Bioinformatics, 33(15), 2017, 2384–2385
doi: 10.1093/bioinformatics/btx212
Advance Access Publication Date: 22 April 2017
Applications Note

OXFORD

Genome analysis

modSaRa: a computationally efficient R package for CNV identification

Feifei Xiao[1], Yue Niu[2], Ning Hao[2], Yanxun Xu[3], Zhilin Jin[3] and Heping Zhang[4,*]

[1]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29201, USA, [2]Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA, [3]Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD 21218, USA and [4]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 15, 2017; revised on April 4, 2017; editorial decision on April 5, 2017; accepted on Month 0, 0000

Abstract

Summary: Chromosomal copy number variation (CNV) refers to a polymorphism that a DNA segment presents deletion or duplication in the population. The computational algorithms developed to identify this type of variation are usually of high computational complexity. Here we present a user-friendly R package, modSaRa, designed to perform copy number variants identification. The package is developed based on a change-point based method with optimal computational complexity and desirable accuracy. The current version of modSaRa package is a comprehensive tool with integration of preprocessing steps and main CNV calling steps.

Availability and Implementation: modSaRa is an R package written in R, C++ and Rcpp and is now freely available for download at http://c2s2.yale.edu/software/modSaRa.

Contact: heping.zhang@yale.edu

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

1 Introduction

Copy number variation refers to a polymorphism that a segment of DNA sequence presents deletion or duplication in the population. Modern technologies now increasingly allow accurate identification of copy number variants (CNVs) and many studies have explored the association between them and risk of complex diseases (Fanale *et al.*, 2013; Poultney *et al.*, 2013; Zahnleiter *et al.*, 2013). Detection of CNVs has been proven essential in understanding the influence in human complex diseases biologically and epidemiologically. With the development of modern technologies, various complex data have been generated hence efficient statistical methods and computational tools are highly demanded. Methods based on change-points searching has been investigated for CNV analysis for a long time, including binary segmentation method (Vostrikova, 1981), circular binary segmentation (CBS) (Olshen *et al.*, 2004), and exhaustive search methods (Braun *et al.*, 2000; Yao, 1989). However, most of these popular methods limit their usage in modern big data analysis with computational complexity usually at least $O(n^2)$.

More recently, a screening and ranking algorithm (SaRa) was proposed with computational complexity$O(n)$ (Niu and Zhang, 2012). However practical issues arise from the application of this method in array-based data, including over-sensitivity and robustness of the normal based model. To address these issues, a modified SaRa method (modSaRa) was proposed as an endeavor to detect CNVs with optimal performance in practice (Xiao *et al.*, 2015). Multiple bandwidths are used in change-point search to ensure better coverage. The candidate segments are clustered to CNVs via a Gaussian mixture model based clustering step. More details of the algorithm and applications implemented in modSaRa are described in the method paper (Xiao *et al.*, 2015). A usable code of modSaRa was made available to the public.

Here, we present a comprehensive package of the modSaRa method. Compared to the originally release, this version includes

useful new features as follows. We make use of Rcpp for a seamless integration of C++ and R which greatly increases the computational speed. A user-friendly manual is embedded in the installation package that facilitates readers' usage. The two preprocessing procedures are implemented for greater convenience than the original release. Importantly, CNVs are known to be important risk factors in cancer research. In this application note, we present an overview of the package and include a real dataset to illustrate the usage of modSaRa. The Additional file 1 contains a manual (the vignette for the package).

2 Methods and implementation

modSaRa formulates the analysis of array-based data as a problem of detecting change points. Let $\boldsymbol{Y}$ be the data. A high dimensional normal mean based model was proposed.

$$Y_i = \mu_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ 1 \leq i \leq n \qquad (1)$$

where $\boldsymbol{\mu}$ is the vector for the underlying mean. A vector $\boldsymbol{\tau}$ consists of change-point locations. The major task is to search for the number and the exact locations of change points. Next, modSaRa adopts screening and ranking algorithm with essential implementations to accurately identify change points, including Gaussian mixture model based clustering approach to filter false positives.

We then developed and integrated a set of functions in an R package called modSaRa. modSaRa is executable for both Windows and Linux. The package consists of two main components, preprocessing procedures and main CNV calling step (Supplementary Information). We implement the preprocessing steps in Rcpp (an R package to seamlessly integrate C++ and R) which greatly improves the speed. The modSaRa package is freely available at http://c2s2.yale.edu/software/modSaRa.

3 Example usage

Two input files are required: a signal intensity file and a map file. The signal intensity file contains information for one marker per line. Each column represents data for one sequence. The map file provides information about marker name, chromosome ID and coordinates. Detailed instruction and output format can be found in the Additional file 1 and Supplementary information.

Figure 1 displays the plot of signal intensities of two CNVs identified by modSaRa with a melanoma study (Amos *et al.*, 2011). Shown in the figure are two CNVs identified from two individuals on the same chromosomal region. LRRs usually cluster around zero when no duplications or deletions are present. In the identified CNV region, LRRs are obviously below zero, which indicates that these are deletions. Technically, the clustering pattern of BAFs also provides subtle information about the state. For deletions, they usually cluster around 0 and 1 (without cluster around 0.5). For the normal state, they usually congregate around 0, 0.5 and 1. Combining the information from the LRR and BAF distributions, the two CNVs were valid calling results.

4 Significance and conclusions

Recent advances in sequencing technology greatly reduce cost in next generation sequencing (NGS), and therefore, require more computationally efficient methods such as modSaRa. For NGS data, the problem becomes to find the break points along the sequences which are in read counts. Extension of the modSaRa to NGS data will be complicated yet plausible if the discreteness of the intensities is appropriately modeled, which warrants further development.

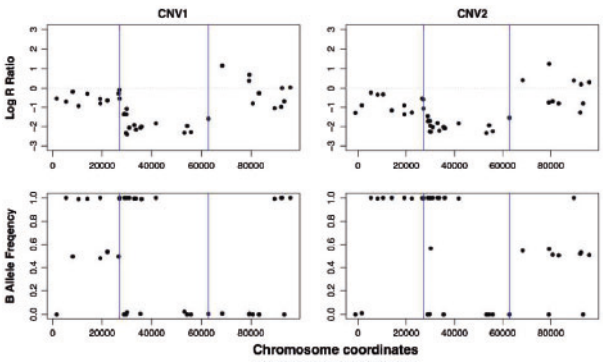

Fig. 1. CNVs identified by modSaRa from a melanoma study. The region between two vertical lines are identified CNVs. The horizontal axis indicates relative map locations on the chromosome and the vertical axis stands for LRR values (upper panel) and B allele frequency (BAF) (bottom panel)

Acknowledgements

We acknowledge Dr. Christopher I. Amos from Dartmouth College for providing data in the example usage.

Funding

This research is supported in part by grant R01DA016750 from the National Institute on Drug Abuse, National Science Foundation Grant DMS-1309507 and the internal fund from the University of South Carolina.

Conflict of Interest: none declared.

References

Amos,C.I. *et al.* (2011) Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum. Mol. Genet.*, **20**, 5012–5023.

Braun,J.V. *et al.* (2000) Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, **87**, 301–314.

Fanale,D. *et al.* (2013) Analysis of germline gene copy number variants of patients with sporadic pancreatic adenocarcinoma reveals specific variations. *Oncology*, **85**, 306–311.

Niu,Y.S. and Zhang,H. (2012) The screening and ranking algorithm to detect dna copy number variations. *Ann. Appl. Stat.*, **6**, 1306–1326.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Poultney,C.S. *et al.* (2013) Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.*, **93**, 607–619.

Vostrikova,L.I. (1981) Detection of the disorder in multidimensional random-processes. *Dokl Akad Nauk Sssr+*, **259**, 270–274.

Xiao,F. *et al.* (2015) Modified screening and ranking algorithm for copy number variation detection. *Bioinformatics*, **31**, 1341–1348.

Yao,Y.C. A.S. (1989) Least-squares estimation of a step function. *Sankhya A*, **51**, 370–381.

Zahnleiter,D. *et al.* (2013) Rare copy number variants are a common cause of short stature. *PLoS Genet.*, **9**, e1003365.