MethylAid-summarized data for Illumina 450k (N=2800) and EPIC (N=2620) arrays

***Davy Cats*[1], *Tyler J Gorrie-Stone*[2], *Bastiaan T Heijmans*[1], *John W Holloway*[3,4], *BIOS Consortium*[1], *Maarten van Iterson*[1], *Faisal I. Rezwan*[4], *and Leonard Schalkwyk*[2]**

[1]Dept. of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[2]School of Biological Sciences, University of Essex, Essex, UK.
[3]Human Development and Health, Faculty of Medicine, University of Southampton, UK
[4]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, UK

May 3, 2022

1 Introduction

MethylAidData contains *MethylAid*-summarized data on 2800 Illumina 450k array samples and 2620 Illumina EPIC/850k array that can be used as reference when processing newly generated methylation array data using *MethylAid*.

The data on 450k arrays is based on 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) [1]. The raw Illumina 450k array data, idat-files, are available through the EGA archive (https://ega-archive.org/dacs/EGAC00001000277).

MethylAid-summarized data for EPIC methylation arrays stems from studies led by the University of Southampton (N=1434) [2] and the University of Essex (N=1186).

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

1. calculation of the median Methylated and Unmethylated 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.*[3] for detailed description of *MethylAid*.

2 Preparation of the data

Using raw idat-files (e.g. EGA accession number EGAC00001000277 after approval by 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*[4] 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[5]. 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. J. et al Bonder. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.*, 49(1):131–138, Jan 2017.

[2] P. G. Burney, C. Luczynska, S. Chinn, and D. Jarvis. The European Community Respiratory Health Survey. *Eur. Respir. J.*, 7(5):954–960, May 1994.

[3] 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.

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

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