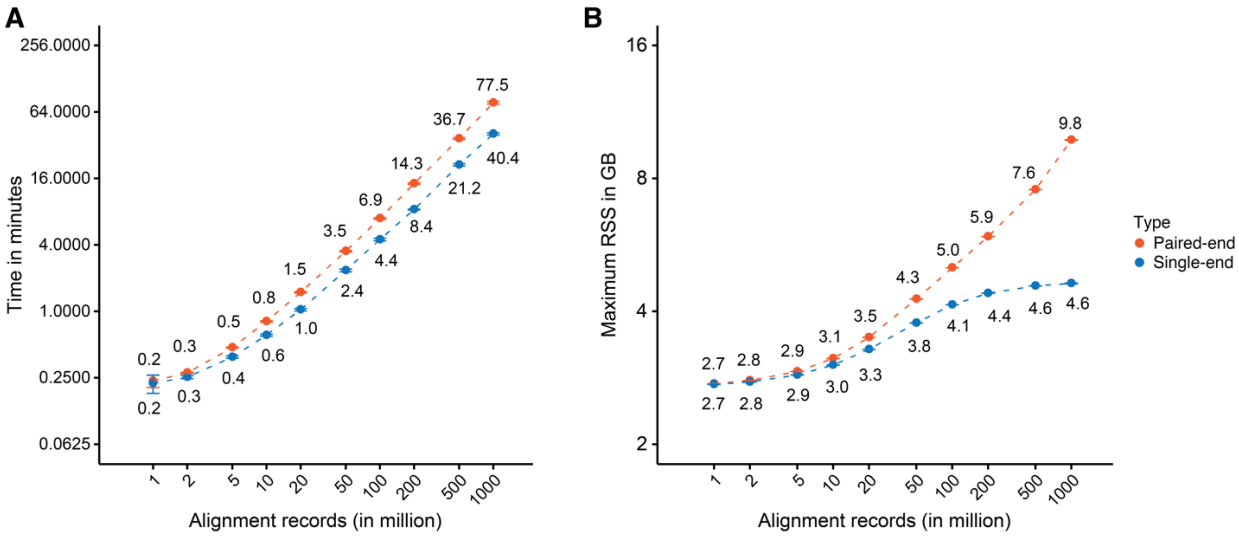
# Supplementary Material

## I Performance of RLM

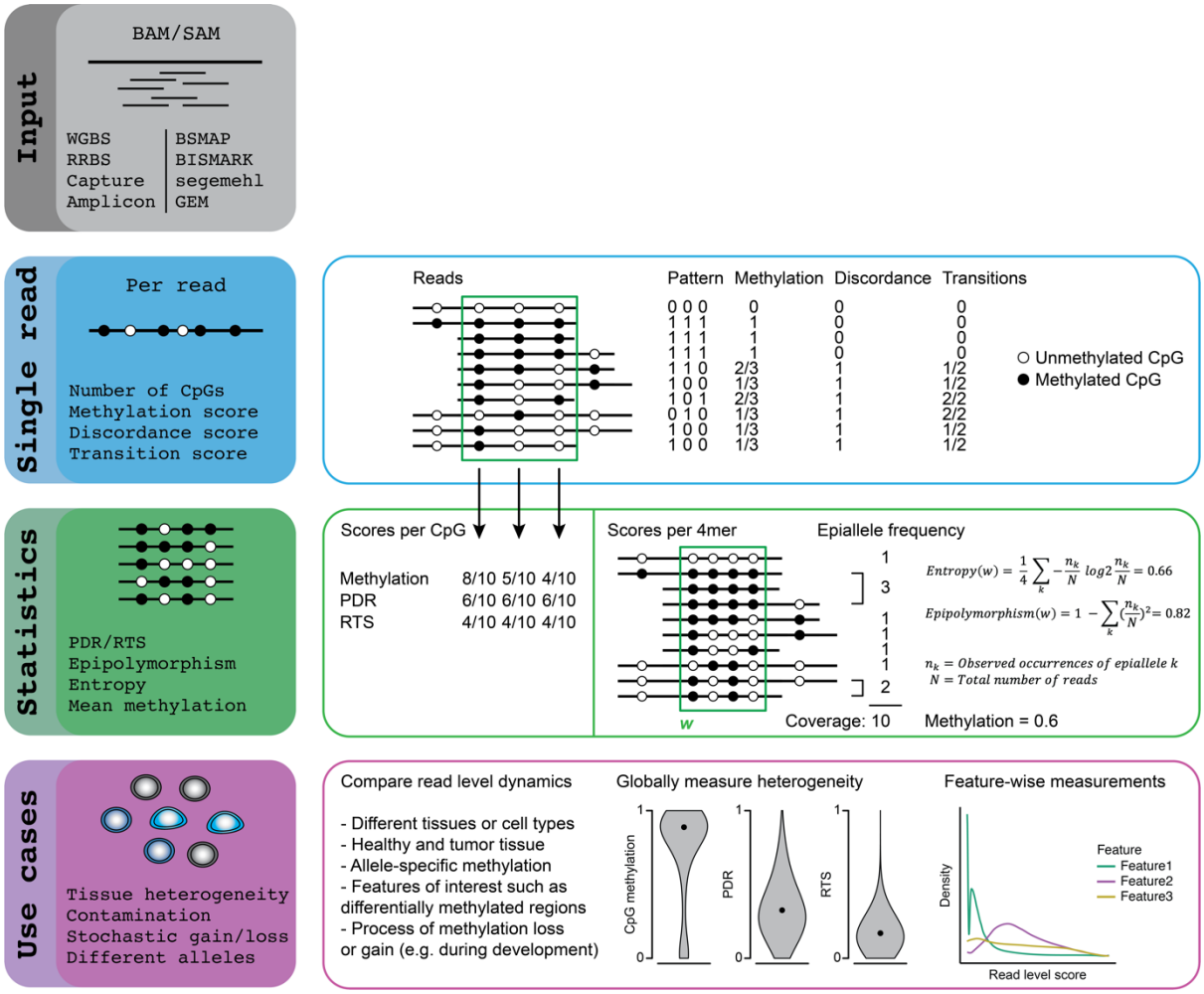



**Supplementary Fig. 1 Feature overview and performance of RLM**
**A)** Runtime measurements using different numbers of BAM records for single- and paired-end mode. For paired-end reads, the corresponding number of pairs in million contained in the test files are 0.5, 1.0, 2.4, 4.9, 9.7, 24.3, 48.7, 97.5, 243.7 and 487.3 respectively. Mean measurements across five runs are reported and whiskers indicate standard deviation.
**B)** Memory measurements using different numbers of BAM records using the maximum resident set size (RSS) for single- and paired-end mode (number of pairs in paired-end mode as in B). Mean measurements across five runs are reported and whiskers indicate standard deviation.

We measured the performance of RLM by randomly sampling increasing numbers of BAM records from a publicly available data set (GSM4075619, aligned to the mouse reference genome mm10 using BSMAP) ranging from the size of an RRBS experiment to a high-coverage WGBS data set. Benchmarks were executed on an Intel Xeon 6242 @ 2.80GHz with test data located on NFS mounted file servers and measurements are reported as the average across five runs with standard deviation. The runtime of RLM scales linearly with the input size for both paired- and single-end modes. The single-end mode is more than twice as fast compared to the paired-end mode (around 40 and 90 minutes for one billion records respectively), which can be attributed to the temporary storage and specific treatment of paired-end reads. Here, reads are kept in memory until the mate has been read, potential overlaps of the two mates are resolved and reads are removed from memory afterwards. This is also reflected in the memory consumption, which stays consistently below five GB for the single-end mode while the peak memory increases exponentially for paired-end reads. However, even for an extremely well covered experiment using a billion reads (roughly 500 million fragments), the maximum memory remains below 32 GB. Typical data sets in the range of RRBS experiments (50 million fragments) run with modest memory requirements. Runtime comparisons with an existing software package and detailed feature comparisons can be found in the next section.

# II Implementation, usage and comparability with other tools

## 1 Implementation details




**Supplementary Fig. 2 Feature overview of RLM**
Structure and features of RLM. Both SAM and BAM files for different experiment types such as target enrichment approaches, reduced representation bisulfite sequencing (RRBS), whole genome bisulfite sequencing (WGBS) with single-end (SE) and paired-end (PE) reads can be used. Statistics per single read as well as aggregated in form of read-level methylation scores can be produced as output.
We also offer a separate, standalone R Markdown script that generates a report based on the RLM output files including summary statistics, distribution of coverage, methylation and read level scores as well as figures visualizing global and per-feature read-level methylation dynamics. Read-level methylation scores can be used for various applications where the underlying population dynamics are important to consider besides the bulk methylation measurements such as tissue heterogeneity, comparing healthy and tumor tissue, allele-specific methylation and population methylation dynamics over time.

### 1.1 General

RLM processes BAM files by streaming over the input records. Generally, reads get excluded if they:

- Are not a primary alignment
- Are QC-failed
- Are PCR or optical duplicates

- Contain indels
- Fail a user-defined mapping quality threshold (default: 30)
- Contain mismatches at CpG positions or less than 3 CpGs (A minimum number of CpGs is required to make a useful statement about heterogeneity or methylation patterns. For entropy and epipolymorphism calculations at least 4 CpGs need to be present.)

For every read, the methylation status of each CpG is detected and read-wise methylation statistics are reported. Additionally, if PDR/RTS scores or entropy/epipolymorphism scores are requested, these are calculated for all CpGs or 4-mers covered by a minimum number of reads defined by the user (default: 10).

**Indels:** Indels complicate read-level analyses because CpGs can potentially be inserted or (partially) deleted which makes it tricky to compute valid statistics for such regions. Therefore, currently reads containing indels are not included in the analysis. However, pairs where one mate contains an indel while the other does not will be partially processed using only the mate without indels.

### 1.2 Paired-end sequencing mode

For paired-end reads, reads get stored in memory until the mate is mapped. Depending on the fragment size, mates are mapped with a specific insertion size but sometimes also overlap each other. This brings confounding factors into the read-level statistics calculation: The two mates belong to the same allele and if both reads would be treated independently, the overlapping positions would be counted twice for the same allele and by that biasing the population heterogeneity measurement. RLM therefore merges overlapping reads of the same pair into one long, contiguous read which is then further processed. This procedure has the additional advantage that more consecutive CpGs on the same read can be examined and increase the genome-wide coverage when looking at scores such as entropy.

### 1.3 RRBS mode

Reduced representation bisulfite sequencing (RRBS) is a specific type of bisulfite sequencing experiments where the DNA gets fragmented using restriction enzymes. The most commonly used enzyme is MspI cutting the sequence 5'-CCGG-3' which enriches for fragments in CG-rich regions (such as CpG islands). During the fragment end-repair after cleavage, artificial cytosines get introduced at the 3' end of reads (or 5' end for non-directional/paired-end reads originating from the reverse complement of the original forward or reverse strand) which do not represent the original methylation status of the read at this position (see the Babraham RRBS guide, https://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf). If the sequencing reads are longer than the fragment size, these artificial CpGs could get incorporated into downstream analysis which is undesirable. Therefore, RLM offers the option to trim 2 bases of the 3' end of reads from the original forward or reverse strand and 2 bases of the 5' end of reads from the reverse complemented original forward or reverse strand (option: -rrbs). You should not set this option if you already accounted for this problem during trimming (e.g. using Trim Galore, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

## 2 Input files

### 2.1 Reads

The main input for RLM consists of a BAM file from a bisulfite sequencing alignment tool. RLM supports BAM files from either BSMAP, BISMARK, segemehl or GEM (e.g. included in gemBS) and the used alignment tool needs to be specified when running RLM using the option -a. All four alignment tools use different tags to report the strand each read originated from and based on this option RLM will choose which tag to look for.

We recommend trimming of low quality bases (and potentially tail trimming for swift libraries) prior to the alignment (e.g. using cutadapt) as well as the removal of technical duplicates after the alignment

(e.g. using Picard MarkDuplicates). This way technical bias is reduced and read-level metrics are not influenced by artefacts. We also recommend sorting the BAM file by position in order to reduce memory consumption while running RLM in paired-end mode but RLM can also process unsorted or name-sorted BAM files.

RLM can process BAM files of different bisulfite sequencing experiments such as whole genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), hybrid capture methods or amplicons with single-end or paired-end reads. For RRBS, we recommend using BSMAP in RRBS mode in order to reduce runtime and avoid mis-mappings.

### 2.2 Reference genome

RLM can be run using any reference genome (also custom genomes or assemblies), however, it should be the same genome (ideally the same file) that was used to align the reads in the BAM file. The order and number of reference sequences in the BAM header and the reference genome FASTA file are compared by RLM and different numbers or order of sequences will result in an error and termination of the program.

## 3 Output files

RLM offers multiple different output files that can be requested all together or separately. Which output files will be computed can be set via the -s option. This option can be set to either “single_read”, “entropy”, “pdr” or “all”.

### 3.1 Single read output

The “single_read” option only writes an output file containing information for every read or read pair that passed all filters and can be considered for read level analysis. This file gets written for every score option since the information it contains needs to be computed for all downstream scores. It has the following tab-delimited format:

```
#chr      start     end       read_name  CpG_pattern          n_CpGs    n_CpGs_methyl
          discordance_score   transitions_score    mean_methylation
chr_test  2384      2518      A00442:HFH2KDSXX190418:HFH2KDSXX:1:1108:19334:12555 gGg       3
          1         1         1         0.333333
chr_test  5286      5421      A00442:HFH2KDSXX190418:HFH2KDSXX:1:1103:6334:22670  GGG       3
          3         0         0         1
chr_test  7418      7553      A00442:HFH2KDSXX190418:HFH2KDSXX:1:1108:17463:25692 gGGG      4
          3         1         0.333333  0.75
chr_test  7444      7579      A00442:HFH2KDSXX190418:HFH2KDSXX:1:1104:11957:9768  gGgg      4
          1         1         0.666667  0.25
```

For every read, 10 different fields are reported:

1. The chromosome the read aligned to.
2. The start position of the read with respect to the chromosome (0-based, half-open intervals).
3. The end position of the read with respect to the chromosome (0-based, half-open intervals).
4. The read name. This will be the same for mates of the same pair.
5. The methylation pattern for all CpGs spanned by the read. Capital G indicates methylation, lower case G refers to unmethylated CpGs.
6. The number of CpGs spanned by the read.
7. The number of methylated CpGs spanned by the read.
8. The discordance score of the read (0 if all CpGs are either unmethylated or methylated, 1 otherwise).
9. The transition score of the read (how often does the pattern switch from methylated to unmethylated for consecutive CpGs normalized by the possible number of transitions n - 1)
10. The mean methylation of the read based on all CpGs spanned by it.

Note: If trimming of reads in the RRBS mode is enabled, the start and end position of the reads will match the sequence considered for RLM and will be truncated either at the 3' end (reads originating

from the original forward/reverse strand) or 5' end (reads originating from the reverse complement of the original forward/reverse strand).

### 3.2 Entropy output

When choosing “entropy” as score option, additionally to the single read output file another file will be written of the following form:

```
#chr       start     end       entropy   epipolymorphism     gggg      gggG      ggGg      ggGG
           gGgg      gGgG      gGGg      gGGG      Gggg      GggG      GgGg      GgGG      GGgg
           GGgG      GGGg      GGGG      mean_methylation    coverage
chr_test   7390304   7390306   0.663386  0.792899  5         2         1         1         1
           1         0         1         0         0         0         1         0         0
           0         0         0.269231  13
chr_test   7390619   7390621   0.42511   0.579882  8         0         1         0         2
           0         0         0         1         0         0         1         0         0
           0         0         0.134615  13
chr_test   7390646   7390648   0.584893  0.764444  6         2         0         0         2
           1         0         2         2         0         0         0         0         0
           0         0         0.233333  15
chr_test   7390665   7390667   0.508735  0.662722  7         0         2         0         0
           0         0         0         1         0         0         1         0         0
           1         1         0.25      13
```

For every 4-mer of consecutive CpGs that are spanned by a user-defined minimum number of reads (default: 10), the following fields are reported:

1. The chromosome.
2. The start position of the first CpG in the 4-mer (0-based, half-open intervals).
3. The end position of the first CpG in the 4-mer (0-based, half-open intervals).
4. The methylation entropy calculated for the 4-mer based on the reads that span the complete 4-mer. For more information on this score see (Xie *et al.*, 2011).
5. The methylation epipolymorphism calculated for the 4-mer based on the reads that span the complete 4-mer. For more information on this score see (Landan *et al.*, 2012).
6. The count of reads for all possible 16 epialleles that underlay the entropy and epipolymorphism calculations (16 columns, the header defines the epiallele per column. Capital G indicates methylation, lower case G refers to unmethylated CpGs).
7. The mean methylation of the 4-mer based on all 4 CpGs across all considered reads. This might slightly deviate from the value that can be calculated by standard methylation calling since RLM excludes certain reads that might be considered by standard methylation callers such as reads with indels, low quality reads, etc.
8. The coverage defined as the number of reads considered that span the complete 4-mer.
9. 4-mers are reported using the first CpG as position in order to allow creating browser tracks but the value refers to the complete 4-mer starting with this CpG.

### 3.3 PDR output

When choosing “pdr” as score option, additionally to the single read output file another file will be written of the following form:

```
#chr       start     end       PDR       RTS       mean_methylation    coverage
chr_test   7390271   7390273   0.727273  0.360606  0.636364  11
chr_test   7390275   7390277   0.727273  0.360606  0.363636  11
chr_test   7390304   7390306   0.655172  0.366667  0.103448  29
chr_test   7390346   7390348   0.619048  0.372222  0.214286  42
```

For every CpG that is spanned by a user-defined minimum number of reads (default: 10), the following fields are reported:

1. The chromosome.
2. The start position of the CpG (0-based, half-open intervals).
3. The end position of the CpG (0-based, half-open intervals).

4. The percent of discordant reads (PDR) calculated based on the reads that span the CpG. The number of discordant reads (neither completely unmethylated nor completely methylated reads) is normalized by the total number of considered reads. For more information on this score see (Landau *et al.*, 2014).
5. The average read transition score (RTS) calculated based on the reads that span the CpG. The transition score per read (see single read output) normalized by the total number of reads spanning the CpG.
6. The mean methylation of CpG across all considered reads. This might slightly deviate from the value that can be calculated by standard methylation calling since RLM excludes certain reads that might be considered by standard methylation callers such as reads with indels, low quality reads, etc.
7. The coverage defined as the number of reads considered that span the CpG.

### 3.4 All output

When choosing “all” as score option, all three output files (single read, entropy and pdr) will be created.

## 4 Comparison with other tools

Existing tools that perform read-level analysis are rare and frequently limited by their universal usability. In the following we provide an overview about the features of each tool as well as runtime benchmarks against the most comparable tool (WSH, R package) (Scherer *et al.*, 2020).

### 4.1 Features

The different tools available for read-level analysis differ substantially in the scores they provide. While DMEAS provides entropy only, CluBCpG exclusively provides a clustering-based read-level analysis (He *et al.*, 2013; Scott *et al.*, 2020). WSH is the only other tool that summarizes a variety of available read-level scores. While all tools support only BISMARK alignments (at least for some scores), none of the existing tools provides an RRBS-specific mode that allows to ignore potential artificial bases.

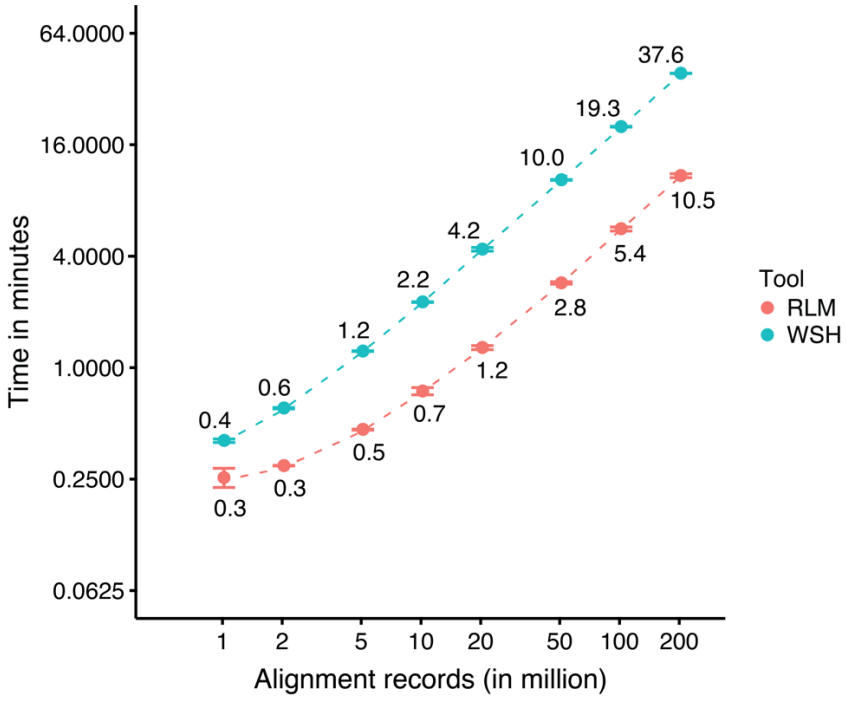
| Tool | Scores | Compatible alignment tools | Reference genome | RRBS mode |
|---|---|---|---|---|
| RLM | Single read discordance and transitions<br>Entropy<br>Epipolymorphism<br>PDR<br>RTS<br>Matching mean methylation per score | BISMARK (Krueger and Andrews, 2011)<br>BSMAP(Xi and Li, 2009)<br>segemehl (Otto *et al.*, 2012)<br>GEM(Marco-Sola *et al.*, 2012) | Any | yes |
| DMEAS | Entropy | BISMARK | Any | no |
| CluBCpG | Clustering-based read-level analysis | BISMARK | Any | no |
| WSH | Entropy<br>Epipolymorphism<br>PDR<br>MHL<br>FDRP<br>qFDRP | BISMARK only for entropy and epipolymorphism | preferably hg38 | no |

RLM filters potentially biasing reads based on multiple criteria and merges overlapping mates of the same pair (see **1**). Additionally, for PDR/RTS and single-read evaluations, only reads with at least three CpGs are considered, while entropy and epipolymorphism calculations require reads with at least four CpGs. This reduces the number of reads considered for the read-level analysis in

comparison to the reads considered for general methylation rate calling. In order to enable combining read-level scores and methylation ratios precisely, RLM provides methylation rates per CpG (PDR/RTS) or per 4-mer (entropy/epipolymorphism) based exclusively on the reads which are used for the read-level analysis.

### 4.2 Runtime comparison

We chose to compare runtime of RLM with WSH since it is the most similar tool, offering a variety of scores that are also provided by RLM. We compared the runtime of both tools for the calculation of entropy since WSH requires additional input for PDR that would need to be computed separately (the exact position of CpG where PDR should be calculated for). RLM was executed in single-end mode since WSH does not offer a specific paired-end mode. We measured the performance by randomly sampling increasing numbers of BAM records from a publicly available data set (GSM3618718) aligned to the reference genome hg38 using BISMARK. Benchmarks were executed on an Intel Xeon 6242 @ 2.80GHz with test data located on NFS mounted file servers and measurements are reported as the average across five runs with standard deviation. For 10 million BAM records and more, RLM finishes entropy calculations more than three times faster than WSH.




### 4.3 Score comparison

Due to the filtering steps embedded in RLM prior to the score calculations, the number of reads considered for read-level scores decreases. Taking one of the examples above (**4.2**) containing around 50 million reads, 8.2 million reads fulfill the criteria of at least three CpGs on one read. Of these, only 82% end up in the RLM single-read output. The remaining reads get filtered out due to e. g. mismatches at CpG positions. WSH does not filter reads that go into the analysis which makes the resulting scores from the same BAM file not comparable between the two tools as different numbers of reads go into each calculation. We therefore do not provide comparisons of the results of both tools here.

In order to ensure correctness of the RLM output, we included extensive tests in the RLM GitHub repository. These tests cover a wide range of use cases such as the calculation of methylation ratio, entropy, epipolymorphism, PDR and RTS but also the correct usage of alignment files from all three supported alignment tools, WGBS and RRBS mode. Additionally, tests for the correct filtering of reads

that should be excluded from the analysis and merging of overlapping mates are included here. All tests are automated using continuous integration.

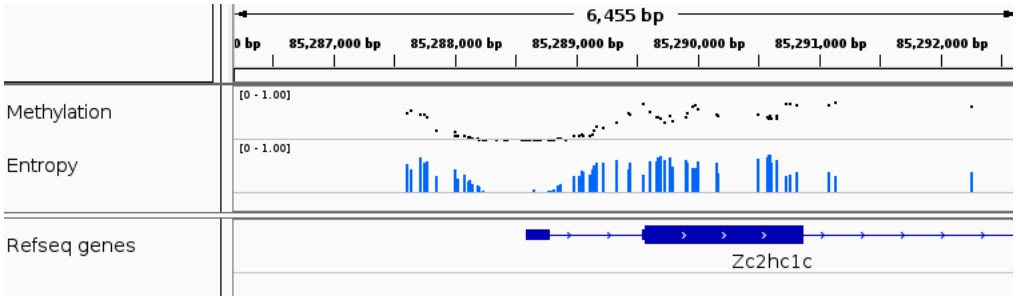
## III Post-processing and use cases

RLM is a standalone application that processes a BAM file and returns output files containing methylation information and read-level metrics for and based on single reads. These output files can be used for downstream processing using common command line tools such as bedtools or UCSCtools and visualized in R. In the following, we describe the processing of an example file using RLM and simple potential steps to visualize the read-level information. We also provide a R Markdown script within this repository that creates global and - if desired - feature-wise summary statistics based on the RLM output.

#### 1. Browser tracks

Entropy, epipolymorphism, PDR and RTS can be visualized per 4-mer or CpG e.g. using UCSCtools:

```
cut -f 1,2,3,4 output_entropy.bed | sed '1d' | sort -k1,1 -k2,2n > entropy.bedgraph
bedGraphToBigWig entropy.bedgraph genome.chrom.sizes entropy.bw
```




Note: Entropy and epipolymorphism will be reported using the coordinates of the first CpG in a 4-mer.

#### 2. Aggregation per feature

Scores reported per CpG or 4-mer can be simply aggregated per feature e.g. using bedtools. The following example code shows how to calculate the mean entropy across a bed file with regions of interest (3 columns: chr, start, end):

```
bedtools intersect -a regions.bed -b output_entropy.bed -wa -wb | \
sort -k1,1 -k2,2n | \
bedtools groupby -i stdin -g 1,2,3 -c 7 -o mean > mean_entropy_regions.bed
```

#### 3. Visualization with R

The R script we provide requires R to be installed including the following packages:

- knitr
- data.table
- GenomicRanges
- RColorBrewer
- vioplot
- ggplot2
- ggpubr

The script can be called the following way:

```
Rscript -e "rmarkdown::render('summarize_read_level_stats.Rmd',
```

```
params=list(
single_read_input_file = '/path/to/output_single_read_info.bed',
pdr_input_file = '/path/to/output_pdr.bed',
entropy_input_file = '/path/to/output_entropy.bed',
sample_name = 'my_sample',
feature_input_file = '/path/to/features.bed'),
output_file = 'my_output.pdf')"
```

The parameter `feature_input_file` is optional and if not provided, no feature-wise figures will be reported. If provided, it should be a bedgraph file of the following format: `<chr> <start> <end> <feature_name>` where `<feature_name>` should be the name of the feature type the region belongs to. For example, a feature file separating the genome into CpG islands (CGIs), CpG island shores (2kb upstream and downstream of a CGI), CpG island shelves (2kb upstream and downstream of the shores) and open water regions (the remaining parts) could look like the following:

```
chr1	0	3527624	OpenWater
chr1	3527624	3529624	CGIshelf
chr1	3529624	3531624	CGIshore
chr1	3531624	3531843	CGI
chr1	3531843	3533843	CGIshore
chr1	3533843	3535843	CGIshelf
chr1	3535843	3666619	OpenWater

...
```

The script will output a report containing basic statistics and figures of the single reads considered for the analysis, the reported PDR/RTS scores and the entropy/epipolymorphism scores. In the following, we will illustrate and explain the type of figures produced by the script using a WGBS sample of mouse epiblast (embryo tissue at embryonic day E6.5) as an example (GSM4075619). **The following figures are directly extracted from the example report.**

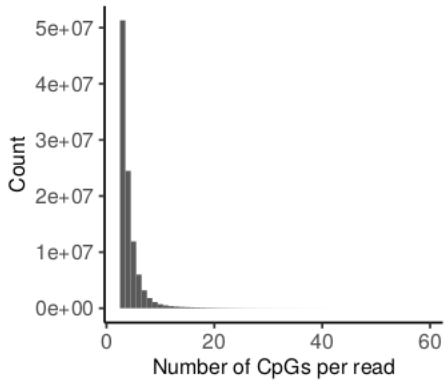
First, the report summarizes the number of reads (total and discordant) that passed all filtering thresholds and were considered for the score calculations. Additionally, statistics regarding the number of CpGs per read (all and methylated), the transitions and the methylation per read are reported. The WGBS sample covers the complete mouse genome with high coverage. The majority of mammalian genomes are highly methylated while CpG-dense regions (CpG islands) usually remain free of methylation. This is reflected in the per read statistics were most reads genome-wide tend to have few CpGs but are highly methylated with few transitions between methylated and unmethylated CpGs.

## Single read summary statistics

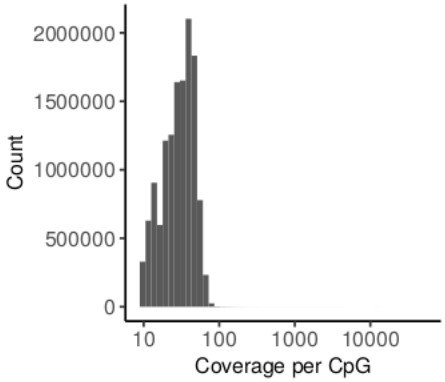
Number of reads considered for the analysis: 103170432.

Number of discordant reads considered for the analysis: 33408946.

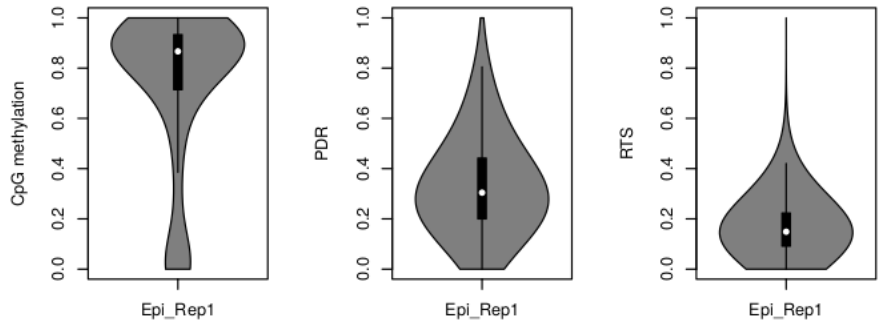
| | Summary CpGs per read | Summary methylated CpGs per read | Summary transitions per read | Summary methylation per read |
|---|---|---|---|---|
| Min. | 3.00 | 0.0 | 0.00 | 0.00 |
| 1st Qu. | 3.00 | 2.0 | 0.00 | 0.67 |
| Median | 4.00 | 3.0 | 0.00 | 1.00 |
| Mean | 4.29 | 3.1 | 0.17 | 0.78 |
| 3rd Qu. | 5.00 | 4.0 | 0.33 | 1.00 |
| Max. | 59.00 | 42.0 | 1.00 | 1.00 |



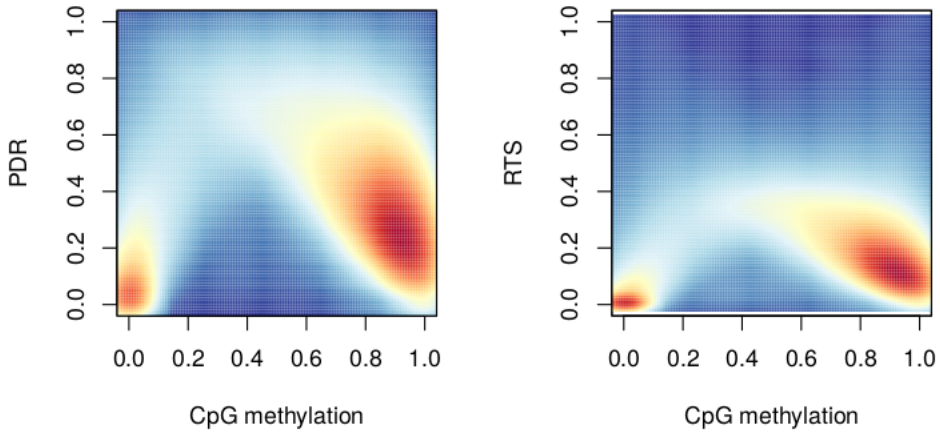

For PDR and RTS scores, the report summarizes the distribution of the coverage of reported CpGs. Additionally, violin plots visualizing the distribution of CpG methylation, PDR and RTS are provided. Some genomic regions artificially tend to get high amounts of reads during mapping which is why removing outlier CpGs with abnormally high coverage could be considered.



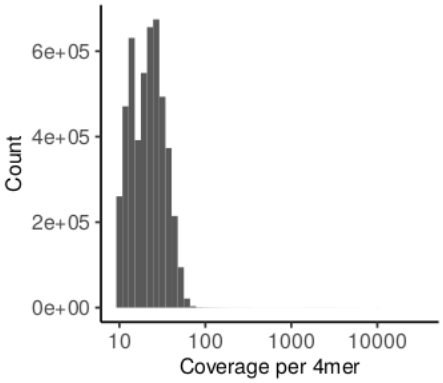

Generally, PDR is a very broad measure and will usually be higher than the RTS because a discordant read can have an arbitrary amount of transitions as long as there is at least one.

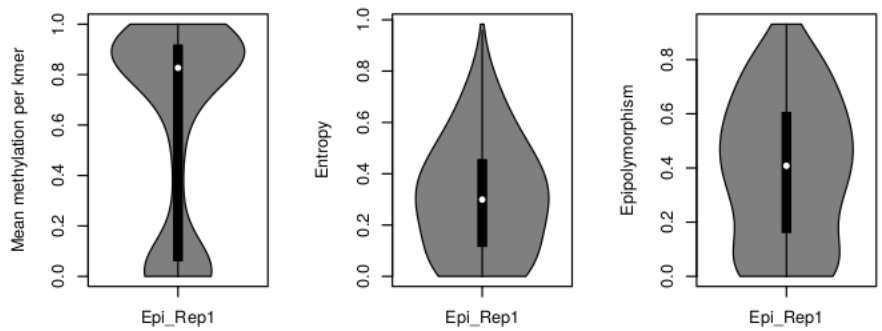



The relationship between methylation and PDR/RTS is summarized using smooth scatter plots where red refers to regions of high density while blue marks regions with low density. Regions with low and high DNA methylation in the genome will tend to have lower read level dynamics (i.e. low PDR or RTS) while intermediately methylated regions tend to have higher heterogeneity.

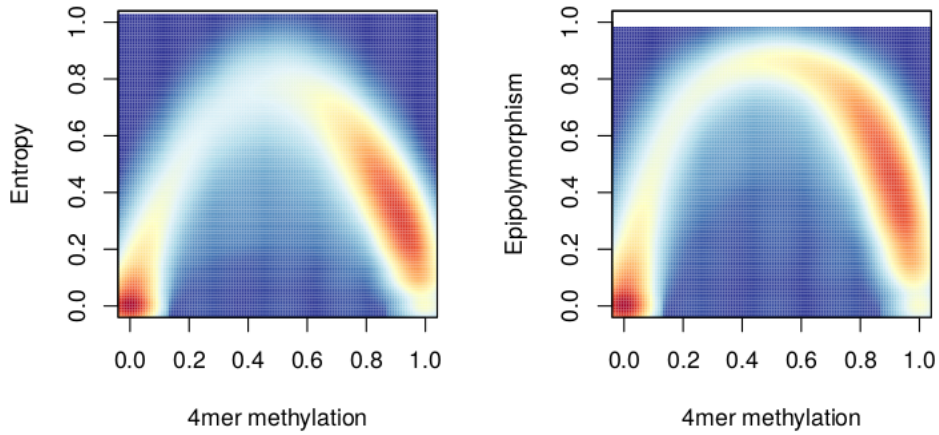



The report offers the same type of figures for entropy and epipolymoprhism scores. Importantly, here methylation and read level scores refer to a 4-mer of CpGs instead of a single CpG. The methylation per 4mer is calculated as the average methylation across all CpGs included. This average methylation will usually be lower compared to the CpG methylation reported PDR/RTS. For entropy and epipolymorphism, 4 CpGs in a row (4mer) are required to be present on a single read in contrast to the minimum of 3 CpGs that is required for PDR/RTS analysis. This biases the analysis towards CpG-rich regions such as CpG islands which tend to be unmethylated in mammalian genomes.

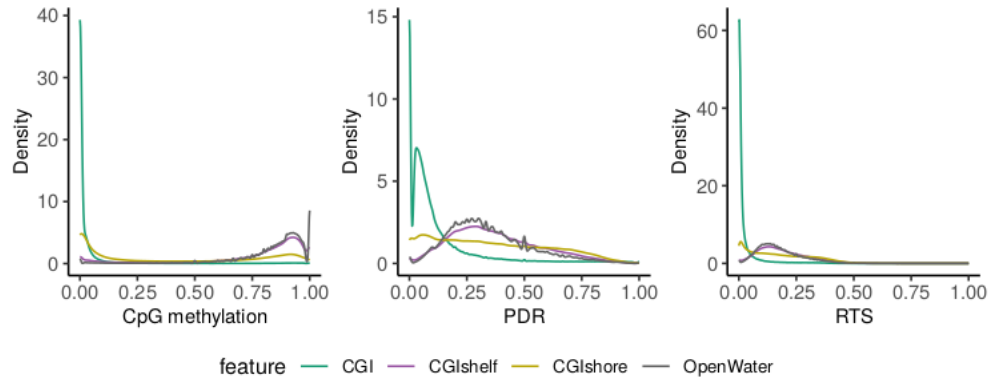


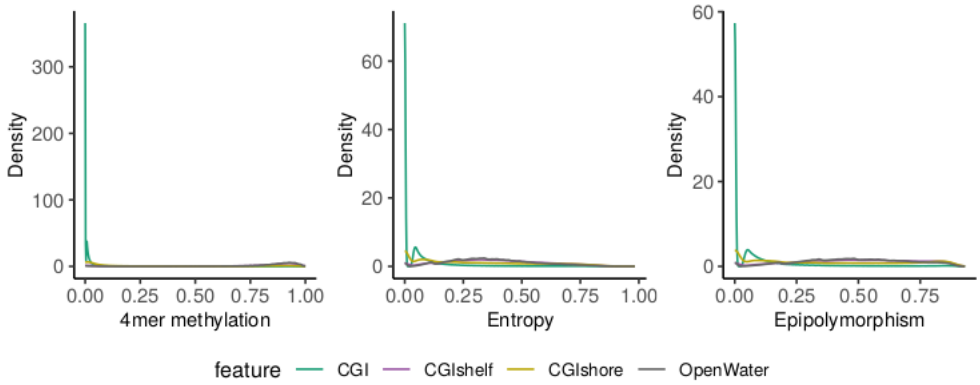



Again, regions with low and high DNA methylation in the genome will tend to have lower read level dynamics (i.e. low entropy or epipolymorphism) while intermediately methylated regions tend to have higher heterogeneity.




The script also offers the possibility to compare the distribution of methylation and read-level scores for different feature groups. This could be genomic regions such as CGIs or promoters but also differentially methylated regions (DMRs), imprinted regions, etc. As an example we compared CGIs with CGI shores, shelves and open water regions. As expected CGIs show low methylation levels which are accompanied by low heterogeneity based on read level scores. However, a group of CpGs/4-mers in CGIs shows higher read level scores indicating a different type of regulation of these regions. The lower the CpG density (i.e. going from CGIs to shores to shelves/open water) the higher the methylation accompanied by a rise in heterogeneity.






**4. Use case**

Read level analyses can be useful in a variety of applications where the pure methylation rates are not informative enough and specific interest regarding the underlying population dynamics exists. Scherer et al. provide a useful summary of potential use cases, applications and concepts. Generally, read level analysis might be useful for questions related to:

Tissue heterogeneity
Comparison of different conditions that are known to differ based on DNA methylation (this could be healthy and tumor tissue, wild-type and knockout contexts, etc.)
Contamination of a cell population
Methylation dynamics over time (such as in early mammalian embryonic development where DNA methylation is almost completely erased and re-established)
Allele-specific methylation (i.e. imprinting)
Comparing different features of interest such as DMRs or other genomic features not only based on average methylation differences but also regarding changes in underlying populations)

He,J. *et al.* (2013) DMEAS: DNA methylation entropy analysis software. *Bioinformatics*, **29**, 2044–2045.
Krueger,F. and Andrews,S.R. (2011) Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
Landan,G. *et al.* (2012) Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.*, **44**, 1207–1214.
Landau,D.A. *et al.* (2014) Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, **26**, 813–825.
Marco-Sola,S. *et al.* (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–1188.
Otto,C. *et al.* (2012) Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, **28**, 1698–1704.
Scherer,M. *et al.* (2020) Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.*, **48**, e46–e46.
Scott,C.A. *et al.* (2020) Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.*, **21**.
Xi,Y. and Li,W. (2009) BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**.
Xie,H. *et al.* (2011) Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.*, **39**, 4099–4108.