metilene - a tool for fast and sensitive detection of differential DNA methylation

Frank Jühling, Helene Kretzmer, Stephan H. Bernhart, Christian Otto,
Peter F. Stadler, Steve Hoffmann

Version 0.28

1 Introduction

metilene is a software tool to annotate differentally methylated regions (DMRs) and differentially methylated cytosine sites (DMCs) from Methyl-Seq data. metilene accounts for intra-group variances and offers different modes de-novo DMR detection, DMR detection within a known set of genomic features, and DMC detection. Various biological data can be used, metilene works for Whole-Genome Bisulfite Sequencing (WGBS), Reduced Representation Bisulfite Sequencing (RRBS), and any other input data, as long as absolute (methylation) levels and genomic coordinates are provided. metilene uses a circular binary segmentation and a 2D-KS test to call DMRs. Adjusted p-values are calculated using the Bonferroni or Benjamini-Hochberg correction (see option `-c`, `--mtc`).

2 Requirements

metilene is available as pre-compiled version for 32-/64-bit linux, or as source code to be built from source. It runs on a normal desktop machine and supports multi-threading. However, the underlying algorithms are efficient enough to run single-threaded as well, if needed. In case of memory issues (almost exclusively for analyzing non-CpG contexts in plants), you can limit the maximum segment length using the option `-G`, `--maxseg`.

3 Installation

If you can not or do not want to use the pre-compiled versions for 32/64bit Linux systems, you can build metilene from source. In both cases, simply download the latest version from from
`http://http://www.bioinf.uni-leipzig.de/Software/metilene/`
and extract it with

```
$ tar -xvzf metilene.tar.gz
```

go to the new directory and type

```
$ make
```

or run the pre-compiled versions directly.

4 Quick start

To do a de-novo annotation of DMRs run

```
$ metilene-a g1 -b g2 <methylation-file>
```

where the input file containing all methylation data is a tab-separated file that **must** be sorted by chromsome and position (e.g. via `sort -k1,1 -k2,2n`) and follows the format below and includes the header line:

chr	pos	g1_xxx	g1_xxx	[...]	g2_xxx	g2_xxx	[...]

or

chr	pos	g2_xxx	g3_xxx	[...]	g1_xxx	g2_xxx	[...]

where the first column refers to the chromosome, the second column to the genomic position of the C or CpG and all following columns to the absolute methylation rates. All methylation rate columns are dedicated to the group described by the prefix in their header, e.g., "g1" or "g2". Options `-a` and `-b` indicate the groups that are considered. The order in the methylation rate columns is of no importance and other groups, e.g., "g3_xxx", can be present and will be omitted for a run with "g1" as fist group and "g2" as second group.

5 DMR de-novo annotation

The default mode of metilene annotates DMRs de-novo without using any prior information on genomic features, e.g., promotor regions. Here, a fast circular binary segmentation approach on the mean difference signal of both groups is used (Siegmund, 1986; Olshen et al., 2004). After additional filter steps are passed, potential DMRs are tested using a two-dimensional Kolmogorov-Smirnov-Test (KS-test)(Fasano and Franceschini, 1987). DMRs are finaly tested through the Mann-Whitney-U test. Note that the multiple testing correction (see option `-c`, `--mtc`) is performed on the p-values of the KS-test.

6 DMR annotation in known features

Instead of annotating de-novo DMRs, metilene can be used to find significant DMRs within a given group of genomc features. Here, the first step calling the circular binary segmentation algorithm is skipped. Instead, statistical tests are performed for each feature, and corresponding p-values are reported in the output. Use the option `-B` option to define the features to be tested by giving a BED-formatted file, which **must** be sorted **equally to the input file**. Note that the multiple testing correction (see option `-c`, `--mtc`) is performed on the p-values of the KS-test.

7 DMC annotation

metilene offers the possibility to test each C or CpG for differential methylation. Hence, the segmentation step is also skipped and the Mann-Whitney-U test is then just calculated for each C or CpG site, and corresponding p-values are reported in the output. Note that the KS-test is skipped and the p-values are just reported as '.' and hence the multiple testing correction (see option `-c`, `--mtc` for setting the method) is performed on the p-values of the Mann-Whitney-U test.

8 Input

The input consists of a single **sorted** (for genomic positions) tab-separated file. It must contain a header line of the format:

chr	pos	g1_xxx	g1_xxx	[...]	g2_xxx	g2_xxx	[...]

or

chr	pos	g2_xxx	g3_xxx	[...]	g1_xxx	g2_xxx	[...]

or any other unsorted order of the columns. The following tab-separated lines contain the data for each C or CpG site, depending on the users choice. The affiliations of samples is assigned through a unique prefix, e.g., “g1”, “g2” , which are passed as arguments when calling metilene. No underscore is required, and names can be labeled completely freely. The input file can contain data of more than two groups, however, only the two selected groups are considered. See section 12 for more details for the group selection when calling metilene.

8.1 Generate an input file from multiple bed files

We offer an easy way to generate an appropriate input file containing all methylation rates. Therefore, the project archive (see section 3) contains the script `generateInput.pl` to generate a sorted tab-separated input file from multiple BED-format files. A basic metilene call for the specific input file is printed to stdout. If you do not like to use `generateInput.pl`, we recommend using `bedtools unionbedg` yourself.

It takes two comma-separated lists of **sorted** BED-format files and creates a metilene input matrix out of it. You can further specify the group affiliation - one for each group, that will show up in the header of the metilene input file. The input wrapper uses `bedtools`, if the executable is not specified, it is assumed to be in `PATH`.

To create the input file for metilene with `generateInput.pl`, please call:

```
$ perl generateInput.pl -in1 <string> -in2 <string> [-out <string>] [-h1 <string>] [-h2 <string>]
[-b <string>]
```

option	description
`--in1`	comma-separated list of **sorted** bed(graph) input files of group 1
`--in2`	comma-separated list of **sorted** bed(graph) input files of group 2
`--out`	path/file of out file (metilene input) (default: metilene_g1_g2.input, “g1” set by `--h1` option, “g2” set by `--h2` option)
`--h1`	identifier of 1st group (default: g1)
`--h2`	identifier of 2nd group (default: g2)
`-b`	path to executable of bedtools (default: in `PATH`)

9 Missing values

metilene can handle missing values, indicated by “-” or “.” in the input file. Missing values are replaced by a random number taken from a beta distribution estimated from the remaining values of the corresponding group the replicate with the missing value belongs to. The default minimal number of provided values is set to 80% of the group sizes, see options `-X`, `--minNoA` and `-Y`, `--minNoB` for further information how to change these two cutoffs for each input group. All input rows that fall below of one of these cutoffs are ignored. See option `-s`, `--seed` for changing the initial seed of the random number generator.

10 Output

The output for the de-novo DMR annotation mode consists of a format similar to the BED-format:

chr	start	end	q-value	mean methyl difference	#CpGs	p-value (MWU)	p-value (2D KS)	mean g1	mean g2

The columns “mean g1” and “mean g2” refer to the absolute mean methylation level for the corresponding segment per group and the difference (g1 - g2) between the group means is given in the 5th column. Single Cs or CpGs are not tested using the 2D KS-test. Here, q-values are based on MWU-test p-values.

All outputs are unsorted when using multiple threads. We recommend to use sort:

```
$ metilene options | sort -k1,1V -k2,2n
```

for a sorted output.

10.1 Filter output file and plot basic DMR statistics

An easy way to filter your already called DMRs is offered by `filterOutput.pl`. Furthermore, it will create some basic statistic plots characterizing your DMRs, i.e., distribution of DMR differences, DMR length in nucleotides and #CpGs, DMR differences vs. q-values, mean methylation group 1 vs. mean methylation group 2 and DMR length in nucleotides vs. length in CpGs (Fig. 1). A version of `R` with the `ggplot2` package installed is required to be in the `PATH`. DMRs can by filtered by q-value, # CpGs, length in nucleotides and mean methylation difference. 3 files are produced: (i) bedgraph file containing the methylation difference for each DMR, (ii) basic statistic pdf and (iii) filtered bedgraph-like file, containing all information already in the metilene output. To filter the metilene output file and plot the basic statistic plots, please call:

```
$ perl filterOutput.pl -q <string>[-o <string>] [-p <n>] [-c <n>] [-d <n>] [-l <n>]
[-a <string>] [-b <string>]
```

option	description
`-q`	path/filename of metilene output
`-o`	path/prefix of filtered output files, i.e. bedgraph file, filtered output file and pdf (default: metilene_qval_0.05.bed, metilene_qval_0.05.pdf)
`-p`	maximum (<) adj. p-value (q-value) for output of significant DMRs (default: 0.05)
`-c`	minimum (>=) CpGs (default: 10)
`-d`	minimum mean methylation difference (>=) (default: 0.1)
`-l`	minimum length of DMR [nt] (>=) (post-processing, default: 0)
`-a`	name of 1st group (default: g1)
`-b`	name of 2nd group (default: g2)

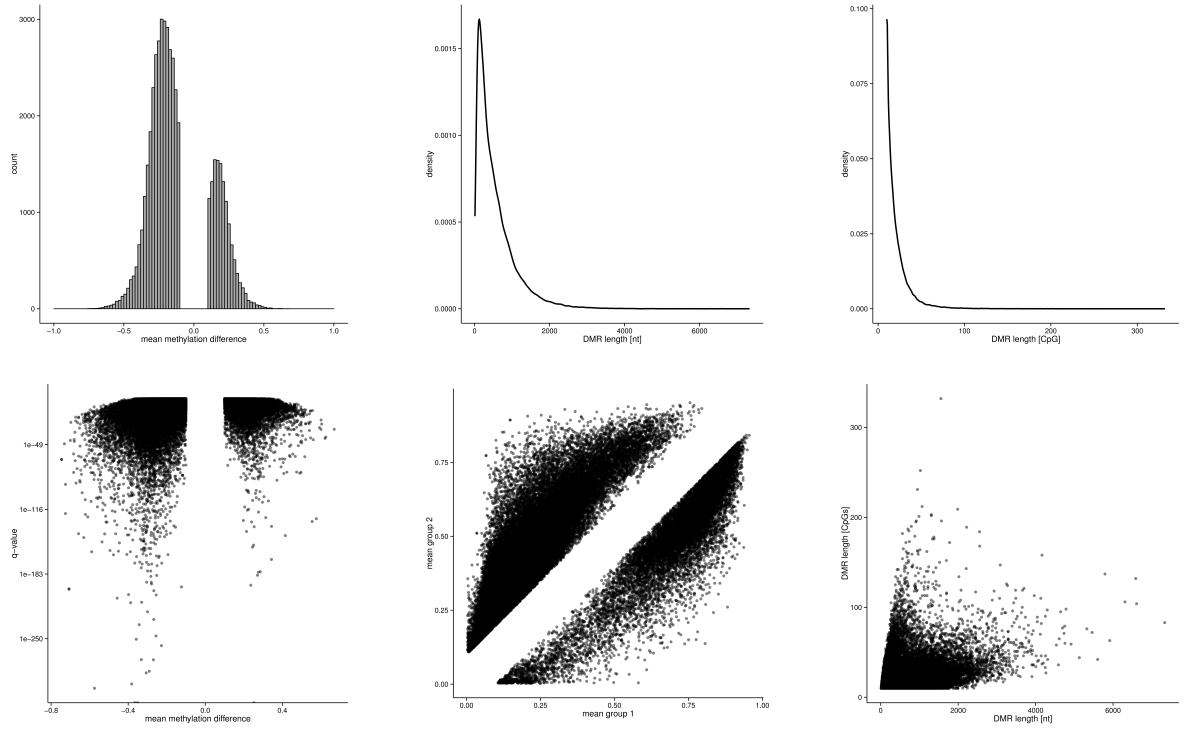

Figure 1: Basic statistics plots produced with `filterOutput.pl`

11 Usage

`metilene [-M <n>] [-G <n>] [-m <n>] [-d <n>] [-t <n>] [-f <n>] [-c <n>] [-a <string>] [-b <string>] [-B <string>] [-X <n>] [-Y <n>] [-s <n>] [-v <n>]` DataInputFile

12 Options

option	unit	default	description
DataInputFile	File	None	A **sorted** file containing the input data
`-M, --maxdist`	Integer	500	Maximum genomic distance (in nt) between two positions (Cs or CpGs) within a DMR
`-G, --maxseg`	Integer	-1	Maximum segment length in case of memory issues
`-m, --mincpgs`	Integer	10	Minimum number of positions (Cs or CpGs) in a DMR
`-d, --minMethDiff`	Double	0.1	Minimum mean methylation difference for calling DMRs
`-t, --threads`	Integer	1	Number of threads
`-f, --mode`	Integer	1	Mode of operation: 1: de-novo, 2: pre-defined regions, 3: DMCs
`-c, --mtc`	Integer	1	Method of multiple testing correction: 1: Bonferroni, 2: Benjamini-Hochberg (FDR)
`-a, --groupA`	String	g1	Prefix of group name of samples of the 1st group
`-b, --groupB`	String	g2	Prefix of group name of samples of the 2nd group
`-B, --bed`	File	None	A **sorted** BED-format file containing regions for mode 2
`-X, --minNoA`	Integer	0.8	Minimum number or fraction of non-missing values for estimating missing values in 1st group*
`-Y, --minNoB`	Integer	0.8	Minimum number or fraction of non-missing values for estimating missing values in 2nd group*
`-s, --seed`	Integer	26061981	Seed for random number generator
`-v, --valley`	Double	0.7	Stringency of the valley filter (0.0 - 1.0)

*If not enough entries are available, the corresponding line is skipped due to too many missing values.

12.1 Option `-M`

metilene works in two steps, first it pre-segments the whole data into windows so that there are no large gaps in genomic coordinates. The option `-M` sets the maximum length in nts between adjacent data points (Cs or CpG) to be considered together. The default value of 300 means, that the whole genome is cut whenever a stretch of 300 nt or more without data (Cs or CpGs) is found. For example, if you do not want to find DMRs that may contain genomic stretches without data points (Cs or CpGs) longer than, e.g., 200nt, the option `-M` should be set to 200.

12.2 Option `-G`

This option sets the maximum segment length in terms of the number of data points in cases where the memory consumption becomes an issue. It occurs almost exclusively when analyzing cytosines from non-CpG contexts (e.g. in plants). By default, the maximum segment length is unlimited (i.e. set to `-1`).

12.3 Option `-m`

The option `-m` sets the minimum number of data points (Cs or CpGs) within a DMR to be reported. As we use a top-down approach, starting with long windows and segmenting them to short significant DMRs, this is also a stop-criteria. Windows that contain a smaller number of data points (Cs or CpGs) are not considered and skipped.

12.4 Option `-d`

The option `-d` sets the minimum mean methylation difference between both groups for a DMR. This prevents to call regions with very small but significant methylation differences. It is set to the default value of 0.1 since for most applications it is reasonable to require at least this difference in the methylation signal to be considered differentially methylated.

12.5 Option `-t`

metilene is implemented with multi-threading, i.e. taking advantage of modern multi-processor machines, and the option `-t` sets the the number of threads. metilene uses multiple threads to search for DMRs within pre-segmented windows (see the option `-M`) in parallel. If you have the possibility to run metilene on a multi-core machine, you should it on as many cores as possible. However, you should consider that reading the input file could be another bottleneck in your environment.

12.6 Option `-f`

This option can be used to specify the mode of operation (by default finding de-novo DMRs). In case of setting the option to 2, it simply tests pre-defined regions from a BED-format file given via the option `-B`. In order to search for differentially methylated positions/cytosines (DMPs/DMCs), you would need to set this option to 3.

12.7 Option `-c`

This option can be used to specify the method that will be used for the multiple testing correction of p-values. By default, the Bonferroni correction is used which controls the family-wise error rate. It can be changed to the method of Benjamini-Hochberg by setting `-c` to 2, which controls the false discovery rate (FDR) and is less stringent. Note that for the modes 1 and 2 (de-novo or pre-defined DMRs; see option `-f`), the correction is based on the p-values of the 2D Kolmogorov-Smirnov test while for mode 3 (DMCs/DMPs) it is based on the p-values of the Mann-Whitney-U test.

12.8 Options `-a` and `-b`

Both options specify the prefixes of group names for columns in order to assign methylation rates of columns to both groups (see section 8).

12.9 Option `-B`

This option specifies a **sorted** (equally to the input data) BED-format file containing regions of interest that are evaluated with respect to differential methylation in case of `-f 2`. Only the first three columns of the BED-file are required and used (i.e., chr, start, and end position).

12.10 Options `-X` and `-Y`

metilene can estimate missing values using the available data of samples in the same group. Both options specify how many samples with non-missing values must be present for a certain position in the first group (`-X`) or the second group (`-Y`) in order to estimate the missing ones. The default value is set to 0.8, which means that at least 80% of the number of samples of each group must have non-missing values. It is also possible to set these options to values $\leq$1 which then refers to the absolute numbers of samples instead of relative ones.

12.11 Option `-s`

This option allows you to set the seed used for initialization of the random number generator to another value.

12.12 Option `-v`

metilene's valley filter prevents to call large regions as a single DMR enclosing a valley in the mean difference signal. The option `-v` sets a cutoff factor for the methylation difference when comparing global and regional methylation differences. Thus, it is forced to segment further until no more valleys are found. The influence of this approach can be reduced by decreasing this factor, or it can be turned off by using -v 0. The effect of parameter settings on resulting DMR calls at different valley sizes/depths within the mean methylation signal is illustrated in Fig. 2.

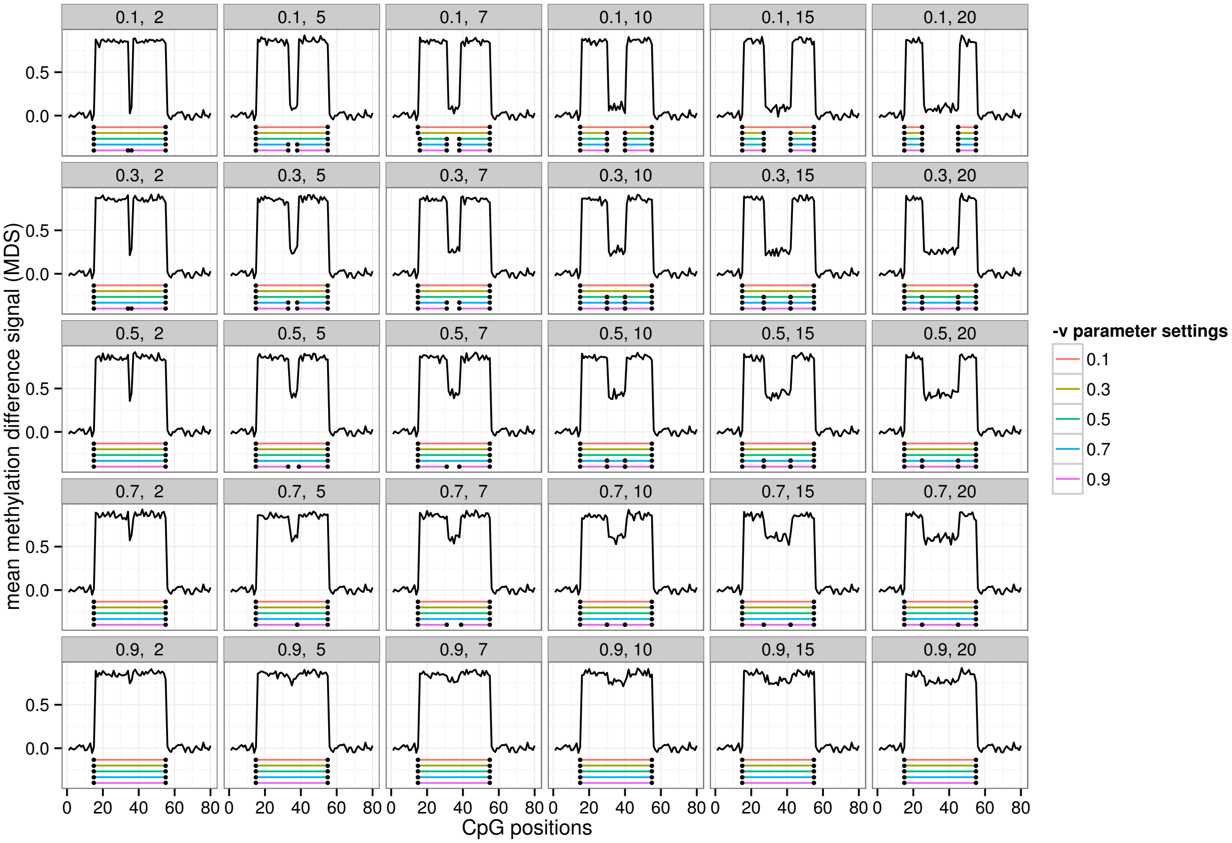

Figure 2: The effect of different settings of the option `-v` on the prediction of DMRs

13 Complaints

All complaints go to [frank,steve] at bioinf dot uni-leipzig dot de.

References

Fasano, G. and Franceschini, A. (1987). A multidimensional version of the kolmogorov–smirnov test. Monthly Notices of the Royal Astronomical Society, **225**(1), 155–170.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics, **5**(4), 557–572.

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. The Annals of Statistics, pages 361–404.