# *MethVisual*- R package for visualization and exploratory statistical analysis of DNA methylation profiles

Arie Zackay and Christine Steinhoff

October 18, 2010

## 1 Background

DNA Methylation is a biochemical modification of DNA which in vertebrates almost exclusively occurs at CpG sites, e.g. a methyl group is added at the 5âĂŹ C position of cytosines. Exploration of DNA methylation and its impact on various regulatory cellular processes has become a very active field of research and comprises cancer, silencing of repetitive elements, development, chromatin remodeling, RNA interference, imprinting, tissue specificity, and evolutionary mutation processes To date the most accurate experimental procedures are based on bisulfite treatment followed by conversion of non methylated cytosines to uracil and sequencing. Analyzing this kind of data is complicated. Several steps, like alignment of bisulfite treated sequence to reference sequence, detection of low conversion rates of C to T in the bisulfite treatment and conversion process and quality control, are necessary before actually extracting methylation profiles for further statistical analysis. However, this procedure is a prerequisite for the investigation of functionality of DNA methylation. MethVisual allows for processing this kind of data as well as several basic exploratory analysis and visualization steps.

*MethVisual* enables intuitive visualization and exploratory analysis of binary DNA methylation data. The package allows the user to import binary methylation sequences (clone sequences) as generated from bisulfite sequencing, aligns them, perform quality control process and execute statistical analysis (Table 1). The package was developed for methylation data but can also be applied on other binary coded data types in a straightforward manner.

This document provides an example for the analysis workflow. It includes the required steps for the analysis of methylation data, using the example data saved in this package. This example data is taken from the program *BiQ-Analyzer*[1].

The analysis steps are:

1. **Reading sequences** sample sequences and reference sequence
2. **Alignment control and quality control**
3. **Computation of methylation status**

[1] http://biq-analyzer.bioinf.mpi-sb.mpg.de/

4. **Exploratory statistics and visualization** including lollipop plot, neighboring cooccurrence- and distant cooccurrence analysis

5. **Further statistical investigations** including hierarchical clustering and correspondence analysis

| **methVisual Functions** | **Description** |
|---|---|
| `Cooccurrence` | Cooccurrence of methylation data |
| `MethAlignNW` | Summary of methylation states |
| `MethDataInput` | Sequence match control |
| `MethLollipops` | Lollipop plot |
| `MethylQC` | Alignment and quality control |
| `cgInAlign` | Amount of CpG sites |
| `cgMethFinder` | Methylation state |
| `coversionGenom` | Sequence conversion |
| `findNonAligned` | Aligned CpG positions |
| `heatMapMeth` | HeatMap over methylation data |
| `makeDataMethGFF` | Processing GFF methylation files |
| `makeLocalExpDir` | Saving example data |
| `makeTabFilePath` | Tab delimited text file |
| `matrixSNP` | Correlation between methylation states |
| `methCA` | Correspondence analysis (CA) methylation states |
| `methData` | BiQ-Analyzer dataset |
| `methFisherTest` | Fisher's exact Test on methylation Data |
| `methWhitneyUTest` | Mann Whitney U-Test on methylation data |
| `plotAbsMethyl` | Plot of the absolute number of methylation |
| `plotMatrixSNP` | Distant cooccurrence of methylation data |
| `readBisulfFASTA` | Read multiple FASTA file |
| `selectRefSeq` | Upload genomic sequence |

Table 1: The functions are available in *methVisual*

In order to start please download the newest version of *methVisual* and load it into R.

```
> library(methVisual)
```

# 2 Reading Sequences

This section demonstrate the analysis of *FASTA* files, that includes the clone sequences and the reference sequence.

First, the example data has to be saved in a directory. Please make sure that you have reading and writing permission under *R.home()* directory. If you do not have permission choose your own path.

Creating BiQ-Analyzer directory in your *R.home()*:

```
> dir.create(file.path(R.home(component = "home"),
+     "/BiqAnalyzer/"))
```

Saving *methVisual* example data in /BiqAnalyzer directory:

```
> makeLocalExpDir(dataPath = "/examples/BiqAnalyzer",
+     localDir = file.path(R.home(component = "home"),
+         "/BiqAnalyzer/"))
```

Upload the list of clone sequences as tab delimited text file:

```
> methData <- MethDataInput(file.path(R.home(component = "home"),
+     "/BiqAnalyzer", "/PathFileTab.txt"))
```

Hereby the sample sequence list is saved in a data frame object `methData` with *PATH* and *FILE* column

```
> methData
```

```
          FILE                                        PATH
1  seq_A.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
2  seq_B.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
3  seq_C.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
4  seq_D.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
5  seq_E.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
6  seq_F.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
7  seq_G.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
8  seq_H.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
9  seq_I.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
10 seq_J.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
```

Read reference sequence into R

```
> refseq <- selectRefSeq(file.path(R.home(component = "home"),
+     "/BiqAnalyzer", "/Master_Sequence.txt"))
```

# 3 Alignment Control and Quality Control

The alignment control (AC) procedure comprises a comparison of the sample sequences to the reference sequence and is performed to prevent false alignment in the further analysis. False alignment can occur because of three reasons, sequences that are reversed, complement or reversed-complement to the genomic sequence. The AC procedure compare the score result computed by pairwise Needleman Wunsch Algorithm for global alignment implemented in the *Biostrings*[2] R package and select the alignment variant with the highest score among

[2] http://www.bioconductor.org/packages/2.2/bioc/html/Biostrings.html

these three possibilities for each clone sequence involved. If changes are necessary, they will be either performed automatically or left to the user to include them. The alignment controlled sequences will be saved. Potential errors during the experimental process of bisulfite sequencing concern mainly bisulfite conversion and sequence identity. However, bisulfite conversion might be incomplete, that means even though non methylated cytosines (Cs) should be converted to uracils (Us) upon bisulfite treatment and subsequent amplification there might be non methylated Cs that have not been converted. In vertebrates we can assume that methylation is restricted to CpG sites. Thus, non converted Cs outside of CpG sites can be regarded as non conversion failure. MethVisual further measures the bisulfite treatment quality by calculating this conversion ratio among Cs in non CpG sites, which is defined as the ratio between the number of unconverted Cs and the sum of all Cs outside CpG sites. *MethVisual* also determines the sequence identity between each sample sequence and genomic sequence by calculating the sequencing error rate and restricting the comparison of sequenced sample versus reference sequence to As, Gs and Ts. The user can define a threshold percentage for rejecting sequences.

All control procedures are implemented under the function *MethylQC()*.

```
> QCdata <- MethylQC(refseq, methData, makeChange = TRUE,
+     identity = 80, conversion = 90)
```

The sample sequence list after AC and QC procedure is saved as a data frame object with *PATH* and *FILE* column.

```
> QCdata
```

```
           FILE                                        PATH
1 QC_seq_B.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
2 QC_seq_C.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
3 QC_seq_E.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
4 QC_seq_G.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
5 QC_seq_H.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
6 QC_seq_I.fasta E:\\biocbld\\BBS-2~1.7-B\\R//BiqAnalyzer/
```

# 4 Calculation of methylation status

Now the user needs to extract the information on methylation states from the quality checked sample sequences. Given the aligned sequences, the function `MethAlignNW()` returns a list object with the following data: Sequence name, methylation state of CpG sites over all clone sequences, start and end position of alignments and the length of reference sequence.

```
> methData <- MethAlignNW(refseq, QCdata)
```

```
Alignment with  QC_seq_B.fasta  done
Alignment with  QC_seq_C.fasta  done
Alignment with  QC_seq_E.fasta  done
```

```
Alignment with  QC_seq_G.fasta  done
Alignment with  QC_seq_H.fasta  done
Alignment with  QC_seq_I.fasta  done
```

```
> methData
```

```
$seqName
[1] "QC_seq_B.fasta" "QC_seq_C.fasta" "QC_seq_E.fasta"
[4] "QC_seq_G.fasta" "QC_seq_H.fasta" "QC_seq_I.fasta"

$alignment
[1] "TTTGGGATTGTTTTTTTAGTAGGTGAAGTTTTGTTATGGATTTTTTTTGTTGGGGTTTTGTGTTGTTTTGTTTGTTTTTAGTTGTTGGTT
[2] "TTCGGGATCGTTTTTTTAGTAGGTGAAGTTTTGTTATGGATTTTTTTCGTTGGGGTTTCGTGTTGTTTTGTTCGTTTTTAGTCGTTGGTT
[3] "TTTGGGATTGTTTTTTTTAGCAGGTGAAGTTTTGTTATGGATTTTTTTTGTTGGGGTTTTGATGTTGTTTCGTTTGTTTCTAGTTGTTGG
[4] "TTTGGGATTGTTTTTTTAGTAGGTGAAGTTTTGTTATGGATTTTTTTTGTTGGGGTTTTGTGTTGTTTTGTTTGTTTTTAGTTGTTGGTT
[5] "TTCGGGATTGTTTTTTTAGTAGGTGAAGTTTTGTTATGGATTTTTTTTGTTGGGGTTTCGTGTTGTTTTGTTCGTTTTTTGTTGTTGGTT
[6] "TTTGGGATTGTTTTTTTAGTAGGTGAAGTTTTGTTATGGATTTTTTTTGTTGGGGTTTTGTGTTGTTTTGTTTGTTTTTAGTTGTTGGTT

$methPos
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    0    0    0    0    0    0     0
[2,]    1    1    0    1    0    1    0    0    1     1
[3,]    0    0    0    0    0    0    0    0    0     0
[4,]    0    0    0    0    0    0    0    0    0     0
[5,]    1    0    0    0    0    1    0    0    1     0
[6,]    0    0    0    0    0    0    0    0    0     0
     [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19]
[1,]     0     0     0     0     0     0     0     0     1
[2,]     0     1     1     1     1     1     1     1     1
[3,]     1     0     0     0     1     1     1     1     1
[4,]     0     0     0     0     0     0     0     0     1
[5,]     1     1     1     1     1     1     1     1     1
[6,]     0     0     0     0     0     0     0     0     0
     [,20] [,21] [,22] [,23] [,24] [,25] [,26] [,27] [,28]
[1,]     0     0     0     0     0     0     0     0     0
[2,]     1     1     1     0     0     0     0     1     1
[3,]     0     0     1     1     1     0     0     0     1
[4,]     0     0     0     0     0     0     0     0     0
[5,]     1     0     1     1     1     1     0     1     1
[6,]     0     0     0     0     0     0     0     0     0
     [,29] [,30] [,31] [,32] [,33] [,34] [,35]
[1,]     0     0     0     0     0     0     0
[2,]     0     1     1     0     0     1     0
[3,]     0     1     0     0     0     0     0
[4,]     0     0     0     0     0     0     0
[5,]     1     0     0     0     1     1     1
[6,]     0     0     0     0     0     0     0

$positionCGIRef
```

```
 [1]   3   9  32  48  51  59  61  69  73  83  96  98 102 104
[15] 110 118 120 125 128 131 136 145 148 153 160 163 165 167
[29] 175 177 198 200 209 217 232

$startEnd
     [,1] [,2]
[1,]    1  233
[2,]    1  233
[3,]    1  233
[4,]    1  233
[5,]    1  233
[6,]    1  233

$lengthRef
[1] 233
```

Now one can proceed with the visualization and statistics of the data!

# 5 Exploratory statistics and visualization

## 5.1 Amount of methylation

Plotting the absolute or relative number of methylation of all CpG positions of all sample sequences provides a global overview of the data set in terms of methylation amounts by genomic position (Figure 1).

```
> plotAbsMethyl(methData, real = TRUE)
```

## 5.2 Lollipop figures

A graphical representation of methylation state can be produced applying *Lollipop* graphs (Figure 2). It allows the user to study the states of CpG sites in sample sequences. Each circle marks a CpG site under study. Full circles display methylated CpG sites and the non filled ones stand for non-methylated CpG states. The examined sequences are aligned with respect to the CpG sites in reference sequence in order to allow an intuitive visualization of methylation states according to their genomic order.

```
> MethLollipops(methData)
```

## 5.3 Neighboring cooccurrence display

The study of cooccurrence of methylated or non methylated CpG sites is frequently investigated. Given a set of bisulfite sequenced samples one would like to detect subgroups where

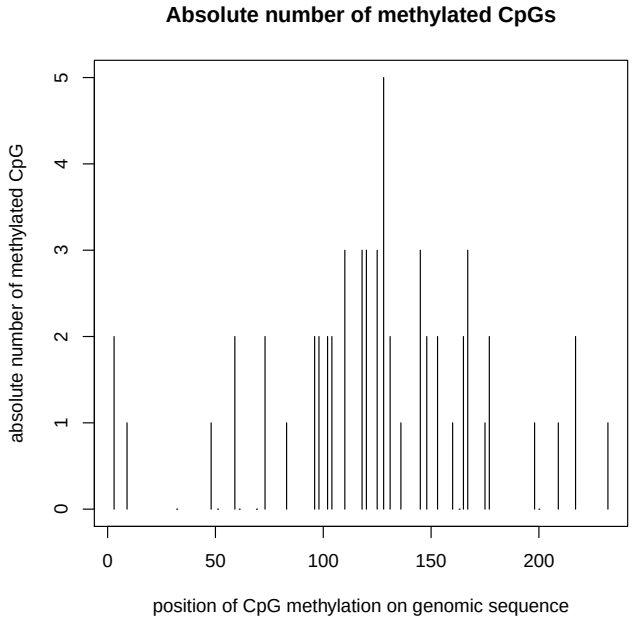


Figure 1: The number of methylated CpG over all 6 analyzed sequences. The x-axis is the genomic CpG position

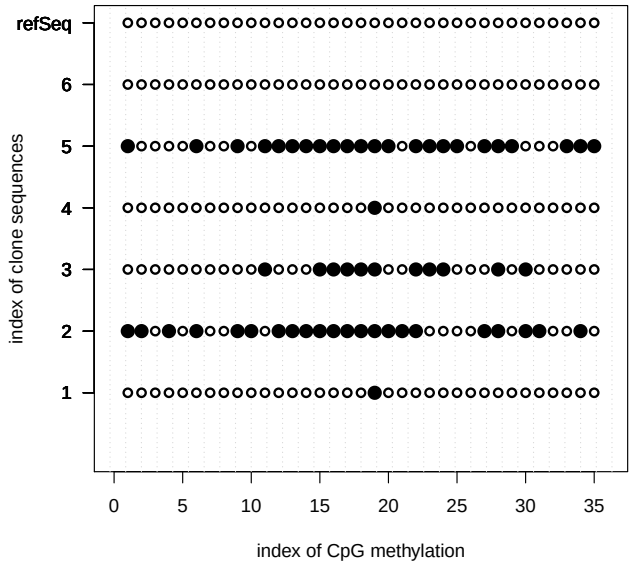


Figure 2: Lollipop display of binary methylation profiles in the genomic context

specific CpG sites always occur coordinately either methylated or non methylated. One drawback of the widely used lollipop representation is that cooccurrence displays are not integrated. We implemented an option to visualize neighbored cooccurrence of methylation patterns. This display is restricted to neighboring cooccurrence of CpG methylation.

```
> file <- file.path(R.home(component = "home"),
+     "/BiqAnalyzer/", "Cooccurrence.pdf")
> Cooccurrence(methData, file = file)
```

## 5.4 Distant cooccurrence display

However, one might also want to explore cooccurrence features in a distant manner, i.e. not directly neighbored CpG sites. Thus, we provide a comprehensive visualization of all pairwise coccurrences of methylation (Figure 3). The correlation structure can be saved and statistically further explored.

```
> summery <- matrixSNP(methData)
> plotMatrixSNP(summery, methData)
```

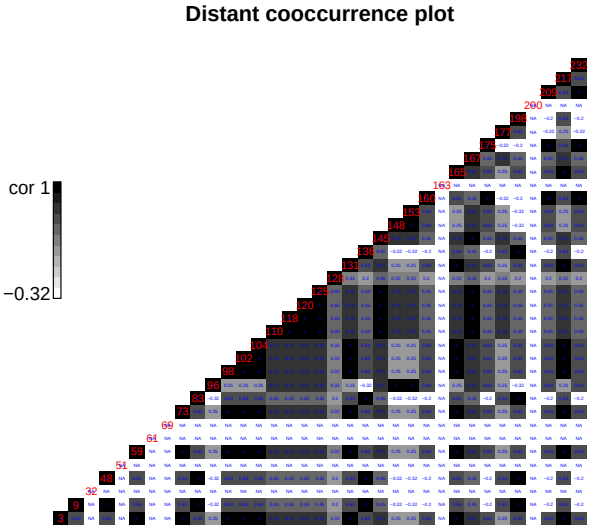


Figure 3: Distant cooccurrence plot. Each pairwise comparison, e.g. neighboring and distant, leads to a correlation value that is displayed in the matrix. Correlation is color coded and the color coding bar is given beside the graph. The numbers in the diagonal give the genomic position of each CpG site.

# 6 Further statistical investigations

## 6.1 Statistical tests

Basic statistical tests options comprise *(i)* testing for independence of each CpG site between two groups (Fisher's exact test) or *(ii)* of entire sets of CpG sites (Mann-Whitney U test). More specifically, for *(i)* given two experimental groups for each CpG site the user can investigate whether there is a dependence of methylation status and class membership of the two groups (Figure 4). In the case of *(ii)* given two experimental groups the hypothesis that the distribution of methylated and non-methylated sites in the profile under study is being tested.

```
> methFisherTest(methData, c(2, 3, 5), c(1, 4, 6))
```

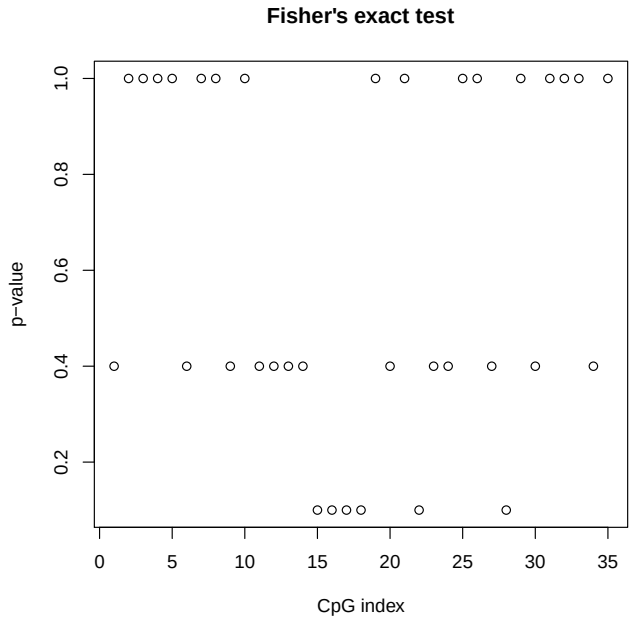


Figure 4: P-values of methylated CpG position

Looking at the Lollipop plot (Figure 2) one can see that the clone sequences *(2,3,5)* seems to have different CpG pattern than clones *(1,4,6)*. The calculated *p-value* using Whithney-U-Test confirm the significant pattern difference.

```
> methWhitneyUTest(methData, c(2, 3, 5), c(1, 4,
+     6))
```

```
[1] 0
```

## 6.2 Clustering

Clustering is a methods for exploring and visualizing groups with similar features. We provide a hierarchical bi-clustering of methylation states. Due to the fact that we analyze binary rather than continuous data the default option for distance is the binary rather than euclidean distance (Figure 5 5).

```
> heatMapMeth(methData)
```

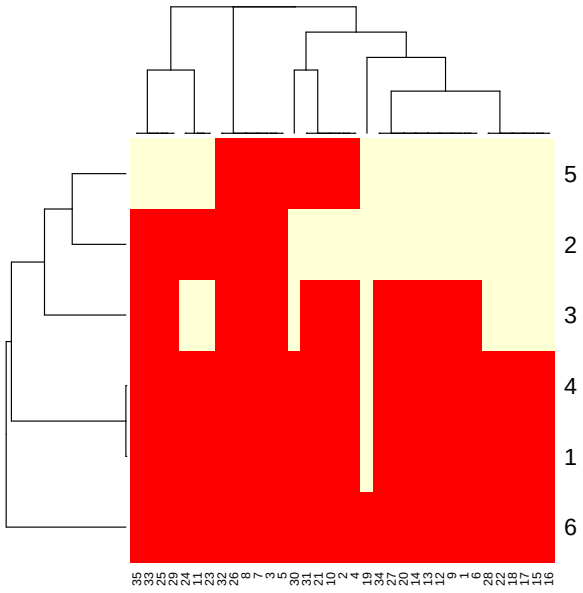


Figure 5: Bi-clustering due to methylated CpG positions of sample sequences

## 6.3 simple correspondence analysis

Using simple correspondence analysis (CA) one can detect clusters of sub-samples that show similar cooccurrence patterns. Based on aligned sequences under study a CA plot displays two way clustering of methylation status of all sequences and all aligned CpG positions (Figure 6).

```
> methCA(methData)
```

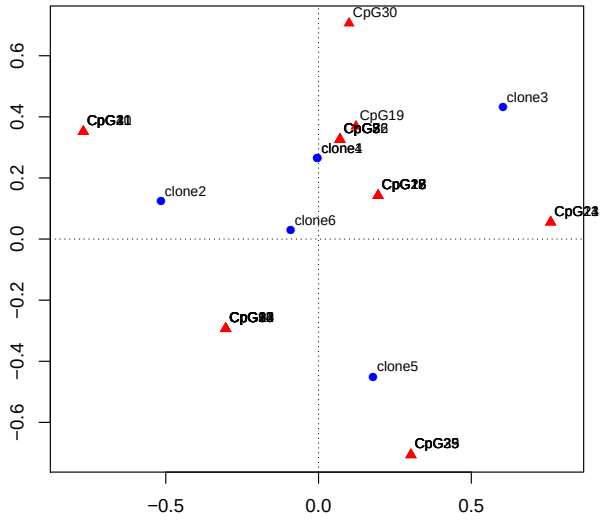


Figure 6: Simple correspondence analysis of methylated CpG positions and samples sequences