INSTITUTE OF BIOINFORMATICS - MEDICAL UNIVERSITY OF INNSBRUCK

meRanTK

Version 1.3.0

User manual

Dietmar Rieder

6/13/2022

Contents

1. Introduction

meRanTK is a versatile high performance toolkit for complete analysis of methylated RNA data.

The toolkit includes five multithreaded programs:

meRanT: bisulfite read aligner using a set of transcripts as reference (e.g. refSeq)
meRanG: bisulfite read aligner using the whole genome as reference
meRanCall: methylation caller for precise identification of m^{5}Cs in RNA-BSseq or Aza-IP
meRanCompare: compare multiple RNA bisulfite datasets to identify differentially methylated m^{5}Cs.
meRanAnnotate: annotation of m^{5}Cs from meRanCall result files.

Together they facilitate transcriptome wide identification of methylated cytosines on RNAs a single base pair resolution.
The aligners, meRanT and meRanG, are designed to work with either single- or paired end sequence reads from strand specific RNA-BSseq libraries. Input files may originate from any high throughput sequencing platform that produces standard FASTQ formatted sequence reads (e.g. Illumina, Ion Proton, Ion Torrent). The BAM or SAM output files serve as input files for the meRanCall methylation caller which aims to precisely identify the positions of methylated cytosines. In order to identify differentially methylated cytosines, methylation call files from multiple experiments can be compared using meRanCompare which also implements replicate handling.

meRanTK is freely available at http://icbi.at/meRanTK (released under GNU general public license). All three programs are written in the Perl programing language and run therefore on a wide variety of computing platforms.

1.1. Purpose of this document

This user manual aims to explain how to install and use meRanTK, what data to use, and how to interpret the output.

1.2. System requirements

meRanTK runs on any (UNIX/Linux) system that supports the Perl programming language version 5.10+. If you do not have Perl installed, please consult your OS documentation how to install it via the OS specific software package manager (e.g. yum, apt). You can also download and compile Perl from http://www.perl.org.

We also provide meRanTK as standalone executables that should run on most of recently released 64Bit Linux systems without the need of Perl and additional Perl module installation. If you decide/need to run meRanTK from source please see Installation instructions for details.

The following third party programs, dependent on which tools you decide to run, are required to be installed on your system:

Tool	Program	Tested versions	Download URL (pre-compiled binaries for 64bit Linux are included in meRanTK)
meRantT	Bowtie2	2.2.9 2.4.5	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
meRanGs	STAR	2.4.0k, 2.4.2a, 2.5.0c, 2.5.2a, 2.5.2b,	https://github.com/alexdobin/STAR/releases

		2.6.1.d, 2.7.10a	
meRanGh	HiSat2	2.0.4, 2.2.1	http://ccb.jhu.edu/software/hisat2/index.shtml

2. Installation

2.1. Downloading meRanTK

To download meRanTK, please visit http://icbi.at/software/meRanTK and click the “Download” tab. There, click on “download” next to the newest version of the package. This should download a ZIP file, containing all the files you need to install and run meRanTK.

2.2. Install meRanTK to run the tools as standalone Linux 64Bit executables (the easy way)

Once you have downloaded meRanTK extract the contents of the ZIP file in the system folder where you want to install meRanTK.

You should now be ready to run the meRan tools ☺

In case you do not want to use the provided versions of the required third party programs (STAR, bowtie2, hisat2, see also 1.2.), please make sure that these programs are installed on your system and can be found in your systems PATH ($PATH). If your system has these tools installed, you should either rename or delete the “./extutils” folder in the meRanTK main folder, this way the third party tools from your system will be used.

Note: *In order to be able to create m-bias plots (see manual) with meRanT/G you will need to install the* ***libgd2*** *on your system. If it is not installed you'll see an error message like the following:*

```
“Can't locate object method "new" via package "GD::Graph::lines" at script/meRanGh.pl line xxxx”
```

2.3. Install meRanTK to run the tools from source (the expert way)

If you need to run the meRanTK tools from the source code, you may need to install a recent version (> 5.10) of the Perl programming language. Please refer to your systems documentation to do so.

Once you have Perl installed (check by running “perl –v”) you may need to install some additional Perl modules:

Bio::DB::Sam
Parallel::ForkManager
GD
GD::Text::Align
GD::Graph::lines
Math::CDF
Text::NSP::Measures::2D::Fisher::twotailed
MCE::Loop (optional)

These modules should be available via CPAN or depending on your OS via the systems package manager (e.g. yum, apt).

On **yum based** systems (e.g. RedHat, CentOS, Fedora) you might need to run:

```
yum install perl-Parallel-ForkManager
yum install perl-GD
yum install perl-GDGraph
yum install perl-GDTextUtil
yum install samtools samtools-devel samtools-libs
```

```
yum install perl-MCE.noarch

cpan Bio::DB::Sam
cpan Math::CDF
cpan Text::NSP::Measures::2D::Fisher::twotailed
cpan MCE::Loop
```

On **apt based** systems (e.g. Debian, Ubuntu, Mint) you might need to run:

```
apt-get install libparallel-forkmanager-perl
apt-get install libbio-samtools-perl
apt-get install libgd-gd2-perl
apt-get install libgd-text-perl
apt-get install libgd-graph-perl
apt-get install libmce-perl

cpan Math::CDF
cpan Text::NSP::Measures::2D::Fisher::twotailed
```

If you want to install these modules via **CPAN** then you might need to run:

```
cpan Parallel::ForkManager
cpan Bio::DB::Sam
cpan GD
cpan GD::Text::Align
cpan GD::Graph::lines
cpan Math::CDF
cpan Text::NSP::Measures::2D::Fisher::twotailed
cpan MCE::Loop
```

Note: Bio::DB:Sam requires the samtools libraries (version 0.1.10 or higher, version 1.0 or higher is not compatible with Bio::DB::Sam, yet) and header files in order to compile successfully.

We provide also a **conda yml** file for creating a **conda environment** with all required software and perl packages, please run:

```
conda env create --name merantk –file=merantk.yml
mv extutil extutil_back
conda activate merantk
```

After you finished installing the required Perl modules, please copy “meRanGs.pl, meRanGh.pl, meRanT.pl, meRanCall.pl, meRanCompare.pl and meRanAnnotate.pl” from the “./src” directory to the main directory.

If you need to use your own installation of the third party tools, install the required programs (STAR, bowtie2, hisat2, see also 1.2.) and make sure that they can be found in your systems PATH ($PATH). Then, either rename or delete the “./extutils” folder in the meRanTK main folder, this way the third party tools from your system will be used.

You should now be ready to run the meRan tools.

3. Running meRanTK

3.1. meRanT – align RNA-BSseq reads to a set of reference transcripts

meRanT aligns directed/strand-specific RNA-BSseq reads to a reference transcriptome e.g. to fasta sequences from the NCBI refSeq database. To do so, meRanT first needs to bisulfite convert the reference database and generate the corresponding database index. This bisulfite conversion and index generation has only to be performed the first time one uses a specific reference, for all following runs that use the same reference transcriptome the bisulfite index can be reused.

3.1.1. meRanT – index generation

Let's assume one wants to align RNA-BSseq reads to a set of transcripts and the sequences of these transcripts are stored in a FASTA-formatted file named "mm.reSeqRNA.fa". To create the bisulfite index for this database use the following command:

```
meRanT mkbsidx -fa mm.refSeqRNA.fa -id /data/mm/BSrefSeqIDX
```

This will create the bisulfite index of the "mm.refSeqRNA.fa" file in the index directory "/data/mm/BSrefSeqIDX" specified by the "-id" option. The index name will be displayed after it is created (e.g. "/data/mm/BSrefSeqIDX/mm.refSeqRNA.C2T"). This index name can then be used in the "-x" option when aligning the reads (see below).

The example above assumes that the Bowtie2 index builder command "bowtie2-build" is found in the systems path "$PATH" or "bowtie2-build" from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, "bowtie2-build" can be specified using the command line option (-bwt2b).

Note: *apart from a single fasta file or a comma separated file list, you can also use an expression pattern to specifiy the genome fasta files: (?, *, [0-9], [a-z], {fa1,fa2,..faX})*

If using an expression pattern, please put single quotes around the "-fa" argument, e.g:

```
-fa '/genome/chrs/chr[1-8].fa'
```

Note: *depending on the size of the data and the computer used this can take a long time, please do not interrupt the index generation step, unless you really need to.*

You may download and combine the mouse refSeq data from:
ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.1.rna.fna.gz
ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.2.rna.fna.gz

3.1.2. meRanT – generate a transcript to gene name map file

The process used by meRanT for selecting the best alignment to a "canonical" transcript (i.e. longest mappable transcript) representing a gene requires a transcript to gene map file. This mapping file must be in the following tab delimited format:

```
#seqID          Genesymbol  sequencelength
[...]
NR_046233.2     Rn45s       13400
NR_003279.1     Rn28s1      4730
NR_003280.2     Rs5-8s1     157
NM_134105.2     Txndc11     1233
[...]
```

This way, each transcript in the transcript database (fasta) is mapped to a Genesymbol. The length of each transcript is stored in order to find the longest mapped sequence. Once a transcript to gene map file has been generated, it can be reused for any meRanT run that uses the same reference sequence database.

A Perl program (mkRefSeq2GeneMap.pl), that automatically generates such transcript to gene map files out of refSeq mRNA fasta files can be found in the "utils" directory of meRanTK. If you have a refSeq mRNA fasta file you can run the following command:

```
zcat mouse.1.rna.fna.gz mouse.2.rna.fna.gz > mm.refSeqRNA.fa
mkRefSeq2GeneMap.pl –f mm.refSeqRNA.fa –m mm.refSeqRNA2GeneName.map
```

The above command generates the "mm.refSeqRNA2GeneName.map" transcript to gene map file from the sequences in the "mm.refSeqRNA.fa" fasta file.

You may download the mouse refSeq data from:
ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.1.rna.fna.gz
ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.2.rna.fna.gz

3.1.3. meRanT – align single end RNA-BSseq reads

Once the tasks described above (3.1.1., 3.1.2.) have been performed, you are ready to align RNA-BSseq reads to the reference transcriptome.

Let's assume you have 3 fastq formatted sequence read files, 01.fastq, 02.fastq and 03.fastq, and you want to align them to a transcriptome database in the "mm.refSeqRNA.fa" file for which you have created the bisulfite index named "/data/mm10/BSrefSeqIDX/mm.refSeqRNA.C2T" (see 3.1.1.). You would then run the following command:

```
meRanT align \
 -o ./meRanTResult \
 -f ./FastqDir/01.fastq,./FastqDir/02.fastq,./FastqDir/03.fastq \
 -t 12 \
 -k 10 \
 -S RNA-BSseq.sam \
 -un \
 -ud ./meRanTunaligned \
 -ra \
 -MM \
 -i2g ./mm.refSeqRNA2GeneName.map \
 -x /data/mm/BSrefSeqIDX/mm.refSeqRNA.C2T \
 -mbp
```

The command above aligns the reads from the three fastq files, separated by commas, to the transcript sequences of the databases in "mm.refSeqRNA.fa", using the index created as indicated in 3.1.1. The process for selecting the best alignment to a transcript representing a gene uses the transcript to gene map file (-i2g mm.refSeqRNA2GeneName.map) created in 3.1.2.

The mapping process will use (-t) 12 CPUs and search for maximum (-k) 10 valid alignments, from which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and meRanT will generate the corresponding sorted BAM file. The program will save the unaligned reads (-un) in (-ud) the directory named "meRanTunaligned" and it will also report ambiguous alignments (-ra) in a separate tab delimited text file. The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the Bowtie2 aligner command “bowtie2” is found in the systems path “$PATH” or “bowtie2” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “bowtie2” can be specified using command line option “-bwt2”.

3.1.3. meRanT – align paired end RNA-BSseq reads

Let’s assume you have 4 fastq formatted sequence read files from 2 paired end sequencing runs, fwd01-paired.fastq, fwd02-paired.fastq, rev01-paired.fastq and rev02-paired.fastq, and you want to align them to a transcriptome database in the “mm.refSeqRNA.fa” file for which you have created the bisulfite index named “/data/mm/BSrefSeqIDX/mm.refSeqRNA.C2T” (see 3.1.1.). You would then run the following command:

```
meRanT align \
 -o ./meRanTResult \
 -f ./FastqDir/fwd01-paired.fastq,./FastqDir/fwd02-paired.fastq \
 -r ./FastqDir/rev01-paired.fastq,./FastqDir/rev02-paired.fastq \
 -t 12 \
 -k 10 \
 -S RNA-BSseq.sam \
 -un \
 -ud ./meRanTunaligned \
 -ra \
 -MM \
 -i2g ./mm.refSeqRNA2GeneName.map \
 -x /data/mm/BSrefSeqIDX/mm.refSeqRNA.C2T \
 -mbp
```

When using paired end reads, one can specify the forward- and reverse reads using the command line options "-f" and "-r" respectively. Multiple files for each read direction files can be specified separated by commas. Not only the sort order of the forward- and reverse reads has to be the same within the fastq files but also the order in which one specifies the forward and reverse read fastq files (see example above).

***Note**: The paired fastq files may not have unpaired reads. If this is the case, one can use for example the "pairfq" (S. Evan Staton) tool to pair and sort the mates.*

The command above aligns paired end reads from the 4 fastq files (2 forward- and 2 reverse read files), to the transcript sequences of the databases in "mm.refSeqRNA.fa", using the index created as indicated in 3.1.1. The process for selecting the best alignment to a transcript representing a gene uses the transcript to gene map file (-i2g mm.refSeqRNA2GeneName.map) created in 3.1.2.

The mapping process will use (-t) 12 CPUs and search for maximum (-k) 10 valid alignments, from which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and meRanT will generate the corresponding sorted BAM file. The program will save the unaligned reads (-un) in (-ud) the directory named "meRanTunaligned" and it will also report ambiguous alignments (-ra) in a separate tab delimited text file. The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the Bowtie2 aligner command “bowtie2” is found in the systems path “$PATH” or “bowtie2” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “bowtie2” can be specified using command line option “-bwt2”.

3.1.4. SAM output

meRanT generates the following SAM output fields:

Column	Field/TAG	Description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string (for fully converted read/reference alignment)
7	RNEXT	Ref. name of the mate/next
8	PNEXT	Position of the mate/next read
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33
>11	ZG	Gene name associated with the transcript in RNAME
>11	AS	Alignment score (<=0 in global mode, <=0>= in local mode)
>11	XS	Alignment score for the best-scoring alignment found other than the alignment reported
>11	YS	Alignment score for opposite mate in the paired-end alignment. Only presentif the SAM record is for a read that aligned as part of a paired-end alignment.
>11	XN	The number of ambiguous bases in the reference covering this alignment.
>11	XM	The number of mismatches in the alignment.
>11	XO	The number of gap opens, for both read and reference gaps, in the alignment.
>11	XG	The number of gap extensions, for both read and reference gaps, in the alignment.
>11	NM	The edit distance; that is, the minimal number of one-nucleotide edits (substitutions, insertions and deletions) needed to transform the read string into the reference string.
>11	YT	Value of \`CP\` indicates the read was part of a pair and the pair aligned concordantly. (There should be no other values, only concordantly aligned reads are reported)
>11	MD	A string representation of the mismatched reference bases in the alignment. See [SAM] format specification for details.

3.1.5. Ambiguous alignments report

When running meRanT with the “-ra” option a tabulator separated text file with information about the ambiguous alignments will be generated. It can have the following entries:

```
MG <tab> ReadID <tab> RNAME <tab> GeneName1 <tab> GeneName2
MP <tab> ReadID <tab> RNAME <tab> POS1 <tab> POS2 <tab> GeneName
```

“MG” (multi-gene) means that the read with “ReadID” aligns to a reference sequence “RNAME” which corresponds to a transcript of the gene “GeneName1”, however it also aligns to a different transcript of the gene “GeneName2”.

“MP” (multi-position) means that the read with “ReadID” aligns at position “POS1” to the reference sequence “RNAME” which corresponds to a transcript of the gene “GeneName”, however it also aligns to the postion “POS2” of the same reference.

3.2. meRanGs – align RNA-BSseq reads to the genome using STAR

meRanGs aligns directed/strand-specific RNA-BSseq reads to a reference genome (e.g. mm10, hg19). To do so, meRanGs first needs to bisulfite convert the reference database and generate the corresponding database index. This bisulfite conversion and index generation has only to be performed the first time one uses a specific reference, for all following runs that use the same reference transcriptome the bisulfite index can be reused.

***Note**: meRanGs requires a lot of memory to hold the reference database indices. For example the mouse genome mm10 bisulfite indices require about 48GB of RAM. In case you don't have a system with enough memory to hold the index + extra memory for the aligning process you should consider to use meRanGh, which only requires a moderate amount of memory.*

3.2.1. meRanGs – index generation

Let's assume one wants to align RNA-BSseq reads to a genome and sequences of its chromosomes are stored in a FASTA-formatted files named "mm10.chr[1..Y].fa". To create the corresponding bisulfite index use the following command:

```
meRanGs mkbsidx \
   -t 4 \
   -fa mm10.chr1.fa,mm10.chr2.fa,[...] \
   -id /data/mm10/BSgenomeIDX \
   -GTF /data/mm10/mm10.GFF3 \
   -GTFtagEPT Parent \
   -GTFtagEPG gene \
   -sjO 99
```

This will generate bisulfite index of a genome database provided as fasta (-fa) file(s) in the index directory "/data/mm10/BSgenomeIDX" specified by the "-id" option. The indexer will use at maximum (-t) 4 threads. A GFF3 (mm10.GFF3) file is used to specify the splice junctions. This example index will be optimized for 100 bp reads by specifying a splice junction overhang (-sjO) of 99 bps on each site. The tag name used as exons' parent transcript for building transcripts is 'Parent' (-GTFtagEPT). The tag name used as exons' parent gene for building transcripts is 'gene' (-GTFtagEPG). The index directory name will again be displayed after it is created (e.g. "/data/mm10/BSgenomeIDX"). This index directory name can then be used in the "-id" option when aligning the reads (see below).

The example above assumes that the STAR aligner command "STAR" is found in the systems path "$PATH" or "STAR" from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, "STAR" can be specified using command line option "-star".

***Note**: If a GFF3 or GTF file is specified, it is important that the sequence identifiers in the GFF of GTF match those from the fasta genome sequence files. (See also the STAR aligner documentation for more details)*

***Note**: If a GFF3 or GTF file is <u>not</u> specified, it can be optionally passed to the aligner on the fly. (See also the STAR aligner documentation for more details)*

***Note**: apart from a single fasta file or a comma separated file list, you can also use an expression pattern to specifiy the genome fasta files: (?, *, [0-9], [a-z], {fa1,fa2,..faX})*

If using an expression pattern, please put single quotes around the "-fa" argument, e.g:

```
-fa '/genome/chrs/chr[1-8].fa'
```

***Note**: depending on the size of the data and the computer used this can take a long time, please do not interrupt the index generation step, unless you really need to.*

3.2.2. meRanGs – align single end RNA-BSseq reads

Once the bisulfite index (see 3.2.1.) has been created, you are ready to align RNA-BSseq reads to the genome.

Let’s assume you have 3 fastq formatted sequence read files, 01.fastq, 02.fastq and 03.fastq, and you want to align them to the genome sequence in the “mm10.[chr1..chrY].fa” files for which you have created the bisulfite index in the directory “/data/mm10/BSgenomeIDX/” (see 3.2.1.). You would then run the following command:

```
meRanGs align \
 -o ./meRanGsResult \
 -f ./FastqDir/01.fastq,./FastqDir/02.fastq,./FastqDir/03.fastq \
 -t 12 \
 -S RNA-BSseq.sam \
 -un \
 -ud ./meRanGsUnaligned \
 -MM \
 -star_outFilterMultimapNmax 20 \
 -id /data/mm10/BSgenomeIDX \
 -bg \
 -mbgc 10 \
 -mbp \
```

The command above maps the reads from the three fastq files, separated by commas, to a genome using the index created as indicated in the section 3.2.1.

The read mapping process will use (-t) 12 CPUs and search for a maximum of 20 valid alignments (-star_outFilterMultimapNmax), of which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and a sorted and indexed BAM file will be created. meRanGs will save the unaligned reads (-un) in the directory (-ud) "meRanGUnaligned". The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. The “-id” option specifies the bisulfite index that will be used. It should be generated as described in 3.2.1 (make sure that the –sjO setting in the index generation process is optimized for your read length [i.e. readLength-1]).
In addition to the SAM/BAM files a bedGraph file (-bg) that reports the read coverage across the entire genome. The coverage will only be reported for genomic positions that are covered by more than 10 reads (-mbgc 10). Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the STAR aligner command “STAR” is found in the systems path “$PATH” or “STAR” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “STAR” can be specified using command line option “-star”.

***Note**: If a GFF3 or GTF file was not specified during index generation, it can be optionally passed to the aligner on the fly by using the “-GTF” option. Moreover, you can specify an additional GTF/GFF3 file to the one that was used during index generation and information of both will be used. (See also the STAR aligner documentation for more details)*

3.2.3. meRanGs – align paired end RNA-BSseq reads

Let’s assume you have 4 fastq formatted sequence read files from 2 paired end sequencing runs, fwd01-paired.fastq, fwd02-paired.fastq, rev01-paired.fastq and rev02-paired.fastq, and you want to align them to the genome sequence in the “mm10.[chr1..chrY].fa” files for which you have created the bisulfite index in the directory “/data/mm10/BSgenomeIDX/” (see 3.2.1.). You would then run the following command:

```
meRanGs align \
 -o ./meRanGsResult \
 -f ./FastqDir/fwd01-paired.fastq,./FastqDir/fwd02-paired.fastq \
 -r ./FastqDir/rev01-paired.fastq,./FastqDir/rev02-paired.fastq \
 -t 12 \
 -S RNA-BSseq.sam \
 -un \
 -ud ./meRanGsUnaligned \
 -MM \
 -star_outFilterMultimapNmax 20 \
 -id /data/mm10/BSgenomeIDX \
 -bg \
 -mbgc 10 \
 -mbp \
```

When using paired end reads, one can specify the forward and reverse reads using the command line options "-f" and "-r" respectively. Multiple files for each read direction files can be specified separated by commas. Not only the sort order of the forward- and reverse reads has to be the same within the fastq files but also the order in which one specifies the forward- and reverse read fastq files (see example above).

***Note**: The paired fastq files may not have unpaired reads. If this is the case, one can use for example the "pairfq" (S. Evan Staton) tool to pair and sort the mates.*

The read mapping process will use (-t) 12 CPUs and search for a maximum of 20 valid alignments (-star_outFilterMultimapNmax), of which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and a sorted and indexed BAM file will be created. meRanGs will save the unaligned reads (-un) in the directory (-ud) "meRanGUnaligned". The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. The “-id” option specifies the bisulfite index that will be used. It should be generated as described in 3.2.1 (make sure that the –sjO setting in the index generation process is optimized for your read length [i.e. readLength-1]).
In addition to the SAM/BAM files a bedGraph file (-bg) that reports the read coverage across the entire genome. The coverage will only be reported for genomic positions that are covered by more than 10 reads (-mbgc 10). Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the STAR aligner command “STAR” is found in the systems path “$PATH” or “STAR” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “STAR” can be specified using command line option “-star”.

***Note**: If a GFF3 or GTF file was <u>not</u> specified during index generation, it can be optionally passed to the aligner on the fly by using the “-GTF” option. Moreover, you can specify an additional GTF/GFF3 file to the one that was used during index generation and information of both will be used. (See also the STAR aligner documentation for more details)*

3.2.4. SAM output

meRanGs generates the following SAM output fields:

Column	Field/TAG	Description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string (for fully converted read/reference alignment)
7	RNEXT	Ref. name of the mate/next
8	PNEXT	Position of the mate/next read
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33
>11	AS	Alignment score
>11	NM	The edit distance; that is, the minimal number of one-nucleotide edits (substitutions, insertions and deletions) needed to transform the read string into the reference string.
>11	NH	Number of reported alignments that contains the query in the current record.
>11	HI	Query hit index, indicating the alignment record is the i-th one stored in SAM.
>11	MD	A string representation of the mismatched reference bases in the alignment. See [SAM] format specification for details.
>11	YG	Bisulfite genome conversion. Can either be CT or GA, for C to T and G to A conversion respectively
>11	YR	Bisulfite read conversion. Can either be CT or GA, for C to T and G to A conversion respectively

3.3. meRanGh – align RNA-BSseq reads to the genome using HiSat2

meRanGh aligns directed/strand-specific RNA-BSseq reads to a reference genome (e.g. mm10, hg19). To do so, meRanGh first needs to bisulfite convert the reference database and generate the corresponding database index. This bisulfite conversion and index generation has only to be performed the first time one uses a specific reference, for all following runs that use the same reference transcriptome the bisulfite index can be reused.

***Note**: In contrast to meRanGs, meRanGh requires only a moderate amount of memory (~10GB vs ~50GB).*

3.3.1. meRanGh – index generation

Let’s assume one wants to align RNA-BSseq reads to a genome and sequences of its chromosomes are stored in a FASTA-formatted files named “mm10.chr[1..Y].fa”. To create the corresponding bisulfite index use the following command:

```
meRanGh mkbsidx \
   -t 4 \
   -fa mm10.chr1.fa,mm10.chr2.fa,[...] \
   -id /data/mm10/BSgenomeIDX
```

This will generate bisulfite index of a genome database provided as fasta (-fa) file(s) in the index directory “/data/mm10/BSgenomeIDX” specified by the “-id” option. The indexer will use at maximum (-t) 4 threads. The index directory name will again be displayed after it is created (e.g. “/data/mm10/BSgenomeIDX”). This index directory name can then be used in the “-id” option when aligning the reads (see below).

The example above assumes that the HiSat2 index builder “hisat2-build” command are found in the systems path “$PATH” or “hisat2-build” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “hisat2-build” can be specified using the command line option (-hs2b).

***Note**: apart from a single fasta file or a comma separated file list, you can also use an expression pattern to specifiy the genome fasta files: (?, *, [0-9], [a-z], {fa1,fa2,..faX})*

If using an expression pattern please put single quotes around the “-fa” argument, e.g:

```
-fa '/genome/chrs/chr[1-8].fa'
```

***Note**: depending on the size of the data and the computer used this can take a long time, please do not interrupt the index generation step, unless you really need to.*

3.3.2. meRanGh – align single end RNA-BSseq reads

Once the bisulfite index (see 3.3.1.) has been created, you are ready to align RNA-BSseq reads to the genome.

Let’s assume you have 3 fastq formatted sequence read files, 01.fastq, 02.fastq and 03.fastq, and you want to align them to the genome sequence in the “mm10.[chr1..chrY].fa” files for which you have created the bisulfite index in the directory “/data/mm10/BSgenomeIDX/” (see 3.3.1.). You would then run the following command:

```
meRanGh align \
 -o ./meRanGhResult \
 -f ./FastqDir/01.fastq,./FastqDir/02.fastq,./FastqDir/03.fastq \
 -t 12 \
 -S RNA-BSseq.sam \
 -ud ./meRanGhUnaligned \
 -un \
 -MM \
 -id /data/mm10/BSgenomeIDX \
 -GTF /data/mm10/mm10-known-transcripts.gtf \
 -bg \
 -mbgc 10 \
 -mbp \
```

The command above maps the reads from the three fastq files, separated by commas, to a genome using the index created as indicated in the section 3.3.1. The “-GTF” option indicates that the program should also search for alignments on known splicesites contained in the specified GTF file.

The read mapping process will use (-t) 12 CPUs and search for valid alignments, of which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and a sorted and indexed BAM file will be created. meRanGh will save the unaligned reads (-un) in the directory (-ud) "meRanGhUnaligned". The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. The

“-id” option specifies the bisulfite index that will be used. It should be generated as described in 3.3.1. In addition to the SAM/BAM files a bedGraph file (-bg) that reports the read coverage across the entire genome. The coverage will only be reported for genomic positions that are covered by more than 10 reads (-mbgc 10). Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the HiSat2 aligner “hisat2” command is found in the systems path “$PATH” or “hisat2” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “hisat2” can be specified using the command line options “-hisat2”.

3.3.3. meRanGh – align paired end RNA-BSseq reads

Let’s assume you have 4 fastq formatted sequence read files from 2 paired end sequencing runs, fwd01-paired.fastq, fwd02-paired.fastq, rev01-paired.fastq and rev02-paired.fastq, and you want to align them to the genome sequence in the “mm10.[chr1..chrY].fa” files for which you have created the bisulfite index in the directory “/data/mm10/BSgenomeIDX/” (see 3.3.1.). You would then run the following command:

```
meRanGh align \
 -o ./meRanGhResult \
 -f ./FastqDir/fwd01-paired.fastq,./FastqDir/fwd02-paired.fastq \
 -r ./FastqDir/rev01-paired.fastq,./FastqDir/rev02-paired.fastq \
 -t 12 \
 -S RNA-BSseq.sam \
 -un \
 -ud ./meRanGhUnaligned \
 -MM \
 -id /data/mm10/BSgenomeIDX \
 -GTF /data/mm10/mm10-known-transcripts.gtf \
 -bg \
 -mbgc 10 \
 -mbp \
```

When using paired end reads, one can specify the forward and reverse reads using the command line options "-f" and "-r" respectively. Multiple files for each read direction files can be specified separated by commas. Not only the sort order of the forward- and reverse reads has to be the same within the fastq files but also the order in which one specifies the forward- and reverse read fastq files (see example above).

Note*: The paired fastq files may not have unpaired reads. If this is the case, one can use for example the "pairfq" (S. Evan Staton) tool to pair and sort the mates.*

The “-GTF” option indicates that the program should also search for alignments on known splicesites contained in the specified GTF file.

The read mapping process will use (-t) 12 CPUs and search for valid alignments, of which the best one will be stored in the (-S) "RNA-BSseq.sam" result file and a sorted and indexed BAM file will be created. meRanGh will save the unaligned reads (-un) in the directory (-ud) "meRanGhUnaligned". The alignments of multi mapping reads (-MM) will additionally be stored in a separate SAM file. The “-id” option specifies the bisulfite index that will be used. It should be generated as described in 3.3.1. In addition to the SAM/BAM files a bedGraph file (-bg) that reports the read coverage across the entire genome. The coverage will only be reported for genomic positions that are covered by more than 10 reads (-mbgc 10). Finally, an m-Bias plot will be generated (-mbp) which may help to detect potential read positional “methylation” biases, that could rise because of sequencing or library problems.

The example above assumes that the HiSat2 aligner “hisat2” command is found in the systems path “$PATH” or “hisat2” from the meRanTK shipped third party programs is used (see Installation 2.2, 2.3). Alternatively, “hisat2” can be specified using the command line options “-hisat2”.

3.3.4. SAM output

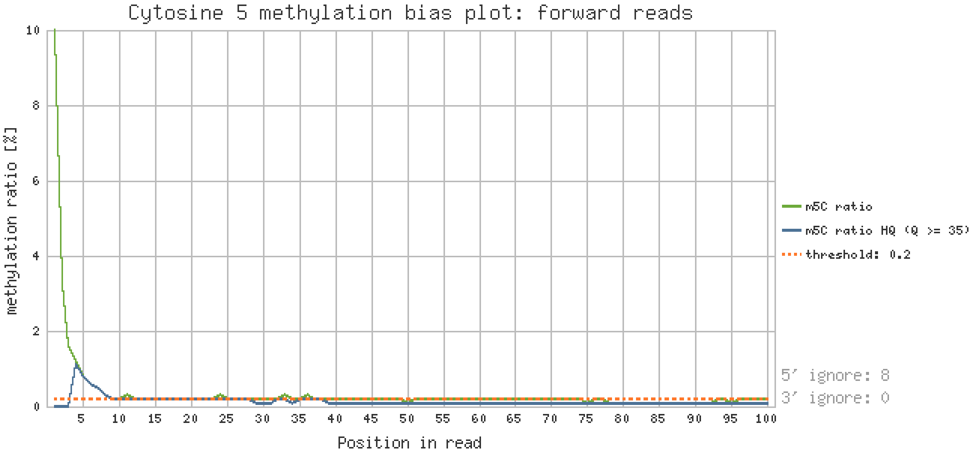
meRanGh generates the following SAM output fields:

Column	Field/TAG	Description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string (for fully converted read/reference alignment)
7	RNEXT	Ref. name of the mate/next
8	PNEXT	Position of the mate/next read
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33
>11	AS	Alignment score
>11	NM	The edit distance; that is, the minimal number of one-nucleotide edits (substitutions, insertions and deletions) needed to transform the read string into the reference string.
>11	NH	Number of reported alignments that contains the query in the current record.
>11	HI	Query hit index, indicating the alignment record is the i-th one stored in SAM.
>11	MD	A string representation of the mismatched reference bases in the alignment. See [SAM] format specification for details.
>11	YG	Bisulfite genome conversion. Can either be CT or GA, for C to T and G to A conversion respectively
>11	YR	Bisulfite read conversion. Can either be CT or GA, for C to T and G to A conversion respectively

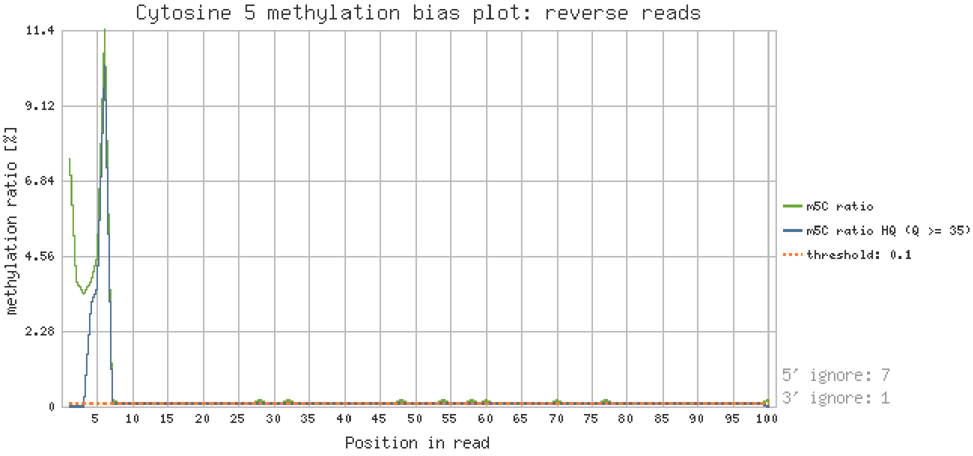
3.4. M-Bias plots

meRanT, meRanGs and meRanGh can all produce m-bias plots which may help detecting potential sequencing or library problems. For each position of each final uniquely aligned read the occurrence of non-converted cytosines is recorded and the fraction N[C]/N[ATG] is plotted as function of the position within the read. In an unbiased dataset this plot should present a flat horizontal line since cytosine methylation is expected to occur independently of the read position.

Example m-bias plot for 5' C-biased forward reads



Example m-bias plot for 5' C-biased reverse reads



In the above shown examples the RNA-BSseq reads had a cytosine bias at the 5' ends which was due to random hexamers used in the library preparation. The green and turquoise graphs in the plots show the C/[ATG] ratio of all uniquely mapped reads at each position within the reads. For the turquoise graph however, only base calls with a Q-value equal or higher than 35 were considered.
The 5' and 3' ignore values are suggesting the number of bases at either end that may be ignored during methylation calling. These numbers correspond to the read positions at which the m5C ratio graph first is under the threshold, which is calculated as the "median + 2 * the median absolute deviation".

3.5. meRanCall – call methylated cytosines (m^{5}C) from the RNA-BSseq alingments

meRanCall is a flexible tool capable of multiprocessing that aims to extract the methylation state of individual cytosines from mapping results produced by meRanT or meRanG (SAM or BAM files, where SAM files will automatically be converted into sorted BAM files).
Methylated cytosines are called based on user supplied thresholds, such as minimum read coverage, minimum non-conversion rate and minimum base quality. Potential PCR duplicates may be filtered by defining a maximum allowed number of identical reads (same- start coordinate, SAM flag and CIGAR string) and potential biased read ends, as determined through the inspection of the M-bias plots, may be excluded from the analysis. meRanCall uses a C-cutoff threshold to filter out reads with unconverted C > C_cutoff, which helps avoiding false positives due to incomplete conversion. It further uses a signal to noise filter that ensures that the fraction of high quality C/T calls (#filter passed / #total calls) at a C position > SNR. If control sequences are included in the data set, meRanCall can determine the overall C→T conversion rate of an experiment, which can then be used for calculating the p-value of the methylation state (Lister et al., 2009) and the p-value of the methylation rate (Barturen et al., 2013) for each "methylated" cytosine. meRanCall can also calculate a gene level C→T conversion rate based on the genes from a GTF gene annotation file, these individual conversion rates can be used as alternative to the global conversion rate in p-value calculations. Besides these p-values meRanCall calculates coverage, C count, methylation rate, 95% confidence intervals and mutation rate. In addition to these metrics, meRanCall reports information about the position, strand, reference base and the sequence context around the methylated cytosine. For alignments obtained from meRanT, associated gene names are reported along with the methylation state data. All data are stored in a simple tab delimited file ready for further analysis. When analyzing SAM/BAM files from meRanG, a BED6 + 3 or narrow peak BED file may be generated that can be used to project the methylation data on a genome browser display.

3.5.1. Determination of the C→T conversion rate of a RNA-BSseq sample

In case you have an un-methylated control sequence spiked into your RNA-BSseq sample you can uses this sequence to determine the C→T conversion rate, which serves later on in the methylation calling process to calculate p-values.

Let's assume you have a spike in control sequence named "UnMethylated_Control", you should first add this sequence to your fasta formatted reference sequence database and then create a bisulfite index using meRanGs, meRanGh or meRanT. When you then align your RNA-BSseq reads using one of the meRanTK aligners, reads that come from your control sequence will also produce alignments entries in the resulting BAM file. meRanCall can extract those and use them to calculate the C→T conversion rate. To do so you would run the following command:

```
meRanCall \
-p 32 \
-fs5 6 \
-fs3 0 \
-rs5 0 \
-rs3 0 \
-s RNA-BSseq_sorted.bam \
-f ./mm10.refSeqRNA.fa \
-rl 100 \
-ccr \
-tref \
-c SeqID UnMethylated_Control
```

The above command then calculates the conversion rate (-ccr) using the specified control sequence identifier (-c, can also be specified multiple times if you have more than one control sequence). In this analysis 6 bases at the 5’ end of the forward reads (-fs5 6) will be ignored (none at the 3’ end of the forward reads [–fs3 0] and non at both end of the reverse reads [–rs5 0, -rs3 0]). The reference sequence file in fasta format is specified by setting –f to “./mm10.refSeqRNA.fa”. We tell meRanCall that the pre-trimmed raw sequence read length was 100 bps.

3.5.2. methylation calling from RNA-BSseq single end reads mapped with meRanT

Let’s assume you have mapped single end RNA-BSseq reads to a transcriptome database in the “mm10.refSeqRNA.fa” fasta file using meRanT (see 3.1.3.). You would then run the following command to call m^{5}Cs from the aligned reads contained in “RNA-BSseq_sorted.bam”:

```
meRanCall \
 -p 32 \
 -o ./meRanCallResult.txt \
 -bam ./RNA-BSseq_sorted.bam \
 -f ./mm10.refSeqRNA.fa \
 -fs5 6 \
 -rl 100 \
 -sc 10 \
 -zg \
 -md 5 \
 -ei 0.1 \
 -cr 0.99 \
 -fdr 0.01 \
 -tref
```

The command above calls methylated C's from reads in "RNA-BSseq_sorted.bam" mapped to the reference transcriptome database (transcripts) in “mm10.refSeqRNA.fa”. The methylation calling process will use (-p) 32 CPUs in parallel.
The mapping was created using meRanT, therefore the "-zg" option is added in the example. This way, the gene names associated with the individual transcripts will be extracted from the BAM file and reported in the methylation calling result file. Since the type of reference for the alignment was a transcript database the "-tref" option has to be used. By specifying “-md 5” we allow for a maximum of 5 potential PCR duplicates (=same- start coordinate, SAM flag and CIGAR string).
Let’s assume that the reads are C biased at the first 6 base positions on the 5' end (this could be estimated from the m-Bias plot produced by meRanT). We tell the meRanCall program to ignore these biased positions by specifying the option "-fs5 6" (forward read skip on 5’ end 6 bases). The original reads had a length of 100 base pairs before any trimming, we tell this by setting “–rl 100”, this way meRanCall ignores only up to 6 bases from the 5’ end of reads that were longer than 93 bps after read trimming (that you did in your QC before aligning the reads).
We want also to get the sequence context 10 bps around the methylated C's (-sc 10).
We set the error interval (-ei 0.1) for calculating the methylation level p-value to 0.1, that means that we calculate the probability that the real methylation level lies within that interval (Barturen et al., 2013). Our C→T conversion rate (-cr 0.99) is 0.99 as we determined from a un-methylated in-vitro transcribed control RNA that was spiked into our sample (see 3.4.1). The false discovery rate of methylated cytosines will be controlled at the specified FDR (-fdr 0.01).

The result (-o) meRanCallResult.txt is a tab separated text file and contains the following data-fields for each potentially methylated C:

1. SeqID : sequence ID from reference database

2. refPos : postion of the methylated C on the reference sequence
3. refStrand : strand (will always be '+' when using a reference transcriptome)
4. refBase : base on the reference sequence
5. cov : coverage (# of reads covering this position)
6. C_count : # of C's counted at this position
7. methRate : methylation rate
8. mut_count : # of non-reference bases at the position
9. mutRate : mutation rate (#non reference bases / coverage)
10. SNR : local signal to noise ratio
11. CalledBase : prevailing base(s) at the position
12. CB_count : CalledBase count
13. state: methylation status (M|MV|UV|V)
    M : methylated C, C on reference
    MV: methylated C, NO C on reference (mutated)
    UV: unmethylated C, NO C on reference (mutated)
    V : mutated base
14. 95_CI_lower : lower bound of the 95% confidence interval (Wilson score interval)
15. 95_CI_upper : upper bound of the 95% confidence interval (Wilson score interval)
16. p-value_mState : p-value of the methylation State (Lister et al. 2009)
17. p-value_mRate : p-value of the methylation Rate (Barturen et al. 2013)
18. Score : methylation call score
19. seqContext : sequence Context arround the mehtylated C
20. geneName : gene name associated with the methylated C
21. candidateName : name assigned to the methylated C candidate

***Note**: The methylation calling process greatly benefits of parallel processing. It nearly scales up linearly and so using twice as many CPUs reduces the runtime to half.*

3.5.3. methylation calling from RNA-BSseq paired end reads mapped with meRanT

For paired end reads a command with analogous options as for single ends can be used. In addition, if you have 3' or 5' "C" biased reverse read ends that you want to be ignored by meRanCall, you can specify this for the reverse reads using the "-rsikp5" and/or "-rsikp3" option.

3.5.4. methylation calling from RNA-BSseq single end reads mapped with meRanGs/meRanGh

Let's assume you have mapped single end RNA-BSseq reads to a genome database in the "mm10.refSeqRNA.fa" fasta file using meRanT (see 3.2.2. or 3.3.2). You would then run the following command to call m^5Cs from the aligned reads contained in "RNA-BSseq_sorted.bam":

```
meRanCall \
 -p 32 \
 -o ./meRanCallResult.txt \
 -bam ./RNA-BSseq_sorted.bam \
 -f ./mm10.allchr.fa \
 -fs5 6 \
 -rl 100 \
 -sc 10 \
 -md 5 \
 -ei 0.1 \
 -cr 0.99 \
 -fdr 0.01 \
 -bed63 \
 -np \
 -gref
```

The command above calls methylated C's from reads in "RNA-BSseq_sorted.bam" mapped to the reference genome database (transcripts) in “mm10.allchr.fa”. The methylation calling process will use (-p) 32 CPUs in parallel.
The mapping was created using meRanGs or meRanGh, therefore the the "-tref" option has to be used. By specifying “-md 5” we allow for a maximum of 5 potential PCR duplicates (=same- start coordinate, SAM flag and CIGAR string).
Let’s assume that the reads are C biased at the first 6 base positions on the 5' end (this could be estimated from the m-Bias plot produced by meRanT). We tell the meRanCall program to ignore these biased positions by specifying the option "-fs5 6" (forward read skip on 5’ end 6 bases). The original reads had a length of 100 base pairs before any trimming, we tell this by setting “–rl 100”, this way meRanCall ignores only up to 6 bases from the 5’ end of reads that were longer than 93 bps after read trimming (that you did in your QC before aligning the reads).
We want also to get the sequence context 10 bps around the methylated C's (-sc 10).
We set the error interval (-ei 0.1) for calculating the methylation level p-value to 0.1, that means that we calculate the probability that the real methylation level lies within that interval (Barturen et al., 2013). Our C→T conversion rate (-cr 0.99) is 0.99 as we determined from a un-methylated in-vitro transcribed control RNA that was spiked into our sample (see 3.4.1). The false discovery rate of methylated cytosines will be controlled at the specified FDR (-fdr 0.01).
The options “-bed63” and “-np” tell meRanCall to generate a BED6 + 3 and a narrow peak BED file that can be used to project the methylation data on a genome browser display.

***Note**: The reference sequence in the fasta formatted file specified via the “-f” option has to contain all chromosome sequences that were uses in the aligning process with meRanGs or meRanGh. If you have each chromosomes sequence in a separate file, please combine these files into one single file.*

***Note**: The methylation calling process greatly benefits of parallel processing. It nearly scales up linearly and so using twice as many CPUs reduces the runtime to half.*

3.5.5. methylation calling from RNA-BSseq paired end reads mapped with meRanGs/meRantGt

For paired end reads a command with analogous options as for single ends can be used. In addition, if you have 3’ or 5’ “C” biased reverse read ends that you want to be ignored by meRanCall, you can specify this for the reverse reads using the “-rsikp5” and/or “-rsikp3” option.

3.5.6. methylation calling over specific regions

If one is only interested in methylation calls for specific regions, one can use the "-region" option and supply a BED file with the regions of interest. meRanCall will then only call methylated C's in these regions.

3.5.7. methylation calling from Aza-IP data sets.

If one needs to analyze Aza-IP data sets, it is recommended to map the sequencing reads using the STAR short read aligner allowing for 10% mismatches. All reads that are not mapped with STAR can then be mapped in a second step using Bowtie2 with the “—very-sensitive-local” mode. The resulting BAM files may be merged with samtools and candidate methylate cytosines may be called using meRanCall in its Aza-IP mode (-aza) an set the conversion rate cutoff to 4% (-mr 0.04). It is important that the same is done with the control data set. After candidate cytosines were called from both alignment files, one needs to run meRanCompare in the Aza-IP mode in order to select for candidates that are enriched (and have a significantly different proportion of C to G conversions) over control in the IP dataset.

3.6. meRanCompare – compare methylated cytosines (m^5C) from different experiments

meRanCompare is a tool designed to identify differentially methylated cytosines in different data sets (conditions). It uses result files from meRanCall and statistically evaluates the individual candidate methylated cytosines. It reports candidates that are unique to either one of the data sets and those that are present, but significant differentially methylated, in both conditions according to user defined thresholds i.e. p-value, fdr, methylation rate fold change.
meRanCall works with experiments that have single or multiple replicates and used a Fisher's exact or a Cochran-Mantel-Haenszel test to assess significant differences between the two conditions. In its Aza-IP mode one can use meRanCall result files from IP and Control and find enriched and statistically different methylated cytosines in the IP sample(s). The user may specify a minimum number of replicates in which an individual m^5C has to be called in order to be analyzed and seen as true call.
For comparing read counts and assessing enrichment in the Aza-IP mode, for each individual position it is important to normalize these counts according to the library size, therefore meRanCompare can take a size factor argument (-size-factors-a, -size-factors-b) for normalizing counts. meRanTK provides a helper tool (estimateSizeFactors) to calculate these library size factors. These calculations are similar to those used in DESeq2.

3.6.1. Comparing two conditions using RNA-BSseq data

Let's assume you have RNA-BSseq data from an experiment with two conditions and 3 replicates for each condition:

```
meRanCompare.pl \
 -fa condArep1_bscall.txt,condArep2_bscall.txt,condArep3_bscall.txt \
 -fb condBrep1_bscall.txt,condBrep2_bscall.txt,condBrep3_bscall.txt \
 -na wildtype \
 -nb knockout \
 -sfa 0.6673,0.6609,0.7347 \
 -sfb 0.9559,1.4098,2.3802 \
 -sig 0.01 \
 -fdr 0.02 \
 -mr 2
```

The command above identifies differentially methylated cytosines from 2 conditions (A,B: wildtype, knockout) with 3 replicates each. It normalizes each condition and replicate by the indicated library size factors and reports only candidates that are either unique to one condition or significantly different ($p < 0.01$ with FDR 0.02) between the two conditions while being present in at least two of the corresponding replicates.

3.6.2. Identify enriched methylated cytosines from Aza-IP data

Let's assume you have Aza-IP data from an experiment with two IP replicates and one control:

```
meRanCompare \
 -fa IPrep1_bscall.txt,IPrep2_bscall.txt \
 -fb CTRL \
 -na Aza-IP \
 -nb Control \
 -sfa 0.6934,0.7937 \
 -sfb 1.5983 \
 -sig 0.01 \
 -fdr 0.02 \
 -mr 2 \
 -fc 3 \
 -aza
```

The command above identifies enriched methylated cytosines from Aza-IP data with 2 replicates and one control. It normalizes each condition and replicate by the indicated library size factors and reports only candidates that are either unique to the IP samples or enriched and significantly different ($p < 0.01$ with FDR 0.02) between the IP and control, while being present in at least two of the corresponding IP replicates.

3.7. meRanAnnotate – annotate cytosines (m^5C)

meRanAnnotate to annotate methylated cytosines from meRanCall result files. It can use either ensembl GTF or NCBI GFF3 files to annotate m^5Cs using selected features like 'mRNA', 'gene', 'ncRNA' and so on. It can also calculate position metrics like distances of the individual m^5Ca to the 5' or 3' end, by respecting the strand information.

```
meRanAnnotate \
 -p 8 \
 -b m5C_bscall.txt \
 -f 'tRNA|rRNA|ncRNA|gene' \
 -g /data/mm10/annotations/refSeq.gff3 \
 -o m5C_bscall_Annotated.txt
```

3.8. Command line options

3.8.1. Command line options for meRanT

```
USAGE: meRanT <runmode> [-h] [-man] [--version]

Required <runmode> any of:
    mkbsidx     :   Generate the Bowtie2 BS index.
    align       :   Align bs reads to transcripts.

Options:
    --version   :   Print the program version and exit.
    -h|help     :   Print the program help information.
    -man        :   Print a detailed documentation.
```

mkbsidx mode:

```
USAGE: meRanT mkbsidx [-fa] [-id] [-h] [-man]

Required all of :
    -fa|fasta           : Fasta file to use for BS index generation.
    -id|bsidxdir        : Directory where to store the BS index.

Options:
    -bwt2b|bowtie2build : Path to Bowtie2 indexer "bowtie2-build".
                                  (default:  bowtie2_build  from  the  meRanTK  installation  or
                          your system PATH)
    -t|threads          : number of CPUs/threads to run
    --version           : Print the program version and exit.
    -h|help             : Print the program help information.
    -man                : Print a detailed documentation.
```

align mode:

```
USAGE: meRanT align [-f|-r] [-x] [-i2g] [-h] [-man]

Required all of:
   -fastqF|-f           : Fastq file with forward reads (required if no -r)
                        This file must contain the reads that align to the
                        5' end of the RNA, which is the left-most end of the
                        sequenced fragment (in transcript coordinates).
   -fastqR|-r           : Fastq file with reverse reads (required if no -f)
                        This file must contain the reads that align to the
                        opposite strand on the 3' end of the RNA, which is
                        the right-most end of the sequenced fragment (in
                        transcript coordinates).
    -id2gene|-i2g       : Transcript to gene mapping file.
                        This mapping file must in in the following tab
                        delimited format:

                        #seqID  Genesymbol  sequencelength

Options:
    -illuminaQC|-iqc    : Filter reads that did not pass the Illumina QC.
                        Only relevant if you have Illumina 1.8+ reads.
                        (default: not set)

    -forceDir|-fDir     : Filter reads that did not pass did not pass the
                        internal directionality check:
                            FWDreads: #C > #G && #C > #T && #A > #G)
                            REVreads: #G > #C && #T > #C && #G > #A)
                        (default: not set)
```

```
-first|-fn           : Process only this many reads/pairs
                    (default: process all reads/pairs)

-outdir|-o             : Directory where results get stored
                    (default: current directory)

-sam|-S                 : Name of the SAM file for uniq and resolved alignments
                    (default: meRanT_[timestamp].sam )

-unalDir|-ud             : Directory where unaligned reads get stored
                    (default: outdir)

                    Note: if -bowtie2un|-un is not set, unaligned reads
                    will not get stored

-threads|-t             : Use max. this many CPUs to process data
                    (default: 1)

-bowtie2cmd|-bwt2     : Path to bowtie2
                    (default: bowtie2 from the meRanTK installation or your
                     system PATH)

-bsidx|-x              : Name of bsindex created in mkbsidx runMode
                    (default: use BS_BWT2IDX environment variable)

-samMM|-MM                : Save multimappers? If set multimappers will be stored
                    in SAM format '$sam_multimappers.sam'
                    (default: not set)

-ommitBAM|-ob         : Do not create an sorted and indexed BAM file
                    (default: not set)

-deleteSAM|-ds        : Delete the SAM files after conversion to BAM format
                    (default: not set)

-reportAM|-ra         : Report ambiguos mappings? If set ambiguos mappings
                    will be stored in '$unalDir/$sam_ambiguos.txt'
                    (default: not set)

-bowtie2mode|-m       : Alignment mode. Can either be 'local' or 'end-to-end'
                    See Bowtie2 documentation for more information.
                    (default: end-to-end)

-max-edit-dist|-e     : Maximum edit distance to allow for a valid alignment
                    (default: 2)

-max-mm-rate|-mmr     : Maximum mismatch ratio (mismatches over read length)
                    [0 <= mmr < 1]
                    (default: 0.05)

-mbiasplot|-mbp       : Create an m-bias plot, that shows potentially biased
                    read positions
                    (default: not set)

-mbiasQS|-mbQS        : Quality score for a high quality m-bias plot. This plot
                    considers only basecalls with a quality score equal or
                    higher than specified by this option.
                    (default: 30)

-fixMateOverlap|-fmo  : The sequenced fragment and read lengths might be such
                    that alignments for the two mates from a pair overlap
                    each other.
                    If '-fmo' is set, deduplicate alignment subregions that
                    are covered by both, forward and reverse, reads of the
                    same read pair. Only relevant for paired end reads.
```

```
                        (default: not set)

    -hardClipMO|-hcmo     : If '-fmo' is set, hardclip instead of softclip the
                        overlapping sequence parts.
                        (default: not set = softclip)

    -bowtie2N|-N          : see Bowtie2 -N option (default: 0)
    -bowtie2L|-L          : see Bowtie2 -L option (default: 20)
    -bowtie2D|-D          : see Bowtie2 -D option (default: 30)
    -bowtie2R|-R          : see Bowtie2 -R option (default: 2)
    -bowtie2I|-I          : Minimum fragment length for valid paired-end
                        alignments. See Bowtie2 -I option.
                       (default: 0)

    -bowtie2X|-X          : Maximum fragment length for valid paired-end
                        alignments. See Bowtie2 -X option.
                       (default: 1000)
    -min-score            : see Bowtie2 -score-min option
                        (default: 'G,20,8' local, 'L,-0.4,-0.4' end-to-end)

    -bowtie2k|-k          : Max. number of valid alignment to consider in mapping.
                        From these the programs will then choose the one with
                        the best score on the longest transcript of the gene
                        to which it maps unambiguosly.
                        see also Bowtie2 -k option
                        (default: 10)

    -bowtie2un|-un        : report unaligned reads. See also -unalDir|-ud
                       (default: not set)

    --version             : Print the program version and exit.
    -h|-help              : Print the program help information.
    -man                  : Print a detailed documentation.

    -debug|-d             : Print some debugging information.
```

3.8.2. Command line options for meRanGs

```
USAGE: meRanGs <runmode> [-h] [-m] [--version]

Required <runmode> any of:
    mkbsidx    :   Generate the STAR BS index.
    align      :   Align bs reads to a reference genome.

Options:
    --version   :   Print the program version and exit.
    -h|help     :   Print the program help information.
    -m|man      :   Print a detailed documentation.
```

mkbsidx mode:

```
USAGE: meRanGs mkbsidx [-fa] [-id] [-sjO] [-GTF] [-h] [-m]

Required all of :
    -fa|fasta          : Fasta file(s) to use for BS index generation.
                      Use a comma separated file list or expression
                      (?, *, [0-9], [a-z], {a1,a2,..an}) if more than one
                      fasta file. If using an expression pattern, please put
                      single quotes arround the -fa argument, e.g:

                          -fa '/genome/chrs/chr[1-8].fa'
```

```
    -id|bsidxdir        : Directory where to store the BS index.


Options:
    -star|starcmd       : Path to the STAR aligner.
                        (default: STAR from the meRanTK installation or your system
                         PATH)

    -t|threads          : number of CPUs/threads to run


    -GTF                 : GTF or GFF3 file to use for splice junction database
                         (highly recommended)

    -sjO                 : length of the 'overhang' on each sede of a splice
                         junction.
                         It should be read (mate) 'length -1'.
                        (default: 100)

    -GTFtagEPT          : Tag name to be used as exons' parents for building
                         transcripts. For GFF3 use 'Parent'


                         see STAR -sjdbGTFtagExonParentTranscript option
                         (default: transcript_id)

    -GTFtagEPG          : Tag name to be used as exons' parents for building
                         transcripts. For GFF3 use 'gene'

                         see STAR -sjdbGTFtagExonParentGene option
                         (default: gene_id)

    -star_sjdbFileChrStartEnd
                       : see STAR -sjdbFileChrStartEnd option
                        (default: not set)

    -star_sjdbGTFchrPrefix
                       : see STAR -sjdbGTFchrPrefix option
                        (default: not set)


    -star_sjdbGTFfeatureExon
                       : see STAR -sjdbGTFfeatureExon option
                         (default: exon)

   -star_limitGenomeGenerateRAM

                             : maximum available RAM (bytes) for genome generation
                               see STAR -limitGenomeGenerateRAM option
                           (default: 31000000000)

    --version           : Print the program version and exit.
    -h|help             : Print the program help information.
    -m|man              : Print a detailed documentation.
```

align mode:

```
USAGE: meRanGs align [-f|-r] [-id] [-h] [-m]

Required all of:
   -fastqF|-f           : Fastq file with forward reads (required if no -r)
                          This file must contain the reads that align to the
                          5' end of the RNA, which is the left-most end of the
                          sequenced fragment (in transcript coordinates).
   -fastqR|-r           : Fastq file with reverse reads (required if no -f)
                          This file must contain the reads that align to the
                          opposite strand on the 3' end of the RNA, which is
```

```
                            the right-most end of the sequenced fragment (in
                            transcript coordinates).

Options:
    -illuminaQC|-iqc      : Filter reads that did not pass the Illumina QC.
                            Only relevant if you have Illumina 1.8+ reads.
                            (default: not set)

    -forceDir|-fDir       : Filter reads that did not pass did not pass the
                            internal directionality check:
                                FWDreads: #C > #G && #C > #T && #A > #G)
                                REVreads: #G > #C && #T > #C && #G > #A)
                            (default: not set)

    -first|-fn            : Process only this many reads per input fastq file
                            (default: process all reads)

    -outdir|-o            : Directory where results get stored
                            (default: current directory)

    -sam|-S               : Name of the SAM file for uniq and resolved alignments
                            (default: meRanGs_[timestamp].sam )

    -unalDir|-ud          : Directory where unaligned reads get stored
                            (default: outdir)

                            Note: if -starun|-un is not set unaligned reads
                            will not get stored

    -threads|-t           : Use max. this many CPUs to process data
                            (default: 1)

    -starcmd|-star        : Path to STAR
                            (default: STAR from the meRanTK installation or your
                             system PATH)

    -id|-bsidxdir         : Path to bsindex directory created in 'mkbsidx' runMode.
                            This directory holds the '+' and '-' strand bs index
                            (default: use BS_STAR_IDX_DIR environment variable)

    -bsidxW|-x            : Path to '+' strand bsindex directory created in
                            'mkbsidx' runMode
                            (default: use '-id' option or BS_STAR_IDX_DIR
                                      environment variable)
    -bsidxC|-y            : Path to '-' strand bsindex directory created in
                            'mkbsidx' runMode
                            (default: use '-id' option or BS_STAR_IDX_DIR
                                      environment variable)


    -dovetail|-dt         : If '-dt' is set, "dovetailing" read pairs in paired end
                            mode are allowed and will not be discordant.

                                  (default: not set = dovetailing reads are not aligned)

    -samMM|-MM            : Save multimappers? If set multimappers will be stored
                            in SAM format '$sam_multimappers.sam'
                            (default: not set)

    -ommitBAM|-ob         : Do not create an sorted and indexed BAM file
                            (default: not set)

    -deleteSAM|-ds        : Delete the SAM files after conversion to BAM format
                            (default: not set)

    -star_outFilterMismatchNmax
                          : Maximum edit distance to allow for a valid alignment
                            (default: 2)
```

```
-star_outFilterMultimapNmax
                    : Max. number of valid multi mappers to report
                      (default: 10)

-starun|-un           : Report unaligned reads. See also -unalDir|-ud
                      (default: not set)

-mbiasplot|-mbp       : Create an m-bias plot, that shows potentially biased
                      read positions
                      (default: not set)

-mbiasQS|-mbQS        : Quality score for a high quality m-bias plot. This plot
                      considers only basecalls with a quality score equal or
                      higher than specified by this option.
                      (default: 30)

-mkbg|-bg             : Generate a BEDgraph file from the aligned reads.
                      ! This can take a while !
                      (default: not set)

-minbgCov|-mbgc       : If '-bg' is set, '-mbgc' defines the minimum coverage
                      that we should consider in the BEDgraph output?
                      (default: 1)

-bgScale|-bgs         : Generate a BEDgraph in log [log2|log10] scale
                      (default: not set, no scaling)

-fixMateOverlap|-fmo  : The sequenced fragment and read lengths might be such
                      that alignments for the two mates from a pair overlap
                      each other.
                      If '-fmo' is set, deduplicate alignment subregions that
                      are covered by both, forward and reverse, reads of the
                      same read pair. Only relevant for paired end reads.
                      (default: not set)

-hardClipMO|-hcmo     : If '-fmo' is set, hardclip instead of softclip the
                      overlapping sequence parts.
                      (default: not set = softclip)

-star_genomeLoad      : see STAR -genomeLoad option
                      (default: NoSharedMemory)


-GTF                  : GTF or GFF3 splice to use for junction database
                      (highly recomended, if not specified during index
                      generation)



-sjO                  : length of the 'overhang' on each sede of a splice
                      junction. It should be read (mate) 'length -1'.
                      (default: 100)


-star_readMatesLengthsIn
                    : see STAR -readMatesLengthsIn option
                      (default: NotEqual)

-star_limitIObufferSize
                    : see STAR -limitIObufferSize option
                      (default: 150000000)

-star_outSAMstrandField
                    : see STAR -outSAMstrandField option
                      (default: None)

-star_outSAMprimaryFlag
```

```
                     : see STAR -outSAMprimaryFlag option (has no effect)
                       (default: AllBestScore)

-star_outQSconversionAdd
                     : see STAR -outQSconversionAdd option
                       (default: 0)


-star_outSJfilterReads
                     : see STAR -outSJfilterReads option
                       (default: All)

-star_outFilterType
                     : see STAR -outFilterType option
                       (default: Normal)

-star_outFilterMultimapScoreRange
                     : see STAR -outFilterMultimapScoreRange option
                       (default: 1)

-star_outFilterScoreMin
                     : see STAR -outFilterScoreMin option
                       (default: 0)

-star_outFilterScoreMinOverLread
                     : see STAR -outFilterScoreMinOverLread option
                       (default: 0.9)

-star_outFilterMatchNmin
                     : see STAR -outFilterMatchNmin option
                       (default: 0)

-star_outFilterMatchNminOverLread
                     : see STAR -outFilterMatchNminOverLread option
                       (default: 0.9)
-star_outFilterMismatchNoverLmax
                     : see STAR -outFilterMismatchNoverLmax option
                       (default: 0.05)

-star_outFilterMismatchNoverReadLmax
                     : see STAR -outFilterMismatchNoverReadLmax option
                       (default: 0.1)

-star_outFilterIntronMotifs
                     : see STAR -outFilterIntronMotifs option
                       (default: RemoveNoncanonicalUnannotated)

-star_outSJfilterCountUniqueMin
                     : see STAR -outSJfilterCountUniqueMin option
                       (default: [ 3, 1, 1, 1 ])

-star_outSJfilterCountTotalMin
                     : see STAR -outSJfilterCountTotalMin option
                       (default: [ 3, 1, 1, 1 ])

-star_outSJfilterOverhangMin
                     : see STAR -outSJfilterOverhangMin option
                       (default: [ 25, 12, 12, 12 ])

-star_outSJfilterDistToOtherSJmin
                     : see STAR -outSJfilterDistToOtherSJmin option
                       (default: [ 10, 0, 5, 10 ])

-star_outSJfilterIntronMaxVsReadN
                     : see STAR -outSJfilterIntronMaxVsReadN option
                       (default: [ 50000, 100000, 200000 ])

-star_clip5pNbases    : see STAR -clip5pNbases option
```

```
                        (default: 0)

 -star_clip3pNbases    : see STAR -clip3pNbases option
                        (default: 0)

 -star_clip3pAfterAdapterNbases
                      : see STAR -clip3pAfterAdapterNbases option
                        (default: 0)

 -star_clip3pAdapterSeq
                      : see STAR -clip3pAdapterSeq option
                        (default: not set)

 -star_clip3pAdapterMMp
                      : see STAR -clip3pAdapterMMp option
                        (default: 0.1)

 -star_winBinNbits     : see STAR -winBinNbits option
                        (default: 16)

 -star_winAnchorDistNbins
                      : see STAR -winAnchorDistNbins option
                        (default: 9)

 -star_winFlankNbins
                      : see STAR -winFlankNbins option
                        (default: 4)

 -star_winAnchorMultimapNmax
                      : see STAR -winAnchorMultimapNmax option
                        (default: 50)

 -star_scoreGap        : see STAR -scoreGap option
                        (default: 0)

 -star_scoreGapNoncan  : see STAR -scoreGapNoncan option
                        (default: -8)

 -star_scoreGapGCAG    : see STAR -scoreGapGCAG option
                        (default: -4)

 -star_scoreGapATAC    : see STAR -scoreGapATAC option
                        (default: -8)

 -star_scoreStitchSJshift
                      : see STAR -scoreStitchSJshift option
                        (default: 1)

 -star_scoreGenomicLengthLog2scale
                      : see STAR -scoreGenomicLengthLog2scale option
                        (default: -0.25)

 -star_scoreDelBase    : see STAR -scoreDelBase option
                        (default: -2)

 -star_scoreDelOpen    : see STAR -scoreDelOpen option
                        (default: -2)

 -star_scoreInsOpen    : see STAR -scoreInsOpen option
                        (default: -2)
 -star_scoreInsBase    : see STAR -scoreInsBase option
                        (default: -2)

 -star_seedSearchLmax  : see STAR -seedSearchLmax option
                        (default: 0)

 -star_seedSearchStartLmax
```

```
            : see STAR -seedSearchStartLmax option
              (default: 50)
-star_seedSearchStartLmaxOverLread
            : see STAR -seedSearchStartLmaxOverLread option
              (default: 1)

-star_seedPerReadNmax : see STAR -seedPerReadNmax option
              (default: 1000)

-star_seedPerWindowNmax
            : see STAR -seedPerWindowNmax option
              (default: 50)
-star_seedNoneLociPerWindow
            : see STAR -seedNoneLociPerWindow option
              (default: 10)

-star_seedMultimapNmax
            : see STAR -seedMultimapNmax option
              (default: 10000)

-star_alignEndsType   : see STAR -alignEndsType option
              (default: Local)

-star_alignSoftClipAtReferenceEnds
            : see STAR -alignSoftClipAtReferenceEnds option
              (default: Yes)

-star_alignIntronMin  : see STAR -alignIntronMin option
              (default: 21)

-star_alignIntronMax  : see STAR -alignIntronMax option
              (default: 0)


-star_alignMatesGapMax
            : see STAR -alignMatesGapMax option
              (default: 0)

-star_alignTranscriptsPerReadNmax
            : see STAR -alignTranscriptsPerReadNmax option
              (default: 10000)

-star_alignSJoverhangMin
            : see STAR -alignSJoverhangMin option
              (default: 5)

-star_alignSJDBoverhangMin
            : see STAR -alignSJDBoverhangMin option
              (default: 3)

-star_alignSplicedMateMapLmin
            : see STAR -alignSplicedMateMapLmin option
              (default: 0)

-star_alignSplicedMateMapLminOverLmate
            : see STAR -alignSplicedMateMapLminOverLmate option
              (default: 0.9)

-star_alignWindowsPerReadNmax
            : see STAR -alignWindowsPerReadNmax option
              (default: 10000)

-star_alignTranscriptsPerWindowNmax
            : see STAR -alignTranscriptsPerWindowNmax option
              (default: 100)

-star_chimSegmentMin  : see STAR -chimSegmentMin option
              (default: 0)
```

```
    -star_chimScoreMin
                        : see STAR -chimScoreMin option
                          (default: 0)

    -star_chimScoreDropMax
                        : see STAR -chimScoreDropMax option
                          (default: 20)

    -star_chimScoreSeparation
                        : see STAR -chimScoreSeparation option
                          (default: 10)

    -star_chimScoreJunctionNonGTAG
                        : see STAR -chimScoreJunctionNonGTAG option
                          (default: -1)

    -star_chimJunctionOverhangMin
                        : see STAR -chimJunctionOverhangMin option
                          (default: 20)

    -star_sjdbScore       : see STAR -sjdbScore option
                          (default: 2)

    --version            : Print the program version and exit.
    -h|help              : Print the program help information.
    -m|man               : Print a detailed documentation.

    -debug|-d            : Print some debugging information.
```

3.8.3. Command line options for meRanGh

```
USAGE: meRanGh <runmode> [-h] [-m] [--version]

Required <runmode> any of:
    mkbsidx    :   Generate the HiSat2 BS index.
    align      :   Align bs reads to a reference genome.

Options:
    --version   :   Print the program version and exit.
    -h|help     :   Print the program help information.
    -m|man      :   Print a detailed documentation.
```

mkbsidx mode:

```
USAGE: meRanGh mkbsidx [-fa] [-id] [-h] [-m]

Required all of :
    -fa|fasta           : Fasta file(s) to use for BS index generation.
                        Use a comma separated file list or expression
                        (?, *, [0-9], [a-z], {a1,a2,..an}) if more than one
                        fasta file. If using an expression pattern, please put
                        single quotes arround the -fa argument, e.g:

                            -fa '/genome/chrs/chr[1-8].fa'

    -id|bsidxdir        : Directory where to store the BS index.


Options:

    -hisat2build|hs2b   : Path to the HiSat2 "hisat2-build" program.
                        (default: hisat2-build from the meRanTK installation
                         or systems PATH)
```

```
    -t|threads          : number of CPUs/threads to run

    --version           : Print the program version and exit.
    -h|help             : Print the program help information.
    -m|man              : Print a detailed documentation.
```

align mode:

```
USAGE: meRanGh align [-f|-r] [-id] [-h] [-m]

Required all of:
   -fastqF|-f           : Fastq file with forward reads (required if no -r)
                         This file must contain the reads that align to the
                         5' end of the RNA, which is the left-most end of the
                         sequenced fragment (in transcript coordinates).
   -fastqR|-r           : Fastq file with reverse reads (required if no -f)
                         This file must contain the reads that align to the
                         opposite strand on the 3' end of the RNA, which is
                         the right-most end of the sequenced fragment (in
                         transcript coordinates).


Options:
            -GTF                 : GTF gene model annotations and/or known
                                  transcripts for building a transcriptome index.

            -illuminaQC|-iqc    : Filter reads that did not pass the Illumina QC.
                                  Only relevant if you have Illumina 1.8+ reads.
                                  (default: not set)

            -forceDir|-fDir     : Filter reads that did not pass did not pass the
                                  internal directionality check:
                                      FWDreads: #C > #G && #C > #T && #A > #G)
                                      REVreads: #G > #C && #T > #C && #G > #A)
                                  (default: not set)

            -first|-fn          : Process only this many reads per input fastq file
                                  (default: process all reads)

            -outdir|-o          : Directory where results get stored
                                  (default: current directory)

            -sam|-S             : Name of the SAM file for uniq and resolved alignments
                                  (default: meRanGh_[timestamp].sam )

            -unalDir|-ud        : Directory where unaligned reads get stored
                                  (default: outdir)
                                  Note: if -hisat2un|-un is not set unaligned reads
                                  will not get stored

            -threads|-t         : Use max. this many CPUs to process data
                                  (default: 1)

            -hisat2cmd|-hisat2  : Path to hisat2
                                  (default: use hisat2 from the meRanTK installation or
                                   in your system PATH)

            -id|-bsidxdir       : Path to bsindex directory created in 'mkbsidx' runMode.
                                  This directory holds the '+' and '-' strand bs index
                                  (default: use BS_HISAT2_IDX environment variable)

            -samMM|-MM          : Save multimappers? If set multimappers will be stored
                                  in SAM format '\$sam_multimappers.sam'
                                  (default: not set)
```

```
-ommitBAM|-ob       : Do not create an sorted and indexed BAM file
                  (default: not set)

-deleteSAM|-ds      : Delete the SAM files after conversion to BAM format
                       (default: not set)

-deleteBAMus|-dbus  : Delete the unsorted BAM files after sorting BAM.
                       (default: not set)


-hisat2un|-un       : Report unaligned reads. See also -unalDir|-ud
                       (default: not set)

-mbiasplot|-mbp     : Create an m-bias plot, that shows potentially biased
                  read positions
                  (default: not set)

-mbiasQS|-mbQS      : Quality score for a high quality m-bias plot. This plot
                  considers only basecalls with a quality score equal or
                  higher than specified by this option.
                  (default: 30)

-mkbg|-bg           : Generate a BEDgraph file from the aligned reads.
                  ! This can take a while !
                  (default: not set)

-minbgCov|-mbgc     : If '-bg' is set, '-mbgc' defines the minimum coverage
                  that we should consider in the BEDgraph output?
                  (default: 1)

-fixMateOverlap|-fmo: The sequenced fragment and read lengths might be such
                  that alignments for the two mates from a pair overlap
                  each other.
                  If '-fmo' is set, deduplicate alignment subregions that
                  are covered by both, forward and reverse, reads of the
                  same read pair. Only relevant for paired end reads.
                  (default: not set)

-hardClipMO|-hcmo   : If '-fmo' is set, hardclip instead of softclip the
                  overlapping sequence parts.
                  (default: not set = softclip)

-dovetail|-dt       : If '-dt' is set, "dovetailing" read pairs in paired end
                  mode are allowed and will not be discordant.
                  (default: not set = dovetailing reads are not aligned)

-bgScale|-bgs       : Generate a BEDgraph in log [log2|log10] scale
                  (default: not set, no scaling)

-max-edit-dist|-e   : Maximum edit distance to allow for a valid alignment
                  (default: 10)

-max-mm-rate|-mmr   : Maximum mismatch ratio (mismatches over read alignement
                  length) [0 <= mmr < 1]
                  (default: 0.05)

-hisat2_trim5
                : see hisat2 manual for -trim5 option
                  (default: 0)

-hisat2_trim3
                : see hisat2 manual for -trim3 option
                  (default: 0)
-hisat2_n-ceil
                : see hisat2 manual for -n-ceil option
                  (default: L,0,0.15)
```

```
-hisat2_ignore-quals
                    : see hisat2 manual for -ignore-quals option
                      (default: not set)
-hisat2_nofw
                    : see hisat2 manual for -nofw option
                      (default: not set)

-hisat2_norc
                    : see hisat2 manual for -norc option
                      (default: not set)

-hisat2_pen-canonsplice
                    : see hisat2 manual for -pen-canonsplice option
                      (default: 0)

-hisat2_pen-noncanonsplice
                    : see hisat2 manual for -pen-noncanonsplice option
                      (default: 12)

-hisat2_pen-canintronlen
                    : see hisat2 manual for -pen-canintronlen option
                      (default: G,-8,1)

-hisat2_pen-noncanintronlen
                    : see hisat2 manual for -pen-noncanintronlen option
                      (default: G,-8,1)

-hisat2_min-intronlen
                    : see hisat2 manual for -min-intronlen option
                      (default: 20)

-hisat2_max-intronlen
                    : see hisat2 manual for -may-intronlen option
                      (default: 500000)

-hisat2_novel-splicesite-outfile
                    : see hisat2 manual for -novel-splicesite-outfile option
                      (default: not set)

-hisat2_novel-splicesite-infile
                    : see hisat2 manual for -novel-splicesite-infile option
                      (default: not set)

-hisat2_no-temp-splicesite
                    : see hisat2 manual for -no-temp-splicesite option
                      (default: not set)

-hisat2_no-spliced-alignment
                    : see hisat2 manual for -no-spliced-alignment option
                      (default: not set)

-hisat2_rna-strandness
                    : see hisat2 manual for -rna-strandness option
                      (default: FR)

-hisat2_transcriptome-mapping-only
                    : see hisat2 manual for -transcriptome-mapping-only
                      option
                      (default: not set)

-hisat2_remove-chrname
                    : see hisat2 manual for -remove-chrname option
                      (default: not set)

-hisat2_add-chrname
                    : see hisat2 manual for -add-chrname option
                      (default: not set)
```

```
    -hisat2_omit-sec-seq
                  : see hisat2 manual for -omit-sec-seq option
                    (default: not set)

    -hisat2_offrate
                  : see hisat2 manual for -offrate option
                    (default: not set)

    -hisat2_ma
                  : see hisat2 manual for -ma option
                    (default: not set)
    -hisat2_mp
                  : see hisat2 manual for -mp option
                    (default: 6,2)

    -hisat2_sp
                  : see hisat2 manual for -sp option
                    (default: 1,2)

    -hisat2_np
                  : see hisat2 manual for -np option
                    (default: 1)

    -hisat2_rdg
                  : see hisat2 manual for -rdg option
                    (default: 5,3)

    -hisat2_rfg
                  : see hisat2 manual for -rfg option
                    (default: 5,3)

    -hisat2_min-score
                  : see hisat2 manual for -min-score option
                    (default: L,0.0,-0.2)

    -hisat2_k
                  : see hisat2 manual for -k option
                    (default: not set)

    -hisat2_minins
                  : see hisat2 manual for -minins option
                    (default: 0)

    -hisat2_maxins
                  : see hisat2 manual for -maxins option

              (default: 500)


    -hisat2_met-file
                  : see hisat2 manual for -met-file option
                    (default: not set)

    --version          : Print the program version and exit.
    -h|help           : Print the program help information.
    -m|man             : Print a detailed documentation.

  -debug|-d          : Print some debugging information.
```

3.8.4. Command line options for meRanCall

```
USAGE: meRanCall [options] [-h] [-man] [--version]

Required options any of:
    -fasta|-f            : Reference sequence FASTA file.

    -sam|-bam|-s         : Sequence read alignment file in SAM or BAM format.

    -result|-o           : Result file where to store the metylation calls.

    -genomeDBref|-gref   : SAM/BAM file was genereated by aligning bs-reads to
                         a genome reference (DNA database): e.g. mouse
                         genome using meRanG.
                         If set, a BED6 + 3 file will be created in
                         addition to the standard result file.
                         (default: not set)

    -transcriptDBref|-tref : SAM/BAM file was genereated by aligning bs-reads to
                         a transcript reference (Transcript database): e.g.
                         mouse refSeqRNA using meRanT.
                         (default: not set)


Options:

    -procs|-p              : Number or processors (CPUs) to use in parallel.
                         Setting this option significantly reduces the
                         processing time. E.g. when set to "-p 16" 16 sequences
                         (e.g. chromosomes) will be processed in parallel.
                         (default: 1)


    -regions|-bi           : BED file with regions to scan for m5Cs. If specified
                         meRanCall will only call m5Cs in the regions present
                         in the BED file.
                         (default: not set, scan entire SAM/BAM file)


    -fskip5|-fs5           : number or bases to ignore on the 5' end of a forward
                         read. This helps to aviod biased results. See m-bias
                         plot from meRanG or meRanT output to get
                         an estimate for this number.
                         (defualt: 0)

    -fskip3|-fs3           : number or bases to ignore on the 3' end of a forward
                         read. This helps to aviod biased results. See m-bias
                         plot from meRanG or meRanT output to get
                         an estimate for this number.
                         (defualt: 0)

    -rskip5|-rs5           : number or bases to ignore on the 5' end of a reverse
                         read. This helps to aviod biased results. See m-bias
                         plot from meRanG or meRanT output to get
                         an estimate for this number.
                         (defualt: 0)

    -rskip3|-rs3           : number or bases to ignore on the 3' end of a reverse
                         read. This helps to aviod biased results. See m-bias
                         plot from meRanG or meRanT output to get
                         an estimate for this number.
                         (defualt: 0)
```

```
-readLength|-rl       : If set to the original read length, then the 3' end
                        skipping will be adjusted for 3' trimming. In other
                        words: if you trimmed some of your reads before
                        mapping, than the number of trimmed bases on the 3'
                        end will be treated as already skipped.
                        This has no effect if fskip5, fskip3, rskip5 or rskip3
                        is 0.
                        (default: 100)

-minMethR|-mr         : Minimum methylation ratio of a single C, that is
                        needed to consider this C as potentially methylated
                        (default: 0.2)

-minMutR|-mutR        : Minimum ratio (bases on reads at a given reference
                        position different from reference base) above which
                        a base will be considered as mutated in respect to
                        the base on the reference sequence.
                        (default: 0.8)


-minBaseQ|-mBQ        : Minimum read base quality (phred score) to condsider
                        for methylation calling.
                        (default: 30)

-minCov|-mcov         : Minimum coverage at a given reference position above
                        which methylation calling will be performed.
                        (default: 10)

-minC|-mc             : Minimum C count at a given reference position above
                        which methylation calling will be performed.
                        (default: 3)

-signalToNoise|-snr   : Only call m5Cs on position with minimum signal to noise
                        ratio. (clean reads / reads before filtering)
                        (default: 0.9)

-maxDup|-md           : Maximum number of read duplicates covering a given
                        position. Read duplicates have the same start positon
                        on the reference and map to the same sequence.
                        (default: 0, do not filter duplicates)

-fisherTest|-fet      : Use Fisher exact Test to statistically test the methylation
                        state against background non conversion levels at the given
                        coverage.
                        (default: not set, default in future)

-C_cutoff|-Cco        :  C cutoff. Number of Cs per read, if this cutoff is exceeded
                        the read will be ignored.
                        (default: 3)

-conversionRate|-cr   : Global C->T conversion rate (0 < cr <= 1). Will be used for
                        calculating p-values.
                        (default: 1)

-geneConvRate|-gcr    : Calculate and use gene specific C->T conversion rates. Will
                        be used for calculating p-values. This CR is specific to each
                        gene. Needs a GTF file (see below).
                        (default: not set)

-gtf                  : GTF file for calculating the gene specific conversion rates.
                        (default: not set)

-errorInterval|-ei    : Error interval for methylation rate p-value
                        calculation
```

```
                 (default: 0)

-fdr             : Control the false discovery rate of methylated
                 cytosines at the specified FDR (0 < fdr < 1).
                 (default: not set)


-fdrRate           : Use the probability that the real methylation level or
                 rate instead of the methylation state p-value to
                 control the false discovery rate at –fdr FDR(0 < fdr <
                 1).
                 (default: not set)


-calcConvRate|-ccr     : Caluclate the C->T conversion rate from an
                 unmehtylated control sequence.
                 (default: not set)

-controlSeqID|-cSeqID  : Control sequence ID for C->T conversion rate
                 calculation. Can be specified multiple times for
                 multiple control sequences.
                 (default: not set)

-excludeSeqID|-exSeqID : Sequence ID(s) to exclude from methylation calling.
                 Can be specified multiple times for multiple control
                     sequences. E.g. -exSeqID chr1 -exSeqID chrUn_gl000220
                 (default: not set)

-reportUP|-rUP        : report unmethylated mutated bases?
                (default: not set)

-bed63             : Generate a BED6 + 3 file - only relevant for genome
                 mapped data!
                (default: not set)
-narrowPeak|-np       : Generate a narrowPeak BED file - only relevant for
                 genome mapped data!
                (default: not set)

-seqContext|-sc       : If set to a number, this number of bases 5' and 3' of
                 the methylated C will be displayed in the result file.
                 (default: not set)

-havZG|-zg           : If set, the methylation caller will look for the "ZG"
                 custom SAM tag and use it a gene name associated with
                 the methylated positon in the result file.


                 Note:
                 meRanT adds this tag to the SAM entries.


                 meRanG does not, however you can use the BED6 + 3
                 file and run the "meRanAnnotate" tool from meRanTK to
                 associate methylated C's with gene(transcript) names.


                 (default: not set)




-azaMode|-aza         : If set, the methylation caller will run in the Aza-IP
                 mode and enables methylation calling from Aza-IP data
                 by looking for C->G conversions, which are character-
                 istic for Aza-IP data.



                 (default: not set)



--version          : Print the program version and exit.
-h|help            : Print the program help information.
```

```
-man                 : Print a detailed documentation.
-debug|-d              : Print some debugging information.
```

3.8.5. Command line options for meRanCompare

```
USAGE: meRanCompare [options] [-h] [-man] [--version]

Required options all of:
    -condition-files-a|-fa : meRanCall result files from condition A. All result
                             files from the first condition can be specified as
                             a comma separated list.
                             (default: not set)

    -condition-files-b|-fb : meRanCall result files from condition B. All result
                             files from the second condition can be specified as
                             a comma separated list.
                             (default: not set)

Options:

    -condition-name-a|-na  : Name of the condition A. This name is used in the
                             file names for the meRanCompare results.
                             (default: ConditionA)

    -condition-name-b|-nb  : Name of the condition B. This name is used in the
                             file names for the meRanCompare results.
                             (default: ConditionB)

    -size-factors-a|-sfa   : Library size factors for samples in condition A/B
    -size-factors-b|-sfb     specified as comma separated list. These size factors
                             are used to calculate normalized counts.

                             e.g. -sfa 0.6673,0.6609,0.7347
                                  -sfb 0.9559,1.4098,2.3802

                             The ordering of the individual size factors has to
                             match the ordering of the meRanCall result files for
                             the corresponding conditions (see -fa and -fb option).

                             The size factors can be calculated using the meRanTK
                             tool "estimatSizeFactors.pl". Alternatively one can
                             use htseq-count and DESeq2 "estimateSizeFactors"

                             (default: not set, no normailzed counts are reported)

    -minRep|-mr            : number of replicates a m5C candidate has to be present
                             so that it is considered as high confidence call.
                             (default: 2)

    -sig|-s                : p-value below which the differential methylation will
                             be reported as significant.
                             (default: 0.01)

    -minFC|-fc             : minimum fold change above which the differential
                             methylation will be reported. In bisulfite mode
                             (default) the foldchange will be calculated as ratio
                             of the methylation rate in condition A and B. In "aza"
                             mode this will be the ratio of the (normalized)
                             coverage at the specific positon in condition A and B.

                             (default: not set, report all significant (see -sig)
                                       changes)

    -fdr                   : FDR, false discovery rate
                             (default: 0.01)
```

```
-azaMode|-aza         : run meRanCompare in Aza-IP mode. In this mode the IP
                       enrichment (A) over control (B) is calculated for each
                       candidate m5C which is compared.
                       (default: not set)

--version             : Print the program version and exit.
-h|help               : Print the program help information.
-man                  : Print a detailed documentation.

-debug|-d             : Print some debugging information.
```

3.8.6. Command line options for meRanAnnotate

```
USAGE: meRanAnnotate [options] ][-h] [--version]

Required options any of:
   -tab|-t             : meRanCall/meRanCompare result file to intersect with
                      gff.
                      Format:
                      <chr><tab><position><tab><strand>[<tab><field>]...
    OR

   -bed|-b              : BED file to intersect with gff.
                      Format:

          <chr><tab><start><tab><end><tab><name><tab><score><tab><strand>[<tab><field>...]

   -gff|-g             : Sorted or unsorted GFF3/GTF file.




Options:

   -ensGTF|-gtf          : Annotation file is a GTF file
                     (default: not set, assuming it is GFF3)

   -feautre|-f           : GFF3 features you want to intersect with your meRanTK
                      or BED file,

                      e.g. if you are interested in mRNA and ncRNA use

                           -f 'mRNA|ncRNA'

                      (default: 'gene|mRNA|transcript|ncRNA' for NCBI GFF3
                              'gene|transcript' for Ensembl GTF)

   -outfile|-o           : Result file where to store the intersecting/overlapping
                      features.
                      (default: STDOUT)

   -parallel|-p           : Number of CPUs to use. Running in parallel mode is way
                      faster but it requires the MCE Perl module to be
                      installed.
                      (default: 1, no parallel processing)

   -chrPrefix|-cp         : Prepend this string to chromosome/sequence name from
                      the meRanCall/meRanCompare result of BED file. The
                      chromosome/sequence name has to match the one used in
                      the GFF3/GTF file.

                      e.g.
                      If you have m5C calls or BED ranges that have
                      chromosome names of the type '1, 2, 3, ...' and want to
                      use a GFF3/GTF file that has chromosome names
                      of the type 'chr1, chr2, chr3, ...' then you might use
                      -cp 'chr'

                     (default: not set, assuming chromosome names are
                             matching)

   -reportDist|-rd        : Calculate distances to features 5' and 3' ends.
                      The following distances are calculated:

                            query 5' to feature 5'
                            query 3' to feature 3'
```

```
                     query 5' to feature 3'
                     query 3' to feature 5'
                     query center to feature 5'
                     query center to feature 3'

                 All distances are reported in stranded mode.

                (default: not set, not distances are reported)

-relativeDist|-reld   : Calculate relative distances. Distances will be
                 reported as %-feature length, that is genomic feature
                 3' - genomic feature 5' end coordinate.

                (default: not set, distances are reported in # of bp)

-expandResults|-er    : Expand results. Report each GFF-feature match as a
                 separate line.

                (default: not set, forced if -rd is set)

--version             : Print the program version and exit.

-h|help               : Print the program help information.


                                        Acknowledgements:
                                           Thanks go to Pedro Silva for his binary range
                                           search algorithm, which was a useful in-
                                           spiration for this program.
```