Supplements to MethRaFo: MeDIP-Seq methylation estimate using Random Forest Regressor

Jun Ding[1] and Ziv Bar-Joseph *[1]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Contents

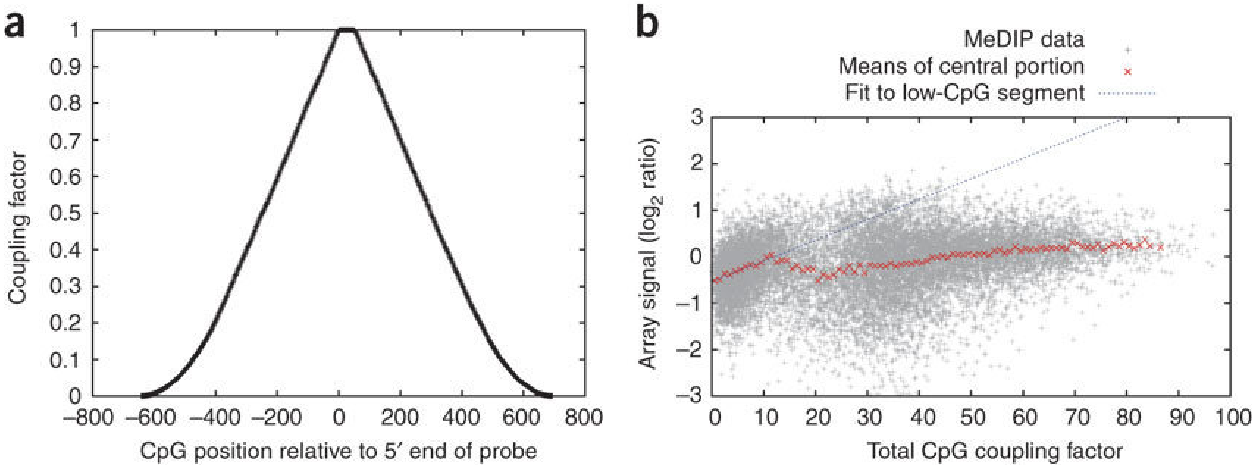
*To whom correspondence should be addressed. Ziv Bar-Joseph, Email:zivbj@cs.cmu.edu

Supporting Methods

Why use MethRaFo?

Profiling genome wide methylation is now used widely when studying development, cancer and several other biological processes. Although whole genome Bisulfite Sequencing provides high-quality methylation measurement at the resolution of nucleoteides, it's relatively costly ($4.5k-$5k). One of the most widely used low cost alternatives is MeDIP-Seq ($1.5k-$1.7k). However, MeDIP-Seq is biased for CpG enriched regions and thus its result need to be corrected to determine accurate methylation level.

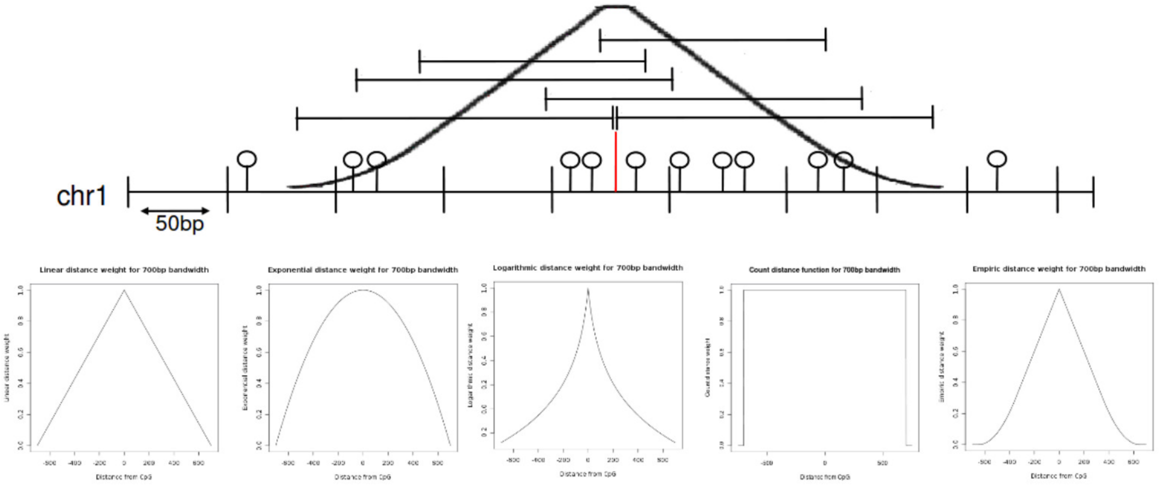
A number of MeDIP-Seq methylation correction methods have been developed in the past few years. BATMAN([Down *et al.*,2008]) utilized a linear function to normalize the methylation based on local CpG density (Supplementary Figure 1) (weight functions for nearby CGs). To infer such linear relationship, BATMAN used a bayesian based model.



Supplementary Figure 1: Linear relationship used by BATMAN([Down *et al.*,2008])

MEDIPS([Chavez *et al.*,2010])is following a similar strategy as BATMAN by considering several different relationships: Linear, Exponential, Logarithmic,count distance function and empiric distance weight (Supplementary Figure 2). BayMeth([Riebler *et al.*,2014]) is also using a linear relationship similar as BATMAN.

Although those relationships (weight functions for nearby CGs) can somehow represent the impact of local CpG density on the methylation at current position, such impact in practice is too complicated to be represented simply using simple linear/exp functions. With the advance of technology, we have more and more useful methylation experiment data available such as Roadmap epigenome database ([Bernstein *et al.*,2010]). In roadmap epigenome database, there are dozens of datasets with both MeDIP-Seq and Bisulfite-Seq ('ground-truth' methylation) datasets available, which provide us a great opportunity to study the impact of local CpG density on methylation. Instead of using simple relationships, here we used a random forest regressor model to learn such hidden knowledge. Such model is more close to the reality if compared with simple functions such as Linear or exponential. This can be illustrated by the following example. we all know that MeDIP-Seq is biased for dense


chr1
50bp
Linear distance weight for 700bp bandwidth
Linear distance weight
Distance from CpG
Exponential distance weight for 700bp bandwidth
Exponential distance weight
Distance from CpG
Logarithmic distance weight for 700bp bandwidth
Logarithmic distance weight
Distance from CpG
Count distance function for 700bp bandwidth
Count distance weight
Distance from CpG
Empiric distance weight for 700bp bandwidth
Empiric distance weight
Distance from CpG

Supplementary Figure 2: relationships used by MEDIPS([Chavez *et al.*,2010])

CG regions. If this is true, we can definitely observe this in the training dataset. In other words, in most case, if the CpG density is high in a region, the MeDIP-Seq methylation for this region will be larger than the true value (bisulfite-seq methylation). We can use the regression model (random forest regressor) to learn such knowledge. For example, suppose we observe the following in the training dataset:

- local CpG density >20 => MeDIP-Seq methylation 20% larger
- local CpG density > 40 => MeDIP-Seq methylation 40% larger
- local CpG desnity < 5 => MeDIP-Seq methylation 0.5% larger
- ...

This knowledge can be easily learned by the random forest regressor model. For example, we have 2 features, so it's in 2D space (X,Y). X: MeDIP-Seq methylation Y: CG density The forest will be made up by the following decision trees:

- tree 1 : If $Y > 20, Z = X(1-20\%), If\ Y > 40, Z = X(1-40\%)$,...
- tree 2 : If $Y > 20, Z = X(1-21\%), If\ Y > 40, z = X(1-39\%)$,...
- ...
- tree n: If $Y > 20, Z = X(1-19\%), If\ Y > 40, z = X(1-41\%)$,...

Where X denotes the MeDIP-Seq methylation while Z represents the true methylation. Each tree was built from a re-samping of the training dataset. Using the random forest model, we are able to learn the hidden knowledge from the training dataset. Therefore, we don't need to use a specific functions (linear, exp, etc) to represent the CpG density impact. With the trained model, we are able to use the hidden knowledge from training dataset to correct the MeDIP-Seq methylation.

How to use MethRaFo?

==

The following 4 commands were provided by the methrafo package: methrafo.bamScript, methrafo.download, methrafo.train, methrafo.predict.

- 1)methrafo.download
  This command is used to download the genome files.
  Files: chromosome sequences(e.g. chr1.fa.gz), chromosome sizes(e.g. hg19.chrom.sizes).

  ```
  $methrafo.download <reference_genome_id> <output_directory>
  ```
  reference_genome_id represents the ID of the genome (e.g. hg19,mm10, et al.)

  example:
  ```
  $methrafo.download hg19 hg19
  ```

- 2)methrafo.bamScript
  This command is used to convert bam file to bigWig file (RPKM).

  ```
  $methrafo.bamScript <bam_file> <genome size file>
  ```
  bam_file => mapped reads in bam file
  genome size file => chromosome sizes .e.g. hg19.chrom.sizes (This file can be found in the downloaded reference genome folder e.g. hg19)

  example:
  ```
  $methrafo.bamScript example/example_raw.bam hg19/hg19.chrom.sizes
  ```

- 3)methrafo.train
  This command is used to train the model.
  ```
  $methrafo.train <downloaded_reference_genome_folder> <MeDIPS.bigWig> <Bisulfite.bigWig>
  <output_prefix>
  ```

  <downloaded_reference_genome_folder>: The downloaded genome reference folder (output folder of command methrafo.download, e.g. hg19)
  <MeDIPS.bigWig> : The bigWig file representing the RPKM on each position of the genome.
  <Bisulfite.bigWig>: The bigWig file representing the Bisulfite-Seq methylation level.

Note: if your input is .bam file, please use methrafo.bamScript to convert it to bigWig format.

example:

```
$methrafo.train hg19 example/example_medip.bw example/example_bisulfite.bw trained_model
```

- 4)methrafo.predict
  This command is used to predict the methylation level based on MeDIP-Seq data

  ```
  $methrafo.predict <downloaded_reference_genome_folder> <train_model> < output_prefix>
  ```

  <train_model>: the trained model from methrafo.train

  example:

  ```
  $methrafo.predict hg19 example/example_medips.bw trained_model.pkl example_out
  ```

The following presents the instructions of how to use the above commands

- 1) methRafo accept the following formats as the input:
  .bam => mapped reads from MeDIP-Seq result.
  .bw => bigWig file from MeDIP-Seq (representing the RPKM value).

  If the input is in .bam format, you need to use methrafo.bamScript to convert it to bigWig format. Please refer to methrafo.bamScript command above for conversion instruction.

  You might need to use methrafo.download the download corresponding reference genome. But you are also allowed to download the reference genome yourself. The reference genome folder needs to contain the following files: all chromosome sequences in fasta format; chromosome size file containing the length information of the chromosome.

- 2) We provided a trained model on human (Based on breast luminal cells dataset from roadmap database). We tested it on a few other datasets on human and it shows pretty good performance in terms of running time and correlation with Bisulfite- Seq data. You can use the trained model we provided or you can use the methrafo.train script to build your own model by specifying MeDIP-Seq input and Bisulfite-Seq output. Please refer to methrafo.train command for complete details.

- 3) With the provided trained model (or your own trained model), we are using methrafo.predict to predict the genome wide methylation level. The output file is a Wiggle (.wig) format file. You can visualize it using IGV or UCSC genome browser. You can also get the methylation level for any given genomic location easily from the generated wig file.

The following is an example about how to use methrafo to predict methylation for a given bam file examle.bam in human (hg19)

- First, run methrafo.download to download reference genome folder and chromosome size files $methrafo.bamScript example/example_raw.bam hg19/hg19.chrom.sizes

- Second, run methrafo.bamScript to process bam file.

  ```
  $methrafo.bamScript example/example_raw.bam hg19/hg19.chrom.sizes
  ```

  => processed bam in bigwig format example_raw.bam.bw

- Third, run methrafo.predict to predict on the bigwig file.

  ```
  $methrafo.predict hg19 example/example_raw.bam.bw trained_model.pkl example_out
  ```

References

[Bernstein *et al.*,2010] Bernstein, B. E. et al. (2010). The NIH roadmap epigenomics mapping consortium. Nature biotechnology, 28(10), 1045-1048.

[Chavez *et al.*,2010] Chavez, L. et al. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome research, 20(10), 1441-1450.

[Down *et al.*,2008] Down, T. A. et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature biotechnology, 26(7), 779-785.

[Riebler *et al.*,2014] Riebler, A. et al. (2014). BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach. Genome biology, 15(2), R35.