Methylation Analysis of Next Generation Sequencing

Muhammad Ahmer Jamil [*,†]
Prof. Holger Frohlich[*]
Priv.-Doz. Dr. Osman El-Maarri[†]

[*]Bonn-Aachen International Center for Information Technology
[†]Institute of Experimental Haematology and Transfusion Medicine, University of Bonn

Contents

1 Methylation Analysis

1.1 Sequence Alignment

Sequences can be aligned to the reference sequence using Local, Global, Global-local and overlap algorithm. Default method used is Smith-Waterman Alogirthm (Local-Alignment). Sequence files were loaded first in the R before alignment. Following files are available in package.

- Healthy.fasta (Healthy patient sample)
- Tumor.fasta (Tumor patient sample)
- Reference.fasta (reference sequence from ncbi of the given patient sample)

```
> library(MethTargetedNGS)
> healthy = system.file("extdata", "Healthy.fasta", package = "MethTargetedNGS")
> tumor = system.file("extdata", "Tumor.fasta", package = "MethTargetedNGS")
> reference = system.file("extdata", "Reference.fasta", package = "MethTargetedNGS")
```

After loading the sequences in R, we can perform sequence alignment using function methAlign.

```
> hAlign = methAlign(healthy,reference)

Time difference of 3.18 secs

> tAlign = methAlign(tumor,reference)

Time difference of 3.85 secs
```

Time difference is the time used for the processing of the methAlign function. This time depends on the number of sequences in the fasta file. This function gives maxtrix contains methylation profile of the given fasta file. Rows correspond to the sequences present in the sample and column represent numbers of CpG sites in the reference sequence.

To view the methylation pattern in the samples, we can use methHeatmap for generating heatmap for the given methylation data.

```
Percentage Result is 99.71
```

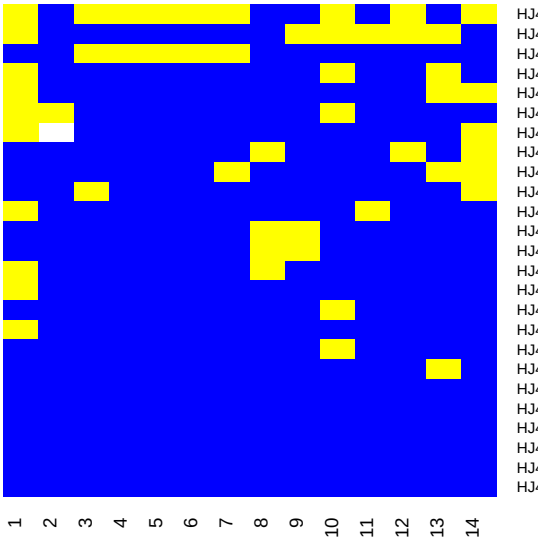

Figure 1: Heatmap

```
> hHeatmap = methHeatmap(hAlign,plot=TRUE)

Percentage Result is 99.71
```

1.2 Methylation Average

Methylation average of a CpG site is the percentage of unmethylated cytosine or methylated cytosine in a particular CpG site. Determining methylation average is the most traditional way of analyzing the DNA methylation level.[1] Methylation average explains the hypomethylation and hypermethylation in the sample. Hypomethylation or hypermethylation caused in tumor cells can have change in the average methylation at particular CpG site. This measure of variability in combination of healthy and tumor can be used to explain the methylation level difference between healthy and tumor samples. The methylation average of a particular CpG site was calculated by number of cytosine divided by sum of total number of methylated and unmethylated cytosine at particular CpG site in a group of reads. To calculate methylation average it was assumed that all

sequences align perfectly with no mismatch or insertion/deletion, although it is an ideal case.

$$average = N_C/(N_C + N_T) \quad (1)$$

In MethTargetedNGS, methylation average can be calculated by passing the matrix obtained from methAlign to methAvg. It will return a vector of methylation averages in percentages.A plot of methylation average will be generated if plot=TRUE.

```
> hAvg = methAvg(hAlign,plot=FALSE)
> hAvg
```

```
 [1] 60.00 95.83 88.00 92.00 92.00 92.00 88.00 84.00 88.00 76.00 92.00 88.00
[13] 80.00 76.00
```

1.3 Methylation Entropy

Determination of average methylation is the traditional way of analyzing DNA methylation data. Such way is unable to dissect DNA methylation pattern. DNA methylation pattern is defined as the combination of methylation statuses of contiguous CpG dinucleotides in a DNA strand. Methylation entropy can be caluclated by using the formula from Xie et. al.[2] which was "implies" to the function methAlign

```
> hEnt <- methEntropy(hAlign)
```

```
Time difference of 0.05 secs
```

```
> hEnt
```

```
 [1] 0.33 0.30 0.37 0.37 0.30 0.27 0.30 0.39 0.31 0.30 0.39
```

This function return vector of methylation entropy values using sliding window of 4.

1.4 Identify significant CpG sites

Fisher exact test is a test to calculate the statistical significance using contingency table. It was used to find the statistically significant differences in the methylation status of one particular CpG site between healthy and tumor sample. Contingency matrix was created for each CpG site in the following pattern. Fisher exact test was performed using a predefined function "fisher.exact" in R. P-value was corrected for multiple testing using Benjamini-Hochberg method to calculate False Discovery Rate (FDR).[3] To calculate the statistically significant CpG sites, pass the healthy and tumor aligned matrix from methAlign to fishertest_cpg.

```
> SigCpGsites = fishertest_cpg(hAlign,tAlign,plot=FALSE)
> SigCpGsites
```

```
 [1] 0.59 2.31 1.72 2.02 2.02 2.31 2.02 1.66 2.02 1.34 2.16 2.02 1.63 1.61
```

1.4.1 Log odd ratio

Log odd ratio defines the hypomethylation and hypermethylation of a sample comparison to other sample. To calculate odd ratio we can pass healthy and tumor aligned matrix from methAlign to odd_ratio

```
> odSamps = odd_ratio(hAlign,tAlign,plot=FALSE)
> odSamps
```

```
 [1] 0.81 3.05 1.91 2.36 2.36 2.68 2.07 1.74 1.99 1.39 2.52 2.07 1.63 1.56
```

```
Time difference of 0.05 secs
Time difference of 0.06 secs
```

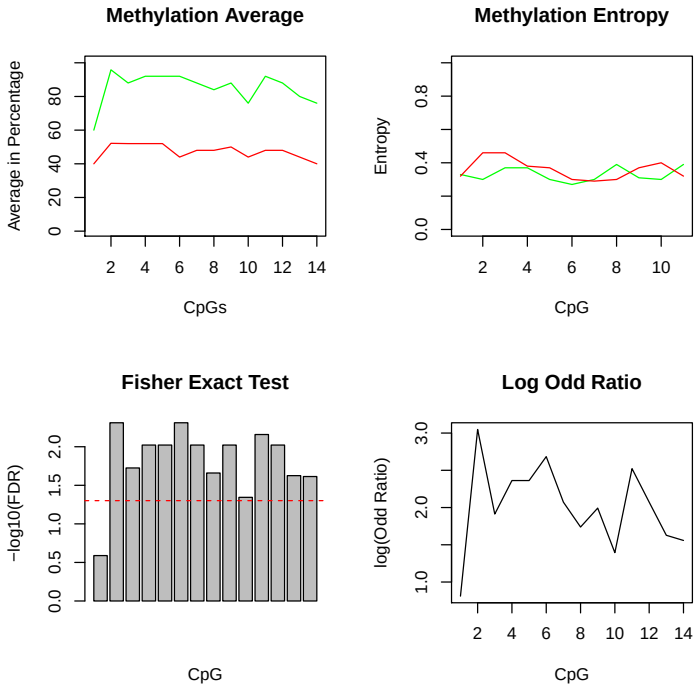

Figure 2: Methylation Analysis between Healthy and Tumor

1.5 Complete Pipeline

Methylation average, methylation entropy and identifying statistically signficant CpG sites can be calculated by using compare_samples function by passing the aligned matrices of healthy and tumor samples from methAlign function. Complete methylation analysis shown in figure 2.

```
> compare_samples(hAlign,tAlign)

Time difference of 0.05 secs
Time difference of 0.06 secs
```

2 Profile Hidden Markov Model

Profile Hidden Markov Model is a method of transforming multiple sequence alignment into a position specific scoring system. It gives the probabilistic model for multiple sequence alignment. Profile HMM allows to compute a likelihood, which can define the similarities between a new sequence and the model.

To create a Profile HMM hmmer software is required;

```
> msa = system.file("extdata", "msa.fasta", package = "MethTargetedNGS")
> if (file.exists("/usr/bin/hmmbuild"))
+ hmmbuild(file_seq=msa,file_out="hmm",pathHMMER = "/usr/bin/")
```

To calculate the likelihood of pool of sequences with HMM model;

```
> if (file.exists("/usr/bin/nhmmer")){
+ res <- nhmmer("hmm",tumor,pathHMMER = "/usr/bin/")
+ res$Total.Likelihood.Score
+ }
```

References

[1] Singer, H., Walier, M., Nusgen, N., Meesters, C., Schreiner, F., Woelfle, J., Fimmers, R.,Wienker, T., Kalscheuer, V.M., Becker, T. et al. (2012) *Methylation of L1Hs promoters is lower on the inactive X, has a tendency of being higher on autosomes in smaller genomes and shows inter-individual variability at some loci.*. Hum Mol Genet, 21, 219-235

[2] Xie, H., Wang, M., de Andrade, A., Bonaldo, M.d.F., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T. and Soares, M.B. (2011) *Genome-wide quantitative assessment of variation in DNA methylation patterns.* Nucleic Acids Research, 39, 4099-4108.

[3] Benjamini, Yoav; Hochberg, Yosef (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, Series B 57 (1): 289 - 300.MR 1325392.