BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 10 2014, pages 1484–1485
doi:10.1093/bioinformatics/btu028

Gene expression
Advance Access publication January 21, 2014

CAFE: an R package for the detection of gross chromosomal abnormalities from gene expression microarray data

Sander Bollen[1,2,*], Mathias Leddin[3], Miguel A. Andrade-Navarro[1] and Nancy Mah[1,*]

[1]Computational Biology and Data Mining Group, Max Delbruck Center for Molecular Medicine, 13125 Berlin, Germany, [2]Graduate School of Life Sciences, Utrecht University, Universiteitsweg 98, 3584 CG, Utrecht, the Netherlands and [3]Roche Diagnostics GmbH, 82377 Penzberg, Germany

Associate Editor: Janet Kelso

ABSTRACT

Summary: The current methods available to detect chromosomal abnormalities from DNA microarray expression data are cumbersome and inflexible. CAFE has been developed to alleviate these issues. It is implemented as an R package that analyzes Affymetrix *.CEL files and comes with flexible plotting functions, easing visualization of chromosomal abnormalities.

Availability and implementation: CAFE is available from https://bitbucket.org/cob87icW6z/cafe/ as both source and compiled packages for Linux and Windows. It is released under the GPL version 3 license. CAFE will also be freely available from Bioconductor.

Contact: sander.h.bollen@gmail.com or nancy.mah@mdc-berlin.de

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

Received on August 8, 2013; revised on January 10, 2014; accepted on January 14, 2014

1 INTRODUCTION

Gross chromosomal abnormalities are a hallmark of cancers (Hanahan and Weinberg, 2011) and are frequently acquired by cultured cells as an adaptation to cell culture conditions (Baker *et al*., 2007). Recently, it has been recognized that induced pluripotent stem cells often feature gross chromosomal duplications or deletions (Laurent *et al*., 2011).

Various methods exist to detect chromosomal gains or losses. Traditional karyotyping relies on careful examination of Giemsa-stained metaphase chromosomes. Newer techniques like spectral karyotyping have increased ease of analysis but nevertheless feature low resolution. For high-throughput and high resolution analysis of gross chromosomal abnormalities, array-based Comparative Genomic Hybridization (a-CGH) is often used. This a-CGH approach is based on the detection of a quantitative difference of DNA content. Whole-genome and SNP-based sequencing approaches have also been developed.

Although not initially developed for this purpose, it is possible to use gene expression microarray data for the detection of copy number abnormalities. This approach is not based on the measurements of DNA content but rather on mRNA expression levels. A protocol to use expression microarrays to ‘karyotype’ samples was recently published by Benvenisty and coworkers but requires the manual use of different tools (Ben-David *et al*., 2013).

Here, we present CAFE—Chromosomal Aberration Finder in Expression data—as an R package for the detection of gross chromosomal gains and losses from expression microarrays, with a resolution up to cytoband level. CAFE follows the expression-based karyotyping workflow (e-karyotyping) and greatly simplifies and speeds up the detection analysis of chromosomal aberrations from expression DNA microarrays.

2 FEATURES AND METHODS

The starting point of a CAFE analysis is a set of gene expression microarrays from samples whose e-karyotype will be computed and another (larger) set of microarrays representing controls. The controls define a normal e-karyotype against which the altered e-karyotype will be defined. We recommend choosing a dataset of at least 10 controls for 2–3 test samples. More controls may be required depending on the particular case. CAFE is implemented as an R package and relies on several Bioconductor packages. It runs on version 2.10 or newer of R. The analyses are performed using Affymetrix *.CEL files as input. Using the `ProcessCels()` function, a list object is created from these *.CEL files and returned to the user. This object contains normalized and relative expression levels, along with several mappings of probesets to chromosomes, chromosomal arms, cytobands and chromosomal locations. The output can be further filtered so as to exclude multiple probesets that map to the same gene or to the same location. CAFE can then be used to perform several enrichment tests for the detection of duplications or deletions of chromosomes, chromosomal arms and cytobands. Furthermore, several plot functions are available to visualize any detected aberrations.

2.1 Enrichment testing

CAFE contains three statistical functions that determine enrichment or depletion of a given chromosome/chromosomal region. One function exists for each of the three resolutions: `chromosomeStats()`, `armStats()` and `bandStats()`, corresponding to chromosomes, arms and cytobands, respectively. The ability of CAFE to detect aberrations within chromosome, arm or cytoband is dependent on the density of the microarray probesets within these areas. Areas that are gene-poor are not likely to be detected, as expression microarrays are designed to detect transcribed genes. The user defines two thresholds as a ratio of median expression values, ‘over’ and ‘under’, for which probesets are called over- and under-expressed. The threshold for genomic DNA hybridized onto a comparative genomic hybridization array would be $\pm\log_2(2)$ ratio to detect a chromosome gain or loss, respectively. Because CAFE uses mRNA expression as a surrogate for DNA copy number and because the levels of mRNA expression are variable and do not strictly reflect

*To whom correspondence should be addressed.

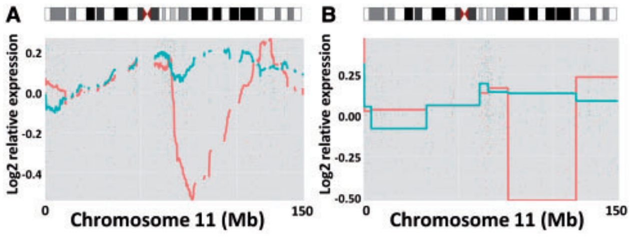

Fig. 1. Samples GSM652238 (orange) and GSM652239 (blue) from GEO dataset GSE26526 were plotted. GSM652238 is known to have a deletion in Chromosome 11q. (**A**) The output of `slidPlot()`. (**B**) The output of `discontPlot()`

DNA copy number, a less restrictive default threshold is recommended as a starting point for analysis: ±$\log_2(1.5)$ ratio. Using these thresholds, enrichment of under- and over-expressed probes in chromosomes or chromosomal regions is computed using a Fisher's exact test or a χ^2 test. *P*-values are Bonferroni-corrected by default.

2.2 Graphics

For graphics, CAFE uses the ggplot2 plotting system (Wickham, 2009). There are four different plotting functions available: (i) `rawPlot()`: the 'raw' log-transformed relative expression values are plotted along the chromosome of interest; (ii) `slidPlot()`: a moving average smoother is applied to the log-transformed relative expression values before plotting the values along the chromosome of interest; (iii) `discontPlot()`: a discontinuous smoother is applied to the log-transformed relative expression values and the values are plotted along the chromosome of interest; (iv) `facetPlot()`: all chromosomes are plotted in one horizontally aligned graph, with relative expression values along each chromosome. This function can be used in conjunction with a moving average smoother.

For all plot functions except `facetPlot()`, it is possible to add a chromosome idiogram over the chromosome plot. This allows easy visualization of chromosomal abnormalities. See Figure 1 for example plots. All plots are printed to the file system and returned as ggplot2 objects. Plot parameters, such as labels and scales, can be modified to the user's liking by altering the ggplot2 object.

2.3 Comparison with other tools

To the best of our knowledge, there are currently no other Bioconductor packages that are designed to identify chromosomal copy number abnormalities from mRNA expression microarray data. However, there are other packages that perform similar functions, such as processing comparative genomic hybridization arrays (aCGH, snapCGH) or identifying differentially regulated regions on a chromosome from expression microarrays (MACAT). The initial choice of datasets for analysis is critical and must be completed by hand, regardless of the package used. Once this is done, CAFE is able to normalize and preprocess the *.CEL files in one step. Although aCGH and snapCGH were designed to analyze CGH arrays, one can use CAFE to preprocess the *.CEL files and then subsequently reformat the preprocessed data for input into these packages. Both aCGH and snapCGH use hidden Markov models to predict state changes (i.e. changes in chromosomal copy number). Plotting functions then show the course of state changes over the chromosome. MACAT uses a modified *t*-statistic and permutation to score regions of the chromosome that are differentially regulated. The scores for a selected chromosome are shown as a static html page. To compare the performance, CAFE and the other three packages were used to analyze two test datasets with known chromosomal aberrations (see Supplementary Data). CAFE was able to detect copy number abnormalities just as well as or better than the other packages.

3 DISCUSSION

Karyotyping by expression microarrays, as described by Ben-David *et al.* (2013), extends the utility of expression microarray data by providing some limited information on the status of chromosomal aberrations in a sample. However, the original e-karyotyping method is a tedious process, using four different programs and requiring an estimated 15 h to analyze only 15 samples. At the time of writing, there are no Bioconductor packages to specifically carry out e-karyotyping from raw microarray data. The CAFE package simplifies the e-karyotyping protocol. Starting from the *.CEL files, CAFE can do the same analysis in minutes and requires no more than basic R knowledge. Bioconductor packages for CGH analysis can be used to perform an e-karyotyping analysis, but the data preprocessing steps must be carried out manually and the resulting graphs are static and not user-configurable. In contrast, CAFE processes the expression data from *.CEL files and all plotting functions return a ggplot2 object that can be modified by the end-user to his or her specific needs. CAFE will become a Bioconductor package, to be freely available for anyone to use, distribute, modify and easily integrate into an R-workflow for automatic analysis.

Currently, only Affymetrix *.CEL files from 3'IVT arrays can be seamlessly preprocessed in CAFE; functions for processing raw microarray data from other platforms may be added in the future if there is demand for this feature. Data import from other platforms is still possible, as CAFE data are represented by a simple R list structure. CAFE analysis currently works with three different resolutions: whole chromosome, chromosomal arm and cytoband. Therefore, it is not suited for smaller deletions or duplications. In addition, it is not readily possible to detect insertions or translocations using this technique, as the probesets will be mapped to the original chromosome or location. Therefore, CAFE is most suited for the detection of numerical abnormalities, rather than structural abnormalities. CAFE has not been designed to replace existing karyotyping techniques but rather to gain information when data from more specific approaches is not yet available.

Funding: German Research Foundation Priority Programme 1463 (Project No. AN 735/2-1) (to N.M.).

Conflict of Interest: none declared.

REFERENCES

Baker,D.E. *et al.* (2007) Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat. Biotechnol.*, **25**, 207–215.

Ben-David,U. *et al.* (2013) Virtual karyotyping of pluripotent stem cells on the basis of their global gene expression profiles. *Nat. Protoc.*, **8**, 989–997.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Laurent,L.C. *et al.* (2011) Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell Stem Cell*, **8**, 106–118.

Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer Science + Business Media, New York, USA.