EUSIPCO 2013 1569738823

PROBABILISTIC BOUNDS FOR ESTIMATES OF GENOME DNA COPY NUMBER VARIATIONS USING HR-CGH MICROARRAY

Jorge Muñoz-Minjares, Jesús Cabal-Aragón, Yuriy S. Shmaliy

Department of Electronics Engineering, Universidad de Guanajuato
Ctra. Salamanca-Valle, 3.5+1.8km, Palo-Blanco, 36855, Salamanca, Mexico
phone: + 52 (464) 647-99-40, email: shmaliy@ugto.mx,
web: www.ingenierias.ugto.mx

ABSTRACT

Estimation of the genome copy number variations (CNVs) measured using the high-resolution array-comparative genomic hybridization (HR-CGH) microarray is provided in the presence of large measurement white Gaussian noise having typically different segmental variances. Jitter inherent to the breakpoints of such signals can be approximated with the discrete skew Laplace distribution. Referring to and aimed at sketching a more clear picture about possible chromosomal changes, we have justified the estimate UB and LB probabilistic masks in the three-sigma sense to guarantee an existence of true changes with the probability of 99.73%. Some real measurements are tested by these mask and practical conclusions are provided.

1. INTRODUCTION

The disease such as cancer is often accompanied with structural changes called copy-number variations (CNVs) in the deoxyribonucleic acid (DNA) of a genome essential for human life. 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 [1]. A brief survey of types of chromosome alterations involving copy number changes is given in [2]. A commonly accepted unit of measurement in molecular biology is kilobase (kb) equal to 1000 base pairs of DNA [3]. The human genome with 23 chromosomes is estimated to be about 3.2 billion base pairs long and to contain $20000-25000$ distinct genes [7]. Each CNV may range from about one kb to several megabases (Mbs) in size [1].

The array-comparative genomic hybridization (aCGH) is one of the most modern techniques employing chromosomal microarray analysis to detect the CNVs at a resolution level of 5–10 kbs [4]. It was reported in [5] 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 $\log_2 R/G = \log_2$ Ratio, where R and G are the fluorescent Red and Green intensities, respectively [6]. The Ratio is highly contaminated by noise which intensity does not always allow for correct visual identification of the breakpoints and copy numbers if the number of segmental reads is small. A 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 [5] to map a 622-bp heterozygous deletion.

From the standpoint of signal processing, the following properties of the CNVs function were recognized [2]:

- It is piecewise constant (PWC) and sparse with a small number of alterations on a long base-pair length.
- Its constant values are integer, although this property is not survived in the $\log_2$ Ratio.
- The measurement noise in the $\log_2$ Ratio is highly intensive and can be modeled as additive white Gaussian.

The CNVs estimation problem is thus to predict the breakpoints locations and the segmental levels with a maximum possible accuracy and precision acceptable for medical applications. Because of large noise, the CNVs estimates must be accompanied with probabilistic upper bound (UB) and lower bound (LB) masks.

2. PROBABILISTIC UB AND LB MASKS

Below, we consider the jitter in the breakpoints and segmental errors and specify the probabilistic UB and LB masks in the three-sigma sense to guarantee an existence of the CNVs between the masks with the probability of 99.73%.

2.1 Jitter Distribution

It was shown in [8] that jitter in the lth breakpoints n_l of discrete sparse piecewise-constant signals such as the CNVs measured in white Gaussian noise can be approximated with the discrete skew Laplace probability density function (pdf) [9]:

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

where $d_l = e^{-\frac{\kappa_l}{\nu_l}} \in (0,1)$ and $q_l = e^{-\frac{1}{\kappa_l \nu_l}} \in (0,1)$, $\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}$, $\nu_l = -\frac{\kappa_l}{\ln x_l}$,

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

$$\mu_l = \frac{P(A_l)[1-P(B_l)]}{P(B_l)[1-P(A_l)]}, \qquad (3)$$

$$\phi_l = \frac{P(A_l)+P(B_l)-1}{[1-2P(A_l)][1-2P(B_l)]}, \qquad (4)$$

$$P(A_l) = \begin{cases} 1+\frac{1}{2}[\mathrm{erf}(g_l^\beta)-\mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- < \gamma_l^+, \\ \frac{1}{2}\mathrm{erfc}(g_1^\alpha) & , \quad \gamma_l^- = \gamma_l^+, \\ \frac{1}{2}[\mathrm{erf}(g_l^\beta)-\mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- > \gamma_l^+, \end{cases} \qquad (5)$$

$$P(B_l) = \begin{cases} \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- < \gamma_l^+ , \\ 1 - \frac{1}{2}\mathrm{erfc}(h_l^\alpha) & , \quad \gamma_l^- = \gamma_l^+ , \\ 1 + \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- > \gamma_l^+ , \end{cases} \tag{6}$$

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, and

$$\alpha_l, \beta_l = \frac{a_l\gamma_l^- - a_{l+1}\gamma_l^+}{\gamma_l^- - \gamma_l^+} \mp \frac{1}{\gamma_l^- - \gamma_l^+} \times \sqrt{(a_l - a_{l+1})\gamma_l^-\gamma_l^+ + 2\Delta_l^2(\gamma_l^- - \gamma_l^+)\ln\sqrt{\frac{\gamma_l^-}{\gamma_l^+}}} \tag{7}$$

if $\gamma_l^- \neq \gamma_l^+$. For $\gamma_l^- = \gamma_l^+$, set $\alpha_l = \Delta_l/2$ and $\beta_l = \pm\infty$. Here

$$\gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2} \tag{8}$$

are the signal-to-noise ratios (SNRs) in the lth and $(l+1)$th segments related to the change $\Delta_l = a_{l+1} - a_l$ in the n_l breakpoint and the variances σ_l^2 and σ_{l+1}^2 of the segmental white Gaussian noise.

2.1.1 Jitter Bounds

The left jitter bound (LJB) J_l^{L} and the right jitter bound (RJB) J_l^{R} can be determined with respect to the lth breakpoint $\hat{i}_l$ as follows. Consider the jitter distribution (1) for known γ_l^- and γ_l^+. Increase k in (1) from zero until $p_k < 0.27\%$. Accept the relevant value of k as the right jitter k_l^{R}. Next, reduce k from zero until $p_k < 0.27\%$ and accept the relevant value of k as the left jitter k_l^{L}. Form the LJB and RJB as

$$J_l^{\mathrm{L}} \cong \hat{n}_l - k_l^{\mathrm{R}}, \tag{9}$$

$$J_l^{\mathrm{R}} \cong \hat{n}_l + k_l^{\mathrm{L}}. \tag{10}$$

2.2 Segmental Error Probability

Provided the estimate $\hat{n}_l$ of the breakpoint n_l, simple averaging applied on an interval of $N_l = n_l - n_{l-1}$ readings from n_{l-1} to $n_l - 1$ gives the segmental level estimate $\hat{a}_j = \frac{1}{N_j}\sum_{v=n_{j-1}}^{n_j-1} y_v$, which mean value is $E\{\hat{a}_j\} = a_j$ and which variance is $\hat{\sigma}_j^2 = \frac{\sigma_j^2}{N_j}$, where σ_j^2 is the noise variance in the jth segment. The Gaussian pdf of $\hat{a}_j$ is thus

$$p_j(x) = \sqrt{\frac{N_j}{2\pi\sigma_j^2}}\exp\left[-\frac{(x-a_j)^2 N_j}{2\sigma_j^2}\right] \tag{11}$$

and the error probability for $\hat{a}_j$ to exceed a threshold ε around actual a_j can be found as

$$P_E(N_j) = 2\int_{a_j+\varepsilon}^{\infty} p_j(x)dx = \mathrm{erfc}\left(\mu\sqrt{\frac{N_j}{2}}\right), \tag{12}$$

where $\mathrm{erfc}(x)$ is the complementary error function and $\mu = \frac{\varepsilon}{\sqrt{\sigma^2}}$ is the normalized threshold.

2.2.1 Segmental Bounds

In the 3-sigma sense, the UB for segmental estimates can be formed as $\hat{a}_j^{\mathrm{UB}} = E\{\hat{a}_j\} + 3\sqrt{\frac{\sigma_j^2}{N_j}}$. However, neither the 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

$$\hat{a}_j^{\mathrm{UB}} \cong \hat{a}_j + 3\sqrt{\frac{\sigma_j^2}{N_j}}, \tag{13}$$

$$\hat{a}_j^{\mathrm{LB}} \cong \hat{a}_j - 3\sqrt{\frac{\sigma_j^2}{N_j}}. \tag{14}$$

2.3 UB and LB Masks

By combining (9), (10), (13), and (14), the UB and LB masks can be formed to outline the region for true CNVs. The algorithm for computing the UB mask $\mathscr{B}_n^{\mathrm{U}}$ and LB mask $\mathscr{B}_n^{\mathrm{L}}$ is developed in Table 1. Its input is measurements y_n, breakpoint estimates $\hat{n}_l$, allowed error probability $\xi = 0.0027$ (3-sigma), number L of the breakpoints, and number of readings M. At the output, the algorithms produces the masks $\mathscr{B}_n^{\mathrm{U}}$ and $\mathscr{B}_n^{\mathrm{L}}$. The masks have the following basic properties:

- The true CNVs exist between $\mathscr{B}_n^{\mathrm{U}}$ and $\mathscr{B}_n^{\mathrm{L}}$ with the probability of 99.73% (3-sigma).
- If $\mathscr{B}_n^{\mathrm{U}}$ or $\mathscr{B}_n^{\mathrm{L}}$ covering two or more breakpoints is uniform, then there is a probability of no changes in this region.
- If both $\mathscr{B}_n^{\mathrm{U}}$ and $\mathscr{B}_n^{\mathrm{L}}$ covering two or more breakpoints are uniform, then there is a high probability of no changes in this region.

3. EXPERIMENTAL VERIFICATION

The purpose of this section is to test some CNVs estimates by the UB and LB probabilistic masks (Table 1). For clarity, we first compute some useful characteristics of the processes and put them to tables. We base our studies on some HR-CGH array measurements published in [10] and available from [11]. Voluntary, we select the ones associated with large jitter and large segmental errors. The estimates $\hat{i}_l$ of the breakpoint locations are also taken from [11]. For a comparison, we also provide another estimates based on [12, 13]. Note that $\hat{i}_l \cong \hat{n}_l\bar{r}$, where $\bar{r} = 30$ kb is an average probe resolution. Estimates of the segmental levels are found by data averaging between the breakpoints and we notice their good correspondence with [11]. We finally employ the algorithm (Table 1) and plot the UB and LB masks along with the CNVs estimates provided.

3.1 Large Jitter

The first database processed is a part of the 7th chromosome in archive "159A–vs–159D–cut" of [11]. It is shown to have 14 segments and 13 breakpoints (Fig. 1 and Fig. 2). However, there is a high probability that some breakpoints do not exist. Observe Fig. 2a and the characteristics collected in Table 2. 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 for the 2nd-to-6th segments is almost uniform. Thus, there

Table 1: Algorithm for computing the UB mask $\mathscr{B}_n^{\mathrm{U}}$ and LB mask $\mathscr{B}_n^{\mathrm{L}}$ via HR-CGH array CNVs measurements y_n and the breakpoint locations estimates $\hat{n}_l$. Given: allowed error probability $\varkappa = 0.0027$ (3-sigma), number L of breakpoints, and number of readings M.

Input: y_n, $\hat{n}_l$, $\xi = 0.0027$, L, M

$N_{L+1} = M - \hat{n}_L, \quad \hat{n}_0 = 0$

for $j = 1 : L+1$ **do**

$N_j = \hat{n}_j - \hat{n}_{j-1}, \quad \hat{a}_j = \frac{1}{N_j}\sum_{v=\hat{n}_{j-1}}^{\hat{n}_j - 1} y_v$

$\sigma_j = \sqrt{\frac{1}{N_j}\sum_{v=\hat{n}_{j-1}}^{\hat{n}_j - 1}(y_v - \hat{a}_j)^2}$

and for

for $l = 1 : L$ **do**

$\Delta_l = \hat{a}_{l+1} - \hat{a}_l, \quad \gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2}$

α_l by (7) with "$-$" and $a_l = \hat{a}_l$

β_l by (7) with "$+$" and $a_l = \hat{a}_l$

$g_l^\beta = \frac{\beta_l - \hat{a}_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^-}{2}}, \quad g_l^\alpha = \frac{\alpha_l - \hat{a}_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^-}{2}}$

$h_l^\beta = \frac{\beta_l - \hat{a}_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^+}{2}}, \quad h_l^\alpha = \frac{\alpha_l - \hat{a}_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^+}{2}}$

P_l^A by (5), $\quad P_l^B$ by (6), $\quad \phi_l$ by (4)

$\mu_l = \frac{P_l^A(1-P_l^B)}{P_l^B(1-P_l^A)}, \quad x_l$ by (2) , $\quad \kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}$,

$\nu_l = -\frac{\kappa_l}{\ln(x_l)}, \quad d_l = e^{-\frac{\kappa_l}{\nu_l}}, \quad q_l = e^{-\frac{1}{\kappa_l \nu_l}}$

$k_l^{\mathrm{R}} = \left\lfloor \frac{\nu_l}{\kappa_l} \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)} \right\rfloor$ ▷ right jitter

$k_l^{\mathrm{L}} = \left\lceil \nu_l \kappa_l \ln \frac{(1-d_l)(1-q_l)}{\xi(1-d_l q_l)} \right\rceil$ ▷ left jitter

and for

$\mathscr{C}_{L+1} = M - 1, \; \mathscr{D}_{L+1} = M - 1$

for $l = 1 : L$ **do**

$$\mathscr{C}_l = \begin{cases} \hat{n}_l - k_l^{\mathrm{R}} & \textbf{if} \quad \Delta_l > 0 \\ \hat{n}_l + k_l^{\mathrm{L}} & \textbf{if} \quad \Delta_l < 0 \end{cases}$$

$$\mathscr{D}_l = \begin{cases} \hat{n}_l + k_l^{\mathrm{L}} & \textbf{if} \quad \Delta_l > 0 \\ \hat{n}_l - k_l^{\mathrm{R}} & \textbf{if} \quad \Delta_l < 0 \end{cases}$$

and for

for $l = 1 : L$ **do**

$$\mathscr{C}_l = \begin{cases} \mathscr{C}_l & \textbf{if} \quad \mathrm{Im}\,\mathscr{C}_l = 0 \\ \mathscr{C}_{l-1} & \textbf{if} \quad \Delta_l \geqslant 0 \quad \wedge \quad \mathrm{Im}\,\mathscr{C}_l \neq 0 \\ \mathscr{C}_{l+1} & \textbf{if} \quad \Delta_l < 0 \quad \wedge \quad \mathrm{Im}\,\mathscr{C}_l \neq 0 \end{cases}$$

$$\mathscr{D}_l = \begin{cases} \mathscr{D}_l & \textbf{if} \quad \mathrm{Im}\,\mathscr{D}_l = 0 \\ \mathscr{D}_{l+1} & \textbf{if} \quad \Delta_l \geqslant 0 \quad \wedge \quad \mathrm{Im}\,\mathscr{D}_l \neq 0 \\ \mathscr{D}_{l-1} & \textbf{if} \quad \Delta_l < 0 \quad \wedge \quad \mathrm{Im}\,\mathscr{D}_l \neq 0 \end{cases}$$

and for

$l = 1, k = 1$

for $n = 0 : M - 1$ **do**

$$l = \begin{cases} l & \textbf{if} \quad n < \mathscr{C}_l \\ l+1 & \textbf{if} \quad n \geqslant \mathscr{C}_l \quad \wedge \quad \mathscr{C}_{l+1} > C_l \\ l+2 & \textbf{if} \quad n \geqslant \mathscr{C}_l \quad \wedge \quad \mathscr{C}_{l+1} \leqslant C_l \end{cases}$$

$$k = \begin{cases} k & \textbf{if} \quad n < \mathscr{D}_l \\ k+1 & \textbf{if} \quad n \geqslant \mathscr{D}_l \quad \wedge \quad \mathscr{D}_{l+1} > C_l \\ k+2 & \textbf{if} \quad n \geqslant \mathscr{D}_l \quad \wedge \quad \mathscr{D}_{l+1} \leqslant C_l \end{cases}$$

$\mathscr{B}_n^{\mathrm{U}} = \hat{a}_l + 3\sqrt{\frac{\sigma_l^2}{N_l}}$ ▷ UB mask

$\mathscr{B}_n^{\mathrm{L}} = \hat{a}_k - 3\sqrt{\frac{\sigma_k^2}{N_k}}$ ▷ LB mask

and for

Output: $\mathscr{B}_n^{\mathrm{U}}$, $\mathscr{B}_n^{\mathrm{L}}$

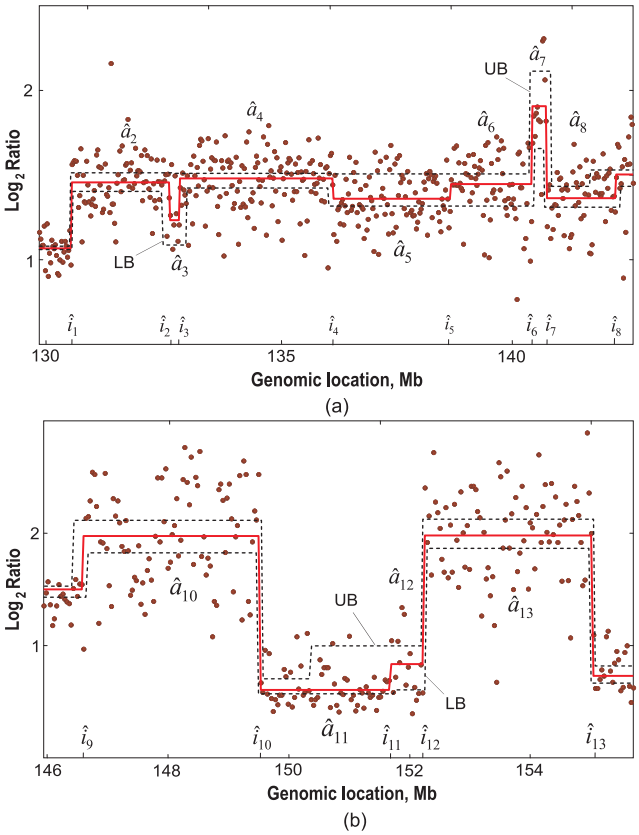

Figure 2: UB mask and LB mask for the estimates of the CNVs shown in Fig. 1: (a) genomic location from 130Mb to 146Mb and (b) genomic location from 146Mb to 156Mb.

is a probability that the 2nd-to-5th breakpoints do not exist. If to follow the LB mask, then locations of the 2nd-to-4th breakpoints can be predicted even with large errors. At least they can be supposed to exist. However, nothing definitive can be said about the 5th breakpoint location and one may suppose that it does not exist. It is also hard to distinguish the true location of the 8th breakpoint.

In Fig. 2b, i_{10}, i_{12}, and i_{13} are well detectable owing to large segmental SNRs. The breakpoint i_9 has a moderate jitter. However, the location of i_{11} is unclear. Moreover, there is a probability that i_{11} does not exist.

3.2 Large Segmental Errors

Another database corresponds to the 2nd chromosome in "159A–vs–159D–cut" of [11] supposedly having 42 segments and 41 breakpoints. This case demonstrates large segmental errors associated with just a few readings. In turn, only a few breakpoints are accompanied here with large jitter. Four specific regions associated with large segmental errors are sketched in Fig. 3. Even a quick look at these figures shows that in almost all of the cases of short chromosomal changes the segmental errors reach tens of percents. Moreover, some segments cannot be estimated at all with a reasonable error. Thus, there is a probability that such changes do not exist.

In fact, errors in the estimates of a_4, a_6, a_{14}, ana a_{16} exceed 30%. A situation is even worse with a_{10}, a_{12}, a_{15}, and a_{18}, where the estimation errors reach (40–50)%. We think

Table 2: Characteristics of the CNVs estimates for measurements of the 7th sample from the archive "159A–vs–159D–cut" of [11] with an average resolution of $\bar{r} = 30$kb. Statistics, bounds, and jitter parameters are given for the Log_2 Ratio. Jitter in $\hat{i}_1$, $\hat{i}_6$, $\hat{i}_7$, $\hat{i}_9$, $\hat{i}_{10}$, $\hat{i}_{12}$, and $\hat{i}_{13}$ is moderate and these breakpoints are well detectable. The breakpoints $\hat{i}_2$, $\hat{i}_3$, $\hat{i}_4$, $\hat{i}_5$, $\hat{i}_8$, $\hat{i}_9$, and $\hat{i}_{11}$ cannot be estimated correctly owing to large jitter. There is a probability that the breakpoints $\hat{i}_2$, $\hat{i}_3$, $\hat{i}_4$, $\hat{i}_5$, and $\hat{i}_{11}$ do not exist. There is a high probability that the breakpoint $\hat{i}_5$ does not exist.

j	jth Segment		Statistics		3-σ Bounds		Jitter parameters			3-σ Jitter	
	n_{j-1}	N_j	$\hat{a}_j$	σ_j^2	$\hat{a}_j^{\mathrm{UB}}$	$\hat{a}_j^{\mathrm{LB}}$	Δ_{j-1}	γ_{j-1}^{-}	γ_{j-1}^{+}	k_{j-1}^{L}	k_{j-1}^{R}
1	3085	602	1.06966	6.301637	1.07937	1.05996	–	–	–	–	–
2	3687	81	1.46235	25.13509	1.51519	1.40950	0.39268	24.4697	6.13482	1	2
3	3768	8	1.27274	26.79370	1.44636	1.09913	– 0.18960	1.43024	1.34171	5	5
4	3776	127	1.46647	27.20261	1.51038	1.42256	0.19373	1.40068	1.37962	5	5
5	3903	97	1.35931	19.86547	1.40224	1.31638	– 0.10716	0.42213	0.57804	17	4
6	4000	67	1.43297	37.85345	1.50428	1.36167	0.07366	0.27315	0.14335	–	–
7	4067	12	1.87920	58.36954	2.08843	1.66997	0.44623	5.26032	3.41139	2	3
8	4079	57	1.37080	22.74918	1.43074	1.31087	– 0.50839	4.42814	11.3617	2	2
9	4136	123	1.48145	31.77443	1.52967	1.43323	0.11065	0.53814	0.38529	4	21
10	4259	84	1.97564	188.6828	2.11782	1.83346	0.49419	7.68614	1.29436	2	5
11	4343	62	0.63597	29.39834	0.70129	0.57064	– 1.33967	9.51186	61.0485	1	1
12	4405	16	0.79879	69.62526	0.99669	0.60089	0.16282	0.90176	0.38076	2	38
13	4421	80	1.99875	156.1977	2.13131	1.86619	1.19996	20.6808	9.21850	1	1
14	4501	48	0.74214	31.54325	0.81905	0.66524	– 1.25660	10.1093	50.0599	1	1

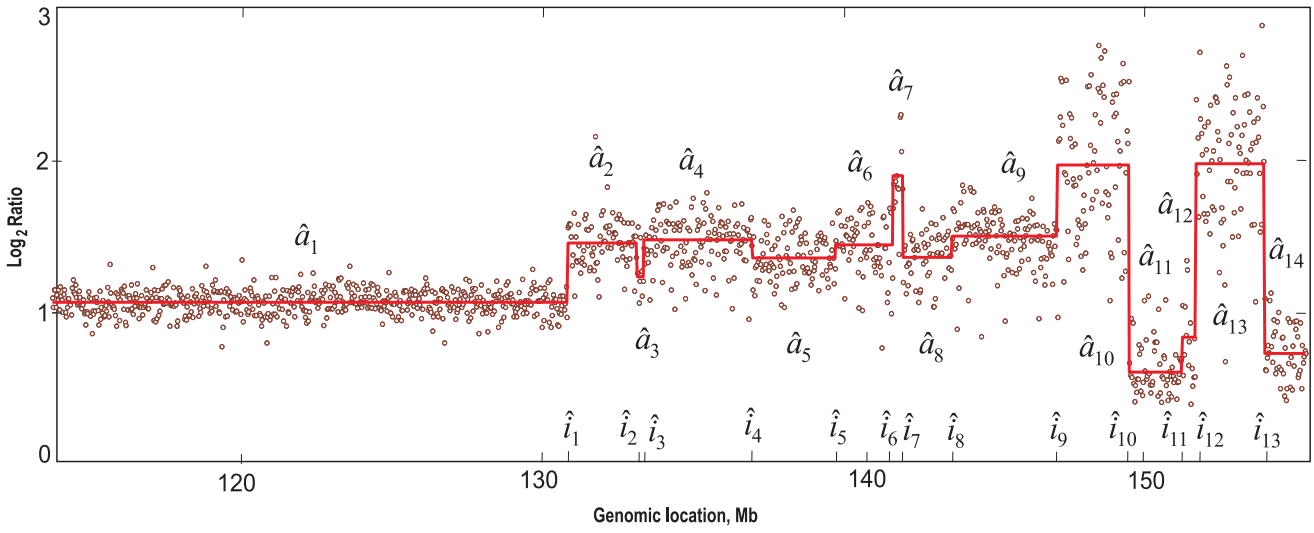

Figure 1: Measurements and estimates of a part of the 7th chromosome taken from the archive "159A–vs–159D–cut" of [11]. Jitter in $\hat{i}_1$, $\hat{i}_6$, $\hat{i}_7$, $\hat{i}_9$, $\hat{i}_{10}$, $\hat{i}_{12}$, and $\hat{i}_{13}$ is moderate and these breakpoints are well detectable. The breakpoints $\hat{i}_2$, $\hat{i}_3$, $\hat{i}_4$, $\hat{i}_5$, $\hat{i}_8$, $\hat{i}_9$, and $\hat{i}_{11}$ cannot be estimated correctly owing to large jitter. There is a probability that the breakpoints $\hat{i}_2$, $\hat{i}_3$, $\hat{i}_4$, $\hat{i}_5$, and $\hat{i}_{11}$ do not exist. There is a high probability that the breakpoint $\hat{i}_5$ does not exist.

that such large errors can hardly be accepted by medical experts. Furthermore, there are two segments a_2 and a_8 which levels cannot be estimated correctly and the question arises about the existence of the predicted changes in these regions.

Another specific of this chromosome is that a part of measurements around the breakpoint i_{19} does not contain enough information for experts. As a consequence, neither a_{19} nor i_{19} can be estimated with a reasonable error.

4. CONCLUSIONS

Measurements of genome changes using the HR-CGH array are available with the probe resolution of 0.2...40 kb in the presence of large white Gaussian noise and segmental SNRs around unity. Under such conditions, estimates of segmental changes and breakpoint locations are often accompanied with large and even unacceptable errors. In order to give medical experts additional information about genomic changes, we have justified the estimation error UB and LB masks. The masks were found in the three-sigma sense to guarantee an existence of true CNVs between UB and LB with the probability of 99.73%. Testing some estimates taken from [11] by the UB and LB masks has revealed large errors exceeding (30...50)% in many segments. It also turned out that jitter in some breakpoints is redundantly large for making any decision about their locations. The masks have also suggested that there is a high probability that some changes and breakpoints do not exist.

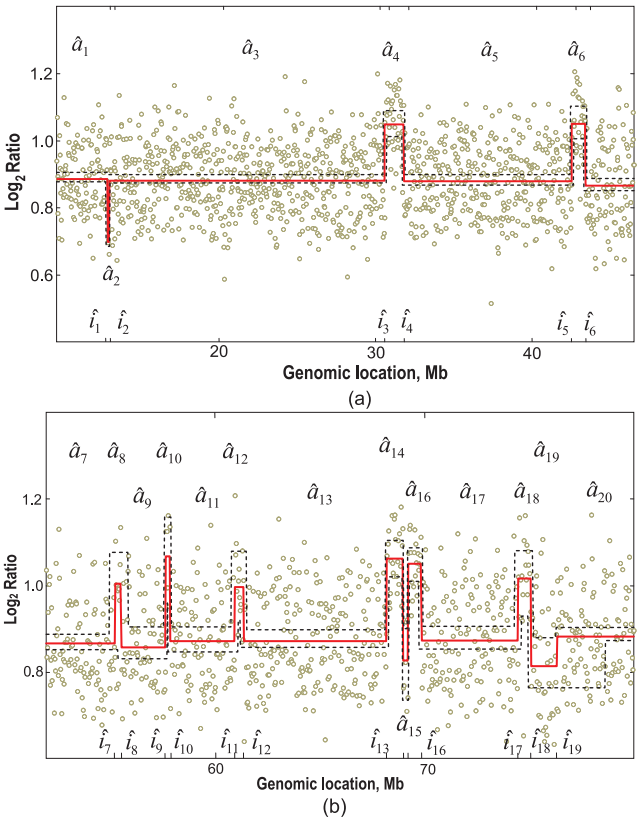

Figure 3: UB and LB masks for the estimates of the CNVs with large segmental errors: (a) genomic location from 10Mb to 45Mb and (b) genomic location from 50Mb to 80Mb. Errors in the estimates of a_1, a_6, a_{14}, and a_{16} exceed 30%. Errors in the estimates of a_{10}, a_{12}, a_{15}, and a_{18} reach 40-50%. There is a probability that changes a_2 and a_8 do not exist.

REFERENCES

[1] 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, 2010.

[2] R. Pique-Regi, A. Ortega, A. Tewfik, and S. Asgharzadeh, "Detection changes in the DNA copy number", *IEEE Signal Processing Mgn.*, vol. 29, pp. 98-107, Jan. 2012.

[3] A. Cockburn, M.J. Newkirk, and R.A. Firtel, "Organization of the RNA genes of dictyostelium: Mapping of the non-trascribed spacer regions", *Cell*, vol. 9, pp. 605-613, Dec. 1976.

[4] H. Ren, W. Francis, A. Boys, A.C. Chueh, N. Wong, P. La, L.H. Wong, J. Ryan, H.R. Slater, and K.H.A. Choo, "BAC-based PCR fragment microarray: High-resolution detection of chromosomal deletion and duplication breakpoints", *Human Mutation*, vol. 25, pp. 476-482, May 2005.

[5] 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, pp. 4534-4539, Mar. 2006.

[6] 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, pp. e15, 2002.

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

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

[9] T. J. Kozubowski and S. Inusah, "A skew Laplace distribution on integers", *Annals of the Inst. of Statist. Math.*, vol. 58, pp. 555-571, Sep. 2006.

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

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

[12] 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)*, Istanbul, Turkey, 2012, pp. 45-50.

[13] 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, pp. 406–413, May 2013.