HumMeth27QCReport: A Package to Generate QC Reports for Infinium Methylation Assay Data

Francesco M. Mancuso, Niles Oien, Guglielmo Roma

July 4, 2012

Contents

1 Introduction

This document describes an R package for generating QC reports. The goal of this project is to create a tool to allow users of Illumina Infinium BeadChip Methylation Assay[1] to quickly assess the data quality of a batch of processed arrays. HumMeth27QCReport works with both the two available Infinium platforms: the HumanMethylation27 BeadChip and the HumanMethylation450 BeadChip. The package makes use of different packages, as *methylumi* or *lumi*, for reading files exported from GenomeStudio software, generating intensity plots and normalizing Beta values. Several new plots are generated and printable pdf files are created. To run properly and generate the summary Excel file, the script needs that a working version of Perl is installed on your machine.

2 Usage

After starting R, the package should be loaded using the following.

This will load *HumMeth27QCReport* as well as the *methylumi*, *lumi*, *IlluminaHumanMethylation27k.db*, *IlluminaHumanMethylation450k.db*, *amap*, *Hmisc*, *gplots*, *plotrix*, *WriteXLS* and *tcltk2* packages and their dependencies.

To generate an example report simply use the method `HumMeth27QCReport` (here is reported an example for the Infinium HumanMethylation27 BeadChip platform).

Dir <- system.file("extdata",package="HumMeth27QCReport")
ImportDataR <- ImportData(Dir)
normMvalues <- HumMeth27QCReport(ImportDataR, platform="Hum27", pval=0.05, ChrX=F, ClustMethod="euclidean", quoteOutput=TRUE, normMethod="quantile")

where:

- **Dir** is a character string containing the location of the directory in which the input file are. All output files will be stored here.

[1] `www.illumina.com/`

- **platform** is the type of Illumina Infinium BeadChip methylation assay. This must be one of "Hum27" (Infinium HumanMethylation27 BeadChip) or "Hum450" (Infinium HumanMethylation450 BeadChip).
- **pval** is the p-value threshold number to define which samples keep for the normalization and the following analysis;
- **ClustMethod** is the distance measure to be used for the clustering. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary", "pearson", "correlation", "spearman" or "kendall";
- **ChrX** ia a logical value indicating whether the CpGs that belong to chromosome X should be deleted from normalization and the following analyses. The default is FALSE;
- quoteOutputLogical value determining if non-numeric items in the output file of normalized data should be in quotes. Default is TRUE
- normMethodCommonly either "quantile" or "ssn". Determines the normalization method. See documentation for lumiMethyN() function in lumi package. Default is "quantile"

3 Inputs files

HumMeth27QCReport takes input three files from GenomeStudio plus an optional text file with the chip control samples to discard from the normalization step:

* **Sample table** (it is compulsory that the file name contains the word "Sample", case sensitive, and not the other reserved words)

* **Control table** (it is compulsory that the file name contains the word "Control", case sensitive, and not the other reserved words)

* **BetaAverage table** (it is compulsory that the file name contains the word "AvgBeta", case sensitive, and not the other reserved words)

* **Discard.txt** (compulsory name)

Sample table - Required columns from GenomeStudio:
- Index
- Sample ID
- Sample Group
- Sentrix Barcode
- Sample Section
- Detected Genes (0.01)
- Detected Genes (0.05)
- Signal Average GRN
- Signal Average RED
- Signal P05 GRN
- Signal P05 RED
- Signal P25 GRN
- Signal P25 RED
- Signal P50 GRN
- Signal P50 RED
- Signal P75 GRN
- Signal P75 RED
- Signal P95 GRN

- Signal P95 RED
- Sample_Well
- Sample_Plate

Control table - Required columns from GenomeStudio (<Sn> = Sample Name):
- Index
- TargetID
- ProbeID
- <Sn>.Signal_Grn
- <Sn>.Signal_Red
- <Sn>.Detection Pval
- ...

Required controls (rows):
* BISULFITE CONVERSION (4 rows)
* EXTENSION (4 rows)
* HYBRIDIZATION (3 rows)
* NEGATIVE (16 rows)
* NON-POLYMORPHIC (4 rows)
* SPECIFICITY (4 rows)
* STAINING (4 rows)
* TARGET REMOVAL

AverageBeta table - Required columns from GenomeStudio (<Sn> = Sample Name):
- Index
- TargetID
- <Sn>.AVG_Beta
- <Sn>.Intensity
- <Sn>.Signal_A
- <Sn>.Signal_B
- <Sn>.BEAD_STDERR_A
- <Sn>.BEAD_STDERR_B
- <Sn>.Avg_NBEADS_A
- <Sn>.Avg_NBEADS_B
- <Sn>.Detection Pval
- ...
- SYMBOL

Discard.txt - Text file containing the name of the samples (the same name present in the Sample table; one sample per row) you want to discard from normalization. i.e. sample controls to see if chips worked properly like un-methylated samples.

4 Figure Details

The analysis consist of three parts: Internal Controls, Quality Check and Explorative Analysis. This section will describe the details of each part and the function call to generate the individual analysis. **An extensive explanation on how to interpret the results can be found in an example analysis at http://biocore.crg.es/wiki/HumMeth27QCReport.**

4.1 Internal Controls

* **getAssayControls** creates histogram plots relative to the internal controls of the Illumina Infinium HumanMethylation BeadChip assay into a pdf file called "InternalControl.pdf".

```
R> Dir <- system.file("extdata/",package="HumMeth27QCReport")
R> ImportDataR <- ImportData(Dir)
R> ControlResults <- getAssayControls(ImportDataR,platform="Hum27")
```

After data import, the method computes simple statistics and generates quality plots for monitoring the Illumina Infinium sample-independent and sample-dependent internal quality controls. For each control, HumMeth27QCReport generates a plot representing the percentage of background on signal. The sample-independent controls allow evaluating the quality of specific steps in the process flow and include:

- DNP staining control;
- Biotin staining control;
- Hybridization control;
- Target Removal control;
- Extension control in green channel;
- Extension control in red channel.

The sample-dependent controls allow evaluating performance across samples and include:

- Bisulfite control in green channel;
- Specificity control (mismatch 1) in red channel;
- Specificity control (mismatch 2) in green channel;
- Negative control;
- Non-Polymorphic control (green channel);
- Non-Polymorphic control (red channel).

In the case of 450k platform 3 more plots for sample-independent will be created:

- Bisulfite control in red channel: the same of the previous Bisulfite conversion but monitored in red channel;

- Bisulfite II control: these controls use Infinium II probe design and single base extension to monitor the efficiency of bisulfite conversion. If the bisulfite conversion reaction was successful, the "A" base will be incorporated and the probe will have intensity in the Red channel. If the sample has unconverted DNA, the "G" base will be incorporated across the unconverted cytosine, and the probe will have elevated signal in the Green channel.
- Specificity II control: these controls are designed to monitor extension specificity for Infinium II probes and check for potential non-specific detection of methylation signal over unmethylated background. Specificity II probes should incorporate the "A" base across the nonpolymorphic T and have intensity in the Red channel. In case of non-specific incorporation of the "G" base, the probe will have elevated signal in the Green channel.

4.2 Quality Check

* **QCCheck** creates all the plots relative to the quality of the samples.

```
R> Dir <- system.file("extdata/",package="HumMeth27QCReport")
R> ImportDataR <- ImportData(Dir)
R> QCresults <- QCCheck(ImportDataR, pval=0.05)
```

HumMeth27QCReport, moreover, generates plots to monitor eventual dye biases in not-normalized data. To this purpose, it makes use of the function plotSampleIntensities of the methylumi package to plot the intensities distribution for each sample. Additionally, HumMeth27QCReport plots the percentage of those CpGs that could not be detected at two different p-value cut-offs (0.01 and 0.05). This plot gives an immediate overview of the global CpG coverage. As further control, HumMeth27QCReport also evaluates the average detection p-value of each sample, and removes those samples with average pvalue cut-off higher than a threshold choosen by the user (see figures from 13 to 15 for examples).

4.3 Explorative Analysis

* **NormCheck** normalize the Beta Values and plot a PCA and a hierarchical Clustering of the samples using the normalized data

```
R> Dir <- system.file("extdata/",package="HumMeth27QCReport")
R> ImportDataR <- ImportData(Dir)
R> normMvalues <- NormCheck(ImportDataR, platform="Hum27", pval=0.05, ChrX=F, ClustMethod="euclidean
```

Principal Component Analysis (PCA) and hierarchical clustering are computed to assess sample similarities using normalized data (see figures 16 and 17 as examples). The users have the possibility to choose the distance method to use in the clustering calculation. As ulterior output, an Excel file is provided. It contains the normalized M-value, a summary of the Internal Controls and of the gene detection and different lists of non-detected CPGs. The methods to normalize the data are described further in the *lumi* package documentation.