MethylExtract Manual (v1.9)

Ricardo Lebrón, Guillermo Barturen, Antonio Rueda, José L. Oliver and Michael Hackenberg

Please contact us for any doubt or consult

MH (mlhack@gmail.com), JLO (oliver@ugr.es), GB (bartg01@gmail.com) or RL (rlebron@ugr.es)

Whole genome methylation profiling at a single cytosine resolution is now feasible due to the advent of high-throughput sequencing techniques together with bisulfite treatment of the DNA. To obtain the methylation values, the sequence reads need to be first aligned to a reference genome, afterwards profiling the methylation level from the alignments. A huge effort has been made to fast and correctly align the reads and many different algorithms and programs do exist. However, the second step is likewise crucial and non-trivial, but much less attention was paid to the final inference of the methylation states. Important error sources do exist like sequencing errors, bisulfite failures, clonal reads and single nucleotide variants. *MethylExtract* implements all these quality control features together. Furthermore, the program is able to detect variation (SNVs) using the same sequence library, which is crucial for many downstream analyses, for example when deciphering the impact of sequence variation on differential methylation.

Note that before running *MethylExtract*, the user needs to align the reads first, using for example *NGSmethPipe* or *Bismark*. Additionally, *Bis-SNP* provides two scripts to ensure the correctly alignment and quality calibration for bisulfite treated reads, that can be used before running *MethylExtract* (as it is proposed in *GATK* tools best practices).

Main features of MethylExtract:

1. The program implements five quality filters: (i) detection and removal of potential PCR duplicates, (ii) detection of putative bisulfite failures, (iii) optional elimination of 5’ and/or 3’ end of the reads, (iv) control of sequencing errors by means of the PHRED Scores, (v) SNVs (single nucleotide variants) profiling (the methylation level will be reassigned to the real sequence context found in the sample).
2. Profiling of all possible cytosine sequence contexts (CG, CHG and CHH; where H is A, T or C).
3. Multiple optional outputs for methylation values and SNVs, like WIG or BED formats for representing the methylation output datasets and VCF format for SNVs.

4. Complete statistics of the entire process, including discarded reads, discarded positions, chromosome data coverage, etc.
5. The memory and CPU needs can be adapted to the user's computer resources.
6. SAM (gzip format or uncompressed) or BAM input files are accepted (to process BAM files, *samtools* must be installed).

1. Installation of MethylExtract (Unix platforms)

- Install the Perl interpreter (http://www.perl.org/) and resolve two dependencies. From the command-line (with super-user mode):
  - ✓ *Bundle::Thread* (perl –MCPAN –e "install Bundle::Thread").
  - ✓ *IO::Uncompress::AnyUncompress* (perl -MCPAN -e "install IO::Uncompress::AnyUncompress).
- Download MethylExtract last version http://bioinfo2.ugr.es/MethylExtract/downloads/MethylExtract.tgz
- Download the example http://bioinfo2.ugr.es/MethylExtract/downloads/Example.tgz (for example to /home/user/Example/). The compressed example file includes all necessary sequences and reads to test the program.
- Download the genome sequences into a folder of your choice (/home/user/seq/assembly/). The genome sequences can be downloaded for example from the UCSC genome browser: http://hgdownload.cse.ucsc.edu/downloads.html. *Note that this step is not necessary to test the program.*

2. Quick start

In the example directory, two folders can be found, 'methDataset' and 'unmethDataset' containing two different simulated datasets based on human contig GL000022.1 (one with all cytosines methylated and the other completely unmethylated). This data is located in the "inDir" folder while the "seq" folder holds the entire sequence of the hg19 chromosome 2. Next, we will give some examples to explain how to use the program:

i. Quick Launch (default Options)

'perl MethylExtract.pl seq=*/home/user/Example/methDataset/seq/* inDir=*/home/user/Example/ methDataset/inDir/*'

This command will run MethylExtract with default parameters. MethylExtract assigns 4 threads by default, but notice that the number of assigned threads depends on the number of chromosomes which is 1 in this case (human chromosome 2).

ii. Quality & SNV detection (minQ=30 varFraction=0.05 maxPval=0.01)

```
'perl MethylExtract.pl seq=/home/user/Example/ methDataset/seq/
inDir=/home/user/Example/ methDataset/inDir/minQ=30 varFraction=0.05
maxPval=0.01'
```

Several quality parameters can be set for the methylation and SNV profiling. In this example, we increase the PHRED score threshold ('minQ') from the default value of 20 to 30 (more stringent). The 'varFraction' is reduced to 0.05 what makes sense if the 'minQ' is increased to reduce the sequencing errors contribution and the 'maxPvalue' is set to 0.01 trying to increase the SNVs positive predictive values (see **Parameters**).

iii. Parallelization Options (p=6 chromDiv=800 memNumReads=500000)

```
'perl MethylExtract.pl seq=/home/user/Example/ methDataset/ seq/
inDir=/home/user/Example/ methDataset/ inDir/ p=6 chromDiv=800
memNumReads=500000'
```

In order to adapt the memory and CPU requirement of the program, three parameters can be tuned: 'p' (number of threads), 'chromDiv' (the number of temporary files per chromosome) and 'memNumReads' (number of reads that are kept in memory). The last two parameters affect the memory requirement of the program. Higher 'chromDiv' will result in lower memory requirements but it will slow down the program, while higher 'memNumReads' will require more memory, increasing the speed (see **Parameters**).

iv. Output Options

```
'perl MethylExtract.pl seq=/home/user/Example/ methDataset/seq/
inDir=/home/user/Example/ methDataset/inDir/pattern=ALL wigOut=Y bedOut=Y'
```

This example shows how to obtain the final methylation levels. The parameter 'context' allows to choose the sequence context which should be processed (in the example all contexts will be processed). In order to activate other output formats as WIG or BED on the example, the user has to specify them (wigOut=Y and/or bedOut=Y).

v. Pair-end reads Option

```
'perl MethylExtract.pl seq=/home/user/Example/methDataset/seq/
inDir=/home/user/Example/ methDataset/inDir/ flagW=99,147 flagC=83,163'
```

Most of the pair-end aligners use at least two tags to specify the reads strand. In order to process this type of reads, the user must include as many tags as needed for each strand, separated by commas. This example shows the tags used by *Bismark*, a widely used bisulfite-aligner.

3. Parameters

The software has some mandatory parameters: the input directory, where the SAM/BAM files are located (**'inDir'**); the fasta sequences directory or multifasta file (**'seq'**) and FLAGs for the Watson and Crick strands (**'flagW'** and **'flagC'**). The parameters are supplied with this format: <*name of the option*>=<*value*>.

- **Mandatory parameters:**
  - **seq** → path to the reference genome sequences folder, containing multiple files in single fasta format *(/home/user/Example/methDataset/seq/)* or the full path to a multifasta file *(/home/user/Example/methDataset/seq/multifasta.fa)*. Note that the fasta files must have the same IDs that were used to align the reads.
  - **inDir** → path to SAM/BAM files *(/home/user/Example/methDataset inDir/)*.
  - **flagW** (or deprecated tagW) → used tags in the SAM file for reads mapped on Watson strand (default 0 for single-end alignments). In case of pair-end reads, the user should use at least two tags comma-separated, for example flagW=99,147. For more information, consult: http://samtools.sourceforge.net/.
  - **flagC** (or deprecated tagC) → used tags in the SAM file for reads mapped on Crick strand (default: 16 for single-end alignments). In case of pair-end reads, the user should use at least two tags comma-separated, for example flagC=83,163. For more information, consult: http://samtools.sourceforge.net/.

- **Optional parameters:**
  - **Quality parameters**
    - **qscore** → PHRED score encoding used in the fastq files. The different options are: *phred33-quals*, *phred64-quals*, *solexa-quals* or *solexa1.3-quals*. The default is *phred33*-quals. This parameter needs to be correct, otherwise, *MethylExtract* will work with false quality values.
    - **delDup** → activates the detection of duplicated reads (*Y*), by default it is deactivated (*N*). (Warning: The de-duplication step is not advised for RRBS methodology). It groups reads by their start position (reads mapped on Watson strand) or end position (reads mapped on Crick strand), and considers duplicated reads those with equal '**simDupPB**' nucleotides at the 5' end. Then the read with the highest number of quality values >= '**minQ**' or the longest read in case of equal

number of high quality bases will be selected. If more than one read show equal quality/length features, one will be picked randomly. Currently, *MethylExtract* removes duplicated pair-end reads as they were single-end, so we recommend to use *picard-tools* to remove duplicates in pair-end protocols.

- **peOverlap** → discards the second mate overlapping segment on pair-end alignment reads (by default N).
- **simDupPb** → number of equal nucleotides at the 5' end of the reads, that map to the same position, in order to consider them putatively duplicated reads. If the reads have been aligned with a seed methodology, it is recommended to use the same length as the alignment seed (by default 32).
- **FirstIgnor** → number of bases to be ignored at the 5' end of the reads. This option is useful for RRBS non-directional methodology; the digestion with MspI restriction enzyme leaves 3 or 2 nucleotides from the restriction site, depending on the orientation. These nucleotides need to be removed in order to avoid bias in the methylation values. See the following guide for more details: http://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf. The parameter can be set to trim any number of nucleotides, in order to be useful for other methodologies and situations, as avoiding M-bias (by default 0).
- **LastIgnor** → number of bases to be ignored at 3' end of the reads (by default 0).
- **minDepthMeth** → minimum number of mapped reads to a given cytosine (sequencing depth) in order to report the methylation value of this position (by default is set to 1).
- **minDepthSNV** → minimum sequencing depth required to check a position for variation (by default is set to 1, the p-value is a better parameter to distinguish fluctuations and real variation).
- **minQ** → minimum PHRED quality value required for methylation and SNV profiling. All positions with lower values are discarded (default minQ=20).
- **methNonCpGs** → methylation outside CpG contexts to detect putative bisulfite failures. The value can be the fraction of methylated non-CpG contexts within a read (values between 0 and 1) or the absolute number of methylated non-CpG contexts (integers higher than 1). Default value is 0.9,

i.e. a read is discarded when more than 90% of its non-CpG contexts are methylated. A value of 0 will turn off the bisulfite check.

- **varFraction** → minimum nucleotide frequency threshold. This parameter acts before the statistical test for each position. It discards those nucleotides at a given position with low frequencies in the sample. By default, varFraction=0.1, which means that the nucleotides with a frequency below 0.1 will not be considered during the variation detection.
- **maxPval** → the p-value threshold for SNV profiling (default 0.05). Fisher's exact test is implemented and the 2x2 contingency table is generated with the reference and non-reference nucleotide counts.
- **maxStrandBias** → filter for SNVs Fisher strand bias, maximum strand bias allowed by default is 0.7.

- **Working parameters**
  - **p** → number of threads to parallelize the process (by default p=4).
  - **chromDiv** → number of divisions made in the chromosomes to sort the reads - it is also the number of temporally files that will be created for that purpose. Increasing the number of divisions will reduce the memory used during the sort process, but will generally decrease the speed. By default, it is set to 400, which corresponds to approximately 50Mb temporary files for a standard 15x experiment in the human genome.
  - **memNumReads** → number of reads that will be kept in memory during sorting and profiling steps. Be careful modifying this parameter, the memory usage could drastically increase for datasets with a high coverage. By default, it is set to 200,000 what maintains the memory usage for each thread below 1.5GB.
  - **chromSplitted** (or deprecated chromSorted) → if the SAM/BAM input is already sorted and split into chromosomes (chromSplitted=Y **deactivates** the sorting procedure).
- **Output parameters**
  - **context** → sequence context used for methylation profiling. The options are: CG, CHG, CHH or ALL (where H is any nucleotide except G) (by default context=CG).
  - **outDir** → output folder (by default all output files will be written into the '**inDir**' directory).

- **bedOut** → extract methylation values in BED format (see http://genome.ucsc.edu/FAQ/FAQformat#format1).
- **wigOut** → extract methylation values in WIG format. (see http://genome.ucsc.edu/goldenPath/help/wiggle.html).

4. Output formats

- Methylation values output:

  The methylation output files will indicate the used sequence context. The output file has a tab-separated format:

  - CHROM → chromosome.
  - POS → sequence context most 5' position on the Watson strand (1-based).
  - CONTEXT → sequence contexts with the SNVs annotated using the IUPAC nucleotide ambiguity code (referred to the Watson strand).
  - Watson METH → number of methyl-cytosines (referred to the Watson strand).
  - Watson COVERAGE → reads covering the cytosine in this sequence context (referred to the Watson strand).
  - Watson QUAL → PHRED score average for the reads covering the cytosine (referred to the Watson strand).
  - Crick METH → number of methyl-cytosines (referred to the Watson strand).
  - Crick COVERAGE → reads covering the guanine in this context (referred to the Watson strand).
  - Crick QUAL → PHRED score average for the reads covering the guanine (referred to the Watson strand).

- SNV output:

  The SNVs results are reported in VCF format. For more information, please see: http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41.

- LogFile output:

  The log file summarizes the used parameters and reports some interesting statistics of the process: elapsed time, number of discarded reads by the bisulfite check, number of positions discarded by means of the Q value threshold, number of deleted duplicated reads, number of homozygous and heterozygous SNVs found and the sequence contexts used for methylation profiling, coverage as a function of chromosome.

5. Additional scripts

Besides the main functions provided by the *MethylExtract* main script, other scripts have been added in order to complete and improve the process of methylation profiling.

- **MethylExtractBSCR**:
    - Function: Allows the user to calculate the bisulfite conversion rate from an unmethylated genome:
      For plant genomes, the script can be run with the reads aligned to the chloroplast genome, while in other organisms an unmethylated genome (as the phage lambda) must be added to the experimental setup and the alignment process.
    - Input parameters
      The mandatory parameters of this script are the SAM file with the unmethylated genome alignments (‘**inFile**’), the fasta file with the reference sequence (‘**seqFile**’) and both strands FLAGs (‘**flagW**’ and ‘**flagC**’). The ‘**minQ**’, ‘**FirstIgnor**’ and ‘**LastIgnor**’ parameters have been included to increase the quality of the conversion rate estimation.
        - How to extract unmethylated chromosome reads from sam/bam files?
          1. Convert SAM to BAM (only if your file has a SAM format)
             ```
             “samtools view –Sb READS.sam > READS.bam”
             ```
          2. Sort BAM file
             ```
             “samtools sort READS.bam READS.sort”
             ```
          3. Index BAM file
             ```
             “samtools index READS.sort.bam”
             ```
          4. Extract unmethylated chromosome reads
             ```
             “samtools view READS.sort.bam chrID> chrID.bam”
             ```

          (Future versions of MethylExtract will include a script to simplify this process)
        - How to extract unmethylated chromosome sequence from multifasta file?

1. Extract chromosome sequence into a single fasta
   "samtools faidx MULTIFASTA.fa chrID > chrID.fa"

- Output

  The output will go to STDOUT, including the number of cytosines analyzed and the bisulfite conversion rate, which will be used as input in the next script.

- **MethylExtractBSPvalue**:

  - Function

    The script calculates the error probability (p-value) for each methylation ratio using the binomial distribution, given an error methylation interval ('**errorInterval**'). In addition, the Benjamini-Hochberg step-up procedure has been implemented to control the false discovery rate. Note that this last step is optional and can be activated by choosing a FDR value with '**FDR**' parameter.

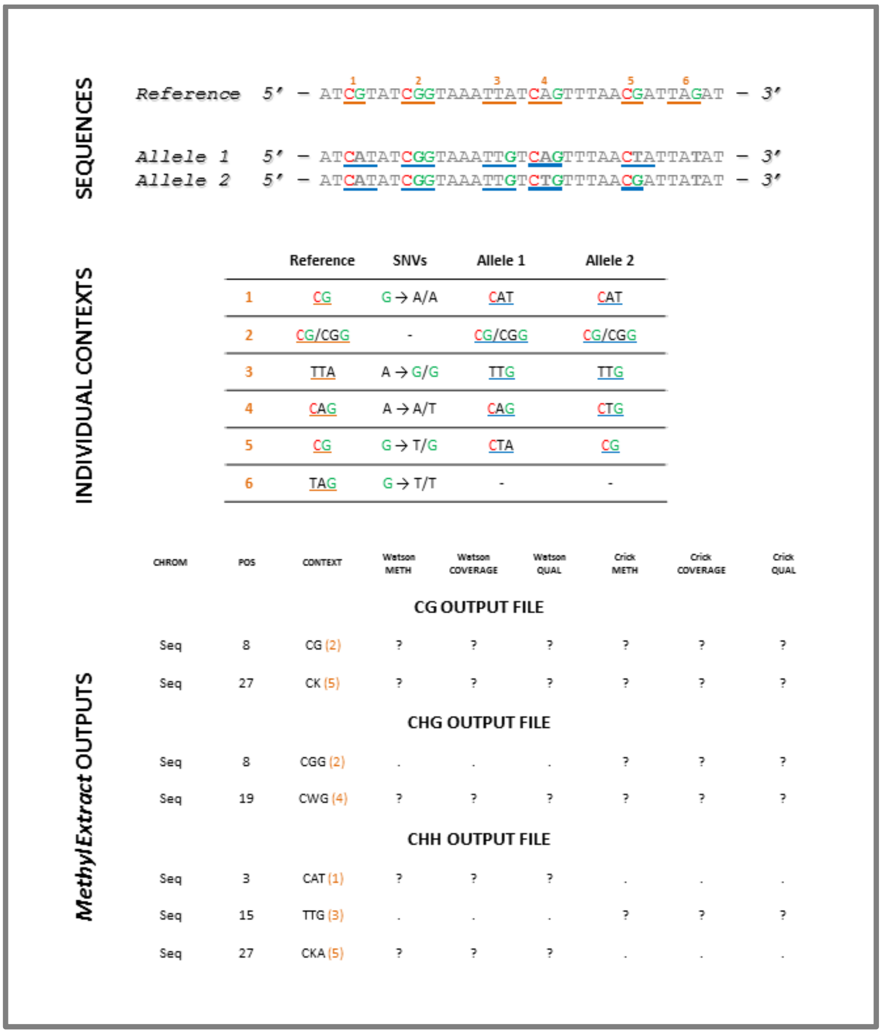
  - Input parameters

    Besides the methylation output file from *MethylExtract* ('**inFile**'), the bisulfite conversion rate ('**BSCR**') given by *MethylExtractBSCR* is the second mandatory parameter in this script.

  - Output

    The output files will have the same structure as the files from *MethylExtract* or *MethylExtractOutMethContexts* (see **Output formats**) with an additional column: the p-value assigned to each methylation value. In case of using the FDR step, the script will retrieve two files: one with the significant positions for the given FDR and another one with the methylation values that must be discarded to maintain the selected FDR in the dataset.

6. How *MethylExtract* manages SNVs within methylation contexts?



```
Reference  5' – ATCGTATCGGTAAATTATCAGTTTAACGATTAGAT – 3'

Allele 1   5' – ATCATATCGGTAAATTGTCAGTTTAACTATTATAT – 3'
Allele 2   5' – ATCATATCGGTAAATTGTCTGTTTAACGATTATAT – 3'
```

	Reference	SNVs	Allele 1	Allele 2
1	CG	G → A/A	CAT	CAT
2	CG/CGG	-	CG/CGG	CG/CGG
3	TTA	A → G/G	TTG	TTG
4	CAG	A → A/T	CAG	CTG
5	CG	G → T/G	CTA	CG
6	TAG	G → T/T	-	-

CHROM	POS	CONTEXT	Watson METH	Watson COVERAGE	Watson QUAL	Crick METH	Crick COVERAGE	Crick QUAL
				CG OUTPUT FILE				
Seq	8	CG (2)	?	?	?	?	?	?
Seq	27	CK (5)	?	?	?	?	?	?
				CHG OUTPUT FILE				
Seq	8	CGG (2)	.	.	.	?	?	?
Seq	19	CWG (4)	?	?	?	?	?	?
				CHH OUTPUT FILE				
Seq	3	CAT (1)	?	?	?	.	.	.
Seq	15	TTG (3)	.	.	.	?	?	?
Seq	27	CKA (5)	?	?	?	.	.	.

Figure 1. SNVs effect on sequence contexts used for methylation profiling. First part of the figure (SEQUENCES) shows a reference example sequence and both chromosomes' Watson strands (Note that cytosines found on Watson strand are red colored and those found on Crick strand are green colored (Guanines on Watson)). Reference sequence contexts are numbered (1-6) and orange underlined, while the "real" methylation sequence contexts are blue underlined. The second part (INDIVIDUAL CONTEXTS) is a table with the reference sequence contexts (Reference), all the SNVs found on the example (SNVs) and the resulting "real" methylation sequence contexts (Allele 1 and Allele 2). The last part is a *MethylExtract* methylation output files example (*MethylExtract* OUTPUTS) for the upper sequence ("?" symbolizes columns with methylation data and "." columns without methylation data for this particular context).

MethylExtract simultaneously performs the methylation and SNVs calling from bisulfite-treated reads. Approximately, two thirds of SNPs (Single Nucleotide Polymorphisms) occur in a CpG context, what apart from masking the real methylation levels, results in different methylation sequence contexts than those on the reference. Figure 1 tries to cover some of the situations that *MethylExtract* could find on real datasets and how it will deal with them:

- Non-reference sequence context: The first context (1) presents a homozygous SNV on the reference guanine, which if it is not taken into account will result in an unmethylated Crick strand cytosine, when there is not even a CG context in the sample sequence (1 real context is CAT, so it appears on the *MethylExtract* CHH methylation output).
- Non-reference allele-specific sequence contexts: The situation gets more complicated, when instead of a homozygous SNV, the sequence context presents a heterozygous SNV (4 and 5). A heterozygous SNV leads to two different sequence contexts, one on each of the two homologous chromosomes. *MethylExtract* uses the ambiguous nucleotide code from the IUPAC to indicate these cases. If both sequence contexts belong to the same methylation context, they will only appear in one output file (4, both belongs to CHG output). However, when the sequence contexts belong to two different methylation contexts, the methylation level will appear twice, one on each output file (5, CKA appears in the CHH output and CK in the CG output; note that only the cytosine methylation level is included in CHH context).
- New methylation sequence context: SNVs could result in the appearance of a new methylation context (3). *MethylExtract* includes them in their corresponding output files. Note that the sequence shown in the output file comes from the Watson strand (TTG) and the methylation values given correspond to the guanine (cytosine on Crick).
- Lost methylation context: the opposite situation is the disappearance of a reference methylation context (6). This is the only case where *MethylExtract* will not report methylation values for this position, because there is no cytosine in the sample sequence.
- Overlapping methylation contexts: another situation that can be found even without a SNV, is the presence of two or more methylation contexts overlapping each other. In these situations, *MethylExtract* reports as many methylation values as different methylation contexts do exist, using ambiguous nucleotide code for each variable nucleotide. A special case is the context CCG or CGG (2), composed by a CG and a CHG context. As the CpG context usually has particular properties, *MethylExtract* will report methylation values for both cytosines in the CG context and just for one cytosine in the CHG context.