measurements using the HR-CGH arrays. As a result, we came up with the following conclusions:

1) The Laplace approximation is reasonably accurate in the lower bound sense if the SNRs exceed unity, $(\gamma_l^-, \gamma_l^+) > 1$. Figure 2 sketches the Laplace pdf and the experimentally found pdf (circled) for the case of $\gamma_l^- = 1.4$ and $\gamma_l^+ = 1.38$ taken from real measurements. Related to the unit change, the approximation error was computed as $\varepsilon, \% = (\text{ML estimate} - \text{Laplace approximation})\times 100$. As can be seen, $\varepsilon_{\max}$ reaches here about 10% at $n = K$ (Fig. 2b). That means that the Laplace distribution fits measurements well for the allowed probability of jitter-free detection of 90%. It narrows the jitter bounds with about ± 2 points for 99%. Observing another example illustrated in Fig. 3 for $\gamma_l^- = 9.25625$ and $\gamma_l^+ = 2.61186$,

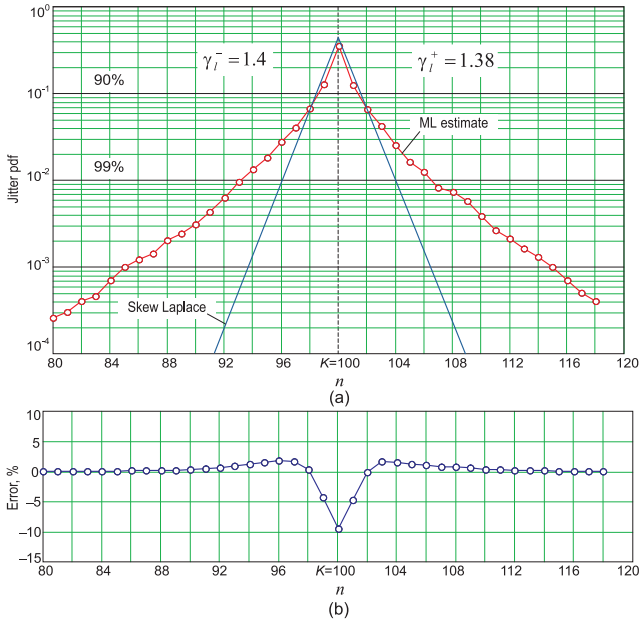

Figure 2: The jitter pdf approximated using the discrete skew Laplace distribution and found experimentally (circled) using a ML estimator over 10^4 runs for $\gamma_l^- = 1.4$ and $\gamma_l^+ = 1.38$: (a) pdfs and (b) approximation errors.

we infer that the Laplace distribution fits the process with very high accuracy if SNR $\gg 1$.

2) The approximation error may be large in the sense of the narrowed jitter bounds if SNR < 1.

3) The jitter bounds commonly cannot be determined correctly for $(\gamma_l^-, \gamma_l^+) \ll 1$.

3 Estimate Bounds

The upper bound (UB) and lower bound (LB) peculiar to the estimate confidential interval can now be found implying segmental white Gaussian noise and accepting the discrete skew Laplace-based jitter distribution in the breakpoints.

Segmental Errors. In white Gaussian noise environment, simple averaging is most efficient between the breakpoints as being optimal in the sense of the minimum produced noise. Provided the estimate $\hat{n}_l$ of the breakpoint location n_l, simple averaging applied on an interval of $N_j = n_j - n_{j-1}$ readings from n_{j-1} to $n_j - 1$ gives the following estimate for the lth segmental change

$$\hat{a}_j = \frac{1}{N_j}\sum_{v=n_{j-1}}^{n_j-1} y_v\,, \tag{25}$$

which mean value is $E\{\hat{a}_j\} = a_j$ and variance is

$$\hat{\sigma}_j^2 = \frac{\sigma_j^2}{N_j}\,. \tag{26}$$

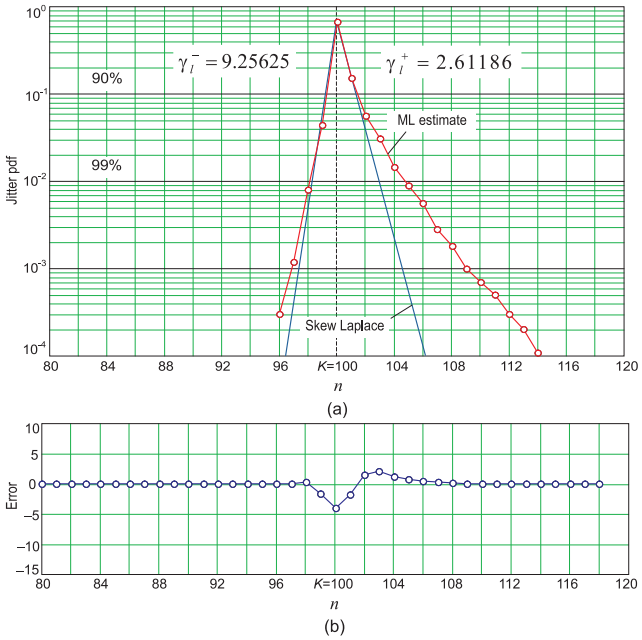

Figure 3: The jitter pdf approximated using the discrete skew Laplace distribution and found experimentally (circled) using a ML estimator over 10^4 runs for $\gamma_l^- = 9.25625$ and $\gamma_l^+ = 2.61186$: (a) pdfs and (b) approximation errors.

The UB for segmental estimates can be formed in the θ-sigma sense as $\hat{a}_j^{\rm UB} = E\{\hat{a}_j\} + \theta\sqrt{\frac{\sigma_j^2}{N_j}}$, where $\theta \geqslant 1$ is commonly integer. However, neither an actual $a_j = E\{\hat{a}_j\}$ nor multiple measurements necessary to approach a_j by averaging are available. We thus specify UB and LB approximately as

$$\hat{a}_j^{\rm UB} \cong \hat{a}_j + \theta\sqrt{\frac{\sigma_j^2}{N_j}}\,, \tag{27}$$

$$\hat{a}_j^{\rm LB} \cong \hat{a}_j - \theta\sqrt{\frac{\sigma_j^2}{N_j}}\,. \tag{28}$$

where $\theta = 1$ guarantees an existence of true changes between UB and LB with the probability of 68.27% or error probability of $\varkappa = 0.3173$ that is 31.73%; $\theta = 2$ of 95.45% or $\varkappa = 0.0555$ that is 5.55% and $\theta = 3$ of 99.73% or $\varkappa = 0.0027$ that is 0.27%.

Jitter Bounds. The jitter left bound (JLB) $J_l^{\rm L}$ and the jitter right bound (JRB) $J_l^{\rm R}$ can be determined with respect to n_l as follows. Because a step is unity with integer k, we specify the jitter probability at the kth point using (5) as

$$P_k(\gamma_l^-, \gamma_l^+) = p[k|d(\gamma_l^-, \gamma_l^+), q(\gamma_l^-, \gamma_l^+)]\,. \tag{29}$$

We then equate (29) to $\varkappa$ and solve it for the right and left jitter to have, respectively,

$$k_l^{\rm R} = \left\lfloor \frac{\nu_l}{\kappa_l} \ln \frac{(1-d_l)(1-q_l)}{\varkappa(1-d_l q_l)} \right\rfloor, \tag{30}$$

$$k_l^{\rm L} = \left\lfloor \nu_l \kappa_l \ln \frac{(1-d_l)(1-q_l)}{\varkappa(1-d_l q_l)} \right\rfloor, \tag{31}$$

where $\lfloor x \rfloor$ means the maximum integer equal to or lower than x. The JLB and JRB can be defined with respect to n_l as $J_l^{\rm L} = n_l - k_l^{\rm L}$ and $J_l^{\rm R} = n_l + k_l^{\rm R}$. Now observe that n_l is unknown and use the estimate $\hat{n}_l$. If it happens that $\hat{n}_l$ lies at the right bound, then the true n_l can be found $k_l^{\rm R}$ points to the left. Otherwise, if $\hat{n}_l$ lies at the left bound, then i_l can be found $k_l^{\rm L}$ points to the right. Approximate JLB and JRB are thus the following

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

$$J_l^{\rm R} \cong \hat{n}_l + k_l^{\rm L}\,. \tag{33}$$

Note that $\varkappa$ in (30) and (31) should be specified in the θ-sense as in (27) and (28).

UB and LB Masks and Algorithm. By combining (27), (28), (32), and (33), the UB mask $\mathcal{B}_n^{\rm U}$ and the LB mask $\mathcal{B}_n^{\rm L}$ can now be formed to outline the region for true genomic changes. The relevant algorithm was designed in [41]. Its inputs are measurements y_n, breakpoints estimates $\hat{n}_l$, tolerance parameter θ, number L of the breakpoints, and number of readings M. At the output, the algorithms produces two masks: $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$.

The UB and LB masks have the following basic applied properties:

- The true CNVs exist between $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ with the probability determined in the θ-sigma sense.
- If $\mathcal{B}_n^{\rm U}$ or $\mathcal{B}_n^{\rm L}$ covering two or more breakpoints is uniform, then there is a probability of no changes in this region.
- If both $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ covering two or more breakpoints are uniform, then there is a high probability of no changes in this region.

We notice again that the jitter bounds in $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ may have enough accuracy if $(\gamma_l^-, \gamma_l^+) > 1$. They may be considered in the lower bound sense if $(\gamma_l^-, \gamma_l^+) < 1$. However, the approximation error is commonly large if $(\gamma_l^-, \gamma_l^+) < 0.5$. For details, see Section 2.2.

4 Applications

In this section, we test some CNVs measurements and estimates by the UB and LB masks computed

ISBN: 978-1-61804-240-8

in the three-sigma sense, $\theta = 3$, using the algorithm [41–43]. Because the algorithm can be applied to any CNVs data with supposedly known breakpoints, we choose the 1st chromosome measured using the HR-CGH array in [28] and available from [44].

The CNVs structure has 34 segments and 33 breakpoints. Most of the segments have the SNRs exceeding unity meaning that the UB and LB masks will have enough accuracy. The SNRs in segments $\hat{a}_{18}$ and $\hat{a}_{21}$ range between 0.5 and unity which means that real jitter can be here about twice larger. The remaining segments $\hat{a}_{23}$, $\hat{a}_{28}$, $\hat{a}_{31}$ and $\hat{a}_{32}$ demonstrate the SNR below 0.5 that means that the jitter bounds cannot be estimated with sufficient accuracy. We just may say that jitter can be much larger in the relevant breakpoints.

Let us consider the CNVs measurements and estimates in more detail following Fig. 4. As can be seen, there are two intervals with no measurements between the breakpoints $\hat{i}_{15}$ and $\hat{i}_{16}$ and the breakpoints $\hat{i}_{28}$ and $\hat{i}_{29}$. A part of measurements covering the breakpoints from $\hat{i}_5$ to $\hat{i}_{14}$ is shown in Fig. 5a. Its specific is that the segmental SNRs are all larger than unity and the masks thus can be used directly for practical applications. The masks suggest that errors in all of the segmental estimates reach tens of percents. In fact, $\hat{a}_5$ and $\hat{a}_{10}$ are estimated with error of about 50%. Error exceeds 30% in the estimates $\hat{a}_7$, $\hat{a}_9$, $\hat{a}_{12}$, and $\hat{a}_{13}$. A similar problem can be observed in the estimates of almost all of the breakpoints in which left and right jitter reaches several points.

A situation even worse with a part of the chromosome covering the breakpoints from $\hat{i}_{17}$ to $\hat{i}_{26}$. The segmental errors exceed 50% here over almost all segments. Furthermore, the UB is placed above LB around $\hat{i}_{17}$, $\hat{i}_{20}$, and $\hat{i}_{22}$. That means that there is a probability that these breakpoints do not exist. On the other hand, estimates in the part covering $\hat{i}_{24}$–$\hat{i}_{26}$ are not reliable. Thus there is a probability of no changes in this region as well.

5 Conclusions

Effect of measurement noise on the HR-CGH array-based estimates of the CNVs naturally results in segmental errors and jitter in the breakpoints due to typically low SNRs. Errors can be so large that medical expert would hardly be able to arrive at correct conclusions about real CNVs structures irrespective of the estimator used. Two rules of thumb for designers of measurement equipment are thus the following: *the higher probe resolution the more segmental accuracy* and *the larger segmental SNRs the lower jitter in the breakpoints*.

Because of large noise, estimates of the CNVs may bring insufficient information to experts and must be tested by UB and LB masks. To form such masks, the jitter distribution must be known. We have shown that jitter in the breakpoints can be modeled using the discrete skew Laplace distribution if the segmental SNRs exceed unity. Otherwise, the approximation errors can be large and more profound investigations of jitter will be required. The UB and LB masks proposed in this paper in the θ-sigma sense outline the region within which the true changes exist with a high probability (99.73% in the three-sigma sense). Provided the masks, information about CNVs is more complete and sometimes can be crucial for medical experts to make a correct decision about true structure. Testing some measurements and estimates by the UB and LB masks has revealed large errors exceeding $(30...50)\%$ in many segments. It was also demonstrated that jitter in some breakpoints is redundantly large for making any decision about their true locations. We finally notice that further investigations must be focused on the jitter statistics at low SNR values that is required to sketch a more correct probabilistic picture of the CNVs.

References:

[1] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-945, Oct. 2004.

[2] P. Stankiewicz and J. R. Lupski, "Structural Variation in the Human Genome and its Role in Disease," *Annual Review of Medicine*, vol. 61, pp. 437-455, Feb. 2010.

[3] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, no. 9, pp. 949-951, Sep. 2004.

[4] J. R. Pollack, T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Borresen-Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proc. Natl Acad. Sci (PNAS)*, vol. 99, no. 20, pp. 12963-12968, Oct. 2002.

[5] R. Pique-Regi, A. Ortega, A. Tewfik, and S. Asgharzadeh, "Detection changes in the DNA copy number," *IEEE Signal Process. Mgn.*, vol. 29, no. 1, pp. 98-107, Jan. 2012.

[6] A. F. Cockburn, M. J. Newkirk, and R. A. Firtel, "Organization of the ribosomal RNA genes of dictyostelium discoideum: mapping of the nontrascribed spacer regions," *Cell*, vol. 9, no. 4, pp. 605-613, Dec 1976.

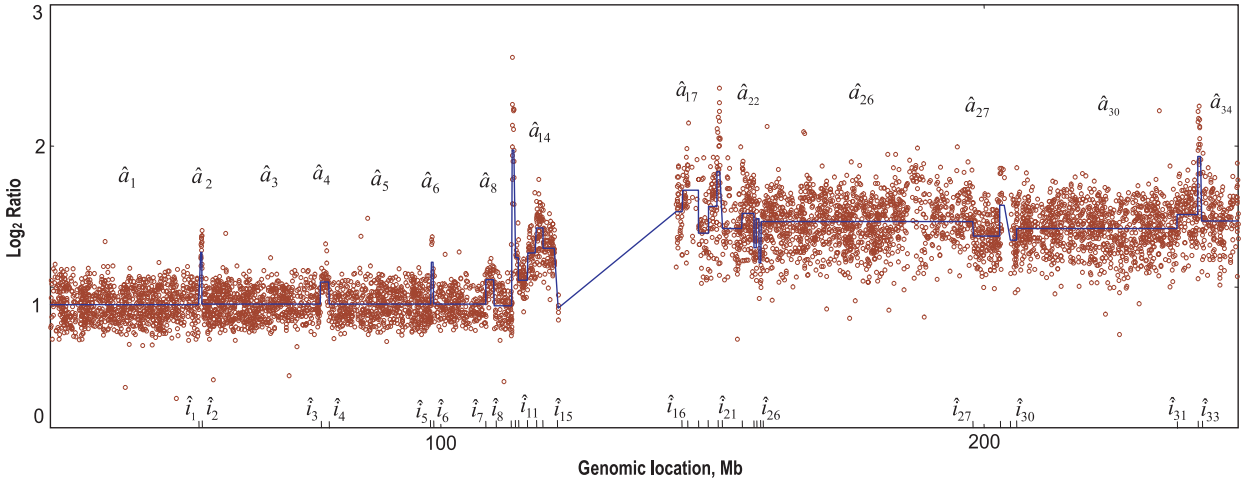

Figure 4: Measurements and estimates of the 1st chromosome changes taken from archive "159A–vs–159D–cut" available in [44].

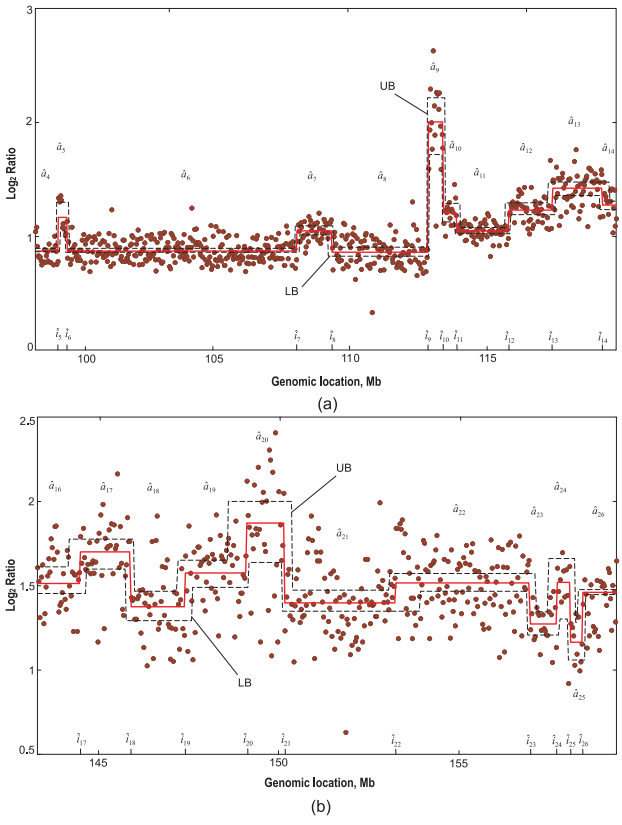

Figure 5: Parts of chromosomal changes (Fig. 4) tested by UB and LB masks: (a) genomic location from about 97Mb to 120Mb and (b) genomic location from 143Mb to 159Mb. Jitter in $\hat{i}_5$, $\hat{i}_6$, $\hat{i}_9$, and $\hat{i}_{10}$ is moderate and these breakpoints are well detectable. Breakpoints $\hat{i}_{17}$, $\hat{i}_{22}$, $\hat{i}_{27}$, $\hat{i}_{30}$, and $\hat{i}_{31}$ cannot be estimated correctly owing to low SNRs. There is a probability that the breakpoints $\hat{i}_{17}$, $\hat{i}_{19}$, $\hat{i}_{20}$, $\hat{i}_{22}$, and $\hat{i}_{24}$ do not exist.

[7] H. Ren, W. Francis, A. Boys, A. C. Chueh, N. Wong, P. La, L. H. Wong, J. Ryan, H. R. Slater, and K. H. Choo, "BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints," *Human Mutation*, vol. 25, no. 5, pp. 476-482, May 2005.

[8] A. E. Urban, J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. V. Popescu, J. F. Cubells, R. Green, B. S. Emanuel, M. B. Gerstein, S. M. Weissman, and M. Snyder, "High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays," *Proc. Natl. Acad. Sci. (PNAS)*, vol. 103, no. 12, pp. 4534-4539, Mar. 2006.

[9] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, pp. 1–10, Feb 2002.

[10] P.J. Campbell, P.J. Stephens, E.D. Pleasance, S. OMeara, H. Li, T. Santarius, L.A Stebbings, C. Leroy, S. Edkins, C. Hardy, J.W. Teague, A. Menzies, I. Goodhead, D.J. Turner, C.M. Clee, M.A. Quail, A. Cox, C. Brown, R. Durbin, M.E. Hurles, P.A.W Edwards, G.R. Bignell, M.R. Stratton, and P.A. Futreal, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing", *Nature Genetics*, vol. 40, no. 6, pp. 722-729, Jun. 2008.

[11] L. Hsu, S.G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter, "Denoising array-based comparative genomic hybridization data using wavelets", *Biostatistics*, vol. 6, no. 2, pp. 211–226, 2005.

[12] E. Ben-Yaacov and Y.C. Eldar, "A fast and flexible method for the segmentation of aCGH data", *Biostatistics*, vol. 24, no. 16, pp. i139–i145, 2008.

[13] V. Katkovnik and V.G. Spokoiny, "Spatial adaptive estimation via fitted local likelihood techniques," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 873–886, 2008.

[14] A. Goldenshluger and A. Nemirovski, "Adaptive de-noising of signals satisfying differential inequalities," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 872–889, 1997.

[15] I.E. Frank and J.H. Friedman, "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, vol. 35, pp. 109-148, 1993.

[16] A.E. Hoerl and R.W. Kennard, R.W., "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55-67, 1970.

[17] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statist. Soc.*, ser. B, vol. 58, pp. 267-288, 1996.

[18] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain, "Hidden Markov models approach to the analysis of array CGH data", *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 132-153, 2004.

[19] J. Chen and Y.-P. Wang, "A statistical change point model approach for the detection of DNA copy number variations in array CGH data", *IEEE/ACM Trans. on Comput. Biology and Bioinform.*, vol. 6, no. 4, pp. 529–541, 2009.

[20] S. H. Chung and R. A. Kennedy, "Forward-backward non-linear filtering technique for extracting small biological signal from noise," *J Neuroscience Meth*, vol. 40, no. 1, pp. 71–86, Nov 1991.

[21] O. Vite-Chavez, R. Olivera-Reyna, O. Ibarra-Manzano, Y. S. Shmaliy, and L. Morales-Mendoza, "Time-variant forward-backward FIR denoising of piecewise-smooth signals," *Int. J. Electron. Commun. (AEU)*, vol. 67, no. 5, pp. 406–413, May 2013.

[22] J. Muñoz-Minjares, O. Ibarra-Manzano, and Y. S. Shmaliy, "Maximum likelihood estimation of DNA copy number variations in HR-CGH arrays data," In *Proc. 12th WSEAS Int. Conf. on Signal Process., Comput. Geometry and Artif. Vision (ISCGAV'12), Proc. 12th WSEAS Int. Conf. on Systems Theory and Sc. Comput. (ISTASC'12)*, Istanbul (Turkey), pp. 45-50, 2012.

[23] J. Huang, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones, and M. H. Shapero, "Whole genome DNA copy number changes identified by high density oligonucleotide arrays," *Hum. Genomics*, vol. 1, no. 4, pp. 287–299, May 2004.

[24] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variantion using high-throughput sequencing," *BMC Bioinform.*, vol. 10, no. 80, pp. 1–9, Mar 2009.

[25] A. K. Alqallaf and A. H. Teqfik, "DNA copy number detection and Sigma filter," In *Proc. GENSIPS*, pp. 1–4, 2007.

[26] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham, and S. Tavare, "CNAseg–A novel framework for identification of copy number changes in cancer from second-generation sequencing data," *Bioinform.*, vol. 26, no. 24, pp. 3051–3058, Dec. 2010.

[27] J. Chen and A. K. Gupta, *Parametric Statistical Change Point Analysis with Applications to Genetics, Medicine, and Finance*, 2nd Ed., Springer, 2012.

[28] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Nquyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler, "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation," *Genome Research*, vol. 13, no. 10, pp. 2291–2305, Oct. 2003.

[29] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, "A statistical approach for array CGH data analysis," *BMC Bioinformatics*, vol. 6, no. 1, pp. 27–37, 2005.

[30] A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigner, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, Oct. 2004.

[31] J. T. Simpson, R. E. Mclntyre, D. J. Adams, and R. Durbin, "Copy number variant detection in inbred strains from short read sequence data," *Bioinformatics*, vol. 26, no. 4, pp. 565–567, Feb. 2010.

[32] L. Wang, A. Abyzov, J. O. Korbel, M. Snyder, and M. Gerstein, "MSB: A mean-shift-based approach for the analysis of structural variation in the genome," *Genomic Res.*, vol. 19, no. 1, pp. 106–117, Jan 2009.

[33] V. Boeva, A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, Jan. 2011.

[34] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. **9**, no. 1, pp. 18–29, Jan. 2008.

[35] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. 24, no. 3, pp. 309–318, Mar. 2008.

[36] X. Gao and J. Huang, "A robust penalized method for the analysis of noisy DNA copy number data," *BMC Genomics*, vol. 11, no. 517, pp. 1–10, Sep. 2010.

[37] O. M. Rueda and R. Diaz-Uriante, "RJaCGH: Bayesian analysis of a aCGH arrays for detecting copy number changes and recurrent regions," *Bioinformatics*, vol. 25, no. 15, pp. 1959–1960, Aug. 2009.

[38] Y. Yuan, C. Curtis, C. Caldas, and F. Markowetz, "A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes," *IEEE Trans. Comput. Biology and Bioinformatics*, vol. 9, no. 4, pp. 947–954, Jul.-Aug. 2012.

[39] J. Muñoz-Minjares, J. Cabal-Aragon, and Y. S. Shmaliy, "Jitter probability in the breakpoints of discrete sparse piecewise-constant signals," *Proc. 21st European Signal Process. Conf. (EUSIPCO-2013)*, 2013.

[40] T. J. Kozubowski and S. Inusah, "A skew Laplace distribution on integers," *Annals of the Inst. of Statist. Math.*, vol. 58, no. 3, pp. 555-571, Sep. 2006.

[41] J. Muñoz-Minjares, J. Cabal-Aragon, Y. S. Shmaliy, "Probabilistic bounds for estimates of genome DNA copy number variations using HR-CGH microarrays," *Proc. 21st European Signal Process. Conf. (EUSIPCO-2013)*, 2013.

[42] J. Muñoz-Minjares, Y. S. Shmaliy, J. Cabal-Aragon, "Confidence limits for genome DNA copy number variations in HR-CGH array measurements," *Biomedical Signal Processing & Control,* vol. 10, pp. 166–173, Mar. 2014.

[43] J. Muñoz-Minjares, J. Cabal-Aragon, Y. S. Shmaliy, "Effect of noise on estimate bounds for genome DNA structural changes," *WSEAS Trans. on Biology and Biomedicine,* vol. 11, pp. 52–61, Apr. 2014.

[44] Representational oligonucleotide microarray analysis (ROMA), http://Roma.cshl.org.