BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 2 2011, pages 268–269
doi:10.1093/bioinformatics/btq635

Genome analysis

Advance Access publication November 15, 2010

Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization

Valentina Boeva[1,2,3,4,*], Andrei Zinovyev[1,2,3], Kevin Bleakley[1,2,3], Jean-Philippe Vert[1,2,3], Isabelle Janoueix-Lerosey[1,4], Olivier Delattre[1,4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, [2]INSERM, U900, Paris, F-75248, [3]Mines ParisTech, Fontainebleau, F-77300 and [4]INSERM, U830, Paris, F-75248 France

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: We present a tool for control-free copy number alteration (CNA) detection using deep-sequencing data, particularly useful for cancer studies. The tool deals with two frequent problems in the analysis of cancer deep-sequencing data: absence of control sample and possible polyploidy of cancer cells. FREEC (control-FREE Copy number caller) automatically normalizes and segments copy number profiles (CNPs) and calls CNAs. If ploidy is known, FREEC assigns absolute copy number to each predicted CNA. To normalize raw CNPs, the user can provide a control dataset if available; otherwise GC content is used. We demonstrate that for Illumina single-end, mate-pair or paired-end sequencing, GC-contentr normalization provides smooth profiles that can be further segmented and analyzed in order to predict CNAs.

Availability: Source code and sample data are available at http://bioinfo-out.curie.fr/projects/freec/.

Contact: freec@curie.fr

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

Received on June 8, 2010; revised on October 28, 2010; accepted on November 9, 2010

1 INTRODUCTION

In many studies that apply deep sequencing to cancer genomes, one has to calculate copy number profiles (CNPs) and predict regions of gain and loss. There exist two frequent obstacles in the analysis of cancer genomes: absence of an appropriate control sample for normal tissue and possible polyploidy. Most current tools do not take these points into account (Supplementary Table 1). For various reasons, sequencing of an appropriate control sample is not always possible. There is therefore a need for a bioinformatics tool able to automatically detect copy number alterations (CNAs) without use of a control dataset. Several programs have been published that allow automatic calculation and analysis of CNPs (Chiang *et al.*, 2009; Xie and Tammi, 2009). However, both CNV-seq (Xie and Tammi, 2009) and SegSeq (Chiang *et al.*, 2009) need datasets for the given tumor and its paired normal DNA. Moreover, both programs predict CNAs without providing information about how many copies were lost or gained. An interesting approach for predicting copy number variants was suggested by Yoon *et al.* (2009), where GC content is used to normalize data. However, to estimate the 'normal' copy number, they rely on the assumption that there are similar percentages of amplified and deleted regions, which is not true in general for cancer cells. Moreover, their tool was designed to analyze normal human genomes and is unable to take into account possible polyploidy.

Here, we propose an algorithm to call CNAs with or without a control sample. The algorithm is implemented in the C++ program FREEC (control-FREE Copy number caller). FREEC uses a sliding window approach to calculate read count (RC) in non-overlapping windows (raw CNP). Then, if a control sample is available, the program normalizes raw CNP using the control profile. Otherwise, the program calculates GC content in the same set of windows and performs normalization by GC content. Since this removes a major source of variability in raw CNPs (Chiang *et al.*, 2009; Yoon *et al.*, 2009), the resulting normalized profile becomes sufficiently smooth to apply segmentation. This is followed by the analysis of predicted regions of gains and losses in order to assign copy numbers to these regions.

2 METHODS

The algorithm includes several steps. First, it calculates the raw CNP by counting reads in non-overlapping windows. If not provided by the user, window size can be automatically selected using depth of coverage information to optimize accuracy of CNA prediction. The second step is profile normalization. If a control is not provided by the user, we compute the GC-content profile. The normalization procedure of RC by GC content (or by control RC) is described below. The third step is segmentation of the normalized CNP. To do this, we implemented a LASSO-based algorithm suggested by Harchaoui and Lévy-Leduc (2008). Segmentation provided by this algorithm is robust against outliers, which makes it suitable for segmentation of deep-sequencing CNPs. The last step involves analysis of segmented profiles. This includes identification of regions of genomic gains and losses and prediction of copy number changes in these regions.

To normalize a raw CNP, we fit the observed RC by the GC content (or the control RC if it is available). We base our fitted model on several assumptions: (i) the sample main ploidy P is provided, (ii) the observed RC in P-copy regions (i.e. regions with copy number equal to P) can be modeled as a polynomial of GC content (or of control RC), (iii) the observed RC in a region with altered copy number is linearly proportional to the RC in P-copy regions and (iv) the interval of measured GC contents (respectively control RCs when a control dataset is available) in the *main ploidy regions* must include the interval of *all* measured GC contents (respectively control RC). The polynomial's degree is a user-defined parameter with a default value of three. We provide an initial estimate of the polynomial's parameters and then optimize these parameters by iteratively selecting data points related to P-copy regions and making a least-square fit on these points only (See Supplementary Methods for more details). The resulting polynomial

*To whom correspondence should be addressed.

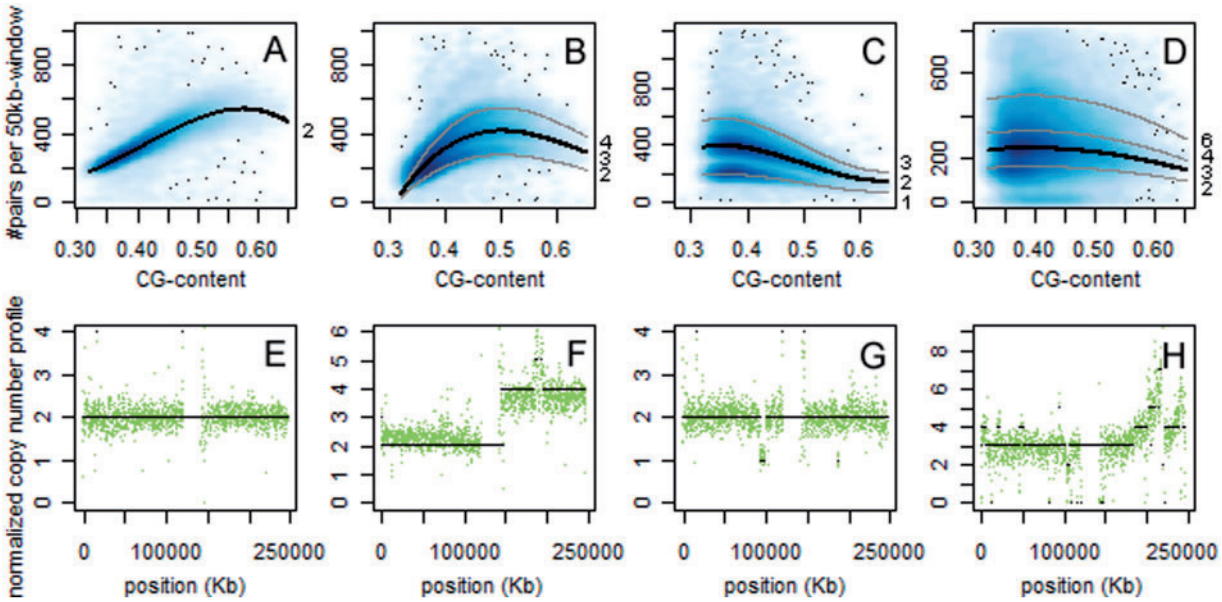

Fig. 1. Normalization of CNPs using only information about average GC content in a window. (**A**–**D**) GC content versus RC in 50 kb windows for COLO-829BL (normal diploid genome), COLO-829, NCI-H2171 and HCC1143, respectively. The result of the least-square fit for *P*-copy regions is shown in black. Curves corresponding to other frequent copy numbers are shown in gray. Values of copy numbers are given at the right of each panel. Chromosomes X and Y were not included. (**E–H**) GC-content normalized CNPs for chromosome 1 for COLO-829BL, COLO-829, NCI-H2171 and HCC1143, respectively. Automatically predicted copy numbers are shown in black.

is then used to normalize the CNP (Fig. 1). The user has an option to include mappability information into the normalization procedure (See Supplementary Methods).

3 RESULTS

We applied the method to predict CNAs in mate-pair datasets for the melanoma cell line COLO-829 and matched normal cell line COLO-829BL (Pleasance *et al.*, 2010), a paired-end dataset for the small-cell lung cancer cell line NCI-H2171 (Campbell *et al.*, 2008) and a single-end dataset for the breast cancer cell line HCC1143 (Chiang *et al.*, 2009). All four samples were sequenced using the Illumina Genome Analyzer platform. The number of reads in samples varied from 14 to 20 million (Supplementary Table 2).

The polynomial fit by GC content explained well the observed RC (Fig. 1A–D). Using CNPs normalized by GC content, we identified regions of gain and loss in the four samples (Fig. 1E–H, Supplementary Fig. 1–4). We also assessed true positive and false positive rate for a normal diploid sample NA18507 (Alkan *et al.*, 2009; Bentley *et al.*, 2008; Supplemenary Table 3).

We compared FREEC with three other existing tools: CNV-seq, SegSeq and RDXplorer (Supplementary Tables 1 and 4). As well as providing other additional functionalities, FREEC understands more input formats than any other tool. It can be used to analyze data produced for any organism and for polyploid genomes. Being implemented in C++, FREEC shows excellent performance and operating system portability.

4 CONCLUSION

We have presented a tool for automatic detection of CNAs and calculation of CNA frequency. FREEC provides more functionalities than existing tools; in particular, it can deal with the situation when no control experiment is available and when the genome is polyploid, frequent problems in cancer studies. The main steps are (i) normalization of the CNP using GC content (or control CNP if available), (ii) segmentation of normalized profiles and (iii) assignment of copy number changes to losses and gains. The program is fast, accurate and freely available.

Funding: The Ligue Nationale contre le Cancer (V.B., A.Z., E.B., I.J.-L. and O.D. are members of a labeled team).

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53–59.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Harchaoui,Z. and Lévy-Leduc,C. (2008) Catching change-points with lasso. *Adv. Neural Inform. Process. Syst.*, **20**, 617–624.

Pleasance,E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.