Analyze Illumina Infinium methylation microarray data

Pan Du[†,*], Gang Feng[‡,†], Spencer Huang[‡,‡], Warren A. Kibbe[‡,§], Simon Lin[‡,¶]

January 12, 2016

[†]Genentech Inc., South San Francisco, CA, 94080, USA

[‡]Biomedical Informatics Center, Northwestern University, Chicago, IL, 60611, USA

Contents

[*]dupan.mail (at) gmail.com
[†]g-feng (at) northwestern.edu
[‡]huangcc (at) northwestern.edu
[§]wakibbe (at) northwestern.edu
[¶]s-lin2 (at) northwestern.edu

1 Overview

This is a tutorial of processing Illumina Infinium 27k and 450k methylation microarray data using lumi package. For the processing of Illumina GoldenGate methylation microarray data, please check the methylumi package. The tutorial will briefly describe the class structure, supported methods and functions in quality and color balance assessment, color balance adjustment, background adjustment, normalization and modeling the methylation status. It will also illustrate why we suggest using M-value instead of Beta-value in the statistics analysis.

2 Major classes of Illumina methylation microarray data

By default, Illumina suggests using the Beta-value to quantify the methylation levels. The Beta-value is a ratio between Illumina methylated probe intensity and total probe intensities (sum of methylated and unmethylated probe intensities). It is in the range of 0 and 1, which can be interpreted as the percentage of methylation. However, we have shown the Beta-value method has severe heteroscedasticity in the low and high methylation range, which imposes serious challenges in applying many statistic models [1]. The M-value, which is the log2 ratio of methylated probe intensity and unmethylated probe intensity, is another method used to measure the methylation level. We have shown the M-value method is approximately homoscedastic in the entire methylation range. As a result, it is more statistically valid in differential and other statistic analysis [1].

For this reason, we defined a new *MethyLumiM* class. *MethyLumiM* is a class inherited from *ExpressionSet* class. The "assayData" slot includes three required data matrix, which are "exprs", "methylated" and "unmethylated". The "exprs" is a matrix of M-values, which is the log2 ratio of methylated and unmethylated probe intensities; "methylated" and "unmethylated" are intensity matrix measured by methylated and unmethylalted probes of Illumina Infinium methylation microarray. "detection" is an optional matrix in "assayData" slot. It is the detection p-value outputted by Illumina GenomeStudio software. To keep the control data information, which is mainly used for the quality control purpose, the *MethyLumiM* class has a "controlData" slot. The "controlData" slot can be either NULL

or a *MethyLumiQC* object. Users are able to plot the control data using `plotControlData` function.

3 Import the methylation data

Illumina microarray has two types of output data files: Illumina IDAT files and Illumina GenomeStudio output files. Illumina IDAT files provide the detailed bead level information, and Illumina GenomeStudio output files keep the probe level summarized information. As expected the Illumina IDAT files is much bigger in file size, but it keeps all data information and the format is consistent across versions. In comparison, the Illumina GenomeStudio output files is portable in size, but it only keeps the summarized information and its format may have slight changes across versions.

3.1 Import bead-level Illumina IDAT files

For convenience of importing IDAT files, we wrote a `importMethyIDAT` function, which is a wrapper of `lumIDAT` in methylumi package. As Illumina organizes the output .idat files by barcodes, `importMethyIDAT` automatically checks the sub-folders in the names of barcodes for .idat files. Users need to a barcode vector or a data.frame with required columns. Please check the help of `importMethyIDAT` function for more details.

One benefit of using `importMethyIDAT` is that it automatically imports the control data information together with other methylation data.

3.2 Import probe-level Illumina GenomeStudio output files

We define a `lumiMethyR` function to read the Illumina methylation data, which is a wrap of the `methylumiR` function in methylumi package. The `lumiMethyR` function coerces the returned object (in *MethyLumiSet* class) as a *MethyLumiM* class object. The annotation library is not required if the GenomeStudio output data has already included the "COLOR_CHANNEL" column.

```
> ## specify the file name
> # fileName <- 'Example_Illumina_Methylation_profile.txt'
> ## load the data
> # example.lumiMethy <- lumiMethyR(fileName, lib="IlluminaHumanMethylation27k.db")  #
```

4 Data preprocessing

The first thing is to load the lumi package.

```
> library(lumi)
```

4.1 Example methylation datasets

The lumi package includes two example datasets. One is a control and treatment dataset (included in `example.lumiMethy` data object), which is measured in two batches. Another one is a titration dataset (included in `example.methyTitration` data object). To save storage space, only a randomly selected subset rows (CpG sites) were kept in the example datasets.

The control and treatment dataset (`example.lumiMethy`) includes four control and four treatment samples together with their technique replicates. The original samples and technique replicates were measured in two batches. Figure 1 and 2 show the overall sample relations of the control and treatment dataset in a PCA plot and Hierarchical cluster tree. We can see the batch difference is the major effect of sample differences.

```
> ## load example data (a methyLumiM object)
> data(example.lumiMethy)
> ## summary of the example data
> example.lumiMethy

MethyLumiM (storageMode: lockedEnvironment)
assayData: 5000 features, 16 samples
  element names: detection, exprs, methylated, unmethylated
protocolData: none
phenoData
  sampleNames: Treat1 Treat2 ... Ctrl4.rep (16 total)
  varLabels: sampleID label
  varMetadata: labelDescription
featureData
  featureNames: cg00002426 cg00012386 ... cg27662379 (5000 total)
  fvarLabels: CHR COLOR_CHANNEL
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:

> ## print sample Names
> sampleNames(example.lumiMethy)

 [1] "Treat1"     "Treat2"     "Treat3"     "Treat4"     "Ctrl1"
 [6] "Ctrl2"      "Ctrl3"      "Ctrl4"      "Treat1.rep" "Treat2.rep"
[11] "Treat3.rep" "Treat4.rep" "Ctrl1.rep"  "Ctrl2.rep"  "Ctrl3.rep"
[16] "Ctrl4.rep"
```

The methylation titration dataset (`example.methyTitration`) includes 8 samples: "A1", "A2", "B1", "B2", "C1", "C2", "D" and "E". They are mixtures of Sample A (a B-lymphocyte sample) and Sample B (is (a colon cancer sample from a female donor) at five different

```
> plotSampleRelation(example.lumiMethy, method='mds', cv.Th=0)
```

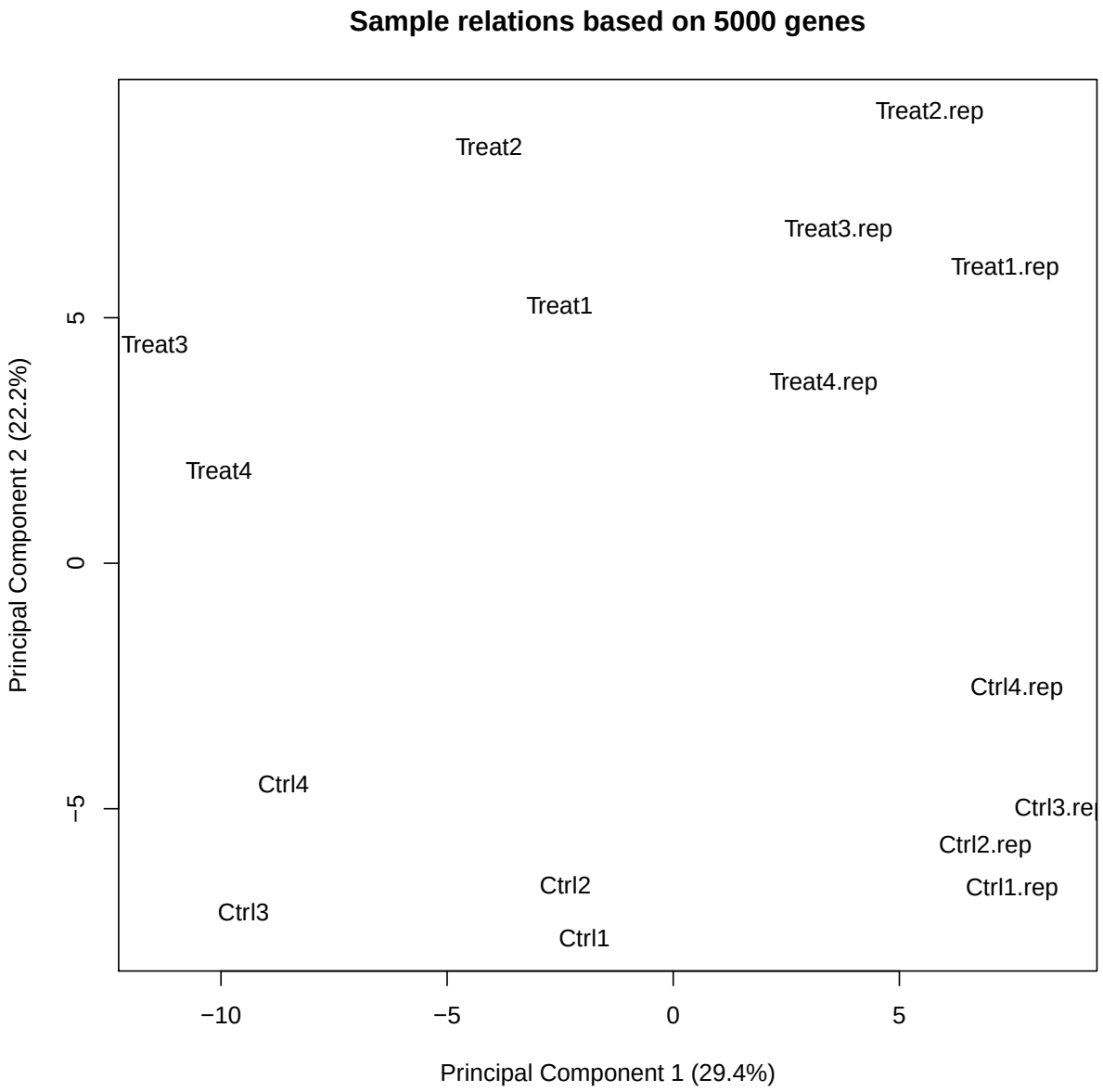

Figure 1: Overall sample relations of the raw data in example.lumiMethy dataset

```
> plotSampleRelation(example.lumiMethy, method='cluster', cv.Th=0)
```

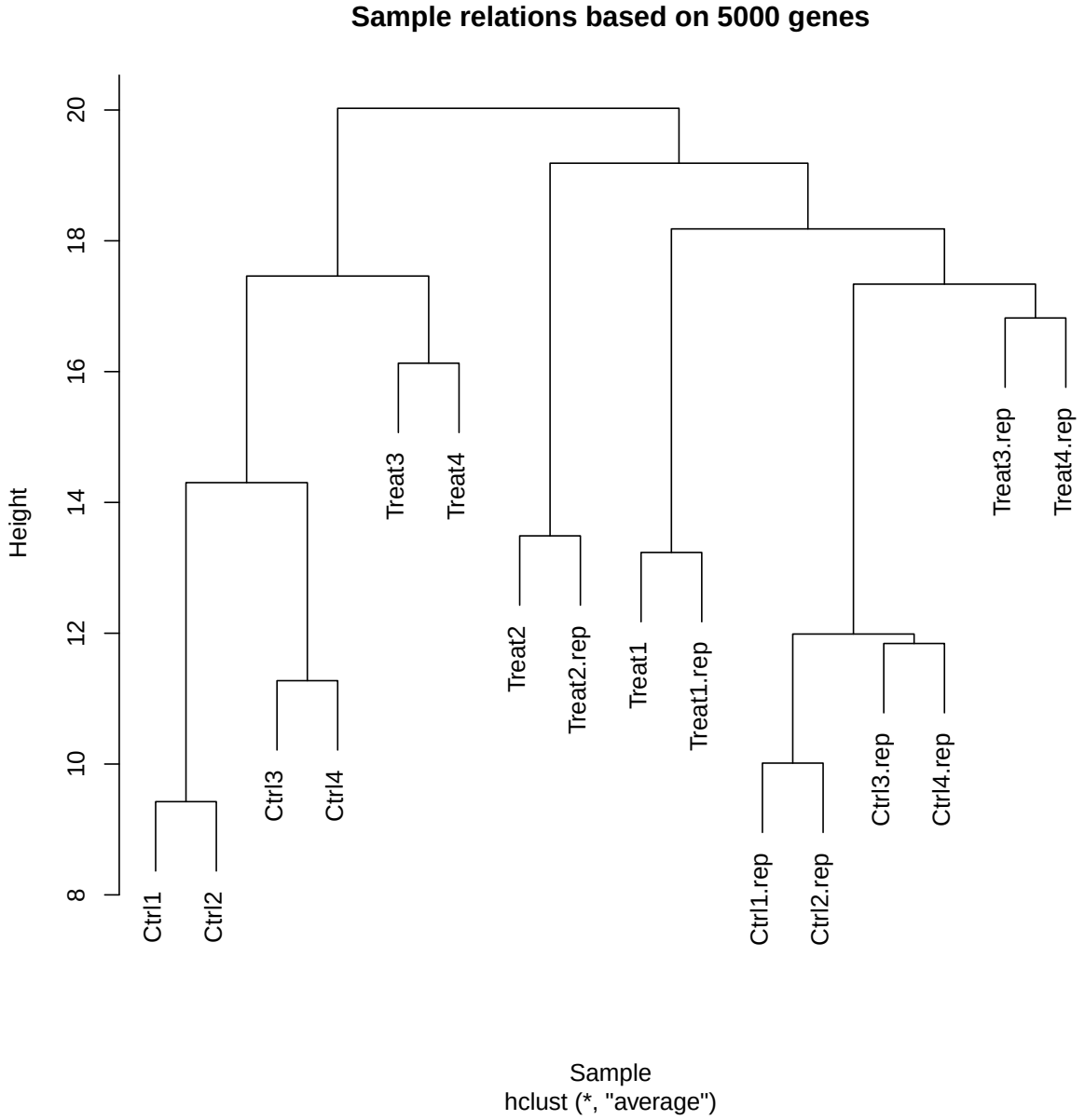

Figure 2: Overall sample relations of the raw data shown as a hierarchical tree

titration ratios: 100:0 (A), 90:10 (C), 75:25 (D), 50:50 (E) and 0:100 (B). Sample A, B and C have technique replicates. Figure 3 shows the overall sample relations of the titration dataset in a PCA plot.

```
> ## load the tritration data (a methyLumiM object)
> data(example.methyTitration)
> ## summary of the example data
> example.methyTitration

MethyLumiM (storageMode: lockedEnvironment)
assayData: 10000 features, 8 samples
  element names: detection, exprs, methylated, unmethylated
protocolData: none
phenoData
  sampleNames: A1 B1 ... E (8 total)
  varLabels: sampleID label
  varMetadata: labelDescription
featureData
  featureNames: cg09234859 cg22654935 ... cg04283392 (10000 total)
  fvarLabels: TargetID SYMBOL COLOR_CHANNEL CHR
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:

> ## print sample Names
> sampleNames(example.methyTitration)

[1] "A1" "B1" "C1" "A2" "B2" "C2" "D"  "E"
```

4.2 Check data distribution

Similar with expression microarray data, the lumi defines density and boxplot to check the overall distribution of the data. Figure 4 and 5 show the density of the example datasets. We can see the sample distributions can be quite different from each other. Because the density plot of M-values usually includes two modes, using the traditional boxplot cannot accurately represent the distribution of the data. We used another type of boxplot (method signature is *MethyLumiM* class) which is capable to show data with multiple modes. It uses different color levels to represent the M-values in different levels of HDRs (highest density regions) as specified in "prob" (probability of density coverage). The M-values locating outside the range of the maximum probability specified in "prob" are plotted as dots, which is similar with the outliers in the regular boxplot. By default, parameter "prob" is set as c(seq(10,90, by=10), 95). This means the M-values outside of 95% of HDR are plotted as outliers. Figure 6 shows the boxplot of M-values of example methylation data. The color depth represents the height

```
> plotSampleRelation(example.methyTitration, method='mds', cv.Th=0)
```

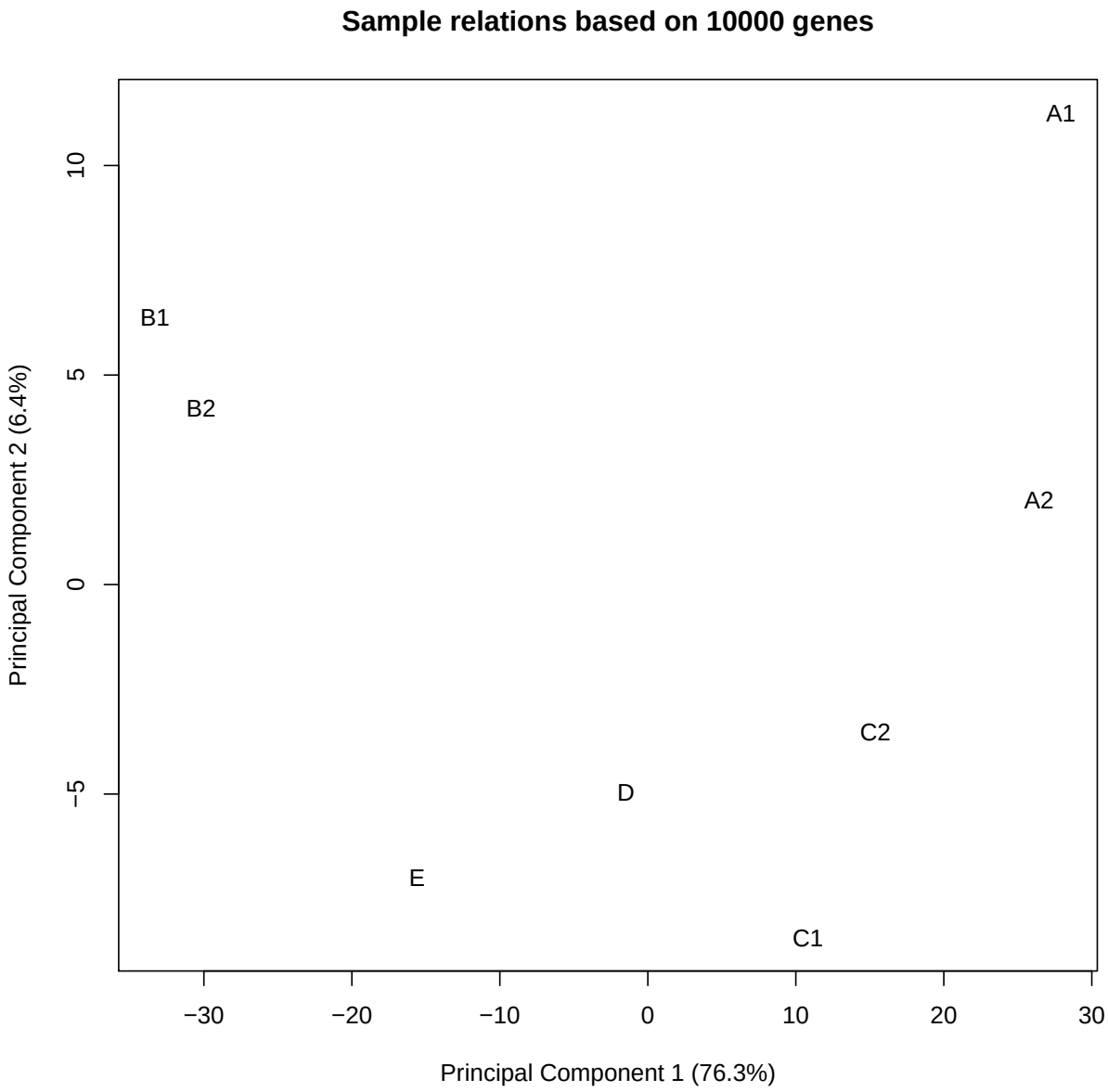

Figure 3: Overall sample relations of the titration dataset before preprocessing

```
> ## plot the density
> density(example.methyTitration, xlab="M-value")
```

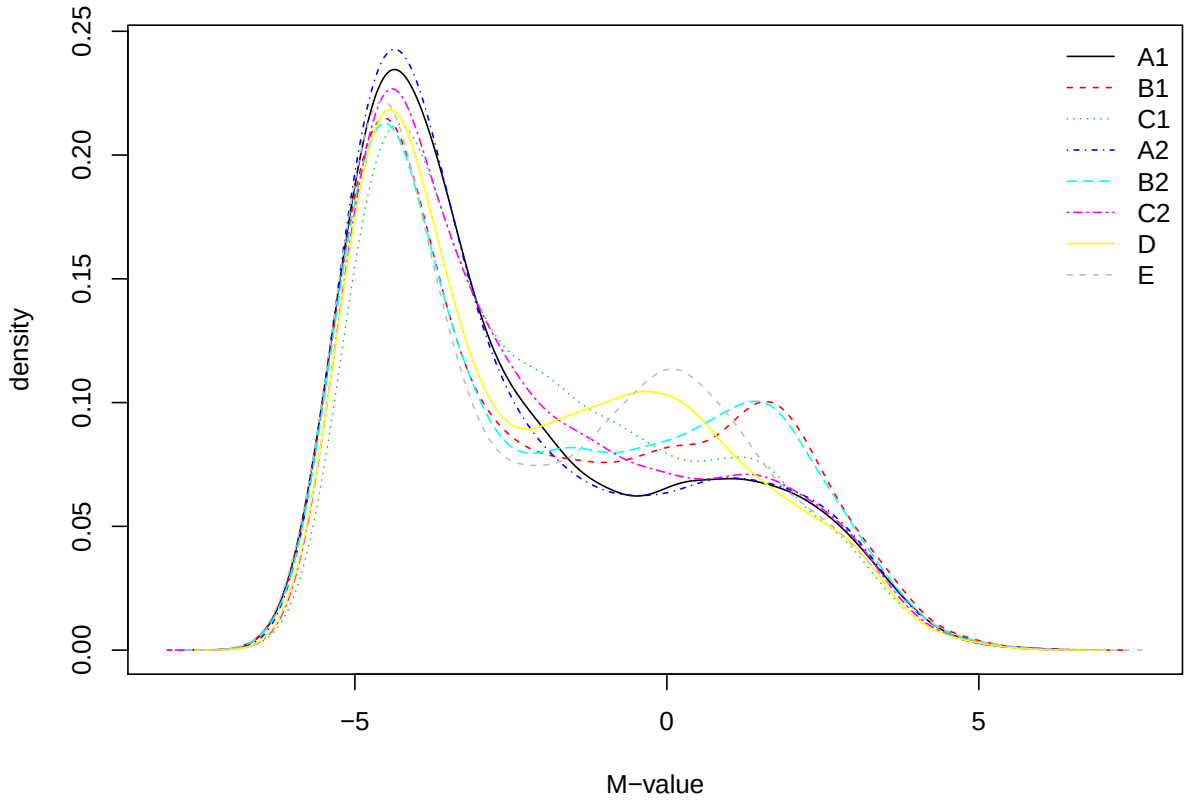

Figure 4: Density plot of M-value methylation levels of titration data before preprocessing

of density. The short horizontal bar at the position with the deepest color corresponds to the maximum position of density distribution of each sample. Based on Figure 6, we can see the distribution difference across samples. Because the methylation levels (measured by M-values or Beta-values) can have big changes across different conditions and treatment, only the methylation level distributions of technique or biological replicates are expected to have consistency. Therefore, we cannot claim a sample has quality issue if its distribution is quite different from others.

4.3 Check color balance

Based on the current Infinium assay design, Illumina uses two colors to label the final extended base following the hybridization of mehtylated or unmethylated probes. As a result,

```
> ## specify the colors of control and treatment samples
> sampleColor <- rep(1, ncol(example.lumiMethy))
> sampleColor[grep("Treat", sampleNames(example.lumiMethy))] <- 2
> density(example.lumiMethy, col=sampleColor, xlab="M-value")        ## plot the densi
```

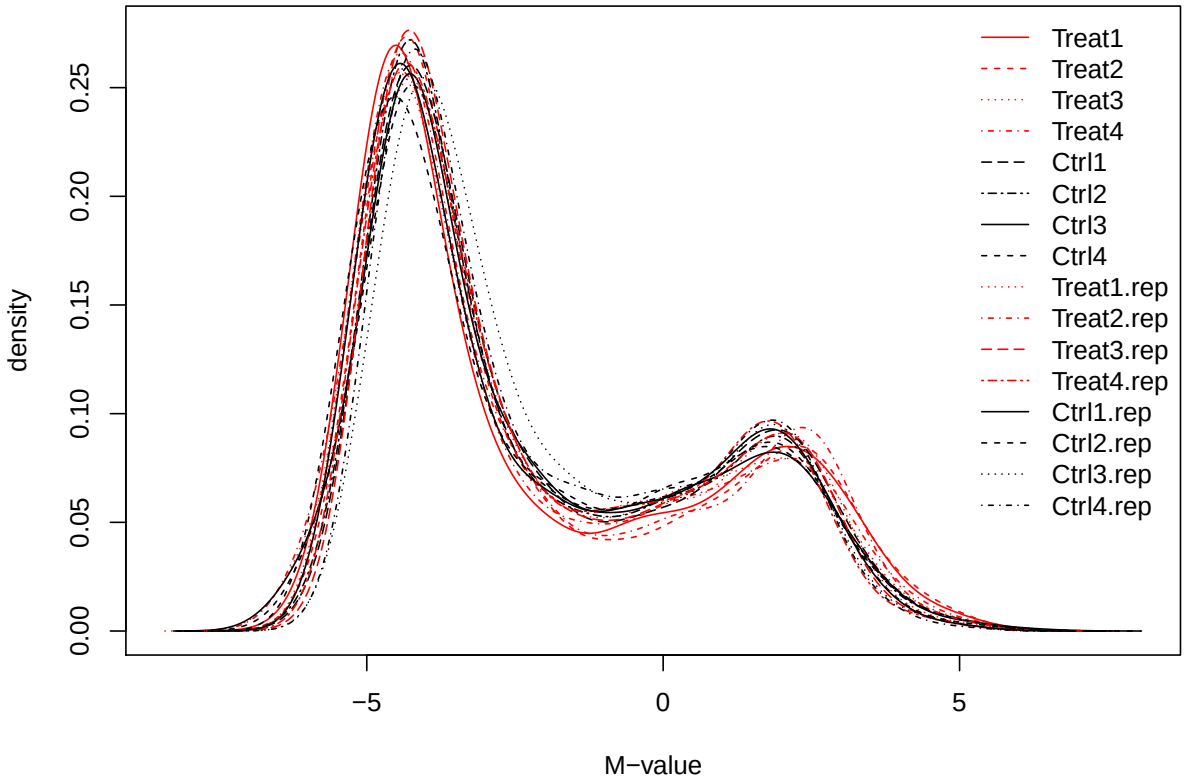

Figure 5: Density plot of M-value methylation levels before preprocessing

```
> ## Because the distribution of M-value has two modes, we use a boxplot different fro
> boxplot(example.lumiMethy)
```

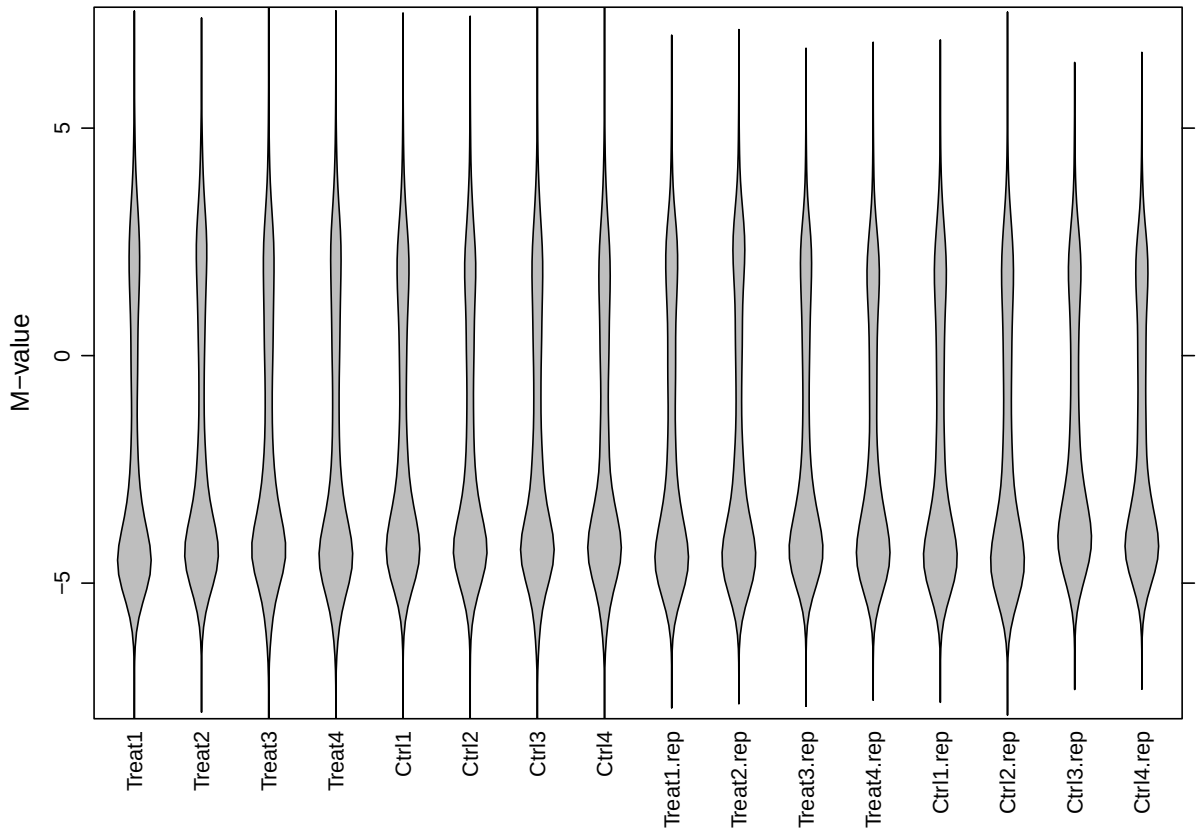

Figure 6: Multi-mode boxplot of M-value methylation levels before preprocessing

some of the CpG sites are measured in the red channel (final extended bases are A or T), whereas others are measured in the green channel (final extended bases are C or G). Note that the methylated and unmethylated probes of the same CpG site have the same color. Due to the difference in labeling efficiency and scanning properties of two color channels, the intensities measured in two color channels might be imbalanced. Because the methylation level is estimated based on the ratio of methylated and unmethylated probe intensities, many probe specific variations (like binding affinity and color balance) can be greatly reduced. However, because of the non-linearity of color effects, especially this effects will be different across different experiment conditions, color imbalance is not ignorable if the color effects are quite different across samples. Therefore, color balance adjustment will be important if the color effect is very inconsistent across samples. Moreover, the inconsistence of color balance is another indicator of sample quality problems. So checking color balance is another important measure of QC.

The lumi package provides three different types of plotting functions to check color balance. Function `plotColorBias1D` plots the density of two color channels separately and shows them in red and green colors. By default, it pools both methylated and unmethylated probe intensities together, as shown in Figure 7. It can also separately plot the methylated and unmethylated probe intensities, as shown in Figure 8 and 9, or plot the CpG-site Intensity (details shown in the next subsection).

Similarly, we defined boxplot function, which plot two color channels separately. Function `boxplotColorBias` separately plots boxplot of the red and green color channels. Like `plotColorBias1D`, we can select to plot methylated and unmethylated probe intensities, as shown in Figure 10 and 11. Function `colorBiasSummary` can produce a quantile CpG-site Intensity summary of each sample. It has the same options of "channel" parameter.

```
> ## summary of color balance information of individual samples
> colorBiasSummary(example.lumiMethy[,1:8], channel='methy')

$red
        Treat1 Treat2 Treat3 Treat4  Ctrl1  Ctrl2  Ctrl3  Ctrl4
Min.     7.238  7.435  6.833  6.870  7.524  7.443  6.954  7.044
1st Qu.  8.200  8.358  7.672  7.516  8.375  8.290  7.607  7.913
Median   8.974  8.966  8.082  7.983  9.093  8.972  8.071  8.435
Mean     9.746  9.780  8.944  8.889  9.753  9.707  8.920  9.277
3rd Qu. 11.180 10.990  9.944 10.030 10.890 10.880  9.970 10.400
Max.    14.810 14.940 14.640 14.430 14.900 14.810 14.710 14.620

$green
        Treat1 Treat2 Treat3 Treat4  Ctrl1  Ctrl2  Ctrl3  Ctrl4
Min.     8.234  8.574  8.388  8.103  8.531  8.304  8.459  8.358
1st Qu.  8.860  9.234  9.079  8.785  9.175  9.236  9.138  9.202
Median   9.472  9.747  9.529  9.340  9.780  9.801  9.631  9.757
Mean    10.320 10.530 10.320 10.150 10.500 10.540 10.420 10.500
```

```
> plotColorBias1D(example.lumiMethy)
```

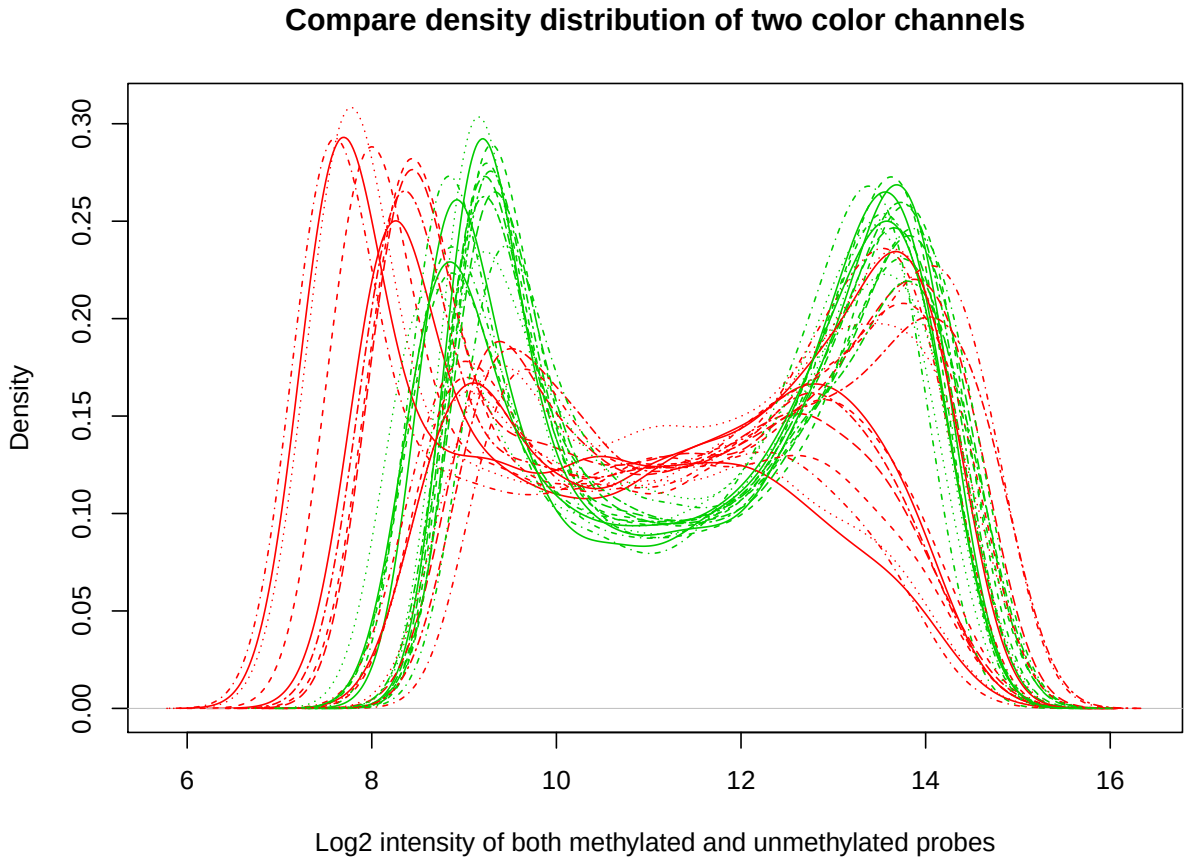

Figure 7: Density plot of two color channels (both methylated and unmethylated probe intensities) before preprocessing

```
> plotColorBias1D(example.lumiMethy, channel='methy')
```

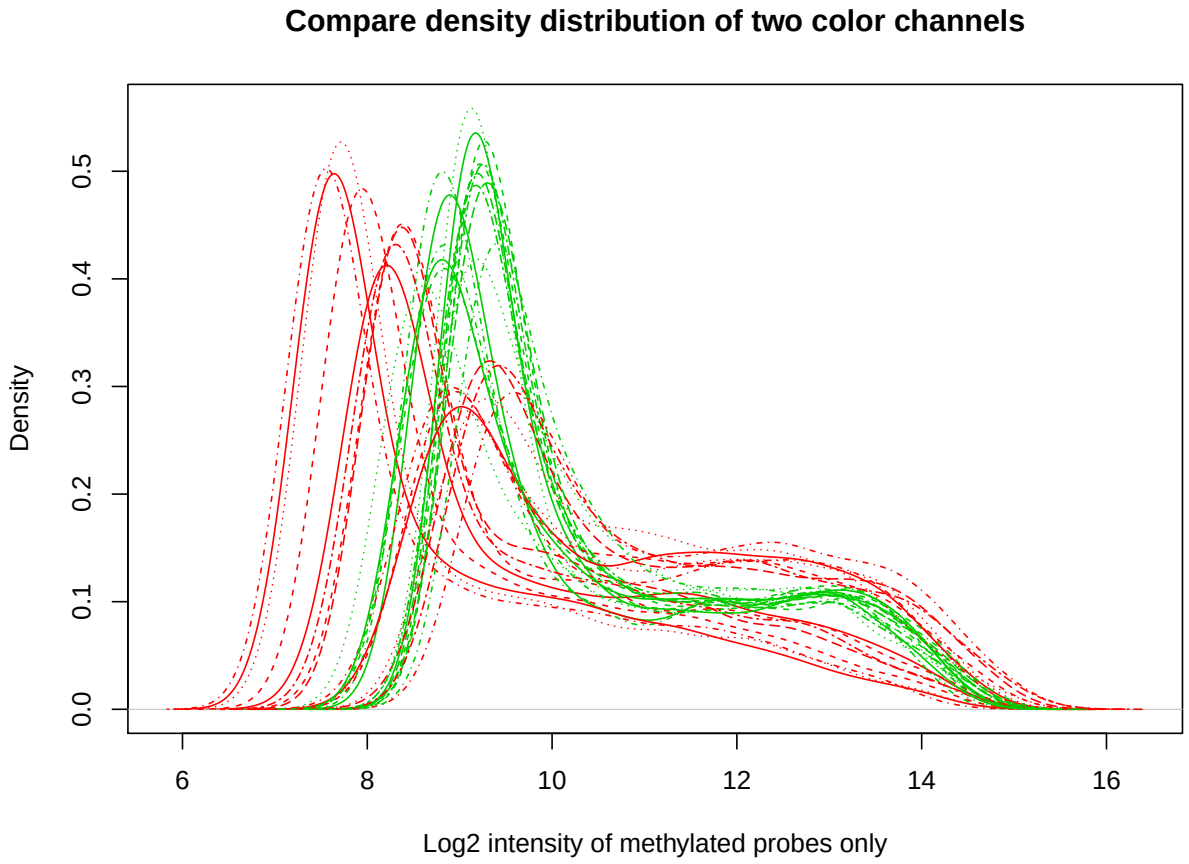

Figure 8: Density plot of two color channels (methylated probe intensities) before preprocessing

```
> plotColorBias1D(example.lumiMethy, channel='unmethy')
```

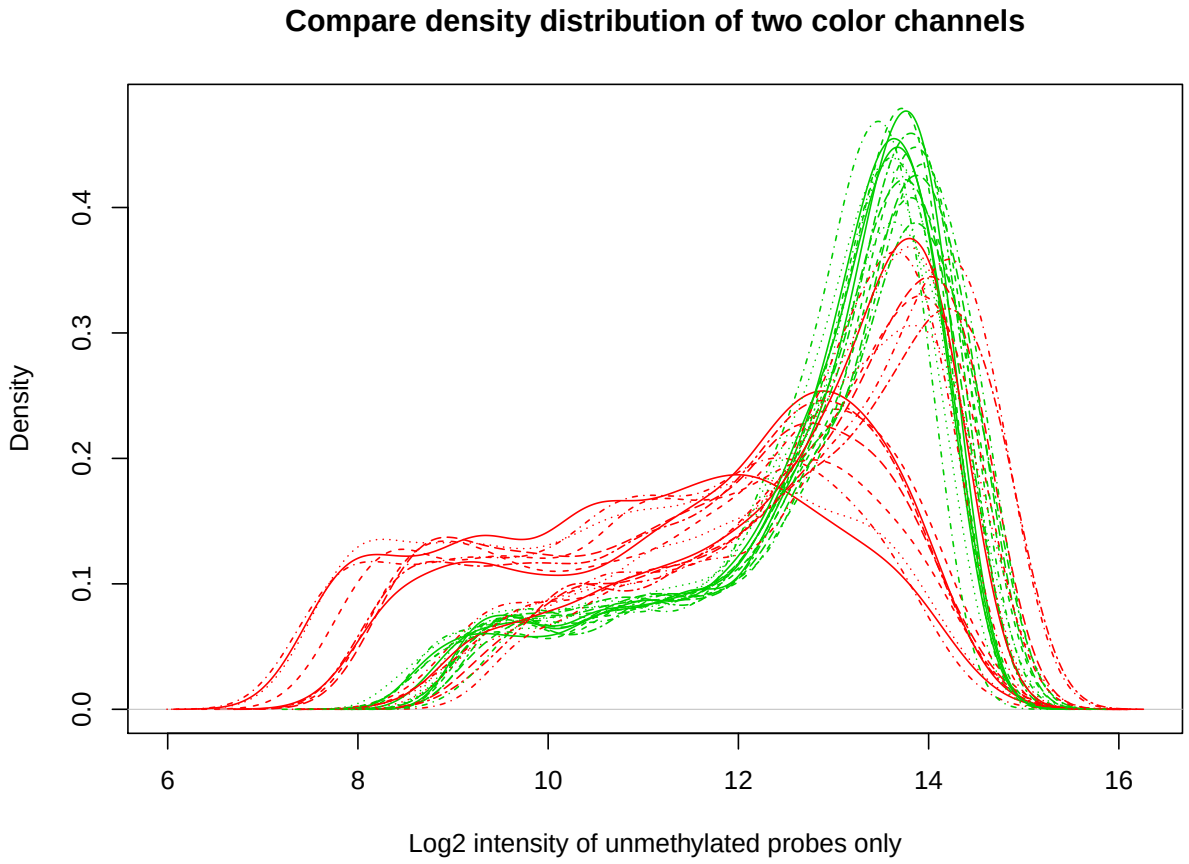

Figure 9: Density plot of two color channels (unmethylated probe intensities) before preprocessing

```
> boxplotColorBias(example.lumiMethy, channel='methy')
```

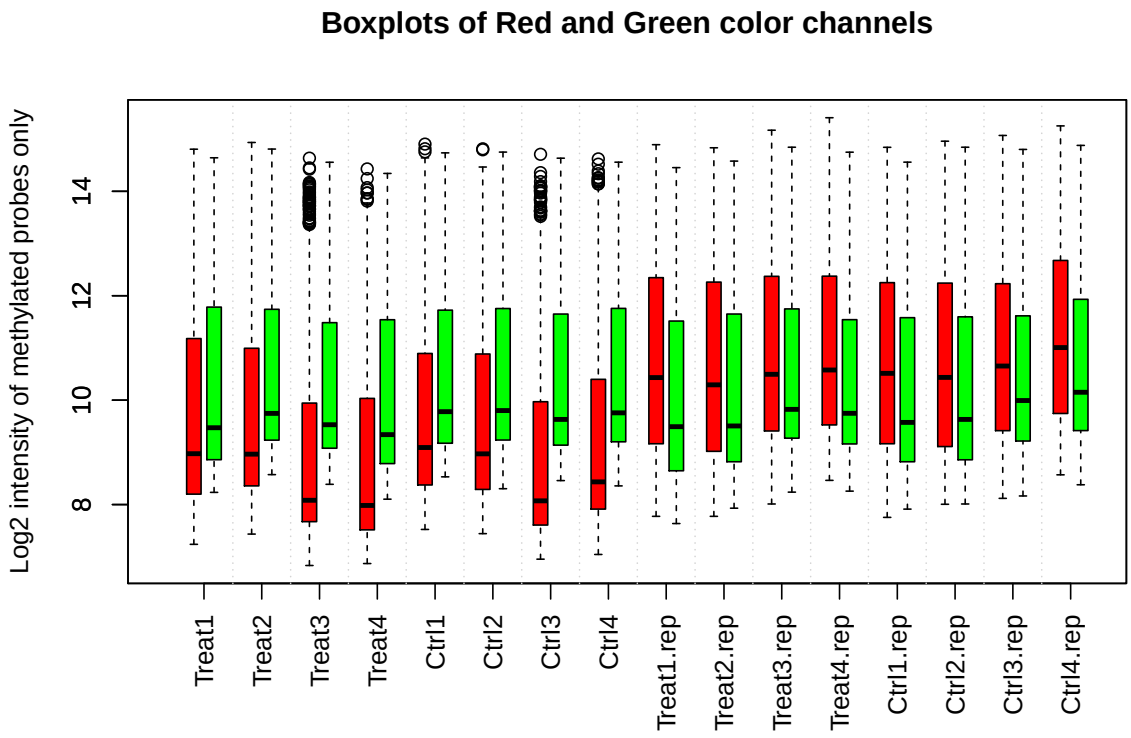

Figure 10: Box plot of two color channels (methylated probe intensities) before preprocessing

```
> boxplotColorBias(example.lumiMethy, channel='unmethy')
```

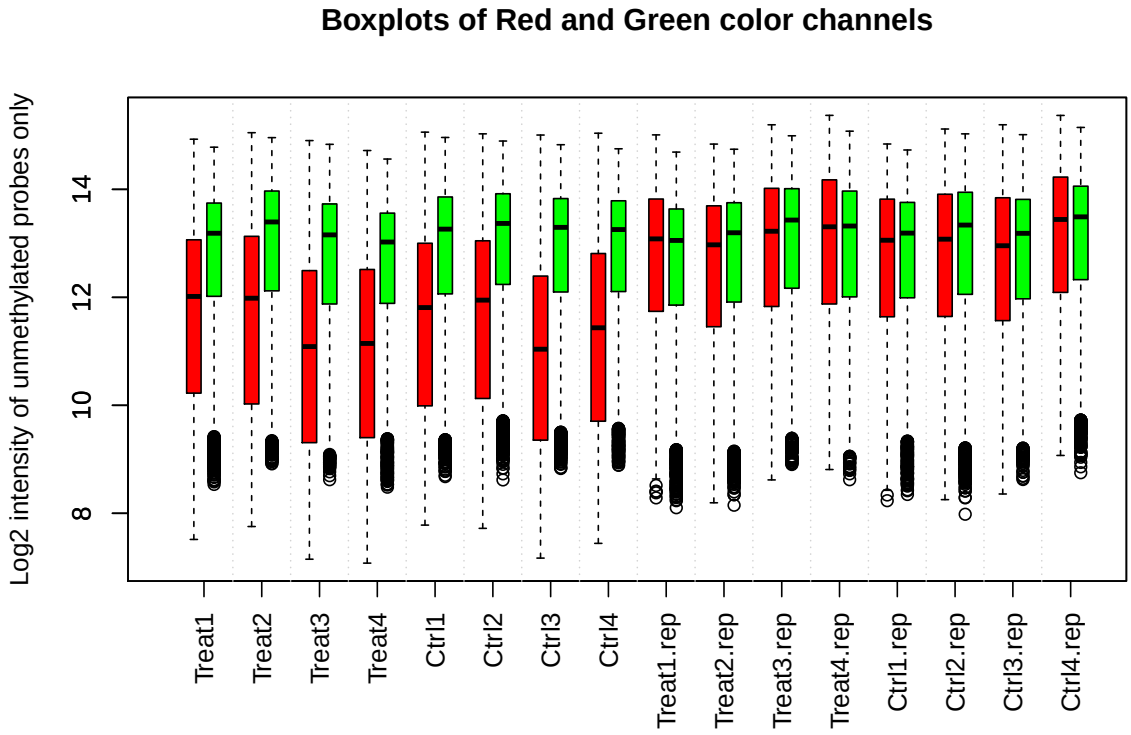

Figure 11: Box plot of two color channels (unmethylated probe intensities) before preprocessing

```
3rd Qu. 11.780 11.740 11.480 11.540 11.720 11.760 11.650 11.760
Max.    14.640 14.810 14.560 14.340 14.740 14.750 14.630 14.560
```

Function `plotColorBias2D` separately plots the methylated and unmethylated probe intensities in a 2-D scatter plot, and shows the interrogated CpG sites in red and green dots based on their color channels. Each time, `plotColorBias2D` can only plot one sample. By comparing the 2D scatter plots, we can easily see the details of color balance difference between samples. Figure 12 shows one example of color imbalance.

4.4 Quality assessment based on the distribution of CpG-site Intensity

Apart from color balance assessment, we propose to assess the sample quality based on the across sample distribution of the measured CpG site intensities. We defined the intensity of a measured CpG site, CpG-site Intensity, as the sum of methylated probe intensity and unmethylated probe intensity. As a single CpG site on one chromosome can only have two methylation status, if methylated probe for this CpG site has higher intensity, then the corresponding unmethylated probe should have lower intensity because only one probe (either methylated or unmethylated probe) can be bound on a particular CpG site on one copy of chromosome. That means the CpG-site Intensity is in proportional to the total copies of CpG sites. It also means that the methylation level (ratio between methylated and unmethylated probe intensities) changes should not have big effects on the CpG-site Intensity. Therefore, the CpG-site Intensity distribution of methylation microarray should still be similar across samples in different conditions if they have only methylation level difference, and we can check the sample quality based on CpG-site Intensity distribution. Figure 13 and 14 show the boxplot and density plot of two color channels. We can see it is much more obvious when using CpG-site Intensity to check color imbalance. Figure 15 and 16 show the CpG-site Intensity distribution of the example data. Because this example data has severe color imbalance, we can see the CpG-site Intensity distributions are similar but still have many differences. We can see the improvements after color balance adjustment.

We can also use other QC plots for expression data to check the sample consistency of methylation data base on CpG-site Intensity. Figure **??** shows pairwise plot of CpG-site Intensities of four samples, Ctrl1, Ctrl1.rep, Treat1 and Treat1.rep. Ctrl1, Ctrl1.rep and Treat1, Treat1.rep are technique replicates. We can clearly see two bands in the pairwise plot, which were caused by the difference of color imbalance between two batches.

4.5 Color balance adjustment

Based on the color balance assessment plots shown in previous two sections, we can see there are clear color balance differences between two batches in the example control and treatment dataset. That means we need to perform color balance adjustment before further analysis. Alternatively, we can separately process two color channels, but will course many inconveniences in the following up analysis. For example, we have to use different p-value

```
> plotColorBias2D(example.lumiMethy, selSample=1, cex=2)
```

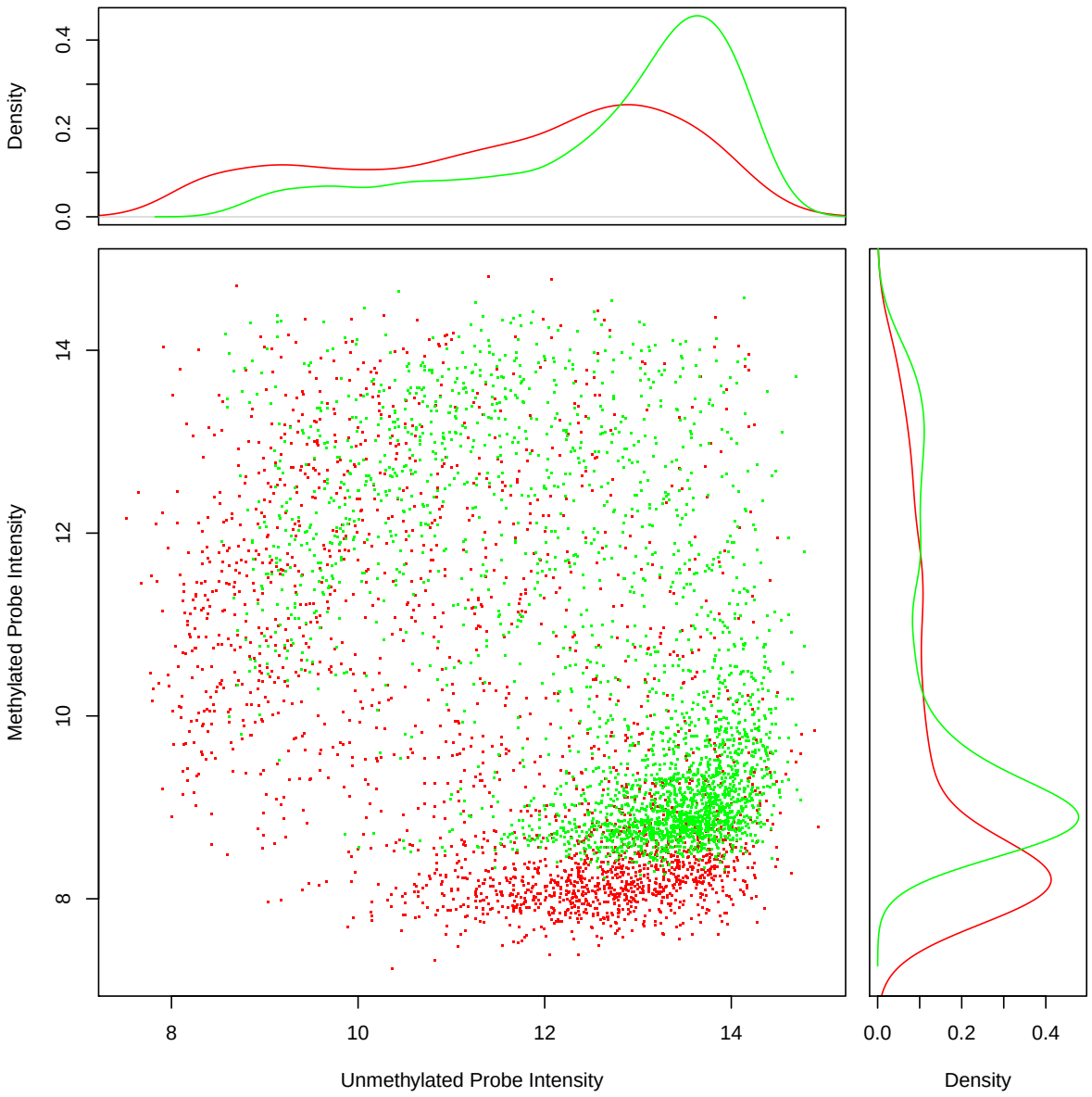

Figure 12: Scatter plot of methylated and unmethylated probe intensities to show color imbalance (example 1)

```
> boxplotColorBias(example.lumiMethy, channel='sum')
```

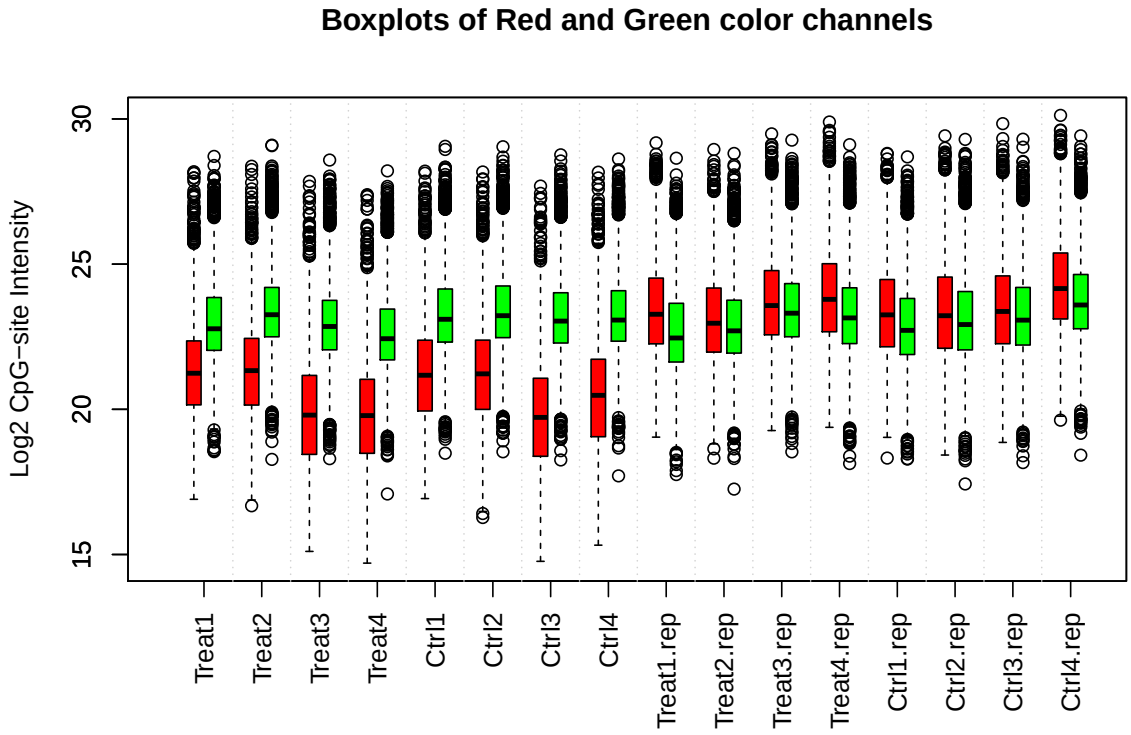

Figure 13: Box plot of of CpG-site Intensity (two color channels were plotted separately) before preprocessing

```
> plotColorBias1D(example.lumiMethy, channel='sum')
```

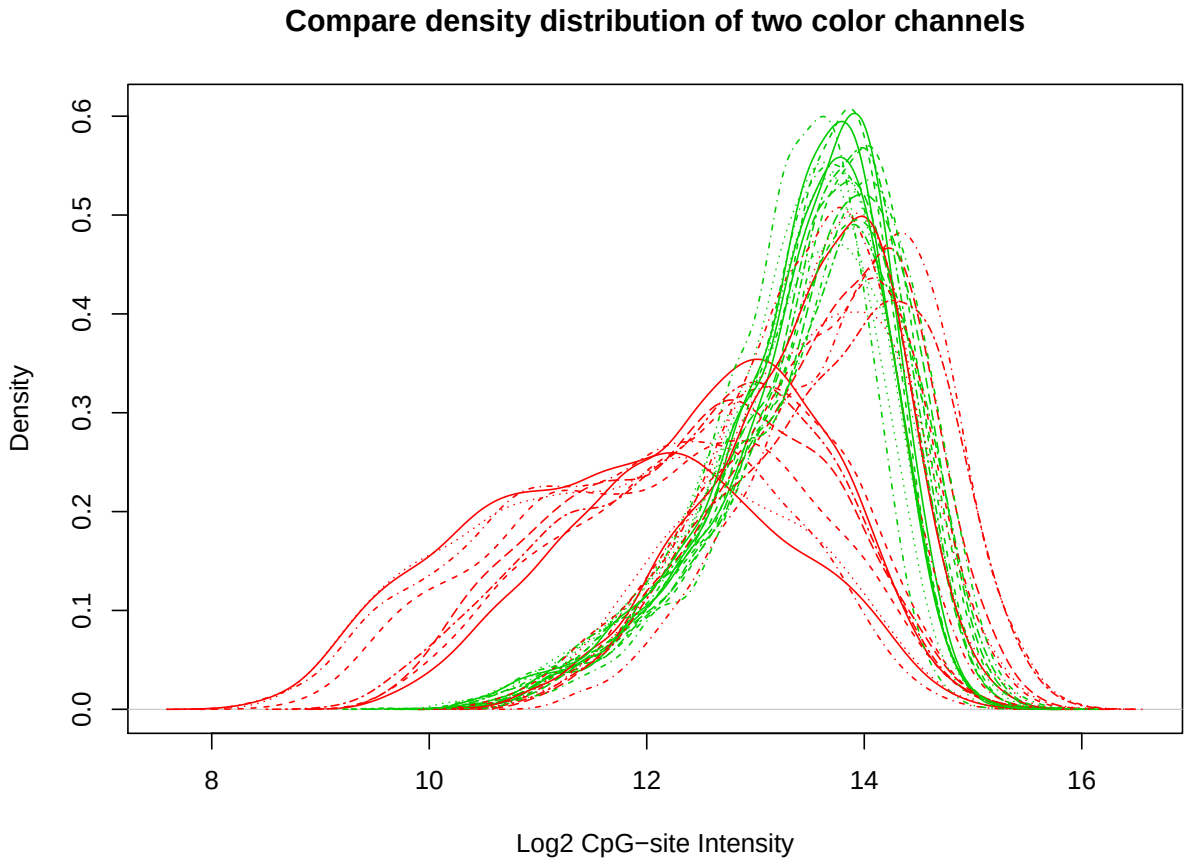

Figure 14: Density plot of CpG-site Intensity (two color channels were plotted separately) before preprocessing

```
> density(estimateIntensity(example.lumiMethy), xlab="log2(CpG-site Intensity)")
```

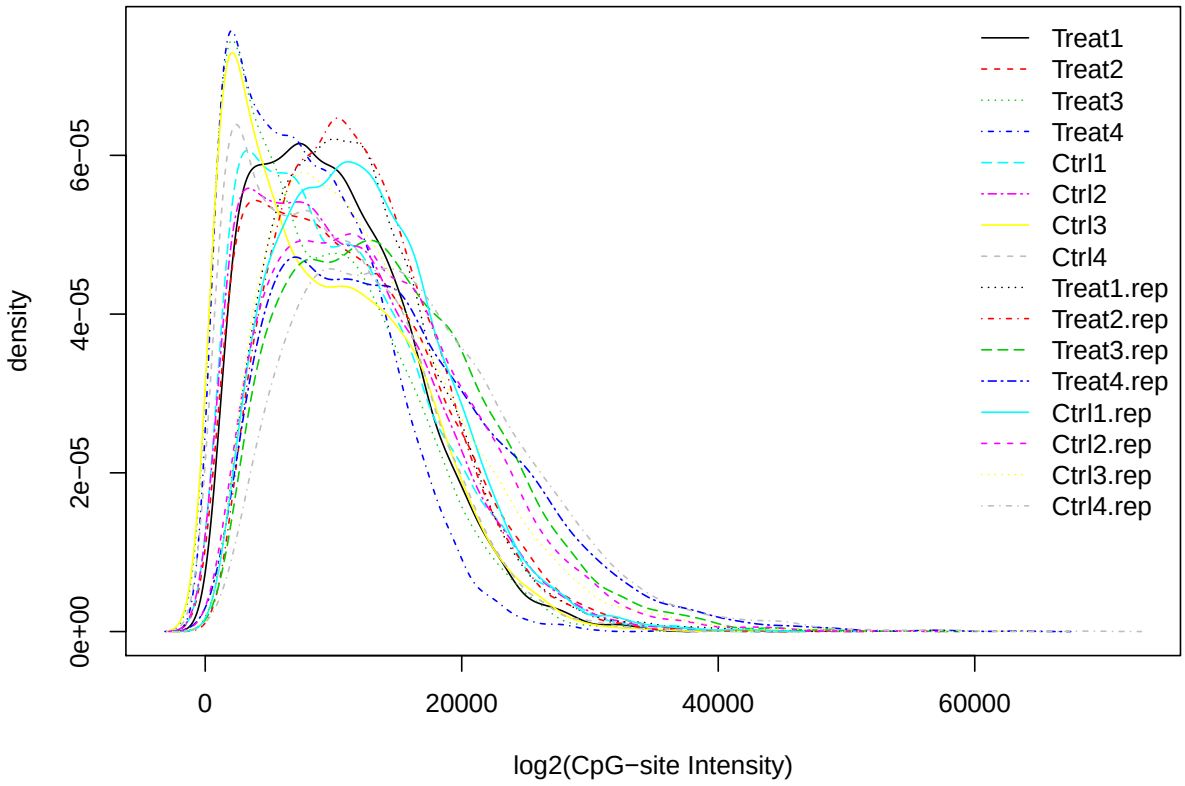

Figure 15: Density plot of CpG-site Intensity before preprocessing

```
> boxplot(estimateIntensity(example.lumiMethy))
```

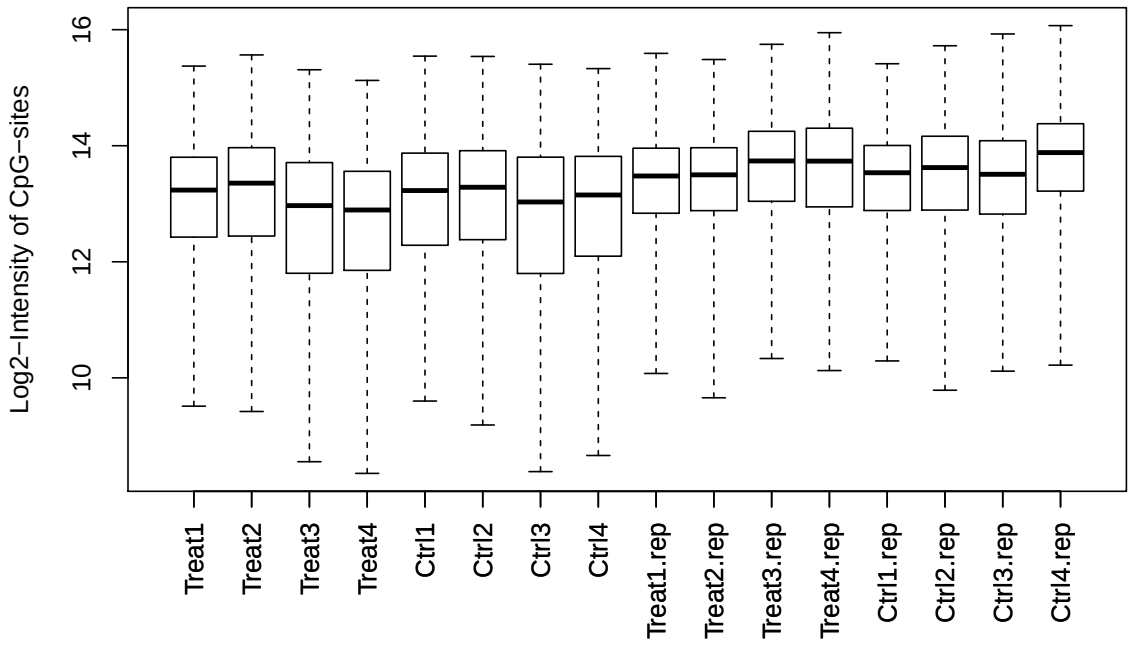

Figure 16: Boxplot of CpG-site Intensity before preprocessing

```
> ## get the color channel information
> colorChannel <- as.character(pData(featureData(example.lumiMethy))[, "COLOR_CHANNEL"]
> ## replace the "Red" and "Grn" as color names defined in R
> colorChannel[colorChannel == 'Red'] <- 'red'
> colorChannel[colorChannel == 'Grn'] <- 'green'
> ## select a subet of sample for pair plot
> selSample <- c( "Ctrl1", "Ctrl1.rep", "Treat1", "Treat1.rep")
> ## plot pair plot with the dots in scatter plot colored based on the color channels
> pairs(estimateIntensity(example.lumiMethy[, selSample]), dotColor= colorChannel, mai
```

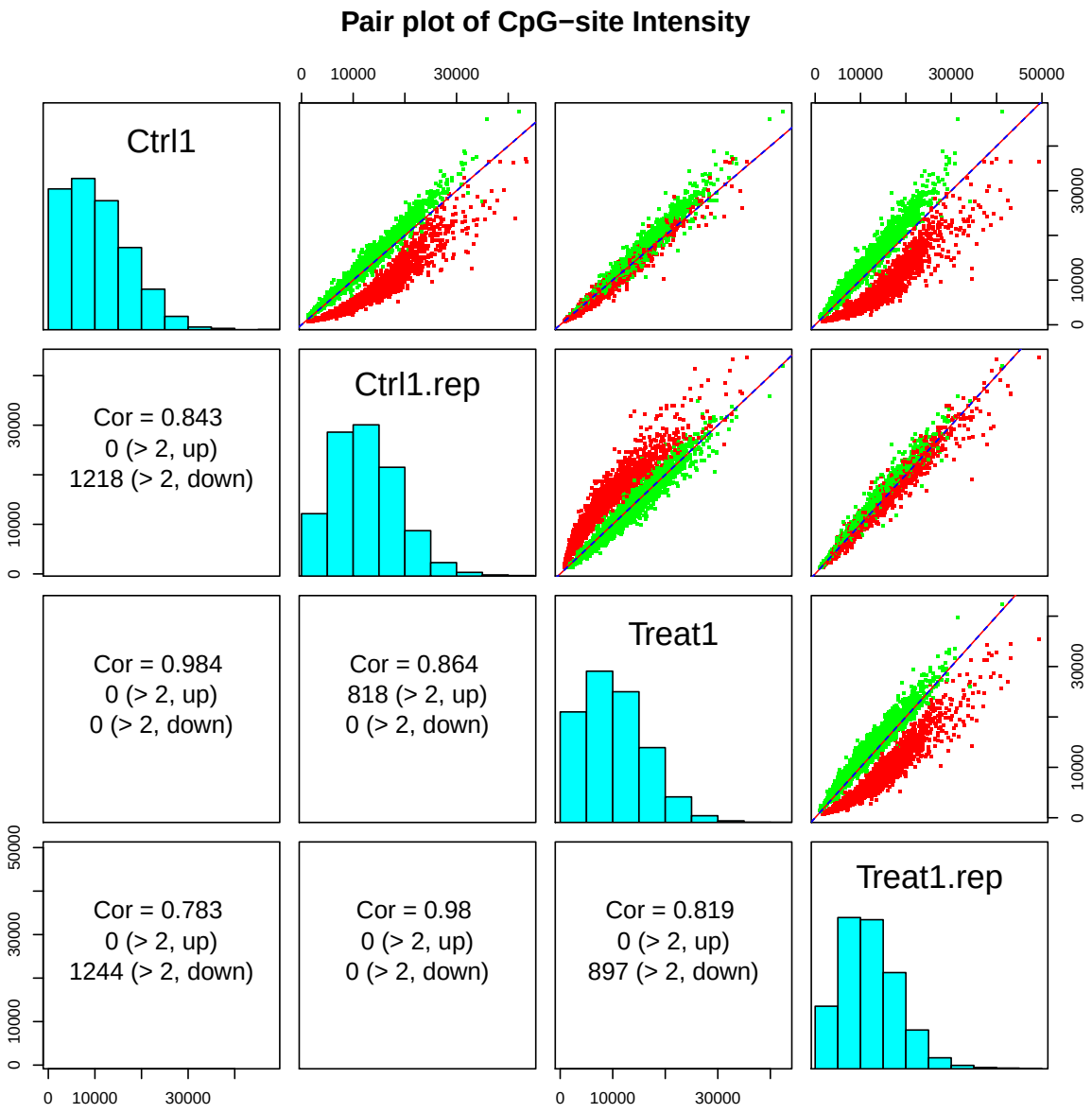

Figure 17: Pair plot of CpG-site Intensity with colors before preprocessing

or fold-change thresholds in identifying the differentially methylated CpG sites. Moreover, considering we may not have access to the color channel information later on, performing color balance adjustment first will make following up analysis much more convenient.

The lumi package provides `lumiMethyC` function for color balance adjustment. The basic idea of color balance adjustment is to treat it as the normalization between two color channels. Function `lumiMethyC` provides two color balance adjustment methods: "quantile" and "ssn". Because two color channels have different number of probes and not match each other, we cannot directly apply regular "quantile" normalization used in expression microarray analysis. To solve this problem, we designed a `smoothQuantileNormalization` method. Basically, it requires the quantile normalization should be smooth. By doing this, the `smoothQuantileNormalization` can further resolve one major withdraw of regular quantile normalization, i.e., quantile normalization doesn't work well in the areas with sparse density, like two extremes of data distribution. Figure 18 and 19 show density and boxplot of two color channels after color balance adjustment using smooth quantile normalization. Figure 20 shows the scatter plots after color balance adjustment using smooth quantile normalization, we can see the distribution becomes very similar between two color channels. And the change of pair plot after color balance is more obvious by comparing Figure 21 and Figure 17.

The Illumina Methylation Infinium 450K uses both Infinium I and II assays with increased breadth of coverage. The Infinium I assay employs two bead types per CpG locus with one for methylated and the other for unmethylated states, and both states are measured with the same fluorescent color. This design is the same as the previous Infinium methylation 27k array. The Infinium II assay uses one bead type and single base extension chemistry with the methylated state determined by the green dye and the unmethylated state by red dye, in which dye bias becomes bigger concern because it is directly related with methylation status. Based on the assumption that the color bias of CpG-sites is independent from the type of probe design, we estimate the color-bias correction curve based on the Infinium I CpG-sites (independent of methylation status), and then apply this correction curve to the CpG-sites with the Infinium II design. Our results show this correction method is effective.

```
> ## summary of color balance information of individual samples
> lumiMethy.c.adj <- lumiMethyC(example.lumiMethy)

Perform quantile color balance adjustment ...
Processing sample Treat1 ...
Processing sample Treat2 ...
Processing sample Treat3 ...
Processing sample Treat4 ...
Processing sample Ctrl1 ...
Processing sample Ctrl2 ...
Processing sample Ctrl3 ...
Processing sample Ctrl4 ...
Processing sample Treat1.rep ...
```

```
> plotColorBias1D(lumiMethy.c.adj, channel='sum')
```

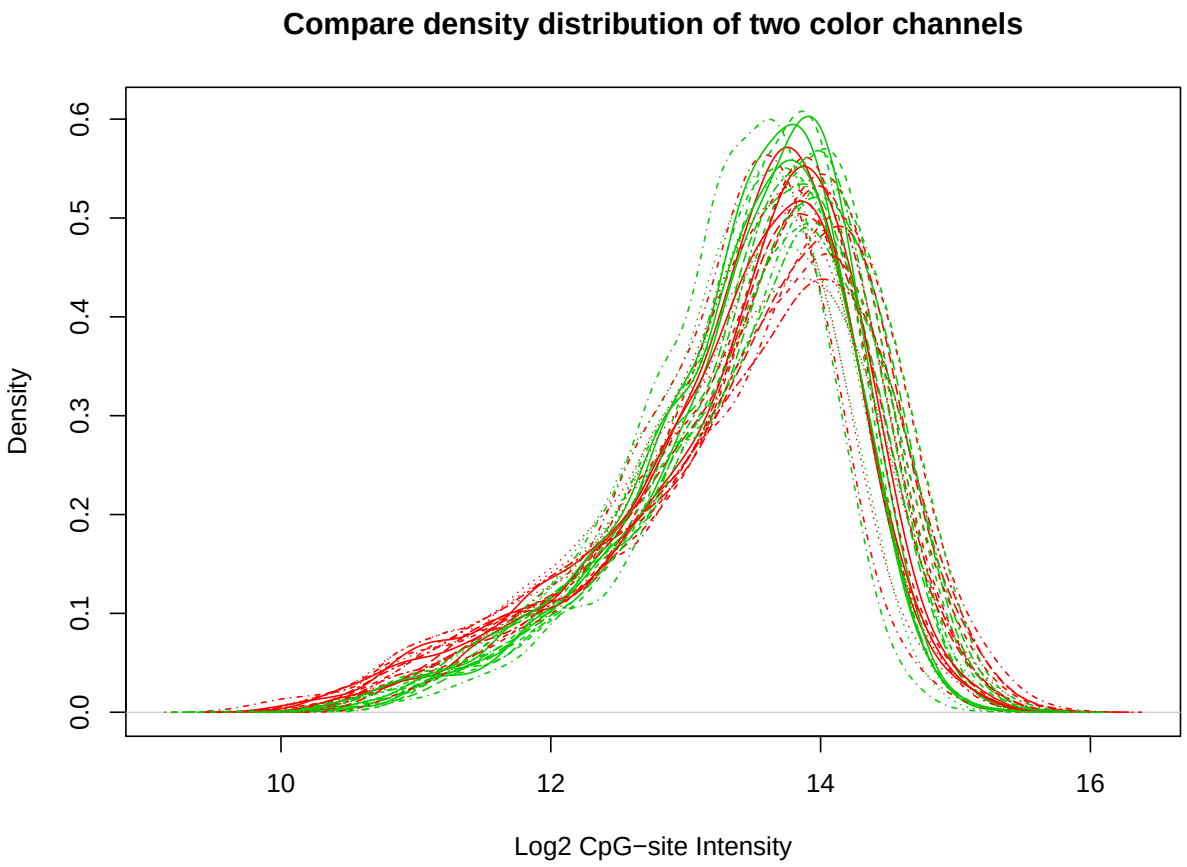

Figure 18: Density plot of CpG-site Intensity (two color channels were plotted separately) after color balance adjustment

```
Processing sample Treat2.rep ...
Processing sample Treat3.rep ...
Processing sample Treat4.rep ...
Processing sample Ctrl1.rep ...
Processing sample Ctrl2.rep ...
Processing sample Ctrl3.rep ...
Processing sample Ctrl4.rep ...
```

4.6 Background level correction

Illumina provides negative control probes for the estimation of background levels. The information of the control probe information is kept in the controlData slot, which is a

```
> boxplotColorBias(lumiMethy.c.adj, channel='sum')
```

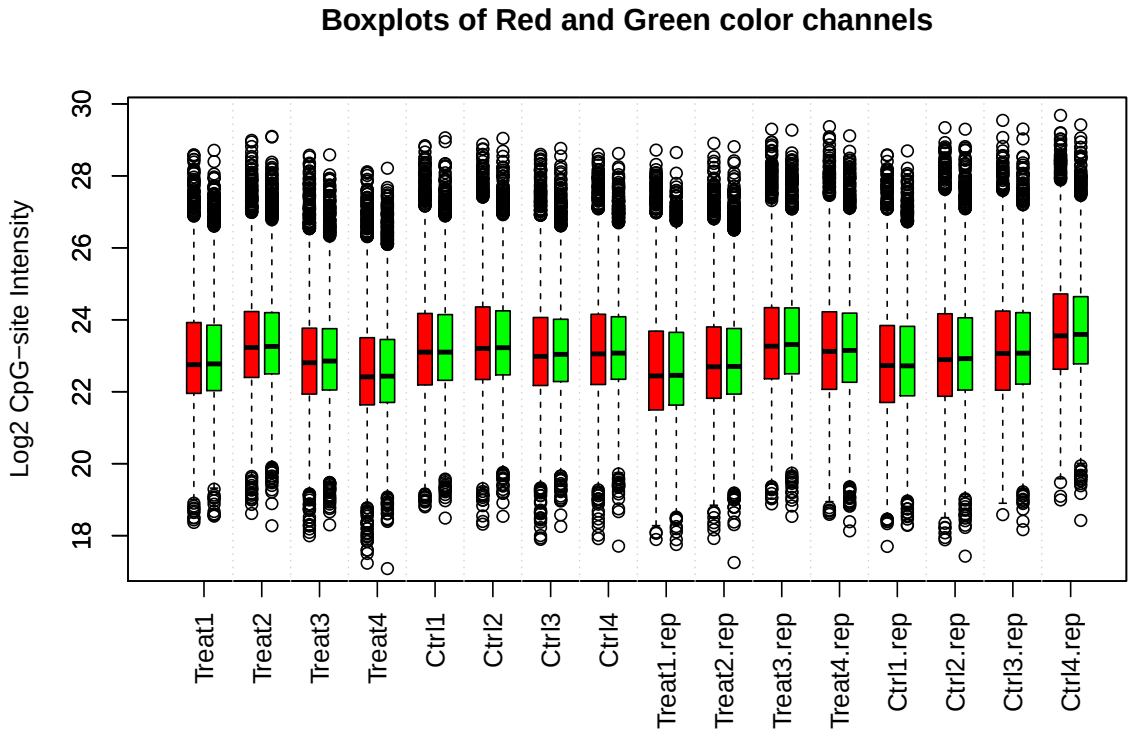

Figure 19: Box plot of of CpG-site Intensity (two color channels were plotted separately) after color balance adjustment

```
> ## plot the color balance adjusted scatter plot of two color channels
> plotColorBias2D(lumiMethy.c.adj, selSample=1, cex=2)
```

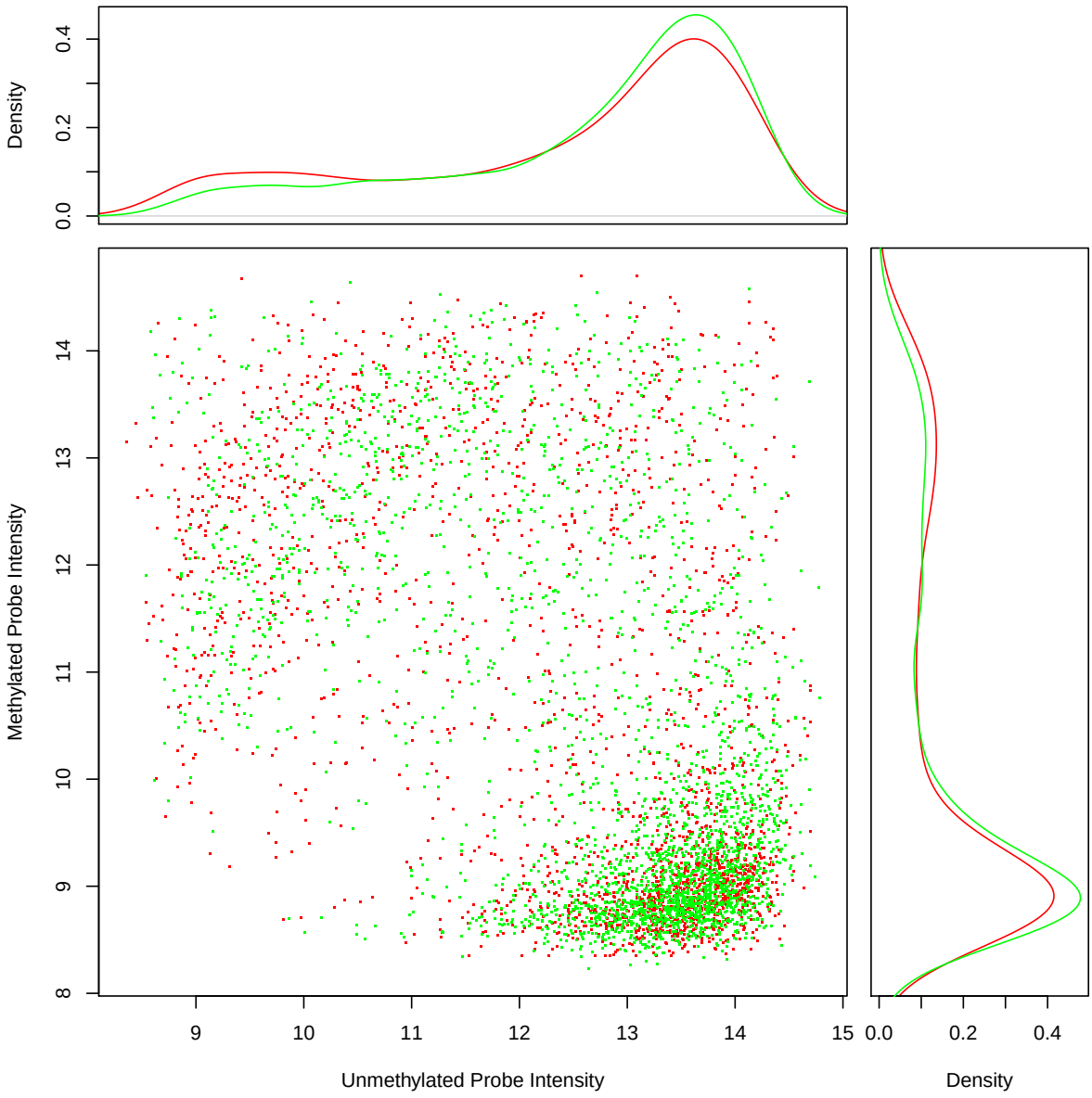

Figure 20: Scatter plot of methylated and unmethylated probe intensities after color balance adjustment (example 1)

```
> ## plot pairwise plot after color balance adjustment
> pairs(estimateIntensity(lumiMethy.c.adj[, selSample]), dotColor= colorChannel, main=
```

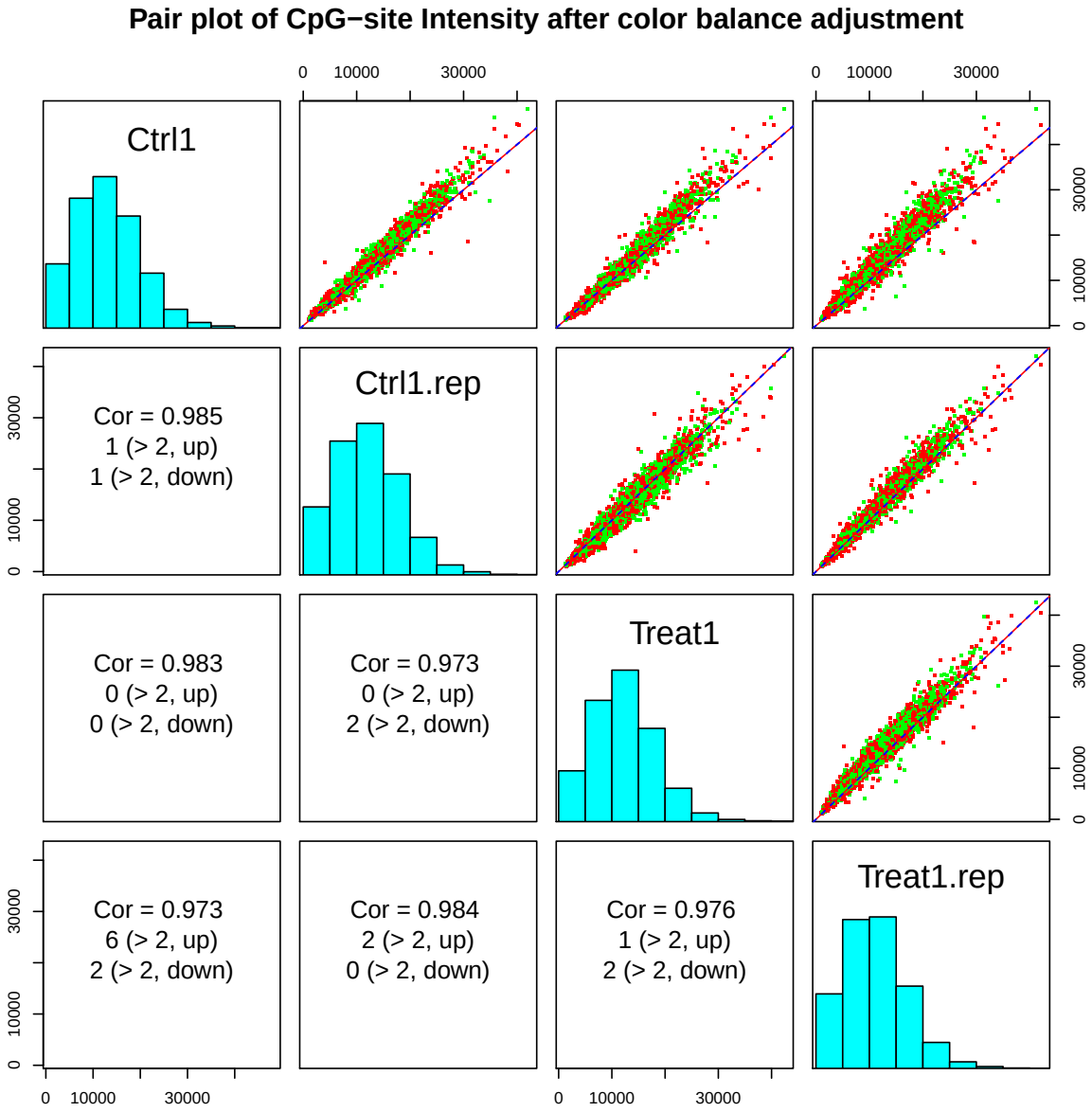

Figure 21: Pair plot of CpG-site Intensity with colors

MethyLumiQC object. When the controlData includes the negative control probe information, the background estimation will be the median of the negative control probes. Red and Green color channels will be estimated separately. However, in many cases the information of negative control probes usually was not outputted together with the methylation data. As a result, we have to find other methods to estimate the background levels of individual samples. As Illumina methylation has two color channels, the background levels of two color channel can be different due to color imbalance. So color balance adjustment is suggested to be performed before background adjustment, or background estimation should be separately performed in each color channel.

The estimation of background level of Infinium methylaton microarray is based on the assumption that the lots of CpG sites are unmethylated, which results in a density mode of the intensities measured by methylated probes. The position of this mode represents the background level. Figure 22 shows the background mode of methylated probe data of first five example samples. We can see the differences between two color channels. Function `estimateMethylationBG` estimates the background level of each sample. Function `bgAdjustMethylation` adjusts the background levels based on the estimation returned by `estimateMethylationBG`. The `lumiMehtyB` function is a general function for background correction, which by default call `estimateMethylationBG` function. If we directly perform background adjustment on the raw data, we have to perform background adjustment separately on two color channels, which can be done by setting the parameter "separateColor" of `lumiMethyB` as TRUE. Figure 23 shows the density plot of methylated probes after background correction with two color channels processed separately. We can see the density modes of two color channels overlapping each other after background correction performed separately performed in each color channel. Alternatively, we can perform background adjustment after color balance adjustment, as shown in Figure 24.

```
> ##separately adjust backgrounds of two color channels
> lumiMethy.b.adj <- lumiMethyB(example.lumiMethy, method="bgAdjust2C", separateColor=

Perform bgAdjust2C background correction ...

> ##background adjustment of individual samples
> lumiMethy.bc.adj <- lumiMethyB(lumiMethy.c.adj, method="bgAdjust2C")

Perform bgAdjust2C background correction ...
```

4.7 Data normalization

Because the total amount of CpG methylation can differ significantly from sample to sample in different conditions, many assumptions used by mRNA expression microarrays are not valid in processing DNA methylation data. So directly applying the normalization methods designed for expression microarray to the methylation data, like M-value or Beta-value, is inappropriate. To circumvent this difficulty, we perform normalization at the probe level, i.e.,

```
> ## plot the background mode of methylated probe data of first five example samples
> plotColorBias1D(example.lumiMethy[,1:5], channel='methy', xlim=c(-1000,5000), logMod
```

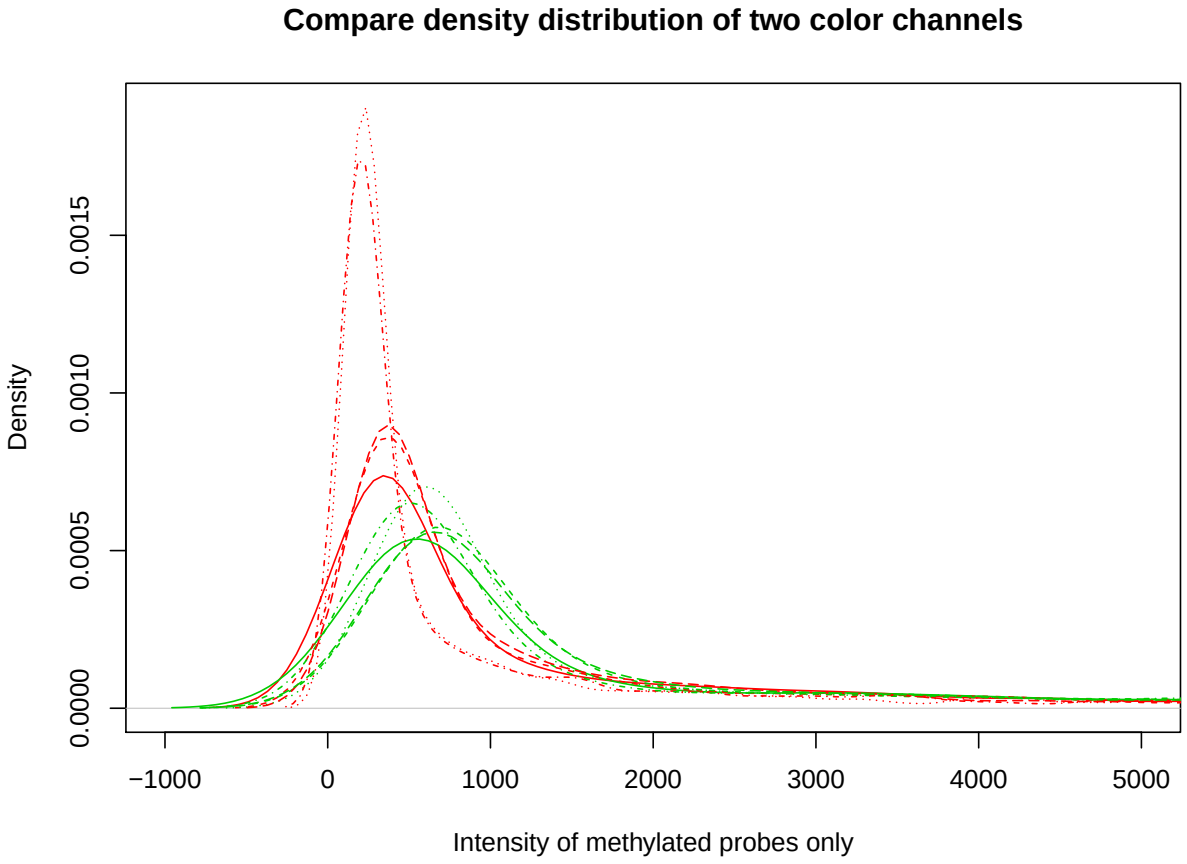

Figure 22: Background mode shown in the density plot of methylated probes before background adjustment

```
> ## plot the background mode of methylated probe data of first five example samples
> plotColorBias1D(lumiMethy.b.adj [,1:5], channel='methy', xlim=c(-1000,5000), logMode
```

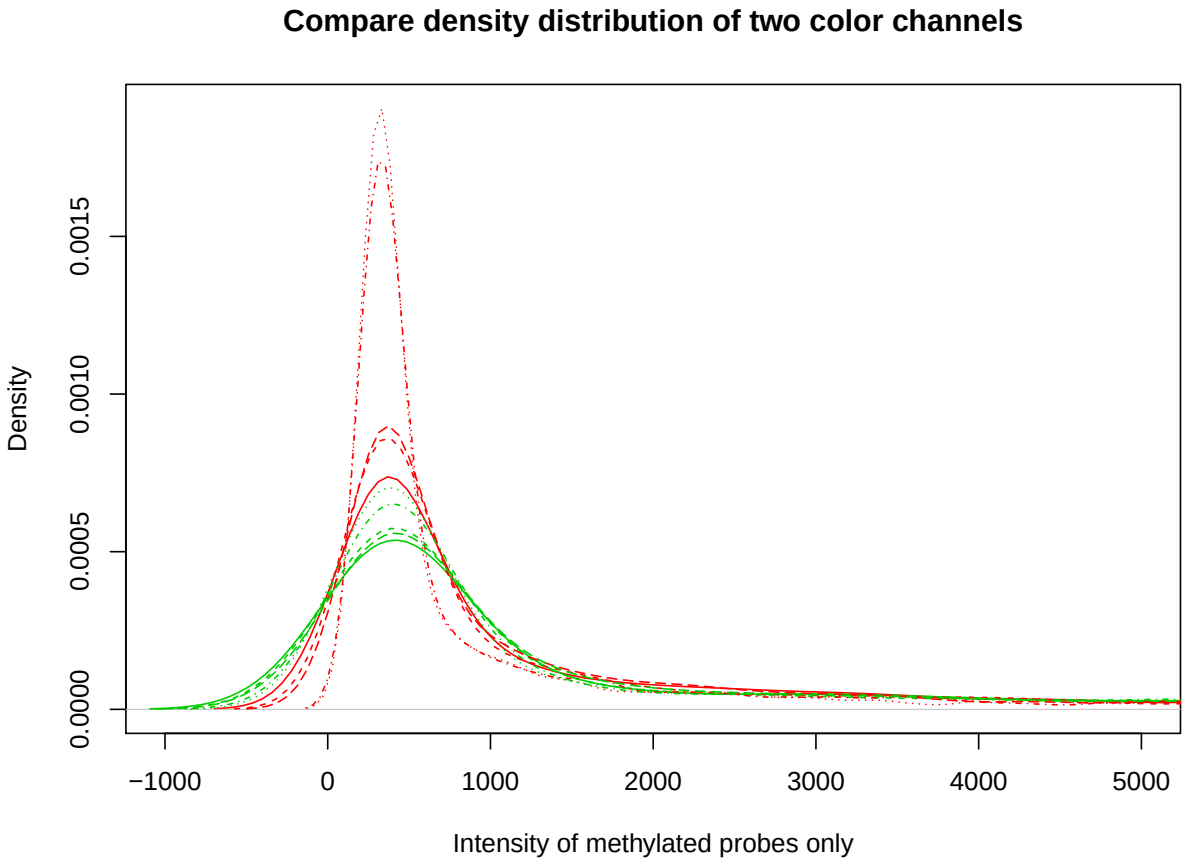

Figure 23: Background mode shown in the density plot of methylated probes after background adjustment (separately in two color channel)

```
> ## plot the background mode of methylated probe data of first five example samples
> plotColorBias1D(lumiMethy.bc.adj [,1:5], channel='methy', xlim=c(-1000,5000), logMod
```

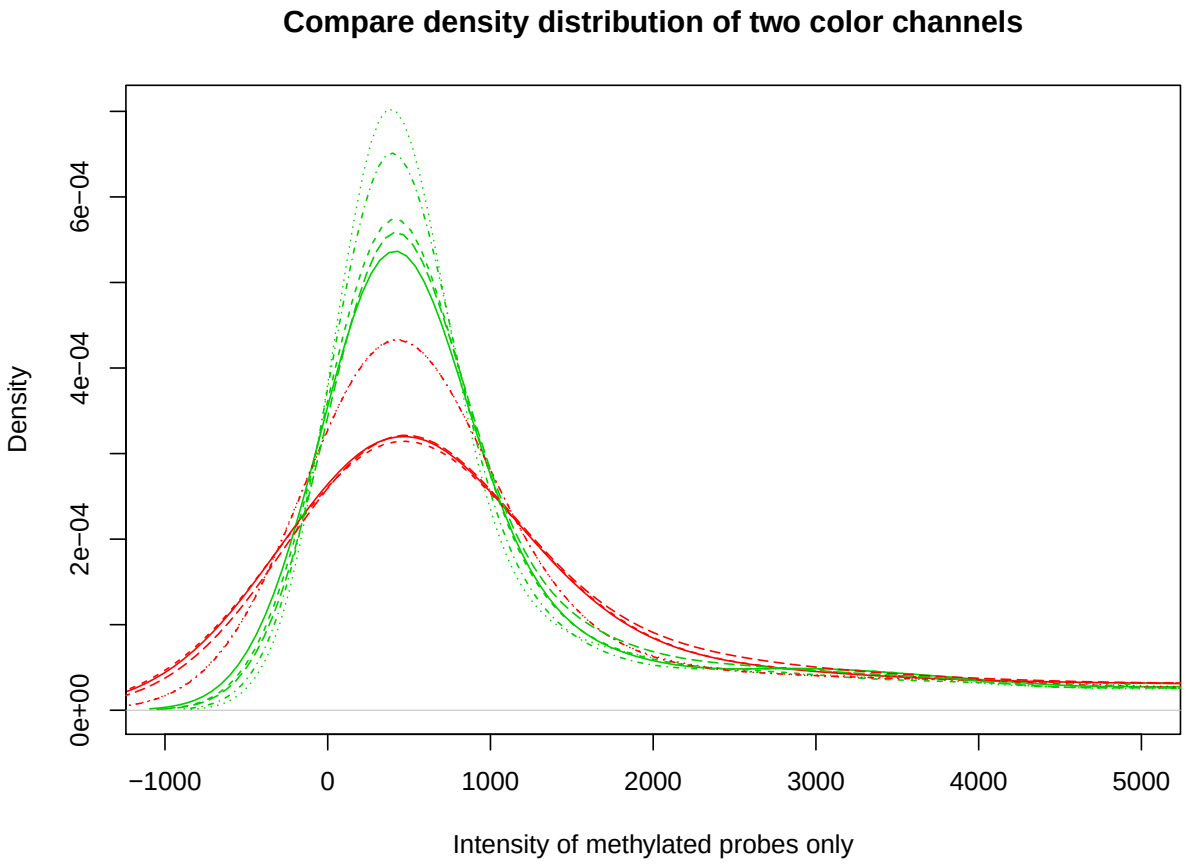

Figure 24: Background mode shown in the density plot of methylated probes after background and color adjustment

normalize the intensities of methylated and unmethylated probes instead of the summarized methylation levels.

The lumi package provides `lumiMethyN` function for probe level normalization. Currently, it supports two methods, "quantile" and "ssn". Users can also provide there own normalization methods, as long as it imports and outputs a data matrix. See subsection " Use user provided preprocessing functions" of for more details. The assumption of "quantile" normalization is that the intensity distribution of the pooled methylated and unmethylated probes, as shown in Figure 7, are the similar for different samples. The quantile normalizaiton is the default method. The "ssn" (simple scaling normalization) method first estimates the background level use function `estimateMethylationBG`, then scale the background subtracted data to the same average intensity level, and finally adjust the background level to a user specified level. The assumption of quantile normalization is more aggressive than SSN, it assumes the distribution of pooled intensities of methylated and unmethylated probes are the same. The selection of SSN and quantile normalization is similar with the case in expression microarray. It depends on the data itself and the purpose of analysis.

To check the effectiveness of normalization, we first check the sample relations after normalization, as shown in Figure 25. We can see most of the technique replicates are clustered together after normalization. Comparing with the sample relations of the raw data shown in Figure 2, the improvement is obvious. Figure 26 show the density plots of methylation levels after quantile normalization. The red and black colors represent "Treatment" and "Control" samples, respectively. We can see the mode positions of normalized data are more consistent across samples while keeping the methylation differences in the middle methylation range.

Figure 27 show the density plot of CpG-site Intensity after normalization. Figure 28 shows the color bias boxplot after preprocessing. We can see they are very consistent across samples. Finally, we show the pair plot of M-value after preprocessing. We expect there is little difference between technique replicates, but more difference between "Treatment" and "Control" samples, this is exactly what is shown in Figure 29. All of these results show the color balance adjustment plus normalization is effective.

```
> ## Perform quantile normalization based on color balance adjusted data
> lumiMethy.c.quantile <- lumiMethyN(lumiMethy.c.adj, method='quantile')

Perform quantile normalization ...

> ## Perform SSN normalization based on color balance adjusted data
> lumiMethy.c.ssn <- lumiMethyN(lumiMethy.c.adj, method='ssn')

Perform ssn normalization ...
```

4.8 Use user provided preprocessing functions

For convenience to the users, the user can specify their own preprocessing function when call `lumiMethyB`, `lumiMethyC` and `lumiMethyN`. The input and output of the user provided function should be a intensity matrix (pool of methylated and unmethylated probe intensities). Following are some examples of using user defined preprocessing functions.

```
> plotSampleRelation(lumiMethy.c.quantile, method='cluster', cv.Th=0)
```

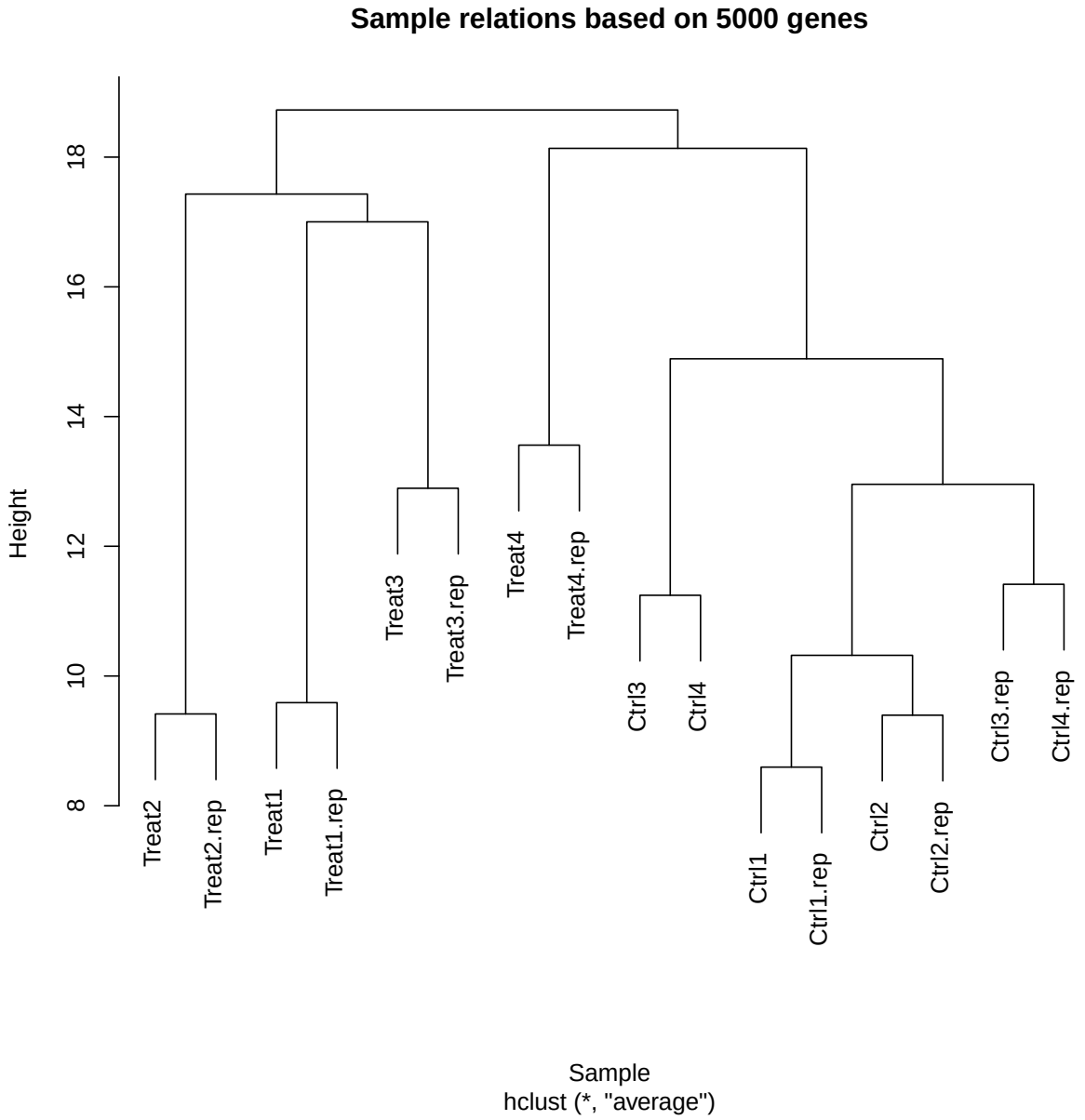

Figure 25: Overall sample relations shown as a hierarchical tree after normalization

```
> ## plot the density of M-values after quantile normalization
> density(lumiMethy.c.quantile, col= sampleColor, main="Density plot after quantile no
```

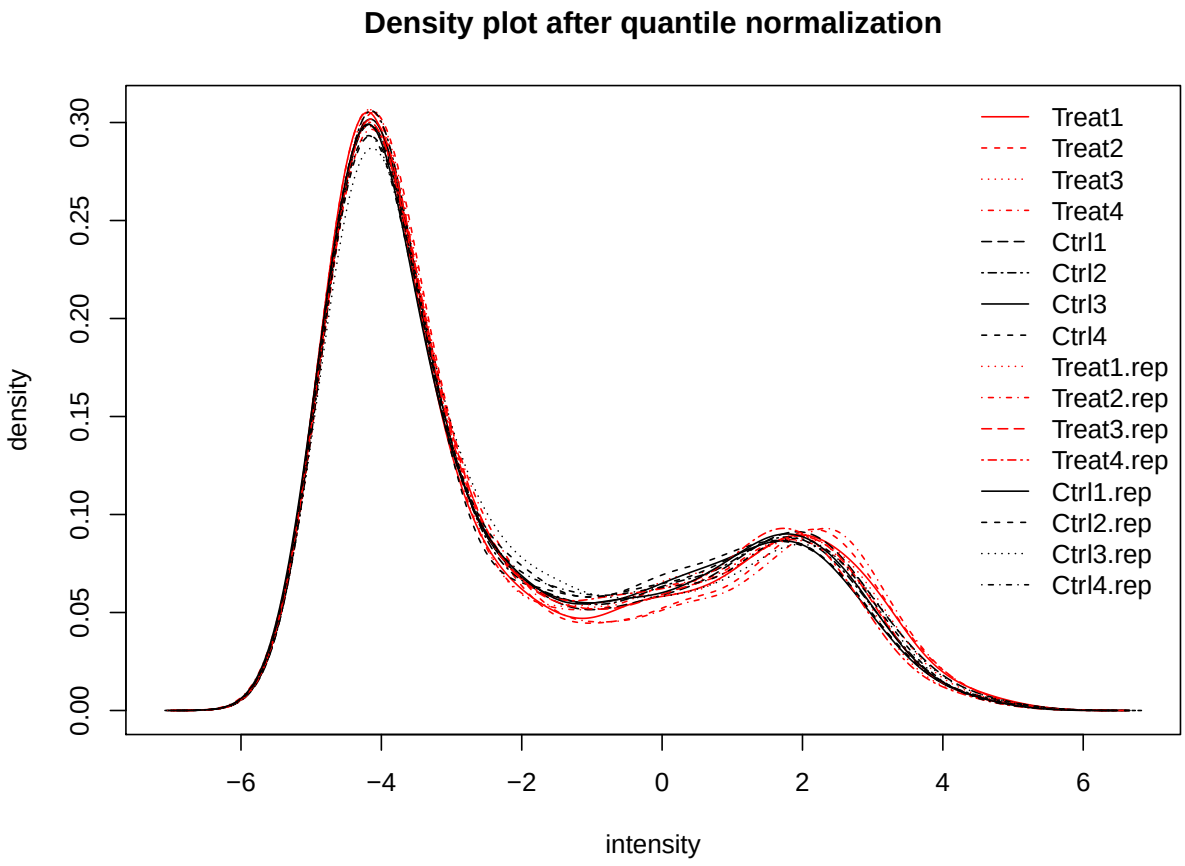

Figure 26: Density plot of M-values after quantile normalization

```
> density(estimateIntensity(lumiMethy.c.quantile), col= sampleColor,  xlab="log2(CpG-s
```

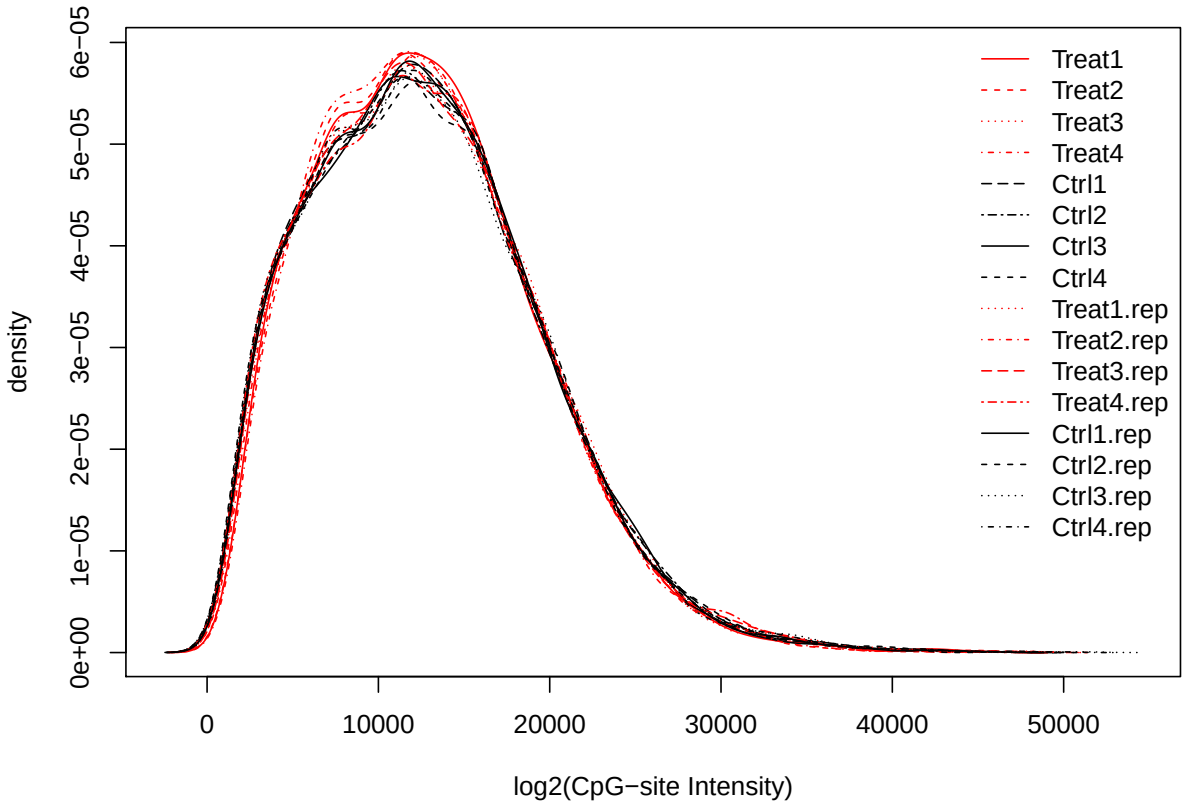

Figure 27: Density plot of CpG-site Intensity after quantile normalization

```
> boxplotColorBias(lumiMethy.c.quantile, channel='sum')
```

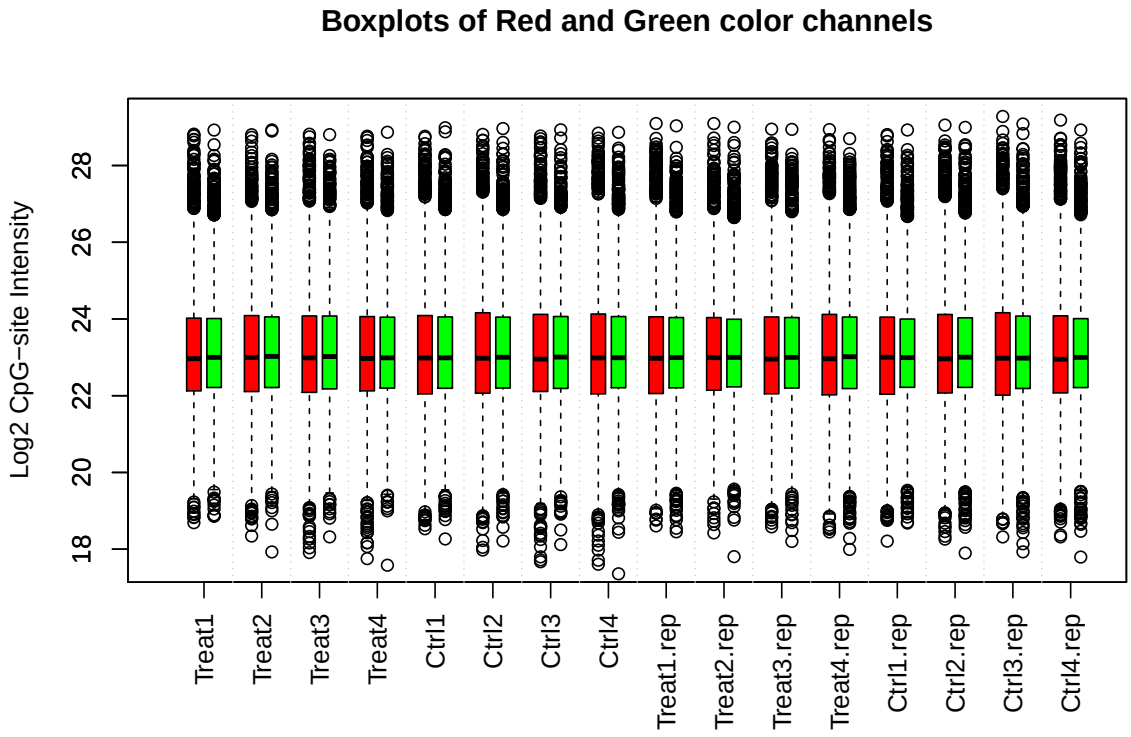

Figure 28: Box plot of of CpG-site Intensity (two color channels were plotted separately) after normalization

```
> ## select a subet of sample for pair plot
> selSample <- c( "Ctrl1", "Ctrl1.rep", "Treat1", "Treat1.rep")
> ## plot pair plot with the dots in scatter plot colored based on the color channels
> pairs(lumiMethy.c.quantile[, selSample], dotColor= colorChannel, main="Pair plot of
```

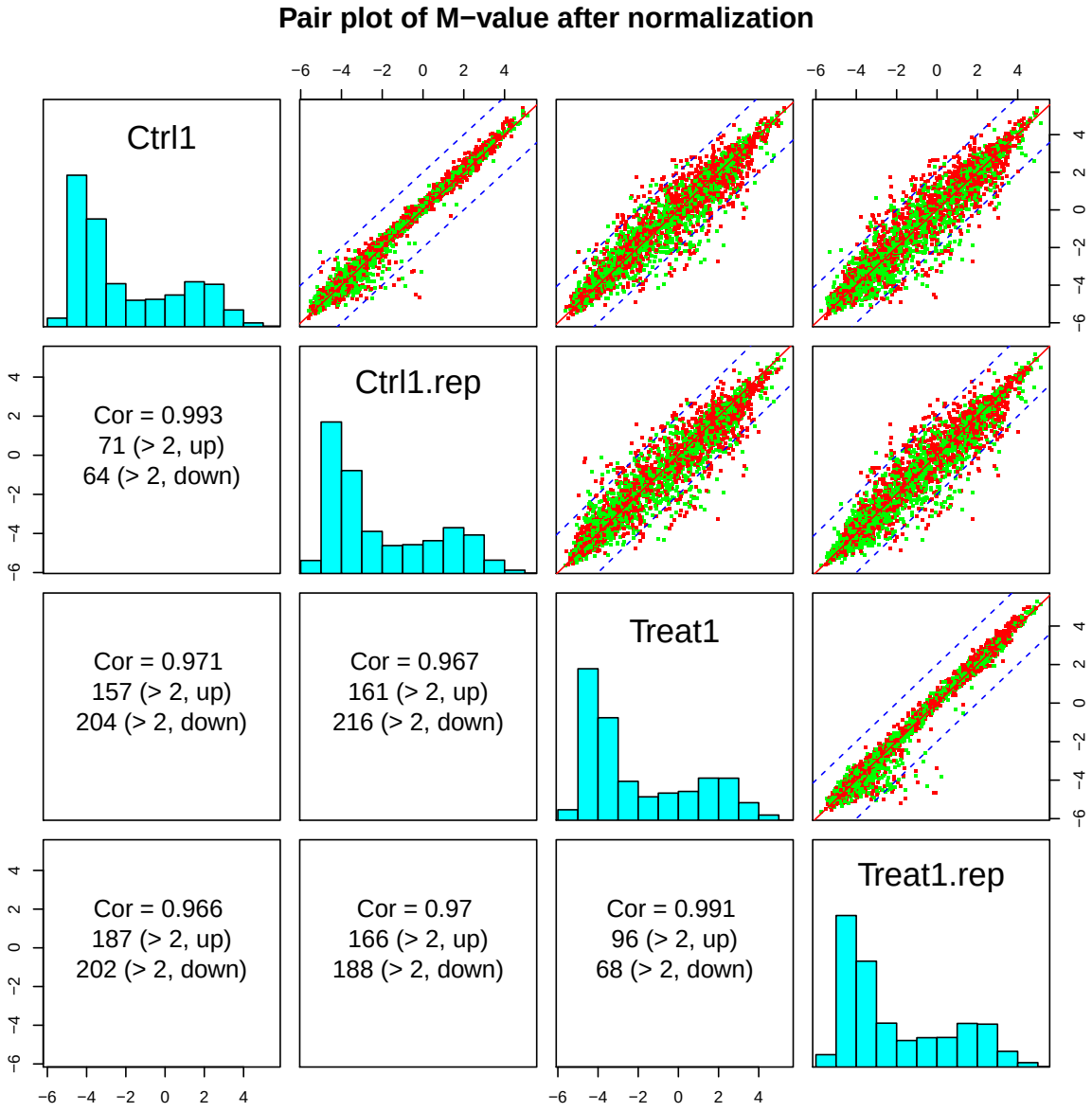

Figure 29: Pair plot of M-value after normalization

```
> ## suppose "userB" is a user defined background adjustment method
>  lumiMethy.b.adj <- lumiMethyB(example.lumiMethy, method=userB, separateColor=TRUE) #
> ## suppose "userC" is a user defined color balance adjustment method
>  lumiMethy.c.adj <- lumiMethyC(example.lumiMethy, method=userC, separateColor=TRUE) #
> ## suppose "userN" is a user defined probe level normalization method
>  lumiMethy.c.n <- lumiMethyN(lumiMethy.c.adj, method= userN, separateColor=TRUE) # n
```

4.9 Options to separately process each color channel

Functions `lumiMethyB`, `lumiMethyC` and `lumiMethyN` also has the option to separately process in each color channel. The presumption is that the data (*MethyLumiM* object) should include the color channel column (with column name "COLOR_CHANNEL") in the featureData and the green and red colors are represented as "Red" and "Grn". If the GenomeStudio (BeadStudio) output the annotation information together with the data (we recommend it), the `lumiMethyR` function will automatically add the annotation information, including color channel information, to the featureData slot. If the data does not include color channel information, users can add it use function `addAnnotationInfo`.

```
> ## retrieve the featureData information
> ff <- pData(featureData(example.lumiMethy))
> ## show the color channel information
> head(ff)

           CHR COLOR_CHANNEL
cg00002426   3           Red
cg00012386   1           Red
cg00013618  22           Grn
cg00014837  12           Red
cg00020533   6           Red
cg00021527  17           Red

>
> ## add user provided color channel information if it is not existed in the featureDa
> # example.lumiMethy <- addAnnotationInfo(example.lumiMethy, lib="IlluminaHumanMethyl
```

Here are examples of separately processing each color channel:

```
> ## suppose "userB" is a user defined background adjustment method
>  lumiMethy.b.adj <- lumiMethyB(example.lumiMethy,  separateColor=TRUE)
> ## suppose "userC" is a user defined color balance adjustment method
>  lumiMethy.c.adj <- lumiMethyC(example.lumiMethy,  separateColor=TRUE)
> ## suppose "userN" is a user defined probe level normalization method
>  lumiMethy.c.n <- lumiMethyN(lumiMethy.c.adj,  separateColor=TRUE)
```

4.10 About the detection p-value of a CpG site

The detection p-values of methylation arrays reflect the strength of DNA hybridization over the background (comparing the CpG-intensity with the intensities of negative control probes). Because all genomic DNA is expect existing in a normal diploid sample, almost all detection p-values are significant. Non-significant detection p-values usually means bad probe design, bad hybridization or possible chromosome abnormalities (like mutations, indels) at the probe matching locations. Given one sample in the titration data set as an example, the 99 percentile of the -log10(detection p-value) is 28.3, which is extraordinarily significant p-value. And there are only 27 probes with detection p-value worse than 0.0001. This is in great contrast to expression microarrays, where up to 40 percent or more of the genes might not be expressed in a given sample, and filtering based on detection p-values can dramatically reduce the false-positive rate for the expression data. However, removing the bad quality CpG measurements is still an important step during preprocessing of Illumina methylation data. Same as Illumina Expression data, we can use function `detectionCall` to estimate the detection call. Please check the help of `detectionCall` for more details of its usage.

```
> ## Estimate the detection call of a CpG site
> presentCount <- detectionCall(example.lumiMethy)
```

5 Gene annotation

The annotation packages, IlluminaHumanMethylation27k.db and IlluminaHumanMethylation450k.db can be downloaded from Bioconductor for HumanMethylation27k and HumanMethylation450K BeadChips, respectively. The usage of this package is the same as the expression microarrays.

Recently, another `GRanges` based package, FDb.InfiniumMethylation.hg19, is available for hg19 human genome. There are `get27k` and `get450k` functions to retrieve the probe information of 27k and 450k BeadChips. For more details, please check FDb.InfiniumMethylation.hg19 package.

6 A use case: from raw data to functional analysis

6.1 Preprocessing the Illumina Infinium Methylation microarray data

```
> library(lumi)
> ## specify the file name
> # fileName <- 'Example_Illumina_Methylation_profile.txt'
> ## load the data
> # example.lumiMethy <- lumiMethyR(fileName, lib="IlluminaHumanMethylation27k.db")
>
```

```
> ## Quality and color balance assessment
> data(example.lumiMethy)
> ## summary of the example data
> example.lumiMethy
> ## preprocessing and quality control after normalization
> plotColorBias1D(example.lumiMethy, channel='sum')
> boxplotColorBias(example.lumiMethy, channel='sum')
> ## select interested sample to further check color balance in 2D scatter plot
> plotColorBias2D(example.lumiMethy, selSample=1)
> ## Color balance adjustment between two color channels
> lumiMethy.c.adj <- lumiMethyC(example.lumiMethy)
> ## Check color balance after color balance adjustment
> boxplotColorBias(lumiMethy.c.adj, channel='sum')
> ## Background adjustment is skipped because the SSN normalization includes backgroun
>
> ## Normalization
> ## Perform SSN normalization based on color balance adjusted data
> lumiMethy.c.ssn <- lumiMethyN(lumiMethy.c.adj, method='ssn')
> ## Or we can perform quantile normalization based on color balance adjusted data
> # lumiMethy.c.q <- lumiMethyN(lumiMethy.c.adj, method='quantile')
>
> ## plot the density of M-values after SSN normalization
> density(lumiMethy.c.ssn, main="Density plot of M-value after SSN normalization")
> ## comparing with the density of M-values before normalization
> density(example.lumiMethy, main="Density plot of M-value of the raw data")
> ## output the normlized M-value as a Tab-separated text file
> write.exprs(lumiMethy.c.ssn, file='processedMethylationExampleData.txt')
```

6.2 Identify differentially methylated genes and other further analysis

We developed a separate package methyAnalysis for further methylation analysis independent of platforms. Please check methyAnalysis package for more details.

6.3 GEO submission of the methylation data

The submission of methylation data is similar with expression microarray data. The submission file will be in the SOFT format. So users can submit all the data in a batch. We also use `produceGEOSampleInfoTemplate` to produce a template of sample information with some default fillings. Some fields have been filed in with default settings. Users should fill in or modify the detailed sample descriptions by referring some previous submissions. No blank fields are allowed. Users are also required to fill in the "Sample_platform_id" by checking

information of the GEO Illumina platform. `produceMethylationGEOSubmissionFile` is the main function of produce GEO submission file including both normalized and raw methylation data information in the SOFT format. By default, the R objects, methyLumiM, methyLumiM.raw and sampleInfo, will be saved in a file named 'supplementaryData.Rdata'. Users can include this R data file as a GEO supplementary data file.

```
## Produce the sample information template
> produceGEOSampleInfoTemplate(methyLumiM, fileName = "GEOsampleInfo.txt")
## After editing the 'GEOsampleInfo.txt' by filling in sample information
> produceMethylationGEOSubmissionFile(methyLumiM, methyLumiM.raw, sampleInfo='GEOsampl
```

7 Session Info

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

- R version 3.2.3 (2015-12-10), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.32.3, Biobase 2.30.0, BiocGenerics 0.16.1, DBI 0.3.1, IRanges 2.4.6, RColorBrewer 1.1-2, RSQLite 1.0.0, S4Vectors 0.8.8, XML 3.98-1.3, annotate 1.48.0, genefilter 1.52.0, limma 3.26.5, lumi 2.22.1, lumiBarnes 1.10.0, lumiHumanAll.db 1.22.0, lumiHumanIDMapping 1.10.0, org.Hs.eg.db 3.2.3, vsn 3.38.0
- Loaded via a namespace (and not attached): BiocInstaller 1.20.1, BiocParallel 1.4.3, Biostrings 2.38.3, GEOquery 2.36.0, GenomeInfoDb 1.6.1, GenomicAlignments 1.6.3, GenomicFeatures 1.22.8, GenomicRanges 1.22.3, KernSmooth 2.23-15, MASS 7.3-45, Matrix 1.2-3, RCurl 1.95-4.7, Rcpp 0.12.3, Rsamtools 1.22.0, SummarizedExperiment 1.0.2, XVector 0.10.0, affy 1.48.0, affyio 1.40.0, base64 1.1, beanplot 1.2, biomaRt 2.26.1, bitops 1.0-6, bumphunter 1.10.0, codetools 0.2-14, colorspace 1.2-6, corpcor 1.6.8, digest 0.6.9, doRNG 1.6, ellipse 0.3-8, foreach 1.4.3, futile.logger 1.4.1, futile.options 1.0.0, ggplot2 2.0.0, grid 3.2.3, gtable 0.1.2, hexbin 1.27.1, igraph 1.0.1, illuminaio 0.12.0, iterators 1.0.8, labeling 0.3, lambda.r 1.1.7, lattice 0.20-33, locfit 1.5-9.1, magrittr 1.5, matrixStats 0.50.1, mclust 5.1, methylumi 2.16.0, mgcv 1.8-10, minfi 1.16.0, mixOmics 5.2.0, multtest 2.26.0, munsell 0.4.2, nleqslv 2.9.1, nlme 3.1-122, nor1mix 1.2-1, pkgmaker 0.22, plyr 1.8.3, preprocessCore 1.32.0, quadprog 1.5-5, registry 0.3, reshape 0.8.5, rgl 0.95.1441, rngtools 1.2.4, rtracklayer 1.30.1, scales 0.3.0,

siggenes 1.44.0, splines 3.2.3, stringi 1.0-1, stringr 1.0.0, survival 2.38-3, tools 3.2.3, xtable 1.8-0, zlibbioc 1.16.0

8 Acknowledgement

We appreciate Dr. Sean Davis from NIH developed methylumi package and built IlluminaHumanMethylation27k.db package for us!

This work was supported in part by the NIH award 1RC1ES018461-01 (to Dr. Lifang Hou), P30CA060553 and UL1RR025741. We would like to thank Lifang Hou for supporting this work, Xiao Zhang for preparing the titration samples, Nadereh Jafari and Vivi Frangidakis for conducting the Illumina BeadChip experiments.

9 References

1. Du P, Kibbe WA and Lin SM: "lumi: a Bioconductor package for processing Illumina microarray" Bioinformatics 2008 24(13):1547-1548

2. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, and Lin SM: "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis", BMC Bioinformatics 2010, 11:587