BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3435–3437
doi:10.1093/bioinformatics/btu566

Genetics and population analysis Advance Access publication August 21, 2014

MethylAid: visual and interactive quality control of large Illumina 450k datasets

Maarten van Iterson*, Elmar W. Tobi, Roderick C. Slieker, Wouter den Hollander, René Luijk, P. Eline Slagboom and Bastiaan T. Heijmans

Department of Molecular Epidemiology, Leiden University Medical Center, 2333 ZC Leiden, the Netherlands

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: The Illumina 450k array is a frequently used platform for large-scale genome-wide DNA methylation studies, i.e. epigenome-wide association studies. Currently, quality control of 450k data can be performed with Illumina's GenomeStudio and is part of a limited number 450k analysis pipelines. However, GenomeStudio cannot handle large-scale studies, and existing pipelines provide limited options for quality control and neither support interactive exploration by the user.

To aid the detection of bad-quality samples in large-scale genome-wide DNA methylation studies as flexible and transparent as possible, we have developed MethylAid; a visual and interactive Web application using RStudio's shiny package. Bad-quality samples are detected using sample-dependent and sample-independent quality control probes present on the array and user-adjustable thresholds. In-depth exploration of bad-quality samples can be performed using several interactive diagnostic plots. Furthermore, plots can be annotated with user-provided metadata, for example, to identify outlying batches. Our new tool makes quality assessment of 450k array data interactive, flexible and efficient and is, therefore, expected to be useful for both data analysts and core facilities.

Availability and implementation: MethylAid is implemented as an R/Bioconductor package (www.bioconductor.org/packages/3.0/bioc/html/MethylAid.html). A demo application is available from shiny.bioexp.nl/MethylAid.

Contact: m.van_iterson@lumc.nl

Received on May 27, 2014; revised on July 23, 2014; accepted on August 16, 2014

1 INTRODUCTION

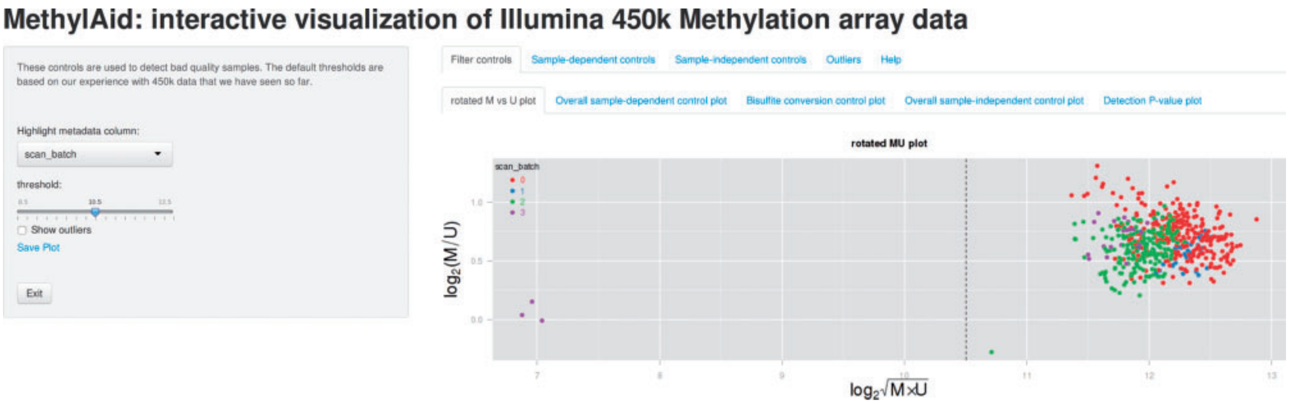
Because the introduction of the Illumina 450k Human Methylation BeadChip (Bibikova *et al.*, 2011), it has become the standard for large-scale genome-wide DNA methylation studies, i.e. epigenome-wide association studies (EWAS; Mill and Heijmans, 2013; Rakyan *et al.*, 2011). The array measures 482,421 CpG dinucleotides, covering 99% of genes (RefSeq; Bibikova *et al.*, 2011) and a broad range of genomic annotations. The principle of the array is that the methylation level of individual CpG dinucleotides is inferred using quantitative 'genotyping' of bisulfite-converted genomic DNA, a chemical reaction that converts unmethylated Cs into Ts, whereas methylated Cs are protected.

To monitor the quality of Illumina 450k data, the array contains 10 different types of quality control probes to evaluate the performance of both samples and specific steps in the process. The most frequently used type of quality control probe is the negative control that represents the background signal level. These are used to determine whether a signal significantly exceeds the background signal (represented as a so-called detection $P \leq 0.01$ or ≤ 0.05). However, current tools do not allow researchers to use the valuable information provided by the other control probes. The Illumina software GenomeStudio visualizes all quality control probes but is limited to small studies (typically <100), while tools developed to handle large studies primarily focus on data processing and analysis and mostly limited their use of the control probe information to the calculation of the detection *P*-values. Moreover, neither have the possibility for interactive exploration by the user. *MethylAid* was developed to address these limitations.

2 INTERACTIVE WEB APPLICATION

The R package *MethylAid* identifies poorly performing samples that are to be removed before processing (i.e. background, dye-bias and probe correction) and further analysis of the data. *MethylAid* takes raw intensity data (idat files) as input, summarizes the data and launches a *shiny*-based (RStudio and Inc., 2014) interactive Web application. The Web application produces multiple interactive diagnostic plots on the basis of the quality control probes present on the 450k array that facilitate the detection of bad-quality samples irrespective of sample size. Five interactive diagnostic plots are available in separate tab-panels to determine bad-quality samples using user-adjustable thresholds (although defaults are provided): a sample median Methylated versus Unmethylated signal intensity plot, a plot showing the efficiency of the bisulfite conversion, plots showing the overall quality of the arrays using sample-dependent and sample-independent control probes (non-polymorphic quality control probes and hybridization quality control probes) and a plot based on the negative quality control probes representing per sample fraction of probes above background. To facilitate the detection of bad-quality samples plots are rotated 45° like a MA-plot used to visualize expression array data (Fig. 1). In addition to the five interactive diagnostic plots, in-depth analysis of all quality control probe (sub)types on the 450k array is supported. In addition, metadata can be used to pinpoint outlying batches and single samples can be selected, which are

*To whom correspondence should be addressed.



MethylAid: interactive visualization of Illumina 450k Methylation array data

These controls are used to detect bad quality samples. The default thresholds are based on our experience with 450k data that we have seen so far.

Highlight metadata column: scan_batch

threshold: 10.5

8.5
12.5
□ Show outliers
Save Plot
Exit
Filter controls
Sample-dependent controls
Sample-independent controls
Outliers
Help
rotated M vs U plot
Overall sample-dependent control plot
Bisulfite conversion control plot
Overall sample-independent control plot
Detection P-value plot
rotated MU plot
scan_batch
● 0
● 1
● 2
● 3
1.0
0.5
0.0
$\log_2(M/U)$
7
8
9
10
11
12
13
$\log_2\sqrt{M \times U}$

Fig. 1. Screen shot *MethylAid* interactive Web application: the Web interface contains three parts; a panel with widgets to control the appearance of the quality control plots, tab-panels for choosing the quality control plot of interest and the interactive plotting area. When the interactive Web application is launched the rotated plot of the median methylated (M) and Unmethylated (U) log2 intensity is shown. Here, using example data from the package on 500 Illumina 450k human methylation samples with selected ‘scan_batches’ using the input selector widget ‘highlight metadata column’. The different ‘scan_batches’ are indicated with different colors. The dashed-vertical line indicated the filter threshold, samples below the threshold should be removed

subsequently highlighted across all other plots. For all plots, multiple views are available, including a sampleplot (as provided in GenomeStudio), a scatterplot or a boxplot. Each plot can be downloaded separately.

3 EXAMPLE

This example shows how to summarize a small set of idat files. The *minfiData* package provides such a set. The package contains 450k DNA methylation data on six samples across two groups. Because *MethylAid* uses internally the *minfi* (Aryee *et al.*, 2014) function `read.450k.exp`, target information should be provided in a similar way.

```
library(minfiData)
baseDir <- system.file("extdata", package =
"minfiData")
targets <- read.450k.sheet(baseDir)
library(MethylAid)
sdata <- summarize(targets)
visualize(sdata) ##this will launch the web
application
```

The function `summarize` performs the summarization of the control probes present on the array. The output produced by `summarize` should be passed on to the function `visualize`, which in turn will launch the interactive Web application.

MethylAid was specifically designed for quality control of large set of DNA methylation data. Summarization can be performed in batches to overcome memory problems by providing the `batchSize` option to the `summarize` function. To prevent excessively long run-times, summarization can be performed in parallel using various computing resources facilitated by the *BiocParallel* (Morgan *et al.*, 2014) package. For example, performing summarization on an 8-core machine can be requested by specifying `MulticoreParam(workers = 8)`. The *BiocParallel* also allows parallelization on a cluster of computers using different job schedulers. For example, using the appropriate configuration file a `BatchJobsParam`-object can be constructed and passed to the `summarize` function.

```
library(BiocParallel)
BPPARAM <- BatchJobsParam(workers = 10,
 conffile = "SGE_config_file.R")
sdata <- summarize(targets, batchSize = 50,
BPPARAM = BPPARAM)
```

The script folder of the package contains example files for parallel summarization on a cluster computer using the Sun Grid Engine job scheduler. For more information on how to set up a configuration file for other job schedulers, see R package *BatchJobs* (Bischl *et al.*, 2011).

4 CONCLUSION

MethylAid provides complete insight in the quality of each sample in a large-scale EWAS and gives the user full control on the selection of bad-quality samples that should be excluded from downstream analyses. Quality assessment of 450k array data is made interactive, flexible and efficient. Therefore, *MethylAid* is expected to be useful for both data analysts and core facilities.

ACKNOWLEDGEMENT

The authors thank Jenny van Dongen for beta testing.

Funding: This work was done within the framework of the Biobank-based Integrative Omics Study (BIOS) Consortium funded by BBMRI-NL, a Research Infrastructure financed by the Netherlands Organization for Scientific Research (NWO 184.021.007) and financially supported by the European Union’s Seventh Framework Program (FP7/2007-2011) under grant agreement No. 259679 (IDEAL).

Conflict of interest: none declared.

REFERENCES

Aryee,M. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Bibikova,M. *et al*. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Bischl,B. *et al*. (2011) Computing on high performance clusters with R: Packages BatchJobs and BatchExperiments. *Technical Report 1*. TU Dortmund University, Germany.

Mill,J. and Heijmans,B. (2013) From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.*, **14**, 585–594.

Morgan,M. *et al*. (2014) *BiocParallel: Bioconductor Facilities for Parallel Evaluation*. R package version 0.6.0.

Rakyan,V. *et al*. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529–541.

RStudio and Inc. (2014) *shiny: Web Application Framework for R*. R package version 0.9.1.