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

Sander Bollen¹,²,* Mathias Leddin³, Miguel A. Andrade-Navarro¹ and Nancy Mah¹,*
¹Computational Biology and Data Mining Group, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany,
²Graduate School of Life Sciences, Utrecht University, Universiteitsweg 98, 3584 CG, Utrecht, the Netherlands and
³Roche Diagnostics GmbH, 82377 Penzberg, Germany

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.

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...
Enrichment of under- and over-expressed probes in chromosomes or 

DiscontPlot()

Data). CAFE was able to detect copy number abnormalities just as 

datasets with known chromosomal aberrations (see Supplementary 

ance, CAFE and the other three packages were used to analyze two test 

chromosome that are differentially regulated. The scores for a selected 

then show the course of state changes over the chromosome. MACAT 

changes (i.e. changes in chromosomal copy number). Plotting functions 

Both aCGH and snapCGH use hidden Markov models to predict state 

arrays, one can use CAFE to preprocess the *.CEL files and then subse-

sequently reformat the preprocessed data for input into these packages. 

For all plot functions except facetPlot(), it is possible to add a 

chromosome idrogram over the chromosome plot. This allows easy visu-

alization 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.

2.2 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 expres-

sion 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. 

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 pack-

ages 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 origi-

nal 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 infor-
mation 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 
ocogenesis 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.