# User’s guide: CALLME, a tool for detection of Clonal Mosaic Events (CME) in a parallelized fashion

## Introduction

We have developed a software that allows to perform all steps of CME detection analysis, from Beadstudio exported raw data to CME calling and plotting. It was designed to parallelize treatments in order to make the analysis faster. It currently works on Unix/Linux platforms.

This program implements five main steps of analysis using different existing programs:

1°) Conversion of Illumina data for analysis with tQN

2°) Thresholded quantile normalization with tQN

3°) GC content adjustment using PennCNV’s utility *genomic wave*

4°) Conversion of normalized data for analysis with R package MAD

5°) CME calling and plotting using MAD

## Requirements

### R

* R software version > 2.10
* R package MAD (sources are supplied in the present package)
* R Bioconductor limma package

### PERL

* Perl version 4.\*
* Perl modules :
  + Getopt::Long
  + Pod::usage
  + Cwd
  + POSIX
  + Carp
  + File::Spec
  + IO::File

### System

This program was successfully tested on three different Linux platforms, with RedHat –like operating systems, 2.6.\* kernels, 64 and 32 bits.

## Installation

MAD R package sources are available in the zip file containing the present package but can also be found at http://www.creal.cat/media/upload/arxius/jr/gada\_0.9-7.tar.gz.

* Open R
* Enter commands :
  + install.packages(“directory\_where\_gada\_sources\_are/gada\_0.9-7.tar.gz”)
  + library(gada)
  + source("http://bioconductor.org/biocLite.R")
  + biocLite("limma")
  + library(limma)
* Unzip file CALLME.zip
* Go to CALLME directory :
  + cd CALLME
* Compile splite\_samples.cc in order to get the split\_samples.exe executable:
  + g++ -O2 –lpthread –o split\_samples.exe split\_samples.cc

The executable file must be named split\_samples.exe .

## Input data

To use CALLME, only one file is needed. It can be exported from Illumina GenomeStudio, as follows.

* Open your project and go to the *Full Data Table* tab.
* Click the *Column chooser* icon
* Display columns *Name*, *Chr* and *Position* (in this order)
* Display subcolumns *X*, *Y* and *GType* (in this order)
* Optionally, but strongly recommended, filter SNPs and individuals (ex: keep only rs markers and individuals with call rate >= 80%)
* Click *Export data to a file* icon (third from the left)
* Choose file name
* When asked whether you would prefer to export the entire table, answer yes.

## This should export the displayed data to a tab delimited text file in the format required for analysis with CALLME.

## Command line execution

Here is an example of command line execution for an Illumina raw data file with the following path: path\_to\_illumina\_file. In this example, the data is extracted from Metabochip arrays.

* Go to CALLME directory
  + cd dir\_where\_CALLME\_is
* Execute CALLME.pl with following options:
  + perl CALLME.pl

--data\_file=path\_to\_illumina\_file

--n\_markers=200000

--rs\_only=true

--n\_procs=12

--T=2

--aAlpha=0.1

--min\_seg\_length=75

### Options :

* data\_file: path to the full data table produced by GenomeStudio export utility. It should be a tab delimited TXT file in the following format,

Name Chr Position Sample1.X Sample1.Y Sample1.GType ... SampleN.X SampleN.Y SampleN.GType

where Sample1 denotes the ID of the first sample.

* n\_markers: number of SNPs in the chip. It must be equal to or greater than the actual number of SNPs (i.e. the number of lines of data\_file + 1).
* rs\_only: if set to true, only SNPs beginning by rs will be kept for analysis, which is recommended for Metabochip arrays (default is false).
* n\_procs: number of process that can be launched in parallel. This option should be set depending on your operating system characteristics, especially the number of processors available. If you have 10 processors, it is advised to set this option to 7 or 8, so that the analysis does not consume all your system resources (default is 1).
* T : significance threshold for a segment to be called as CME (see MAD documentation)
* aAlpha : sensitivity parameter for segmentation algorithm (see MADdocumentation)
* min\_seg\_length : minimum number of consecutive probes for a segment to be called as CME (see MAD documentation)
* sd\_het\_BAF\_thres : threshold for heterozygous BAF standard deviation above which sample is to excluded from the analysis (default is 0.05)
* sd\_LRR\_thres : threshold for Log R Ratio standard deviation above which sample is to excluded from the analysis (default is 0.33)

## Output

All output files, including temporary files, will be written in the directory where Illumina input data file is located. In this directory, you should find:

* data\_check\_all.txt : a tab delimited file containing data quality metrics for each sample with following columns :
  + IID : sample ID
  + sdLRR : Log R Ratio standard deviation
  + sdHetBAF : heterozygous BAF standard deviation
  + callRate : sample overall call rate
* a directory named results\_T\_T\_aAlpha\_aAlpha\_minSegLength\_min\_seg\_length (where T, aAlpha and min\_seg\_length denote chosen options for CME calling) containing two files:
  + segments\_all.txt : a tab delimited file containing information about all segments called during the segmentation analysis with the following columns :
    - IniProbe : position of the first probe in the segment
    - EndProbe : position of the last probe in the segment
    - LenProbe : number of probes in the segment
    - chr : chromosome
    - LRR : mean Log R Ratio in the segment
    - sample : ID of the sample carrying the segment
    - LenKB : length of the segment (in KB)
    - SegID : segment ID
    - Status : segment status (copy-gain, copy-loss or copy-neutral)
    - pctCells : estimated percentage of abnormal cells
    - mu1 : mean BAF value of lower intermediate BAF band
    - mu2 : mean BAF value of upper intermediate BAF band
    - pctProbesG1 : percentage of segment probes located in lower intermediate BAF band
    - pctProbesG2 : percentage of segment probes located in upper intermediate BAF band
    - isMOSAIC : if yes, the segment is likely to be a true CME because it fulfills the following conditions :
      * 0.05 < pctCells <0.95 (sufficient evidence for a mixture of normal and abnormal cells)
      * pctProbesG1 & pctProbesG2 > 0.05 (there are enough probes in intermediate BAF bands)
      * mu1 < 0.5 & mu2 >0.5 (there exists two intermediate BAF bands)
      * min(0.5-mu1,mu2-0.5)/ max(0.5-mu1,mu2-0.5)  > 0.5 (intermediate BAF bands are symmetrically enough distributed around 0.5)
    - Datafile : path to the file containing data for the sample carrying this segment
  + possible\_mosaics.pdf : a pdf file showing representations of Log R Ratio and B Allele Frequency for each segment likely to be a true CME (isMosaic = Yes) and with sufficient data quality (sdLRR <0.33 & sdHetBAF < 0.05). This file is particularly useful to determine visually if a segment is a true CME or not.
* a directory named segmentation containing normalized data in one subdirectory for each sample, for use with MAD, or any further investigation.

## Memory consumption

A brief simulation study of the memory consumption of the program was carried out. Especially, the first step (conversion to individual format) of analysis consumes a lot of RAM, since the whole data file is stocked in a matrix. To avoid any “out of memory” error, an estimation of the RAM needed following the number of samples and the number of probes, using simulations, was performed.

If the memory needed exceeds the memory available, a message is written in the console and the user can choose a much slower – but much less memory-consuming - version of the conversion program.