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

Francesco M. Mancuso and Guglielmo Roma

April 4, 2011

Contents

1 Introduction **1**

2 Usage **1**

3 Inputs files **2**

4 Figure Details **4**
4.1 Internal Controls 4
4.2 Quality Check 17
4.3 Explorative Analysis 17

5 SessionInfo **20**

6 References **20**

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 access 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.

```
> library(HumMeth27QCReport)
```

This will load *HumMeth27QCReport* as well as the *methylumi*, *lumi*, *IlluminaHumanMethylation27k.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")
HumMeth27QCReport(Dir,platform="Hum27",pval=0.03,ChrX=F,ClustMethod="euclidean")
```

[1] `www.illumina.com/`

where:

- **Dir** is a character string containing the location of the directory in which the input file are;
- **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 belongs to chromosome X should be deleted from normalization and the following analyses. The default is FALSE; (WARNING: until the annotation package for Hum450 will not be public, this feature only works with "Hum27");

3 Inputs files

HumMeth27QCReport takes in 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 others reserved words)

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

* **BetaAverage table** (it is compulsory that the file name contains the word "Avg", case sensitive, and not the others 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 Chech and Explorative Analysis. This section will describe the details of each part and the function call to generate the individual analysis. An expample of each part is shown in the following figures.

4.1 Internal Controls

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

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

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.

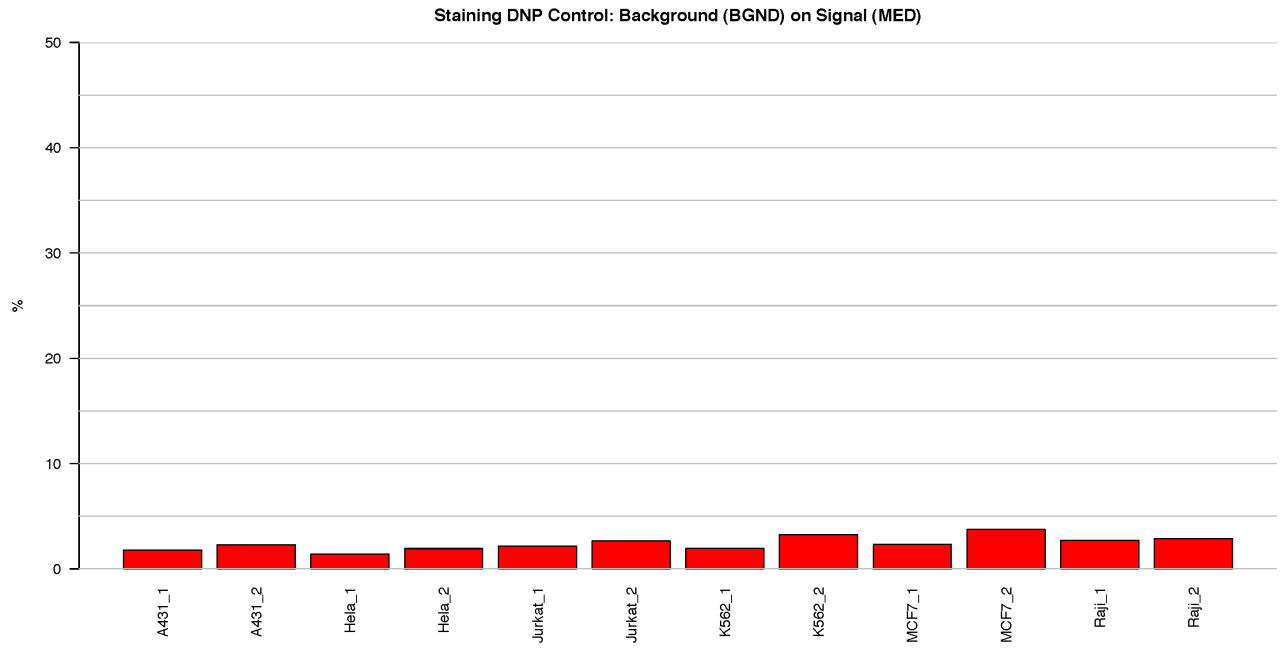

Figure 1: **Barplot of DNP staining control**. This figure represents the ratio (percentage) between background and signal for Staining control in the red channel (DNP). Staining controls are used to examine the efficiency of the staining step in both the red and green channels. Staining controls have dinitrophenyl (DNP) or biotin attached to the beads. The ratios should result in low signal, indicating that the staining step was efficient.

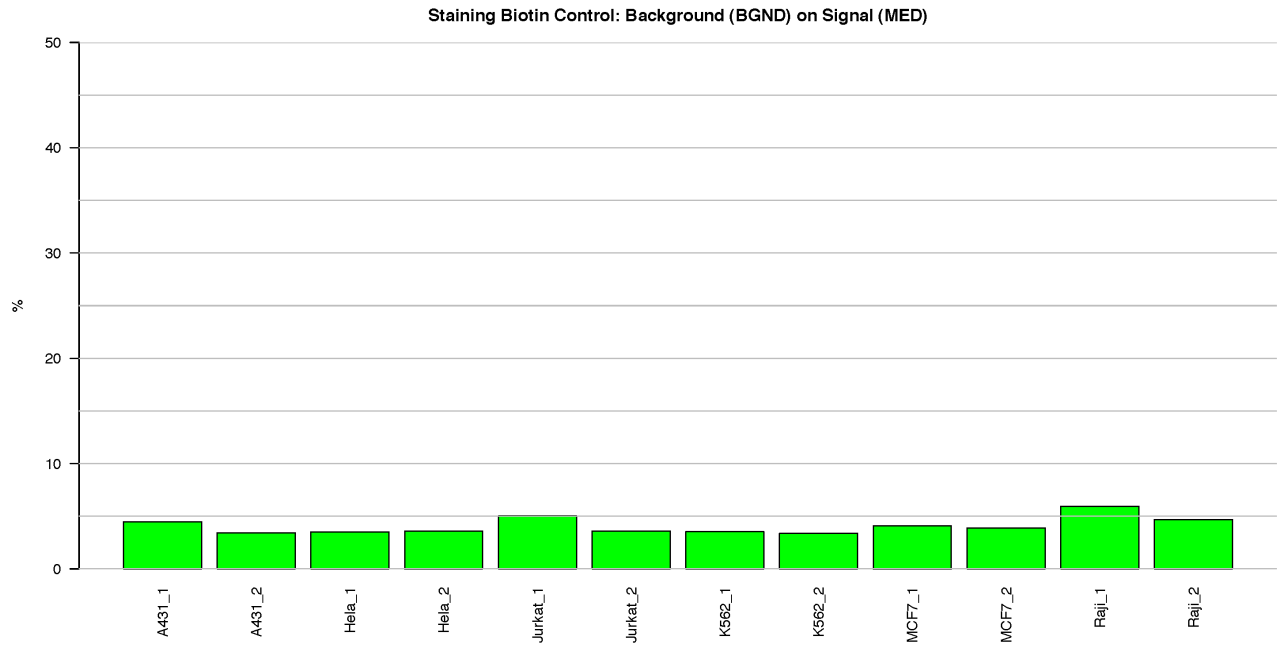

Figure 2: **Barplot of Biotin staining control**. This figure represents the ratio (percentage) between background and signal for Staining control in the green channel (Biotin). These controls are independent of the hybridization and extension step. The ratios should result in low signal, indicating that the staining step was efficient.

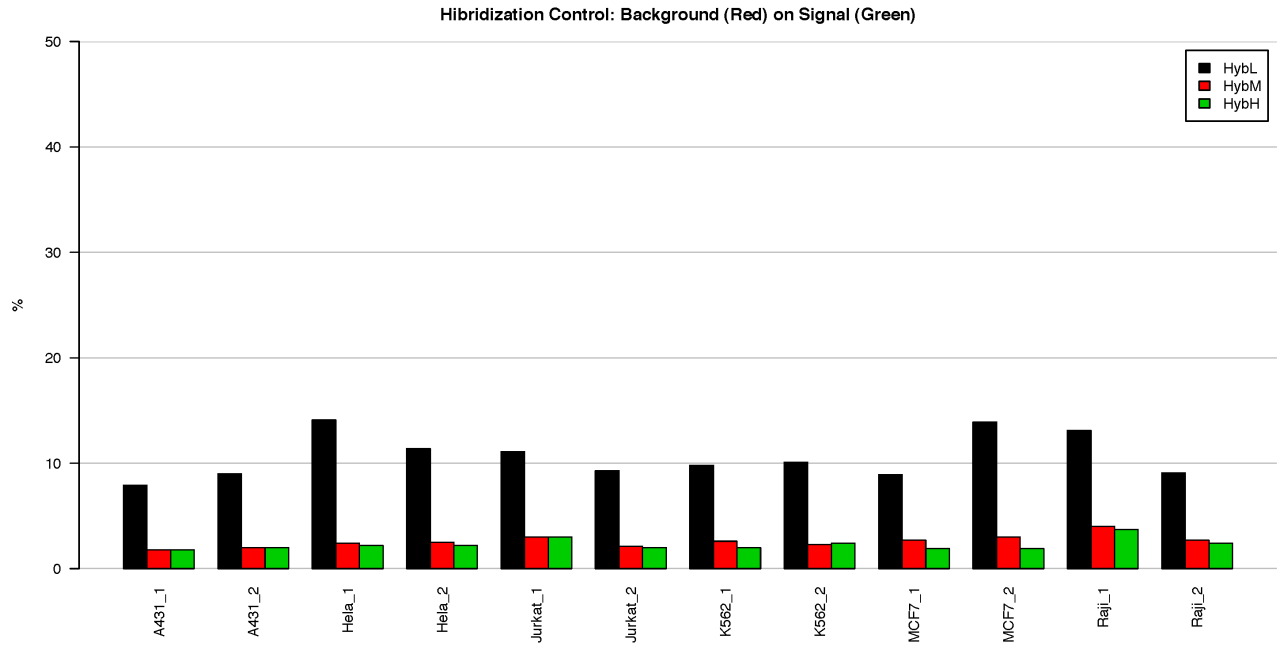

Figure 3: **Barplot of Hybridization control**. This figure represents the ratio (percentage) between background and signal for Hybridization controls in the green channel for three concentrations. The hybridization controls test the overall performance of the entire assay using synthetic targets instead of amplified DNA. These synthetic targets complement the sequence on the array perfectly, allowing the probe to extend on the synthetic target as template. The synthetic targets are present in the hybridization buffer at three levels, monitoring the response from high-concentration (5 pM), medium-concentration (1 pM), and low-concentration (0.2 pM) targets. All bead type IDs should result in signal with various intensities, corresponding to the concentrations of the initial synthetic targets.

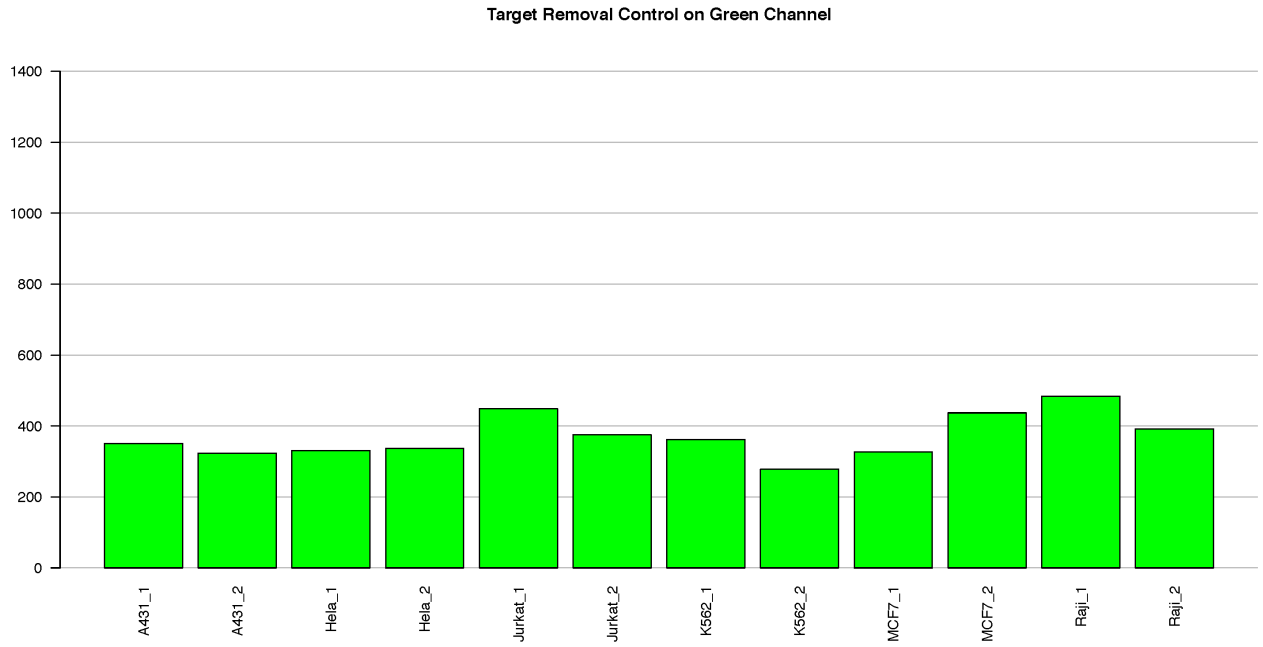

Figure 4: **Barplot of Target Removal control**. This figure represents the intensity value for Target removal controls in the green channel. Target removal controls test the efficiency of the stripping step after the extension reaction. The control oligos are extended using the probe sequence as template. This process generates labeled targets. The probe sequences are designed such that extension from the probe does not occur. All target removal controls should result in low signal, indicating that the targets were removed efficiently after extension. Values < 3400 have been detected (108 samples). There is not a range specified from Illumina, the value is based on previous experiments run in our facility.

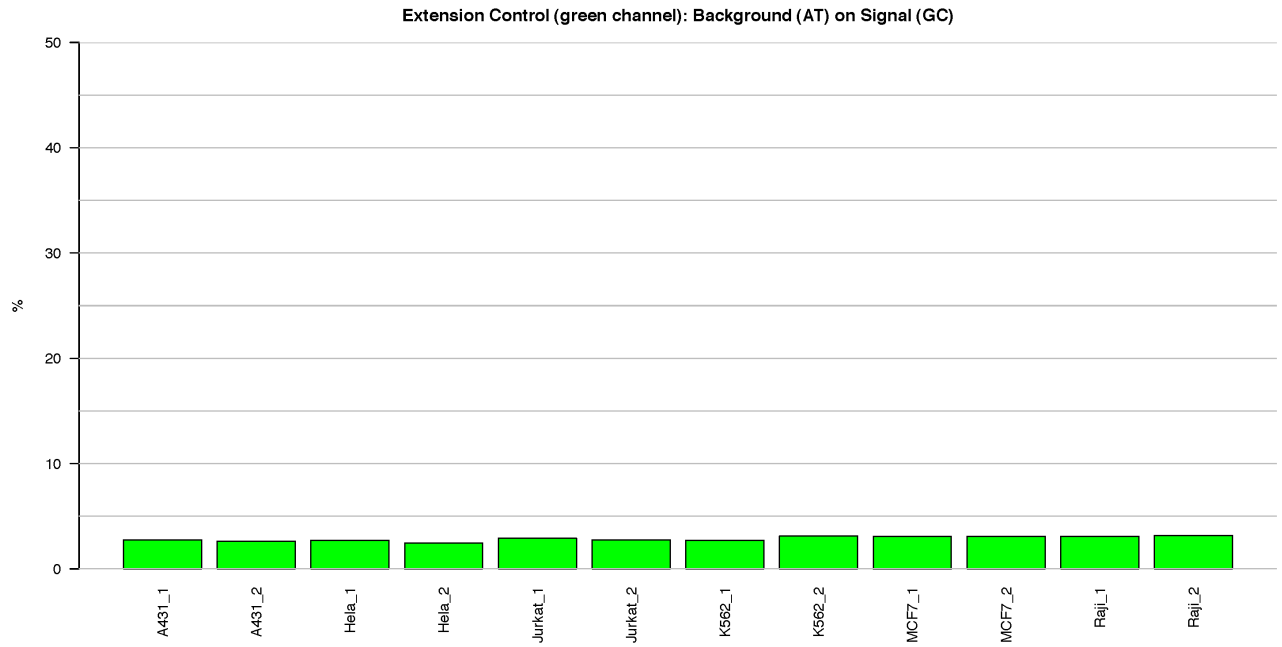

Figure 5: **Barplot of Extension control: green channel**. This figure represents the ratio (percentage) between background and signal for Extension control in the green channel (C,G). Extension controls test the extension efficiency of A, T, C, and G nucleotides from a hairpin probe, and are therefore sample-independent. The ratios should result in low signal, indicating that the extension was efficient.

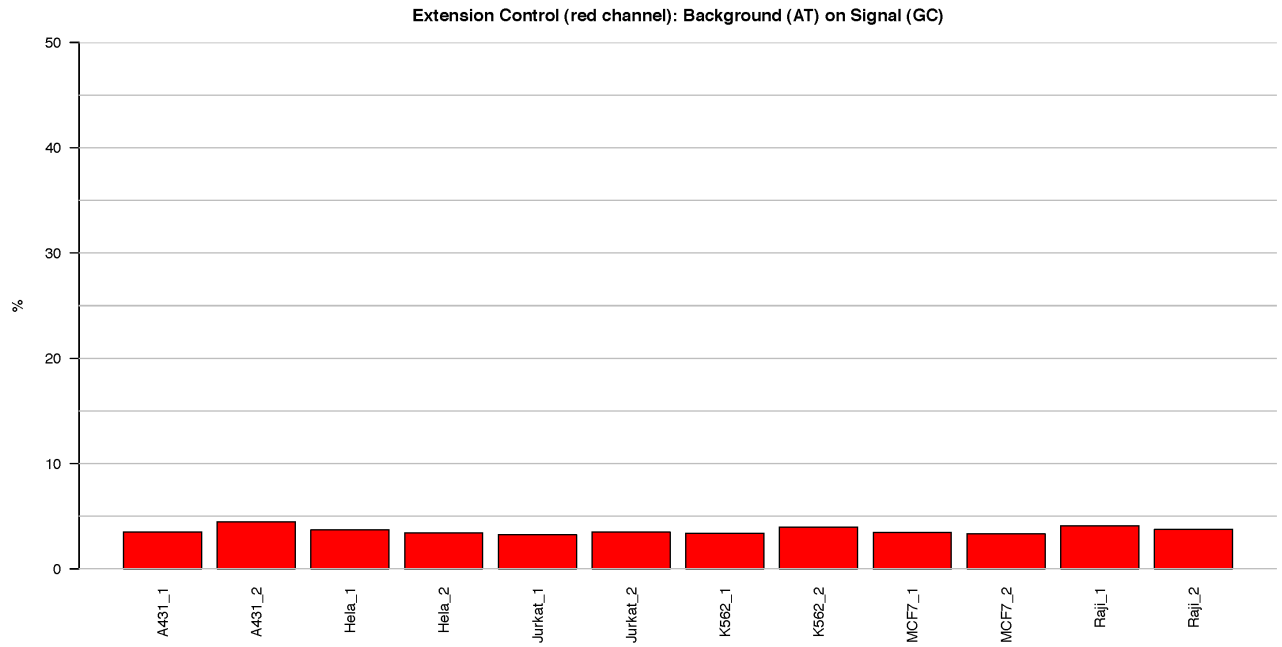

Figure 6: **Barplot of Extension control: red channel**. This figure represents the ratio (percentage) between background and signal for Extension control in the red channel (A,T). The ratios should result in low signal, indicating that the extension was efficient.

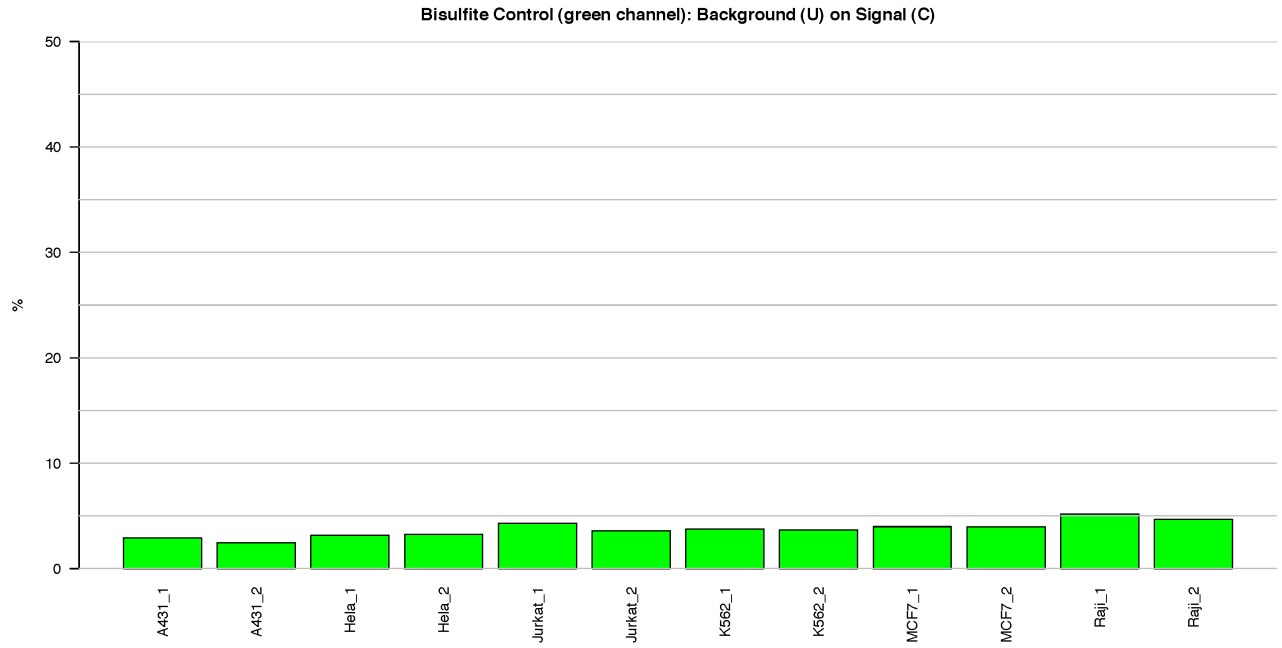

Figure 7: **Barplot of Bisulfite control**. This figure represents the ratio (percentage) between background and signal for Bisultife conversion control. The Bisulfite conversion Control asses the efficiency of bisulfite conversion of the genomic DNA. The Infinium Methylation probes query a [C/T] polymorphism created by bisulfite conversion of two different Hind III sites [AAGCTT] in the genome. If the bisulfite conversion reaction was successful, the "C" (Converted) probes will match the converted sequence and get extended. If the sample has unconverted DNA, the "U" (Unconverted) probes will get extended. There are no underlying C bases in the primer landing sites, except for the query site itself. Performance of bisulfite conversion controls should only be monitored in the Green channel. The ratios should result in low signal, indicating that the Bisulfite conversion was efficient.

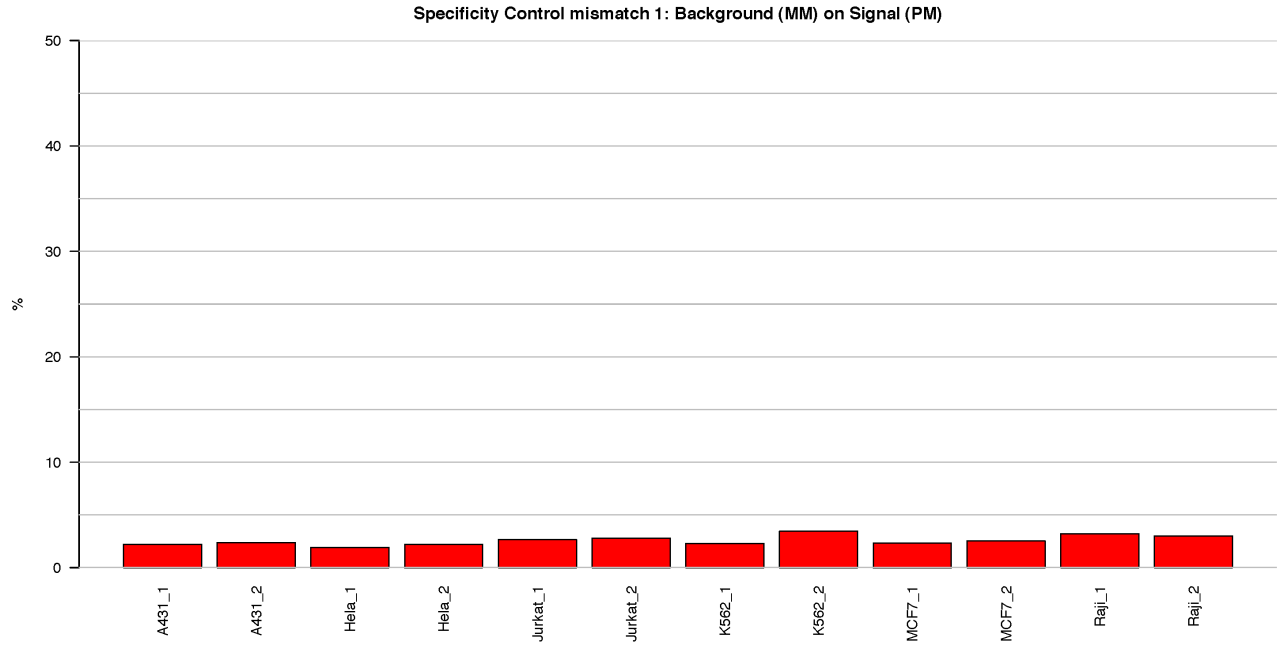

Figure 8: **Barplot of Specificity control (mismatch 1) in red channel**. This figure represents the ratio (percentage) between background (MM) and signal (PM) for Specificity controls in red channel. In the Infinium Methylation assay, the methylation status of a particular cytosine is carried out following bisulfite treatment of DNA by using query probes for unmethylated and methylated state of each CpG locus. In assay oligo design, the A/T match corresponds to the unmethylated status of the interrogated C, and G/C match corresponds to the methylated status of C. G/T mismatch controls check for non-specific detection of methylation signal over unmethylated background. Specificity controls are designed against non-polymorphic T sites. PM controls correspond to A/T perfect match and should give high signal. MM controls correspond to G/T mismatch and should give low signal. The ratios should result in low signal, indicating that the performance of the assay was efficient.

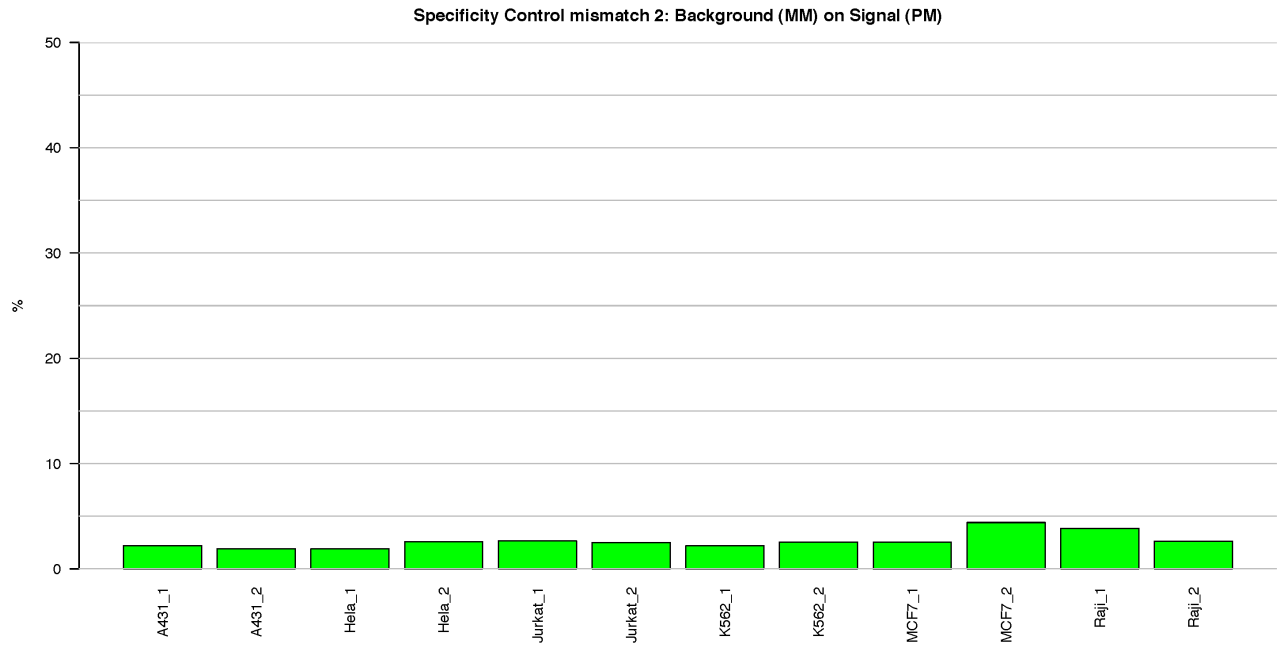

Figure 9: **Barplot of Specificity control (mismatch 2) in green channel**. This figure represents the ratio (percentage) between background (MM) and signal (PM) for Specificity controls in the green channel. PM controls correspond to A/T perfect match and should give high signal. MM controls correspond to G/T mismatch and should give low signal. The ratios should result in low signal, indicating that the performance of the assay was efficient.

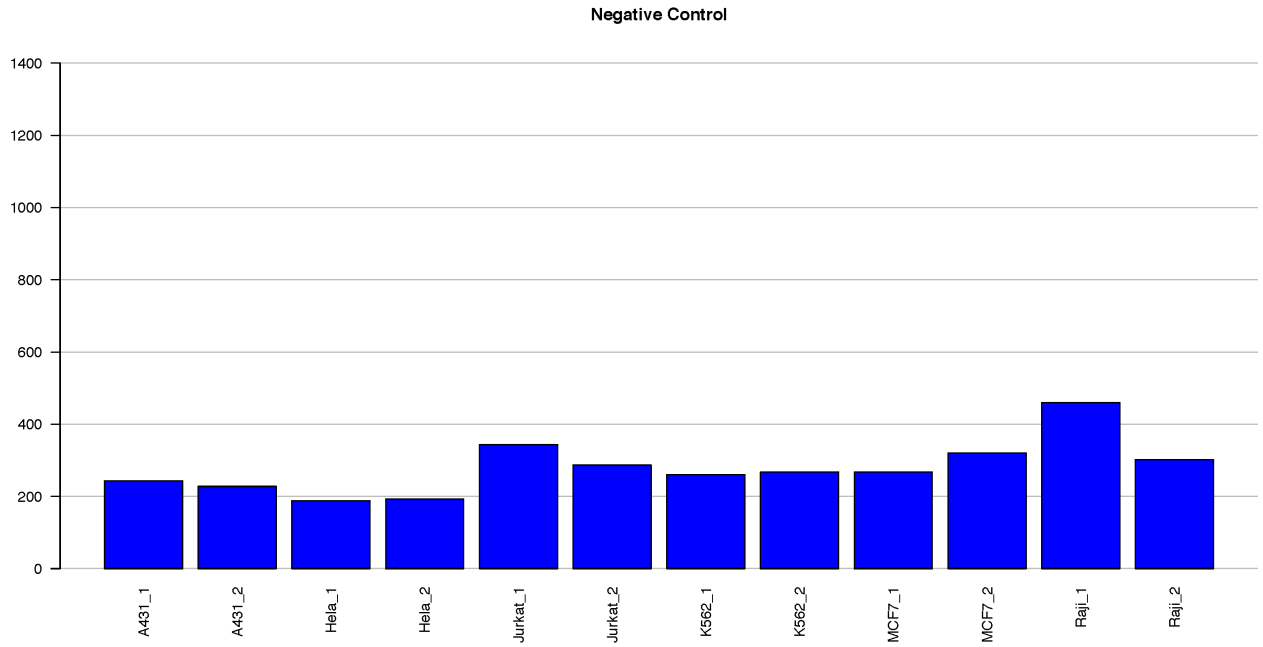

Figure 10: **Barplot of Negative control**. This figure represents the intensity value for the Negative control. Negative control probes are randomly permutated sequences that should not hybridize to the DNA template. Negative controls are particularly important for methylation studies because of a decrease in sequence complexity after bisulfite conversion. The mean signal of these probes defines the system background. This is a comprehensive measurement of background, including signal resulting from cross-hybridization, as well as non-specific extension and imaging system background. All target negative controls should result in low signal. Values < 2500 have been detected (108 samples). There is not a range specified from Illumina, the value is based on previous experiments run in our facility.

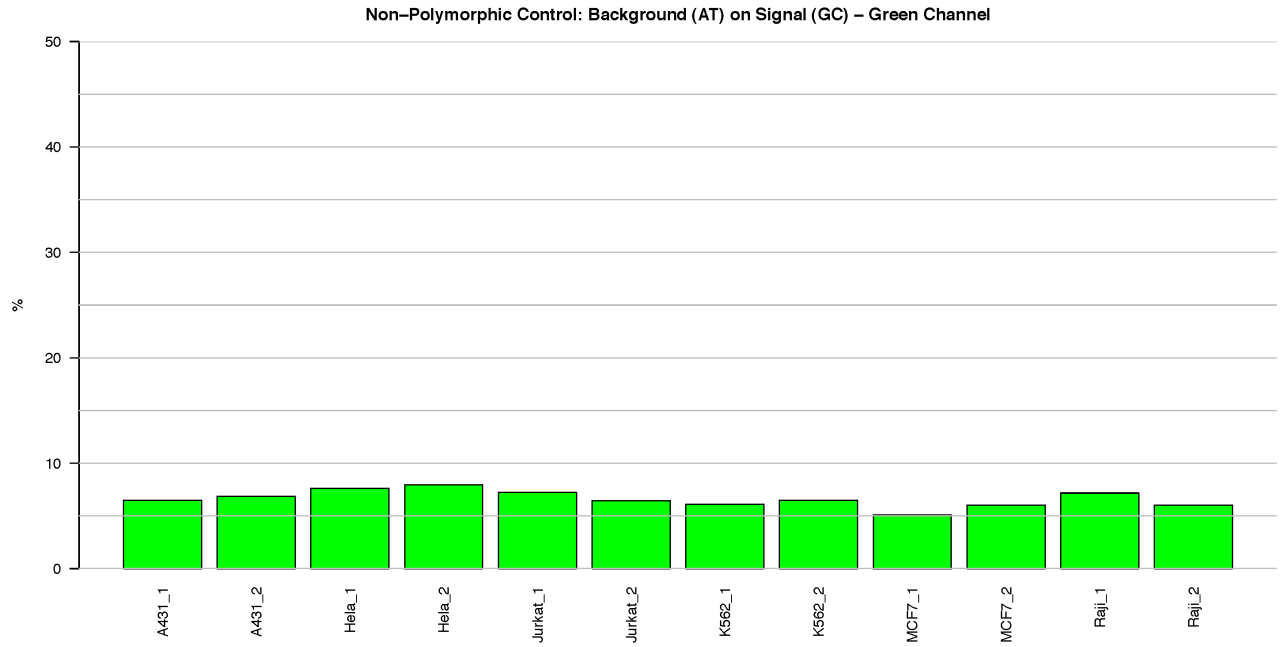

Figure 11: **Barplot for green channel of Non-Polymorphic control**. This figure represents the ratio (percentage) between background and signal for Non-Polymorphic control in the green channel. Non-polymorphic controls test the overall performance of the assay, from amplification to detection, by querying a particular base in a non-polymorphic region of the bisulfite genome. They let compare assay performance across different samples. One non-polymorphic control has been designed to query each of the four nucleotides (A, T, C and G). The target with the C base results from querying the opposite whole genome amplified strand generated from the converted strand. The ratios should result in low signal, indicating that the performance of the assay was efficient.

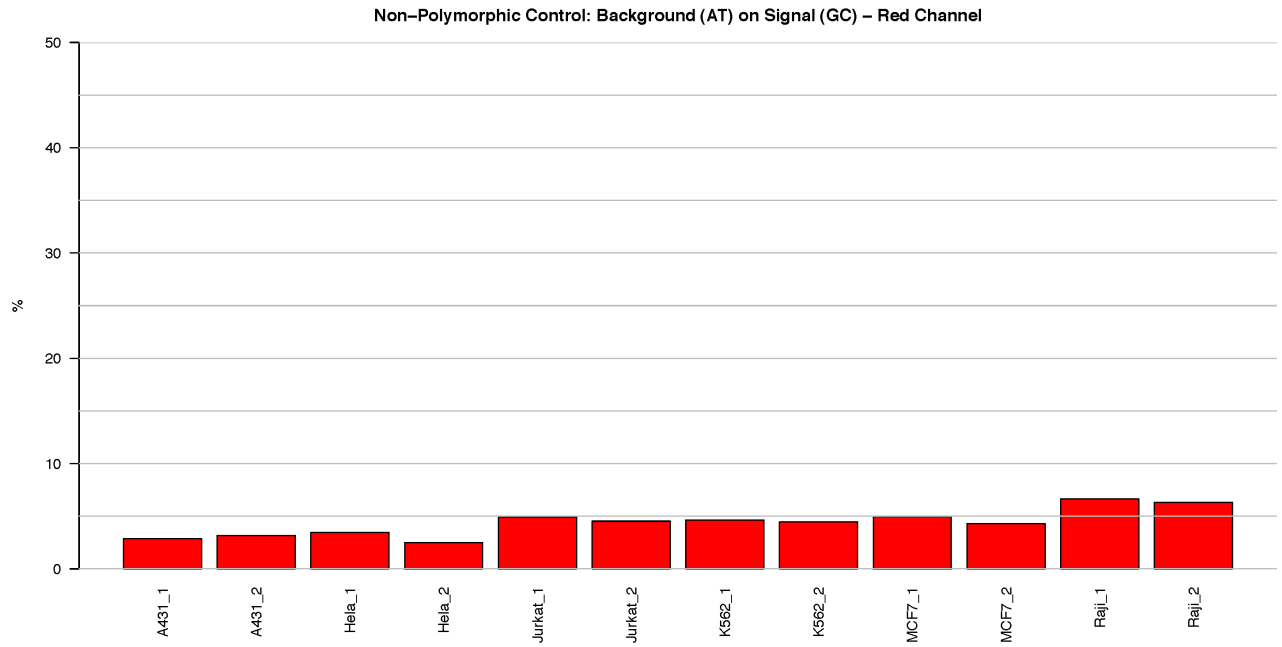

Figure 12: **Barplot for red channel of Non-Polymorphic control**. This figure represents the ratio (percentage) between background and signal for Non-Polymorphic control in the red channel. The ratios should result in low signal, indicating that the performance of the assay was efficient.

- 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> QCCheck(Dir, pval)
```

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.

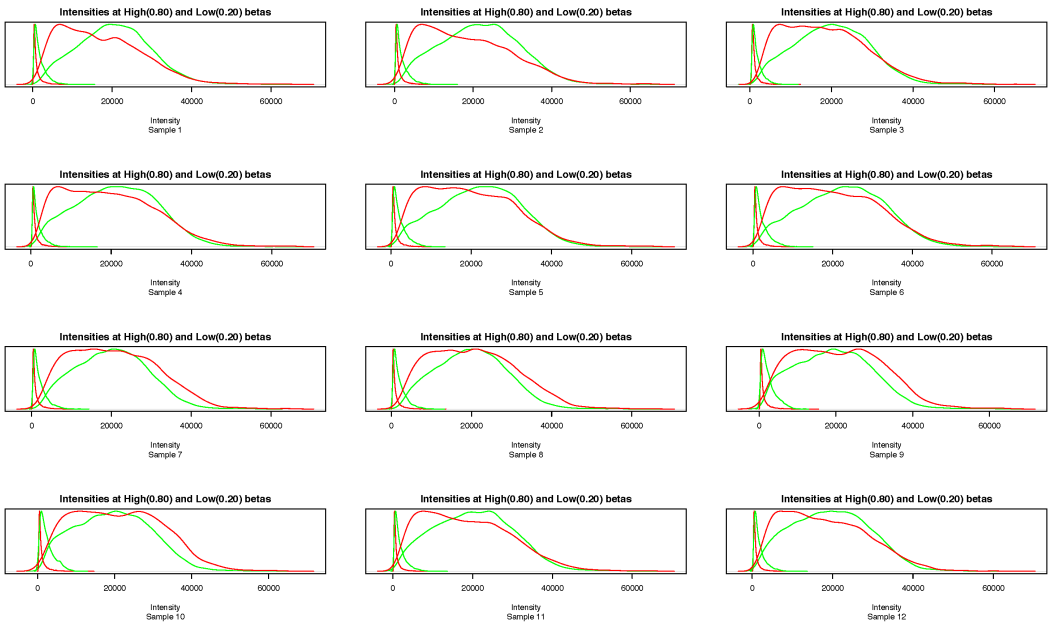

Figure 13: depending on the number of samples there could be more than one figure. For each sample the **intensity at high and low betas** is showed. The intensities as output by the GenomeStudio software often show a considerable amount of dye bias. This is a graphical example of this dye bias. In short, for each of the Cy3 and Cy5 channels, a cutoff in beta is used to calculate which Cy3 and Cy5 values should be plotted at high-methylation and low-methylation status. Any offset between Cy3 and Cy5 when plotted in this way likely represents dye bias and will lead to biases in the estimate of beta.

4.3 Explorative Analysis

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

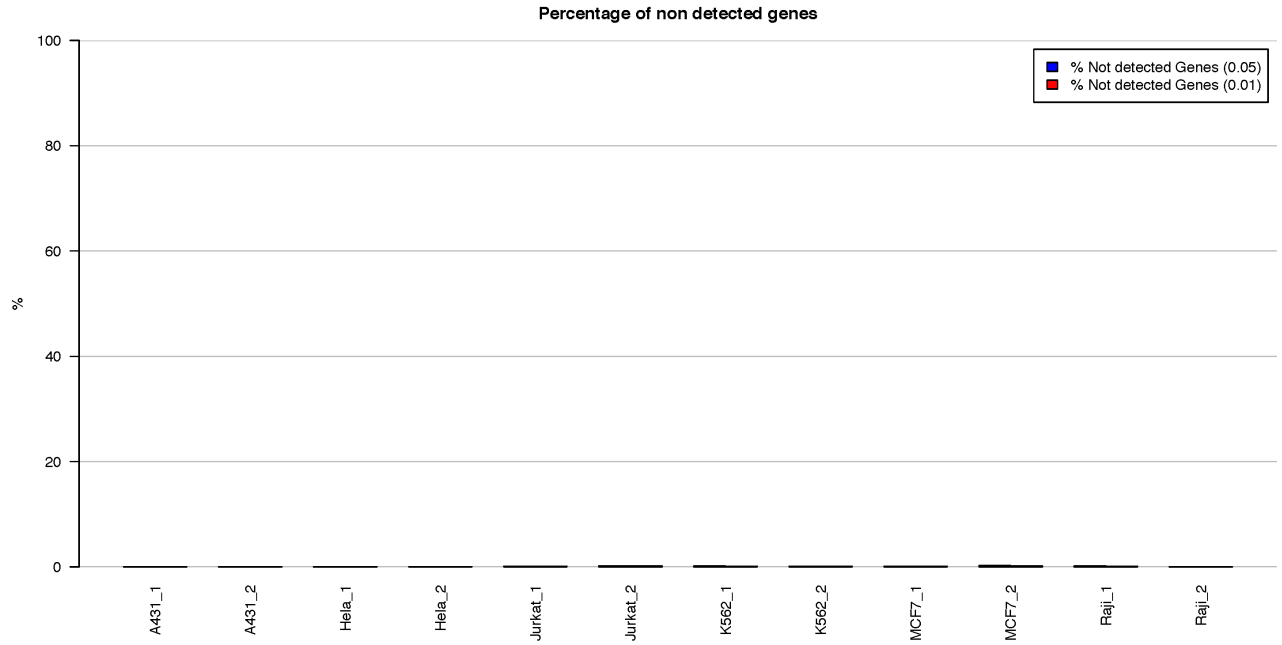

Figure 14: boxplot of the **percentage of non-detected genes** per each sample. Two different thresholds of detection p-value (0.05 and 0.01, blue and red bars, respectively) were set to calculate the percentage.

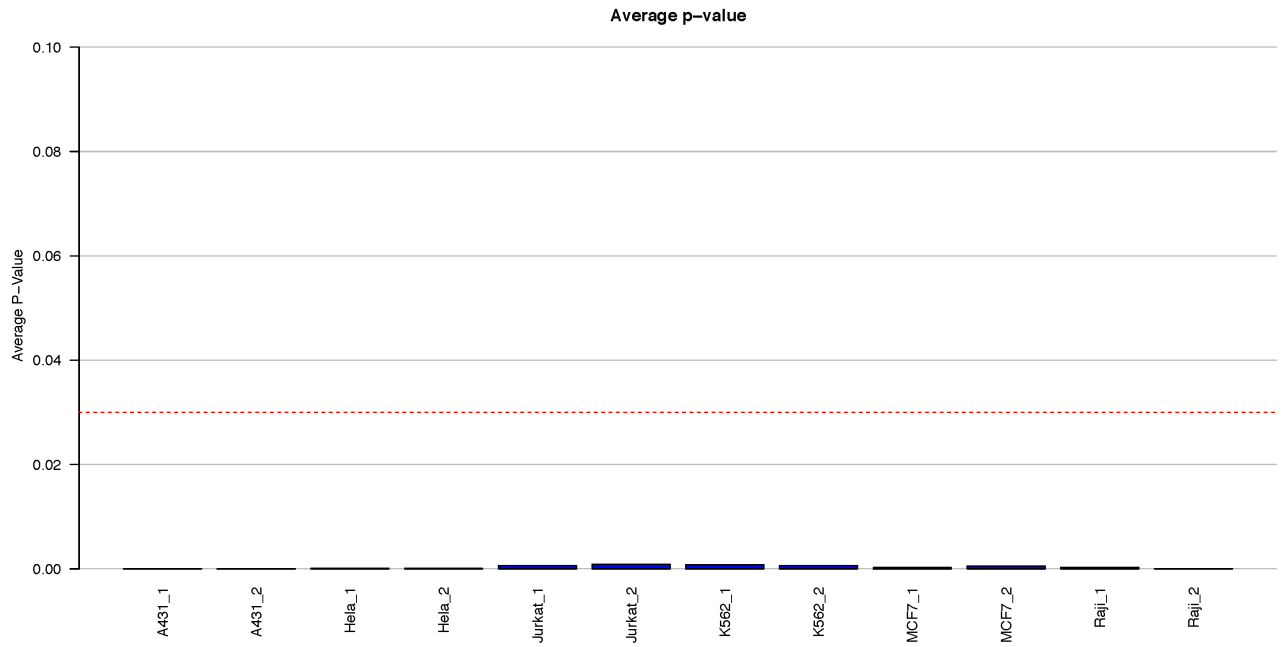

Figure 15: quick evaluation of the detection p-values. The boxplots show the **average p-value** for each sample; the red dotted line is the treshold defined by the user to select the samples for the following analysis.

```
R> NormCheck(Dir, platform, pval, ChrX, ClustMethod)
```

Principal Component Analysis (PCA) and hierarchical clustering are computed to assess sample similarities using normalized data. 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.

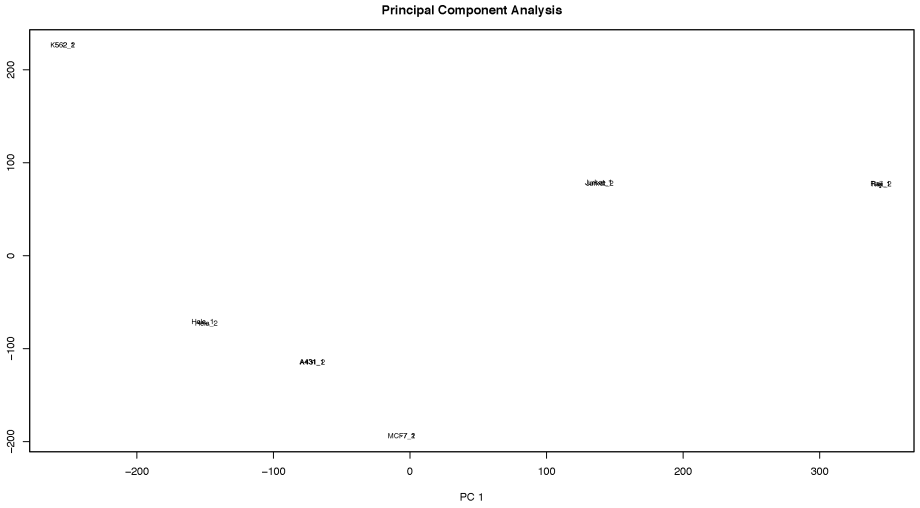

Figure 16: **Principal Component Analysis (PCA)** on filtered and normalized data.

5 SessionInfo

```
> toLatex(sessionInfo())
```

- R version 2.12.2 (2011-02-25), `x86_64-apple-darwin9.8.0`
- Locale: `C/en_GB.UTF-8/C/C/en_GB.UTF-8/en_GB.UTF-8`
- Base packages: base, datasets, grDevices, graphics, grid, methods, splines, stats, tcltk, utils
- Other packages: AnnotationDbi 1.13.17, Biobase 2.10.0, DBI 0.2-5, Hmisc 3.8-3, HumMeth27QCReport 1.2.8, IlluminaHumanMethylation27k.db 1.4.0, RSQLite 0.9-4, WriteXLS 2.1.0, amap 0.8-5, bitops 1.0-4.1, caTools 1.11, gdata 2.8.1, gplots 2.8.0, gtools 2.6.2, lumi 2.2.1, methylumi 1.6.1, org.Hs.eg.db 2.4.6, plotrix 3.1, survival 2.36-5, tcltk2 1.1-5
- Loaded via a namespace (and not attached): KernSmooth 2.23-4, MASS 7.3-11, Matrix 0.999375-49, affy 1.28.0, affyio 1.18.0, annotate 1.28.1, cluster 1.13.3, hdrcde 2.15, lattice 0.19-17, mgcv 1.7-5, nlme 3.1-98, preprocessCore 1.12.0, tools 2.12.2, xtable 1.5-6

6 References

- Du P., Kibbe, W.A., Lin S.M., (2008) 'lumi: a pipeline for processing Illumina microarray', Bioinformatics 24(13):1547-1548

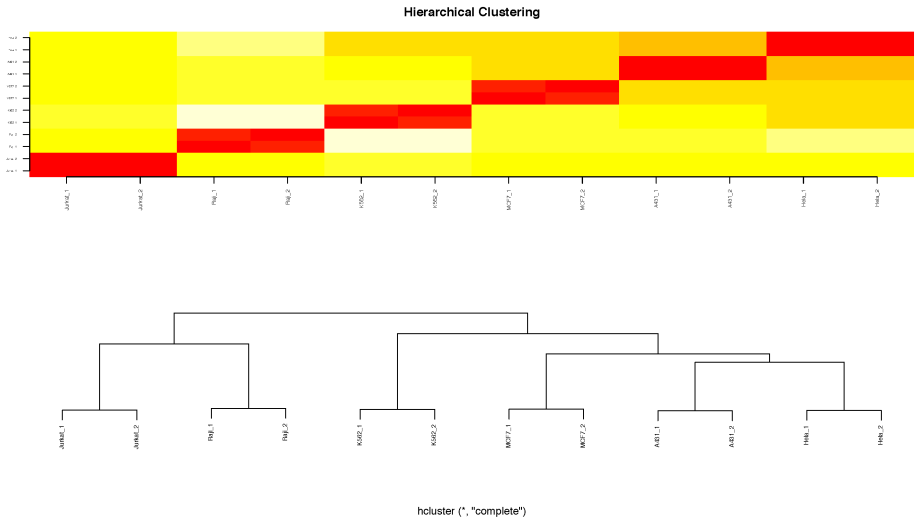

Figure 17: **Hierarchical Clustering** on filtered and normalized data. The distance method is defined by the user.

- Caussinus H., Fekri M.,Hakam S., Ruiz-Gazen A., (2003) 'A monitoring display of multivariate outliers', Computational Statistics and Data Analysis 44:237-252