Research Article

iMedPub Journals
www.imedpub.com

Journal of Genetic Disorders

2017
Vol. 1 No. 1:9

Matching Confidence Masks with Experts Annotations for Estimates of Chromosomal Copy Number Alterations

Minjares JM* and Shmaliy YS

Department of Electronics Engineering, Universidad de Guanajuato, 36885, Salamanca, Mexico

Corresponding author:
Yuriy S Shmaliy

shmaliy@ugto.mx

Department of Electronics Engineering, Universidad de Guanajuato, 36885, Salamanca, Mexico.

Tel: +524731020100

Citation: Minjares JM, Shmaliy YS (2017) Matching Confidence Masks with Experts Annotations for Estimates of Chromosomal Copy Number Alterations. J Genet Disord. Vol. 1 No. 1:9

Abstract

Structural aberrations (SAs), gains or losses in large segments of genomes, are associated with several genetic disorders. The SAs are commonly called the copy number alterations (CNAs) and their identification/classification is required to identify diseases. Many methods have been proposed to estimate the breakpoints and segmental constants in the CNAs with highest precision using the most powerful technologies of hybridization. However, the locations and lengths of CNAs estimated using well-elaborated methods are often contradictive due to extensive variability of measurements and performances of the algorithm. Still much less attention is paid to the estimation accuracy and it is hard to select the best estimator. In this brief, we match the confidence masks earlier designed to improve the estimates with annotations made by several experts. Based upon, we specify the confidence probability for the masks.

Keywords: Structural aberrations (SAs); Copy number alterations; Breakpoints; Confidence Masks

Received: October 25, 2017; **Accepted:** November 13, 2017; **Published:** November 20, 2017

Introduction

Somatic aberrations are abnormal changes in DNA called copy number alterations (CNAs) associated with diverse genetic disorders, mainly with cancer disease [1,2]. Nowadays, a vast amount of technologies have been developed to measure the genome chromosomal structure: Array comparative genomic hybridization (aCGH) [3], high resolution CGH (HR–CGH) [4] and whole genome sequencing (WGS) [5] are among the most known. Nevertheless, the CNAs data obtained using the microarrays and other technologies are still affected by several factors: 1) 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 [6-8]. Consequently, intensive variability in measurements makes an estimator, optimal or robust, unable to produce a reliable estimate [9]. Although the identification of CNAs associated with cancer is very challenging, still no one estimator can guarantee an existence of the detected changes. To guarantee that an estimator has detected chromosomal changes accurately, the interpretation made by medical experts has been considered in [10] as a gold standard, although it has been recognized in [11] that such opinions are time consuming and not necessarily very accurate. Therefore, as has been shown in [8,12,13], testing estimates by confidence masks becomes challenging for medical applications. The masks allow improving the estimates for the required confidence probability by removing some CNAs, which do not match annotations made by experts.

Methods

To guarantee an existence of CNAs with a required probability, an efficient algorithm computing the upper and lower boundary confidence masks has been proposed in [12]. The method is based on using the skew Laplace law representing the jitter distribution in the CNA breakpoints. Later, this algorithm was essentially improved referring to the fact that the Laplace distribution becomes highly inaccurate when the segmental signal-to-noise ratio (SNR) ranges below unity [14]. With respect to the breakpoint, the segmental SNRs are calculated as:

$$\gamma_l^- = \frac{\Delta_l^2}{\sigma^2}, \gamma_l^+ = \frac{\Delta_{l+1}^2}{\sigma_{l+1}^2}$$

Where, $\Delta_l = a_{1+1} - a_l$, is the segmental difference and σ_l^2

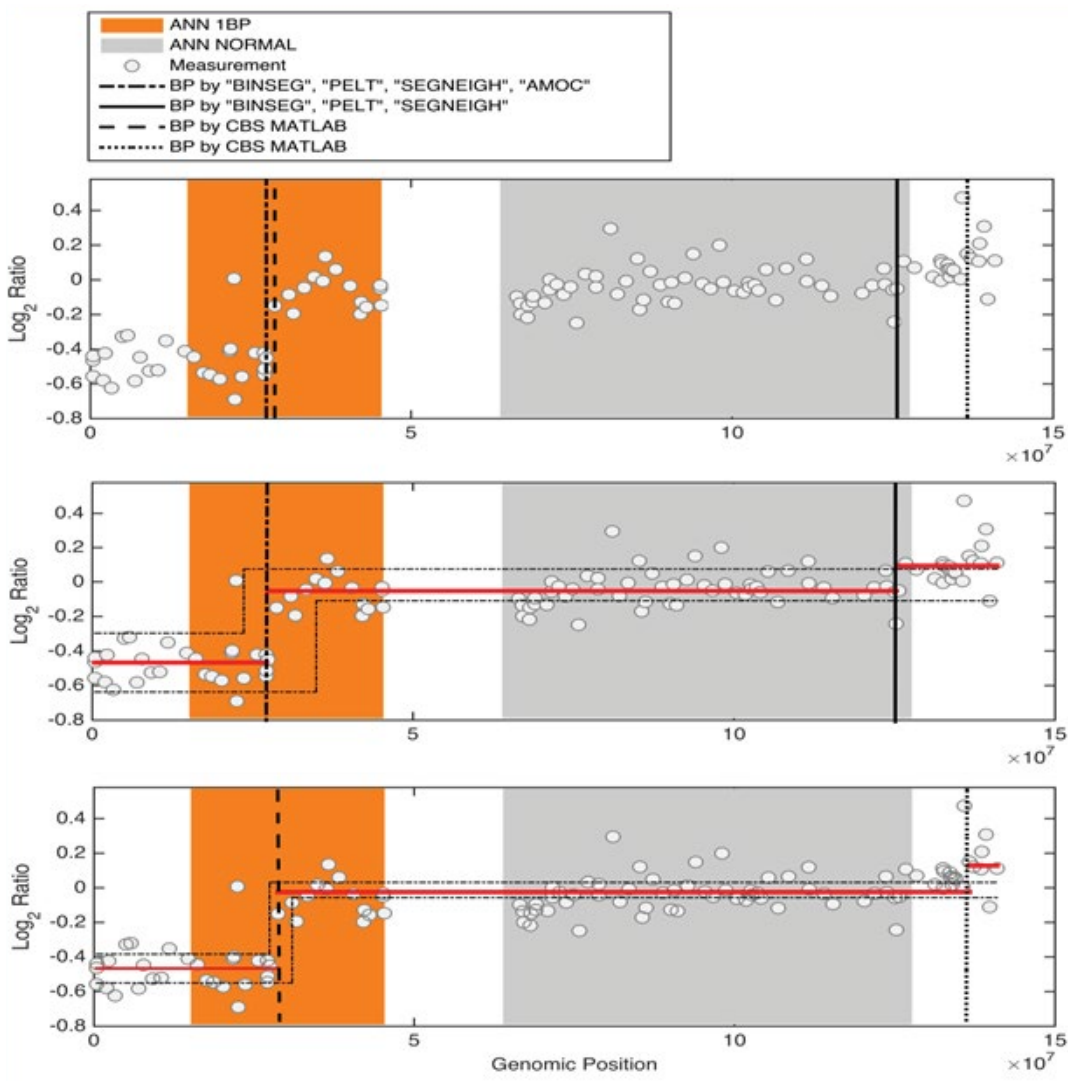

Figure 1 Analysis of annotation regions made by experts and breakpoints found by several algorithms using the confidence masks. a) Breakpoints found by several algorithms: BINSEG, PELT, SEGNEIGH, AMOC and CBS and two annotation regions one breakpoint 1BP and normal. b) By applying the confidence masks to breakpoints found by BINSEG, PELT and SEGNEIGH, the second breakpoint. is removed at a probability of P = 1-4.11e-11. c). The second breakpoint found by CBS algorithm unlikely exist to a lesser probability of P =1-9.66e-04.

and σ^2_{l+1} are the segmental variances, which correspond to measurements to the left (l)-segment and right (l+1)-segment. The confidence masks are designed using statistical properties of the segmental noise and the breakpoint jitter to sketch the upper and low boundaries for the possible estimates. The masks can be formalized for the required probability, which will be associated below with the expert annotations.

Application

To specify the probability for the confidence masks, data used below are taken from the database of 575 annotated neuroblastoma copy number profiles, which are available from the public benchmark for testing new algorithms. We apply the algorithms to detect the CNA breakpoints in Chromosome 9 of Profile 1. The BINSEG, PELT, SEGNEIGH and AMOC algorithms use a penalty value of "0.05*log (n)," where n is the length of probes in the Chomosome 9, while the CBS algorithm detect the breakpoints automatically. The confidence masks applied are based on the asymmetric exponential power distribution with the probability ranging from 0.5 to 1. The detected CNA breakpoints, the SNR levels, and the expert's probabilities listed in (Table 1and Figure 1) illustrate the efficiency of the proposed masks. As can be seen, all estimators detect the first breakpoint. However, the second breakpoint is detected only by BINSEG, PELT, SEGNEIGH and CBS algorithms, although the breakpoint detected by CBS has a different location. The experts have noticed an existence only the first breakpoint shown in (**Figure 1a**). The confidence masks, displayed in **(Figure 1b)**, suggest that the second breakpoint does not exist with the probability of P=1-4.11e-11. For the CBS estimate shown in **(Figure 1c)**, the second breakpoint unlikely exists with the probability less than P=1-9.66e-04. These probabilities should be considered as confidence for the relevant estimators to match the experts' notations.

Table 1 The breakpoints noticed by experts and the relevant probabilities for the confidence masks.

Method	Breakpoint genomic position	Levels of SNR	Matched probability to expert annotation
BINSEG, PELT, SEGNEIGH	27440311	$\gamma-$ = 1.70, $\gamma+$ = 1.46	1-4.11e-11
	125714134	$\gamma-$ = 1.70, $\gamma+$ = 1.46	
CBS	28766431	$\gamma-$ = 11.6, $\gamma+$ = 14.7	1-9.66e-04
	136658606	$\gamma-$ = 1.77, $\gamma+$ = 1.44	
AMOC	27440311	$\gamma-$ = 11.6, $\gamma+$ = 11.9	—

Results

Estimates of the CNAs produced by different algorithms are often inconsistent due to various factors affecting probing. The confidence masks algorithm is appropriate to analyze the CNA measurements because the disturbances can be modeled as Gaussian noise. Accuracy of the CNAs estimates depends mostly on two factors: 1) segmental SNRs and 2) probe data length. For SNR>>1, the breakpoints can be identified in a visual way; therefore, the masks upper and lower boundaries computed for the skew Laplace distribution are highly accurate. Otherwise, when SNR<1, the CNV identification faces difficulties, both by experts and the masks, and the estimators may produce low confidence results. In this work, we have shown that the expert biologists are unable to identify the breakpoints with the confidence probability lesser than 99.9%. But if to accept the expert's probability as the confidence one, then the masks will be able to make conclusions about the estimates matching the experts' opinions.

Note that much subtler chromosomal changes can be efficiently examined using the masks with a smaller confidence probability.

Discussion

The confidence masks are intended to revise the CNVs estimates with the confidence probability matching annotations of medical experts. The propose method accounts for two factors: errors in the confidence masks and the database size. Firstly, the confidence masks still require more accurate approximations for the jitter distribution in the CNA's breakpoints. Because the jitter distribution has been approximated heuristically, an accurate mathematical model is still required to decrease errors. Secondly, an acceptable confidence probability can be justified over a much larger number of annotations made by different qualified experts.

Conclusion

Solving problems mentioned in Discussion may open new horizons in further improvements of the CNVs estimates produced by different estimators Justified a reliable experts-based confidence probability, the confidence masks may play a crucial role in detecting actual chromosomal changes.

References

1 Graham NA, Minasyan A, Lomova A, Cass A, Balanis NG, et al. (2017) Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. Mol Syst Biol 13: 914.

2 Weinberg RA (2007) The biology of Cancer. In: London: Gallard Science Taylor & Francis Group, LLC.

3 Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP, et al. (1997) Genome screening by comparative genomic hybridization. Trends Genet 13: 405-409.

4 Speicher MR, Carter NP (2005) The new cytogenetics: Blurring the boundaries with molecular biology. Nat Rev Genet 6: 782-792.

5 Ng PC, Kirkness EF (2010) Whole genome sequencing. Methods Mol Biol 628: 215-226.

6 Zare F, Dow M, Monteleone N, Hosny A, Nabavi S, et al. (2017) An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics 18: 286.

7 Popova T, Boeva V, Manie E, Rozenholc Y, Barillot E, et al. (2013) Analysis of somatic alterations in cancer genome: From SNP arrays to next generation sequencing. In: Sequence and Genome Analysis I Humans, Animals and Plants. Ltd IP (ed) iConcept Press Ltd.

8 Munoz JU, Shmaliy YS (2017) Estimates of the breakpoints in genome copy number alteration profiles with masks. In: Biomed. Signal Process Contr 10: 238-248.

9 Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207-211.

10 Hocking TD, Schleiermacher G, Janoueeix–Lerosey I, Boeva V, Cappo J, et al. (2013) Learning smoothing models of copy number profiles using breakpoint annotations. BMC Bioinformatics 14: 164.

11 Tibshirani R, Wang P (2007) Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics 9: 18-29.

12 Muñoz JU, Cabal J, Shmaliy YS (2014) Confidence masks for genome DNA copy number variations in applications to HR-CGH array measurements. Biomed Signal Process Contr 13: 337-344.

13 Munoz JU, Shmaliy YS, Cabal J (2014) Confidence limits for genome DNA copy number variations in HR- CGH array measurements. Biomed Signal Process Contr 10: 166-173.

14 Munoz JU, Shmaliy YS (2017) Improving estimates of genome CNVs with confidence masks using SNP array data. Biomed Signal Process Contr 31: 238-248.