# MethylAid-summarized data on 2800 Illumina 450k array samples

***Maarten van Iterson, Elmar Tobi, Roderick Slieker, Wouter den Hollander, Rene Luijk,***

Eline Slagboom and Bas Heijmans
Department of Molecular Epidemiology,
Leiden University Medical Center, Leiden, The Netherlands

**October 31, 2017**

## 1 Introduction

*MethylAidData* contains *MethylAid*-summarized data on 2800 Illumina 450k array samples. These DNA methylation samples are a subset from a large-scale multiple omics study conducted by several Dutch Biobanks; the BIOS consortium (http://www.bbmri.nl/en-gb/activities/rainbow-projects/bios). The raw Illumina 450k array data, idat-files, will be made available through the EGA archive (https://www.ebi.ac.uk/ega/home).

The summarization performed by *MethylAid* entails the following for each sample:

1. calculation of the median Methylated and Unmethylation intensities
2. extraction of all quality control probe intensities
3. construction of quality control metrics e.g. sample-dependent, sample-independent and detection p-values
4. storing everything efficiently to allow fast rendering of the various quality control plots provided by *MethylAid*,

see van Iterson *et al.*[1] for detailed description of *MethylAid*.

# 2 Preparation of the data

The raw Illumina 450k array data, idat-files, will be made available this summer from EGA archive (accession number:EGAS00###). Access to the data must be approved by the Data Access Committee (###). Once the raw idat-files have been downloaded and a targets file is constructed, *MethylAid* can be used to summarize the data and perform quality control using the interactive *shiny*[2] application.

Data sets of this size are preferably summarized in parallel and batches to overcome long run times or memory issues. *MethylAid* provides several options to do this using the *BiocParallel*-package[3]. For example, if multiple cores are available these could be used like this:

```
library(MethylAid)
targets ##constructed from EGA
BPPARAM <- MulticoreParam(workers = 8, verbose=TRUE)
summarize(targets, batchSize = 100, BPPARAM = BPPARAM, file="exampleDataLarge")
```

Another option would be thus use a cluster, see the vignette of *MethylAid* how to set this up.

# 3 Using MethylAidData

The summarized data contained in *MethylAidData* can be used in two ways, 1) to explore a large data set using *MethylAid* and 2) use this data as a background data set on top of own data. Since version 1.1.4, *MethylAid* has the functionality to show as background data set in the filter control plots. As such it can be used as a reference data set and can give guidance to when removing outlying samples. Furthermore, the data gives confirmation of the default thresholds used to determine outlying samples.

Additionally, since *MethylAid*(1.1.4) functionality is added to construct your own background data and several summarizedData-objects can be merged to give one larger summarizedData-object to use as your own reference or to determine filter thresholds, for example for hydroxymethylation data for which there are currently no thresholds available.

# References

[1] M. van Iterson, E. W. Tobi, R. C. Slieker, W. den Hollander, R. Luijk, P. E. Slagboom, and B. T. Heijmans. MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics*, 30(23):3435–3437, 2014.

[2] RStudio and Inc. *shiny: Web Application Framework for R*, 2014. R package version 0.9.1. URL: http://CRAN.R-project.org/package=shiny.

[3] Martin Morgan, Michel Lang, and Ryan Thompson. *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.0.3.