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 0.23

1 Introduction

metilene is a software tool to annotate differentally methylated regions (DMRs) and differentially methylated CpG 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 with 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. Adjsuted p-values are calculated using the Bonferroni correction.

2 Requirements

metilene is available as pre-compiled versions for 32/64-bit linux, or as source code to be built from source. It runs on a normal sesktop machine and supports multi-threading. However, the underlying algorithms are efficient enough to run only single-threaded, if needed.

3 Installation

If you do not want to use the pre-compiled versions for 32/64-bit Linux systems, you can build metilene from source. In both cases, simply download the latest version 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
```

while the input file containing all methylation data is a SORTED tab-separated file with the following format and header:

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 CpG and all following columns to the absolute methylation ratio. All ratio 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 ratio columns order can be mixed, and other groups, e.g., g3_xxx, can be present and will be omitted for a run calling -a g1 and -b g2.

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.

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 "-B *bedfile*" option to define windows through a bedfile SORTED equally to the data input file.

7 DMC annotation

metilene offers the possibility to test each CpG for differential methylation. Statistical tests (KS-test and Mann-Whitney-U test) are calculated for each CpG site, and corresponding p-values are reported in the output.

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 (3) contains a script "generateInput.pl" generating a sorted tab-separated input file from multiple bed-files. A basic metilene call for the specific input file is printed to the command line. If you do not like to use "generateInput.pl", we recommend using bedtools unionbedg yourself.

It takes two comma-separated lists of **sorted** bed 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. Creating an metilene input file using `generateInput.pl`, please call:

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

parameter	description
–in1	comma-seperated list of **sorted** bed(graph) input files of group 1
–in2	comma-seperated 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 group 1 (default: g1)
–h2	identifier of group 2 (default: g2)
-b	path/executable of bedtools executable (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" and "-Y" 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.

10 Output

The output for the de-novo DMR annotation mode consists of a bed-like format:

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

While "mean g1" and "mean g2" refer to the absolute mean methylation level for the corresponding segment in both groups, the difference is given in the 5th column. Single 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 -V -k1,1 -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 is required to be in 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>]
```

parameter	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 group A (default:"g1")
-b	name of group B (default:"g2")

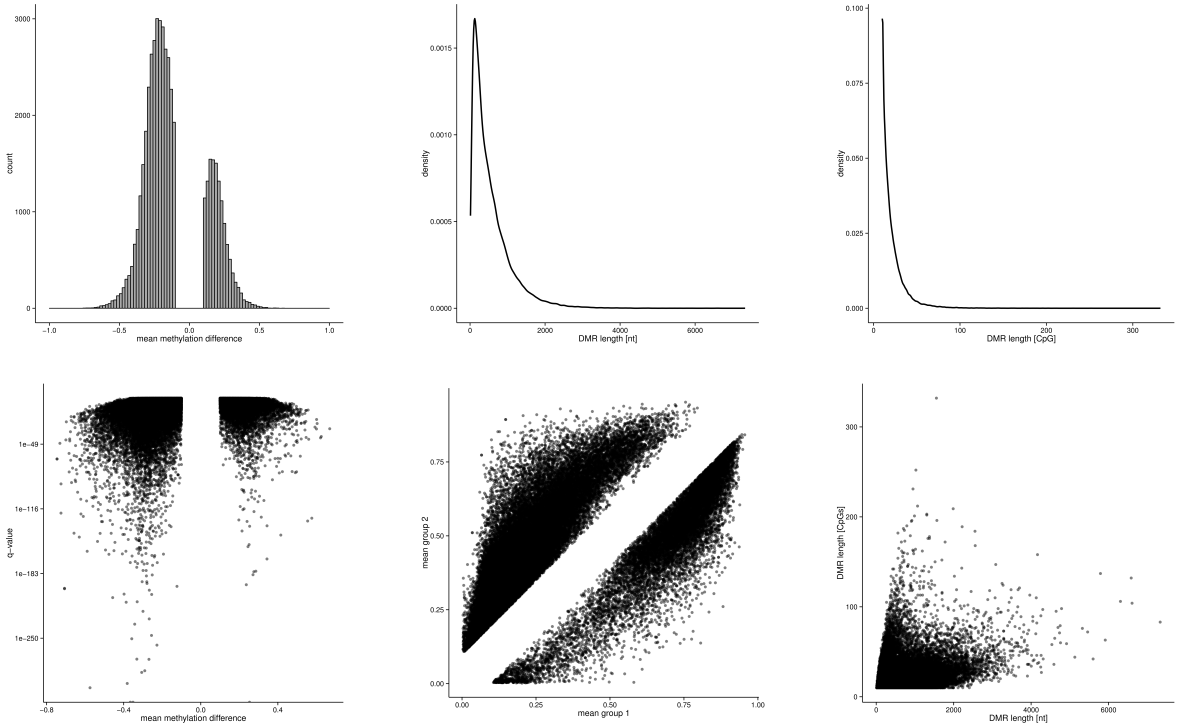

Figure 1: Basic statistics plots produced with filterOutput.pl

11 Usage

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

12 Parameters

parameter	unit	default	description
DataInputFile			a SORTED file containing the input data
-M, –maxdist	Integer	300	The allowed nt distance between two CpGs within a DMR
-m, –mincpgs	Integer	10	The minimum # of CpGs in a DMR
-d, –minMethDiff	double	0.1	The minimum mean methylation difference for calling DMRs
-t, –threads	Integer	1	The number of threads
-f, –mode	Integer	1	The method selection: 1: de-novo, 2: pre-defined regions, 3: DMCs
-a, –groupA	String	g1	The name prefix of replicates in the 1st group
-b, –groupB	String	g2	The name prefix of replicates in the 2nd group
-B, –bed	String		A SORTED (equally to the input data) bed file containing regions for mode 2
-X, –minNoA	Integer	0.8%	Minimal # of non-missing values for estimating missing values in g1*
-Y, –minNoB	Integer	0.8%	Minimal # of non-missing values for estimating missing values in g2*
-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 Parameter -M

metilene works in two steps, first it pre-segments the whole data into windows so that no large gaps without data are possible. The -M parameter sets this length in nts. The default value of 300 means, that the whole genome is cut whenever a stretch of 300nts or more without data (CpGs) is found. E.g., if you the user does not want to find DMRs with stretches without CpGs longer than 200nt, the option “-M 200” should be used.

12.2 Parameter -m

The length parameter -m sets the minimum value of CpGs/data points a DMR need to contain 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 CpGs are not considered and skipped.

12.3 Parameter -d

The option -d sets the minimum mean methylation difference between both groups for a window to be reported as a DMR. This prevents to call regions with very small but significant significant methylation differences. We think that most users do not want to call smaller mean differences than 0.1, as the difference would be too small to term those regions as differentially methylated.

12.4 Parameter -t

metilene is completely multithreaded implemented, the -t parameter sets the number of possible threads. metilene uses multiple threads to search for DMRs within pre-segmented windows (see the -M parameter) 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.

12.5 Parameter -f

This parameter can be used to apply other search methods to the data. If metilene is called using -f 2 it checks pre-defined regions given in a bed file (see parameter -B) for differential methylation. Single differentially methylated CpGs are searched using -f 3.

12.6 Parameters -a and -b

Both parameters specify the prefixes for column names of both groups, see section Input.

12.7 Parameter -B

This parameter specifies a SORTED (equally to the input data) bed file containing regions of interest that should be checked for differential methylation, see -f parameter. Only the first three columns of the bed-file are used (chr–start–stop)

12.8 Parameters -X and -Y

metilene can estimate missing from available data of other replicates. Both parameters specify how many replicates must contain data for a certain CpG position in group 1 (-X) or group 2 (-Y) to estimate missing ones. The default value is set to 80% of the number of replicates of each group. However, when changing this by using these paramters, they are set to absolute numbers of replicates, not to percentages.

12.9 Parameter -v

metilene 's valley filter prevents to call large regions as a single DMR where a valley in the mean difference signal is inside. The -v parameter sets a cutoff factor for the methylation difference when comparing global and regional methylation differences. Thus forces to segment further until no more valleys are found. Its influence 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.

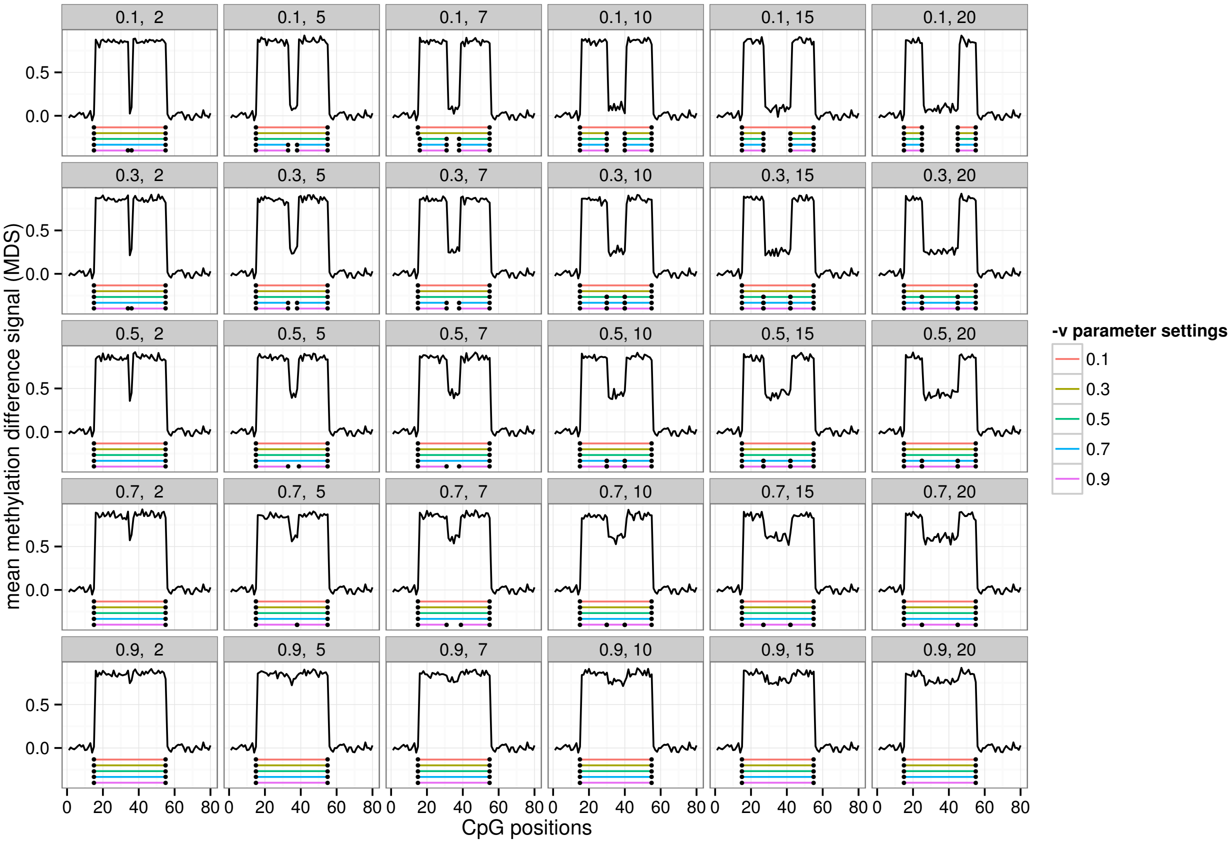

Figure 2: The effect of different -v parameter settings on the prediction of DMRs

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.