Normalization of 450K data

Jonathan A. Heiss

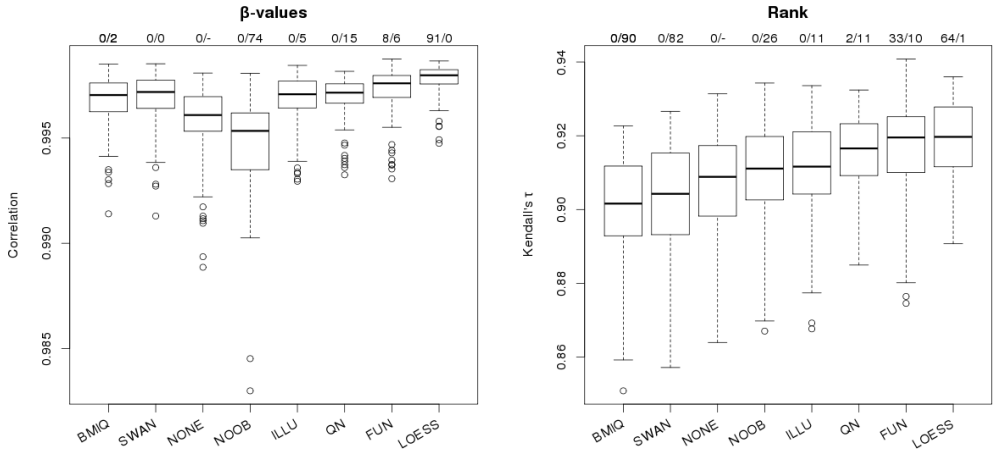
October 30, 2017

This package provides functions to read and normalize data from Illumina Infinium HumanMethylaton450 BeadChips. As input only IDAT files are supported.

```
>  library(normalize450K)
>  library(minfiData) ## this package includes some .idat files
>  library(data.table)
>  path <- system.file("extdata",package="minfiData")
>  samples = fread(file.path(path, 'SampleSheet.csv'),integer64='character')
>  samples[,file:=file.path(path,Sentrix_ID,paste0(Sentrix_ID,'_',Sentrix_Positi
>  ## samples$file is a character vector containing the location of the
>  ## .idat files, but without the suffixes "_Red.idat" or "_Grn.idat"
>
>  raw = read450K(samples$file)
>  none = dont_normalize450K(raw) ## no normalization
>  norm = normalize450K(raw)
```

The normalization method is described in detail elsewhere [1]. In brief, a dye bias correction of signal intensities is performed using the extension control probes, which results in a more symmetrical distribution of beta-values from type II probes. Next, a set of housekeeping CpG sites and a virtual reference array is used to adjust signal intensities of type I/II probes using local regression. M-values are computed based on corrected intensities. Using the set of housekeeping sites and the virtual reference array again, the methylation-related bias, that affects type II probes only, is corrected. At last M-values are transformed to beta-values. No adjustment for probe type bias is performed to avoid the trade-off of precision for accuracy of β-values.

The method was tested extensively on whole blood samples. Below is a benchmark comparing the correlation (of β-values) and rank correlation of 99 pairs of technical replicates after different normalization approaches (BMIQ, SWAN, NOOB - background correction, ILLU - normalization from GenomeStudio, QN - quantile normalization, FUN - functional normalization, LOESS - the method implemented in this package). The two numbers on top of each bar indicate how often a method achieved the highest correlation for a pair of technical replicates and how often correlation decreased compared to the not normalized data (NONE).



How well this method performs on samples from other tissues than blood was not evaluated, but as it is based on the concept of housekeeping genes, performance should be comparable.

If you are using `normalize450K` in a publication, please cite [1].

References

[1] Heiss JA and Brenner H (2015) *Between-array normalization for 450K data.* Front. Genet. 6:92. doi: 10.3389/fgene.2015.00092