BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 3 2011, pages 419–420
doi:10.1093/bioinformatics/btq668

Genome analysis
Advance Access publication December 5, 2010

Rseg—an R package to optimize segmentation of SNP array data

Philippe Lamy[1,2,*], Carsten Wiuf[2] , Torben F. Ørntoft[1] and Claus L. Andersen[1]
[1]Department of Molecular Medicine, Aarhus University Hospital Skejby, Aarhus N and [2]Bioinformatics Research Center, Aarhus University, Aarhus, Denmark

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: The use of high-density SNP arrays for investigating copy number alterations in clinical tumor samples, with intra tumor heterogeneity and varying degrees of normal cell contamination, imposes several problems for commonly used segmentation algorithms. This calls for flexibility when setting thresholds for calling gains and losses. In addition, sample normalization can induce artifacts in the copy-number ratios for the non-changed genomic elements in the tumor samples.

Results: We present an open source R package, Rseg, which allows the user to define sample-specific thresholds to call gains and losses. It also allows the user to correct for normalization artifacts.

Availability: The R package, Rseg, is available at: http://www.cs.au.dk/~plamy/Rseg/ and runs on Linux and MS-Windows.

Contact: plamy@cs.au.dk

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

Received on August 12, 2010; revised on November 24, 2010; accepted on November 25, 2010

1 INTRODUCTION

Affymetrix SNP arrays have been used to estimate genomic copy numbers in tumor samples (Beroukhim *et al*., 2010). Different methods have been developed to segment the genome into contiguous regions that share the same copy number on average. These include CBS, GLAD and more recently Ultrasome (Nilsson *et al*, 2009; Venkatraman and Olshen, 2007). Ultrasome is the fastest method and gives similar output as CBS. However, Ultrasome does not allow to correct for artifacts induced by the normalization of the arrays. In fact, when there is an imbalance between gains and losses, commonly used normalization methods like quantile and invariant set normalization introduce a systematic shift in the copy numbers estimated for normal regions so that they appear as if they have an altered copy number. That is, if a sample contains only two different copy numbers, then normalization will make the regions with no change in copy number appear as being slightly lost or gained (cf patient A in Fig. 1). While this problem could potentially be overcome using a different normalization procedure, an alternative would be to use the intensities distribution to correct for it.

Generally, segmentation algorithms apply a common threshold for calling gains and losses for all investigated samples. However, when applied to tumor samples this often imposes a problem, as several sample specific factors like the level of normal cell contamination, intra tumor heterogeneity, and technical noise affects the resolution with which different copy number levels can be discerned. This calls for an option to set sample specific thresholds.

In this article, we describe an R package that guides the user through the segmentation of each sample. Histograms of the log ratio copy number between a tumor and its matched germline sample can distinguish between different copy numbers (represented by different peaks—cf Fig. 1). Therefore, by modeling each peak and by selecting the peak representing segments with no copy number change and its neighbors, correction for the normalization induced copy number and optimal thresholds for calling gains and losses can be statistically defined.

2 DESCRIPTION

The Rseg package provides different tools to efficiently optimize the segmentation of copy number data produced using SNP arrays (cf Supplementary Figure S1 for an overview). It takes as an input a copy-number file (.cn), for example produced by aroma.affymetrix (Bengtsson *et al*., 2009). This file contains for each probe and for each array the log ratio of the intensity between a tumor sample and a matched germline sample (or an average germline sample).

First Ultrasome is applied to segment the data with parameter $t=0.001$ (other algorithms can be used as well, e.g. CBS). Then, a multi-step protocol is applied to each sample:

- The segmented data is plotted as a histogram of weighted segment values (the weight is the number of markers in the segment).
- The user inputs the number of peaks and marks them manually. This step is difficult to automate due to often noisy nature of the data (cf patient B in Fig. 1).
- The program models each peak using a Laplace or a Gaussian distribution (Fig. 1, center section).
- The user selects the peak representing the segments with no-copy-number change and the two peaks surrounding it.
- The program uses a rigorous statistical procedure to define the shift (difference from the log-ratio = 0 to the mean of the log-ratios of the no-change peak) and the thresholds for calling gains and losses (defined by minimizing the false negative and false positive call rate).

The last step produce the final segmentation taking into account shift and thresholds determined for each sample (Fig.1, right section). The output file is a segmentation file (.seg) and can be

[*]To whom correspondence should be addressed.

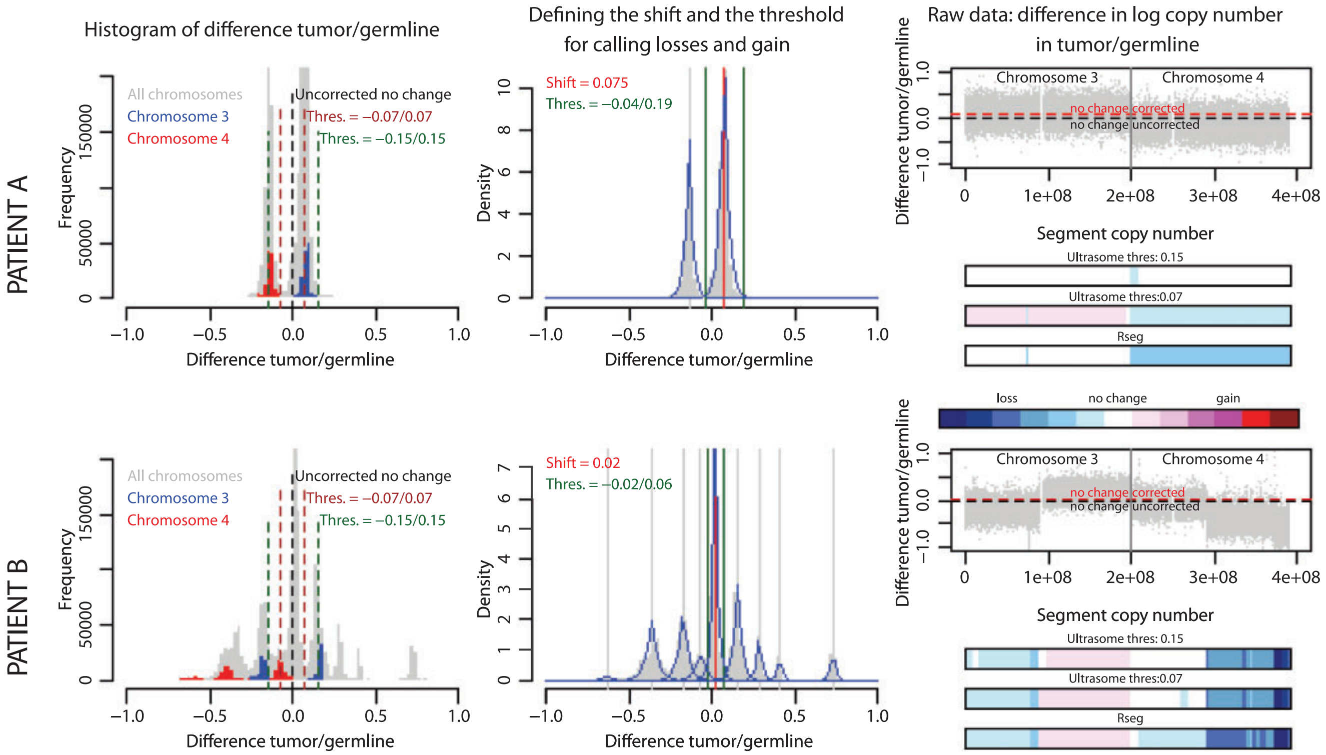

Fig. 1. Optimization of genomic segmentation for two tumor samples using Rseg. Initially data were segmented using Ultrasome with minimal threshold for calling gains and losses. On the left are shown histograms of the initial segment copy numbers (in grey) for all chromosomes, in blue and red are overlain the segment copy number for chromosome 3 and 4, respectively. The histograms distinguish different copy numbers in the sample as different peaks. To understand the effect of choosing a non-sample-specific threshold for calling abnormalities, the uncorrected no-change line and two different thresholds were plotted. In the middle part of the figure, the same histograms were plotted together with the shift of the no-change peak (in red) and the sample-specific thresholds (in green). Each peak was modeled as a Laplace distribution (in blue). The right part of the figure shows the raw data for chromosomes 3 and 4, and the final segmentation using Ultrasome (with two different thresholds: 0.15 and 0.07) and Rseg.

visualized using the Integrative Genomics Viewer (IGV) software (http://www.broadinstitute.org/igv).

3 METHODS

Prerequisites Ultrasome (can be downloaded from the Broad Institute web pages).

Input: Rseg contains different methods. Some requires a segmentation file (.seg) that can be produced by Ultrasome or some other segmentation algorithm and others requires a copy-number file (.cn).

Functions: Before the segmentation, noise reduction e.g. by smoothening the copy-number file (.cn) can be performed using the function *smooth_CNfile*. The segmentation can be done in R by using the function *run_ultrasome* where the arguments are the same as required by Ultrasome. The definition of the sample-specific shift and threshold for calling gains or losses is done by using the function *segment_all* or the function *modelGL* for a single sample using Laplace functions and/or Gaussian functions. The final segmentation is done using the function *apply_coef* or the function *modify_seg* for a single sample. Finally, you can run all these steps in one go by using the function *runRseg* (see the help files for more information).

Output: The main output is the final segmentation file where a combination of sample-specific shift correction and threshold for calling gains and losses has been used. This file can be visualized using IGV. Other outputs include a pdf file for each sample showing the distribution of the weighted segments, its modelization and the shift and thresholds applied and a txt file containing information about the parameters and some statistics.

Funding: This work was supported by the Nordic Center of Excellence, the EU project GENICA, The John and Birthe Meyer Foundation, The Danish Council for Independent Research Medical Sciences, the Lundbeck Foundation, the Danish Cancer Society and The Danish Ministry of the Interior and Health.

Conflict of Interest: none declared.

REFERENCES

Bengtsson,H. *et al*. (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**, 2149–2156.

Beroukhim,R. *et al*. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Integrative Genomics Viewer (IGV). Available at http://www.broadinstitute.org/igv.

Nilsson,B. *et al*. (2009) Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*, **25**, 1078–1079.

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