BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 19 2012, pages 2512–2514
doi:10.1093/bioinformatics/bts453

Genome analysis

Advance Access publication July 18, 2012

VegaMC: a R/bioconductor package for fast downstream analysis of large array comparative genomic hybridization datasets

Sandro Morganella[1,2] and Michele Ceccarelli[1,2,*]

[1]Department of Science, University of Sannio, 82100 Benevento, Italy and [2]Bioinformatics CORE, BIOGEM s.c.a.r.l.,
Contrada Camporeale, 83031 Ariano Irpino, Italy

Associate Editor: Alex Bateman

ABSTRACT

Summary: Identification of genetic alterations of tumor cells has become a common method to detect the genes involved in development and progression of cancer. In order to detect driver genes, several samples need to be simultaneously analyzed. The Cancer Genome Atlas (TCGA) project provides access to a large amount of data for several cancer types. TGCA is an invaluable source of information, but analysis of this huge dataset possess important computational problems in terms of memory and execution times. Here, we present a R/package, called VegaMC (Vega multi-channel), that enables fast and efficient detection of significant recurrent copy number alterations in very large datasets. VegaMC is integrated with the output of the common tools that convert allele signal intensities in log R ratio and B allele frequency. It also enables the detection of loss of heterozigosity and provides in output two web pages allowing a rapid and easy navigation of the aberrant genes. Synthetic data and real datasets are used for quantitative and qualitative evaluation purposes. In particular, we demonstrate the ability of VegaMC on two large TGCA datasets: colon adenocarcinoma and glioblastoma multiforme. For both the datasets, we provide the list of aberrant genes which contain previously validated genes and can be used as basis for further investigations.

Availability: VegaMC is a R/Bioconductor Package, available at http://bioconductor.org/packages/release/bioc/html/VegaMC.html.

Contact: morganella@unisannio.it

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

Received on March 17, 2012; revised on June 8, 2012; accepted on July 12, 2012

1 INTRODUCTION

Genetic alterations, such as deletions, amplification and loss of heterozigosity (LOH), represent an important component of tumor initialization and progression. Technologies based on Array Comparative Genomic Hybridization (aCGH) enabled the study of copy number alterations (CNAs) in cancer and developmental disorders. Recently, high-density single-nucleotide polymorphism (SNP) genotyping arrays (SNP-CGH) using a combination of two genotyping parameters, a normalized intensity measurement and an allelic ratio, have been used to detect CNAs, including LOH, with a high level of accuracy and resolution. Genotyping data consist of two channel intensity data corresponding to the two alleles used to compute normalized intensity and allele frequency. Supplementary Figure S1 shows the typical analysis process that starting from raw data enables the detection of candidate genes. Analysis starts with processing of raw data for extraction of allele signal intensities. These intensity values are successively used to compute the log R ratio (LRR) and B allele frequency (BAF) associated with each probe. LRR and BAF are calculated by comparing the intensities of tumor samples with the intensities of the respective normal samples or alternatively with the intensities of a reference genome (such as, the HapMap reference genome) as also reported in the Supplementary Information. Data are then segmented or smoothed both to decrease the noise level and to aggregate adjacent probes into a single region. Generally, segmentation and data smoothing are performed on one sample at time and results are independent across the samples. The 'adjusted' dataset is the input for the next analysis tool that looks for recurrent CNAs. The distinction between regions functionally related with the tumor (driver CNA) and regions that are detected as altered by biological and experimental noise is a critical step for these tools. In order to deal with this problem, statistical frameworks have been widely used to compute the probability associated with driver mutations. Of course, these frameworks can gain statistical strength by the analysis of a large amount of samples. In this scenario, the TCGA project (Collins and Barker, 2007) represents an invaluable source of information, indeed, the TCGA database provides thousands of copy number samples for common cancer types. Given the resolution of SNP-CGH (with Affymetrix SNP6.0 ~2 millions of probes are observed for each sample), analysis of these large datasets is not a trivial task. In particular, for large datasets, memory and time requirements often represent enormous obstacles for most analysis tools. In the last step of the analysis, available online databases are used to get information on the genes overlapping the identified driver CNA.

Here, we present a R/bioconductor package that, exploiting a novel multichannel segmentation algorithm, identifies the driver genes from LRR and BAF observations (Supplementary Fig. S1). We test VegaMC both on synthetic data which have been previously used to compare performance of other algorithms and on two TCGA datasets: colon adenocarcinoma (COAD) and glioblastoma multiforme (GBM). The analyzed TCGA datasets are composed of 424 samples of COAD and 571 samples of GBM, most of the tumor samples are also matched with normal tissues for a total of 870 samples in the COAD dataset and 992 samples in the GBM dataset, the processing of these

*To whom correspondence should be addressed.

data represents a challenging task. Results are publicly available and they can be used for further investigations.

2 VEGAMC

VegaMC extends the variational segmentation algorithm proposed in Morganella *et al.* (2010) in order to simultaneously perform the segmentation of all the samples. The joint segmentation allows better results to be obtained both in terms of accuracy, since the detection of copy number breakpoints can be based on the evidence from multiple signals, and in terms of computational efficiency as reported in the results. Joint segmentation allows the enhancement of systematic biases leading to the appearance of consistent breakpoints. The segmented regions are then classified, sample by sample, as deletions, amplifications and LOH and a statistical framework is used to assess the statistical significance associated with driver CNA. The statistical framework is based on a conservative permutation test similar to the one previously described in Morganella *et al.* (2011) performed on at region level (see Supplementary Information for more algorithmic details).

One aim of VegaMC is to solve some critical points arising from the analysis process depicted in Supplementary Figure S1. The first issue of this analysis process is the integration of different software tools. Integration is not an easy step: software tools are implemented in different programming languages and output of a tool needs to be adapted so that it can be used as input for the next tool. In order to overcome this problem, VegaMC has been designed so that it is completely compatible with the most used tools that compute LRR and BAF, such as, PennCNV-Affy (Wang *et al.*, 2007) and BeadStudio (www.illumina.com). In addition, VegaMC performs all steps required for a comprehensive analysis of high-density aCGH data: from segmentation of LRR and BAF to the list of driver genes. All the computational demanding steps of VegaMC have been implemented in C, we report later the execution times on huge datasets. Below, we summarize the main features of VegaMC:

- Integration with the output of PennCNV-Affy and Beadstudio.
- Designed to overcome several steps of the analysis process: from LRR and BAF to the list of driver genes.
- Efficient implementation to guarantee execution times which can cope with very large datasets.
- User friendly: implemented as a R/Bioconductor package; web pages are used to allow a rapid navigation of the results; for each driver gene several pieces of information are provided (among them, the user finds the gene symbol, the genomic position, the cytoband, the description and the link to the respective ensembl web page).

3 RESULTS

In order to perform a quantitative evaluation of VegaMC, we used a public synthetic dataset (Morganella *et al.*, 2011) that simulates CNA in different resolution scenarios and in different noise conditions (biological and experimental noise was simulated to perturb the data, see Supplementary Information). Supplementary Tables S1 and S2 show that VegaMC provides good results in all simulated scenarios performance of VegaMC is always comparable to the best performance of other algorithms.

Qualitative evaluation of VegaMC has been performed on two TCGA datasets: COAD and GBM. Datasets are composed of 424 COAD samples and 571 GBM samples, and paired intensities of tumor-normal samples are used to compute LRR and BAF (see Supplementary Information). We report the time required to perform the analyses: ~14′ for COAD and ~18′ for GBM using a Linux server equipped with 2.00 GHz Xeon-Intel E7540 CPU. The results of the analysis can be accessed at the webpage http://www.dsba.unisannio.it/Members/ceccarelli/vegamc/. Results on COAD reveal common cytogenetic mutations, such as, amplification of 8q, 7, 13, 20q, X and deletion on 8p, 15q, 17p and 18q mutations which include several well-known tumor suppressors (*TP53*, *SMAD2*, *SMAD3* and *SMAD4*) and oncogenes (*MYC*, *GNAS*,*LMO7*) (Ashktorab *et al.*, 2010). Also, results on GBM contain many positive controls, such as, cytogenetic deletions of 9p, 10, 17q (including tumor suppressors *PTEN* and *CDKN2A*) and amplification of chromosome 7 (with evident instability of 7p11.2 that includes the oncogene *EGFR*) (McLendon *et al.*, 2008). In order to evaluate the biological coherence of the lists of driver genes, we investigated enriched canonical pathways and functions (Supplementary Fig. S2 and S3); for COAD, we find 'Pancreatic Adenocarcinome Signaling' to be significant and for GBM, we find both 'Glioma Signaling' and 'Glioblastoma Multiforme Signaling'.

Finally, Supplementary Table S3 reports an experimental comparison in terms of running times on synthetic and real dataset. The reported times show that VegaMC outperforms several alternative algorithms aimed at computing recurrent genomic aberrations.

4 CONCLUSIONS

We presented a new bioconductor/package for the analysis of high-density aCGH data. VegaMC is integrated with common LRR and BAF calculation tools and it performs all analysis steps required to obtain the list of driver genes by exploiting a novel multichannel segmentation algorithm characterized by efficiency and accuracy. It also enables an easy navigation of the results by html pages.

We performed analysis of two large TCGA datasets. The obtained gene lists contain several positive controls that highlight important signaling pathways. Results are publicly available for download and they can be used as basis for further biological investigations.

Funding: MiUR (Ministero dell'Universita' e della Ricerca) under (PRIN2008-20085CH22F).

Conflict of Interest: none declared.

REFERENCES

Ashktorab,H. *et al.* (2010) Distinct genetic alterations in colorectal cancer. *PLoS One*, **5**, e8879.

Collins,F.S. and Barker,A.D. (2007) Mapping the cancer genome. *Scientific Am.*, **296**, 5057.

McLendon,R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Morganella,S. *et al.* (2010) VEGA: variational segmentation for copy number detection. *Bioinformatics*, **26**, 3020–3027.

Morganella,S. *et al.* (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics*, **27**, 2949–2956.

Wang,K. *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.