Research &
Pharma Solutions

Methylation Sequencing

Bioinformatic Note

Methylation Sequencing

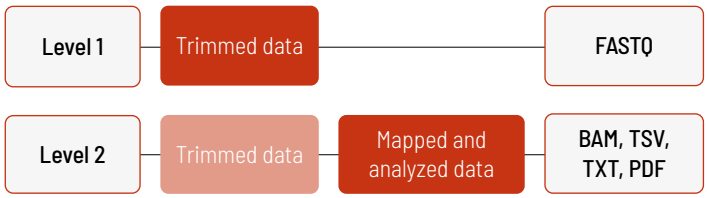
DNA methylation is one of the most common epigenetic modifications that fundamentally influence gene expression, cellular differentiation, and genomic imprinting. Without any change in the DNA sequence itself, gene activity and function can be regulated by DNA methylation. DNA methyltransferases mediate the regulation of gene activity and function by transferring a methyl group to the fifth carbon of the cytosine ring. In mammals, the resulting 5-methylcytosines (5-mC) and 5-hydroxymethylcytosines (5-hmC) occur mainly in cytosine-phosphate-guanine (CpG) dinucleotides. However, methylation can also be found in non-CpG contexts in other organisms.

Changes in the epigenetic signature, especially in DNA methylation, have been reported to happen in normal cell development and aging. However, alterations in DNA methylation are also closely associated with diseases like cancer, metabolic disorders, and neurological diseases. Global hypomethylation and locus-specific hypermethylation of CpG islands have been shown to increase genomic instability and promote tumor progression.

High-quality methylation data analysis can be used for:

- Biomarker discovery
- Clinical studies with methylation-associated treatments or other clinical and scientific applications
- Exploration of cell differentiation mechanisms, characteristic methylation profiles, and specific tissue development

Different levels of bioinformatic data analysis are available:



With increasing bioinformatic level, more data are delivered. The higher level includes the data from the lower level. Thus, in Level 2, trimmed data as well as mapped and analyzed data are provided. In addition to the data, and independent of the analysis level, a project report is generated.

Level 1

If you wish to analyze your data yourself, we recommend Level 1. At this level, trimmed reads in FASTQ format are delivered. This level is provided for every project, regardless of additional bioinformatics analysis purchased.

The additionally generated project report provides information for every sample about the laboratory protocol, including data about quality control of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing. For the trimmed data, the number of sequenced fragments and bases is reported, and the sequence length, quality of the reads and the GC content are illustrated in bar plots for all samples.

Level 2

In Level 2, the trimmed reads are mapped and analyzed. In this example, the reads are mapped to the reference genome hg19 and the positive and negative control (pUC19 and lambda phage). Duplicates are marked and removed.

Mapped Data

In addition to the trimmed reads in FASTQ format, you will receive the mapped reads as BAM files.

Non-methylated cytosines (C) are converted to uracils (U) during the enzymatic treatment of the DNA samples, while methylated Cs are unaffected (Vaisvila *et al.* 2021). Consequently, the mapping of the obtained reads is not straightforward, as reads with high numbers of converted Cs are unlikely to map against the reference (Krueger *et al.* 2011). Therefore, the mapping of reads obtained by a directional protocol is conducted in several steps to increase the overall mapping efficiency. An algorithm is used to determine whether a read stems from the original top (OT) or the original bottom (OB) strand.

The mapping efficiency is then calculated. The mapping efficiency, mapping rate, duplication rate, average genome coverage, and average CpG coverage is indicated in a table in the project report.

Additional mapping metrics are provided in a TSV file. Amongst others, this metrics file contains information about the number of input and unique reads, reads with and without mate sequenced, and information about the mapping quality. Every entry consists of a section description, its value as count or ratio, and, where applicable, the percentage.

A TSV file gives information about the mean coverage of the contigs and chromosomes. In this file, the first column indicates the contig name, the second column the number of bases aligned to that contig (excluding bases from duplicate marked reads, reads with MAPQ=0, and clipped bases), and the third column the estimated coverage. An excerpt of such a TSV file is given in table 1.

Table 1 | Excerpt of the contig mean coverage TSV file (header line added for the sake of clarity

Contig name	Number of aligned bases	Estimated coverage
chr1	6136452339	27.2391
chr2	6520527355	27.3737
chr3	5386644971	27.6526
chr4	5161639638	27.5050

Another TSV file includes information about the coverage metrics. Amongst others, this TSV file lists the number of aligned bases, the average alignment coverage over the genome, and the uniformity of the coverage over the genome. As for the mapping metrics, every entry consists of a section description, its value as count or ratio, and, where applicable, the percentage.

In addition, the percentages of the genome with a specific coverage range are listed, as indicated in table 2. While the coverage metrics file includes additional information about the coverage described above, the TSV format histogram file only provides information about the percentages of bases within a specific coverage range. This information is extracted from the coverage metrics file. Therefore, the file content looks like the one shown in table 2.

Table 2 | Excerpt of the coverage metrics TSV file (header line added for the sake of clarity and only valid for this part of the file).

	Range	Percentage
PCT of genome with coverage	[100x: inf)	0.04
	[50x: inf)	0.25
	[20x: inf)	87.80
	[15x: inf)	94.99
	[10x: inf)	96.70
	[3x: inf)	97.71
	[1x: inf)	98.08
	[0x: inf)	100.00
	[50x:100x)	0.21
	[20x: 50x)	87.55
	[15x: 20x)	7.18

Analyzed Data

After the mapping step, the data are analyzed.

Three cytosine reports are delivered in txt format: (1) the CpG report for cytosines in CpG context, (2) the non-CpG report for cytosines in non-CpG context, and (3) the controls report for cytosines in CpG context for conversion control. The cytosine report contains information on every single cytosine in the genome, considering both strands.

As all three files are structured the same, only an excerpt of such a report file is shown in table 3. The chromosome, position, and strand are given, together with the number of methylated and unmethylated cytosines at this position. Additionally, the cytosine context and the trinucleotide context are given.

When using whole genome methylation sequencing, it is important to assess the accuracy of the conversion efficiency for each sample, answering the question of whether all non-methylated Cs have indeed been converted while all methylated Cs have not. For this, we are using the genomes of the fully unmethylated lambda phage and the fully methylated pUC19 as spike-ins for each sample (Morrison *et al.* 2021). With these spike-ins, the conversion rate and the methylation rate are calculated. Low conversion rates indicate inaccurate conversion, leading to overestimation in the methylation analysis. Low methylation rates indicate inefficient conversion during the conversion step, leading to underestimation in the methylation analysis.

The conversion rates based on the methylation of the lambda phage (negative control) and the methylation rate based on the methylation of pUC19 (positive control) are provided in a TSV file. For every sample, the conversion rate, the methylation rate, and the coverage on the lambda phage and pUC19 are indicated, as shown in table 4. This table can also be found in the project report.

Table 3 | Excerpt of the CpG reports TXT file (header line added for the sake of clarity).

Chromosome	Position	Strand	# of methylated Cs mapping to this position	# of unmethylated Cs mapping to this position	C context (CpG, CHG, CHH)	Trinucleotide context
chrM	33	+	2	801	CG	CGG
	34	-	1	857	CG	CGT
	61	+	0	1825	CG	CGT
	62	-	3	1689	CG	CGA
	78	+	3	2431	CG	CGC
	79	-	3	2185	CG	CGT
	80	+	2	2471	CG	CGA
	81	-	2	2244	CG	CGC
	91	+	4	2750	CG	CGA
	92	-	2	2515	CG	CGC

Table 4 | Conversion rates TSV file.

CeGaT ID	conversion rate (lambda phage)	methylation rate (pUC19)	coverage lambda phage	coverage pUC19
emethylseq_S1163Nr701	99.87	96.49	124.83	104.393

Cytosines in the genome can occur in different contexts: CpG, CHG, CHG, where H stands for A, T, or G. The overall percentage of methylated cytosines in the different contexts is indicated in a table in the project report. The cytosine coverage in these three contexts is also provided as TSV file. An excerpt is shown in table 5. For various coverages, the percentages of cytosines in the CpG, CHH, and CHG context are indicated. This data is additionally visualized in a coverage plot in PDF format, as shown in figure 1.

Table 5 | Excerpt of the cytosine coverage data in TSV format.

coverage	CpG	CHH	CHG
1	2.633715	2.616423	2.711778
5	0.223761	1.372066	1.438117
50	0.009644	0.001524	0.001580
100	0.133771	0.029987	0.031012
>100	0.055492	0.007974	0.007284

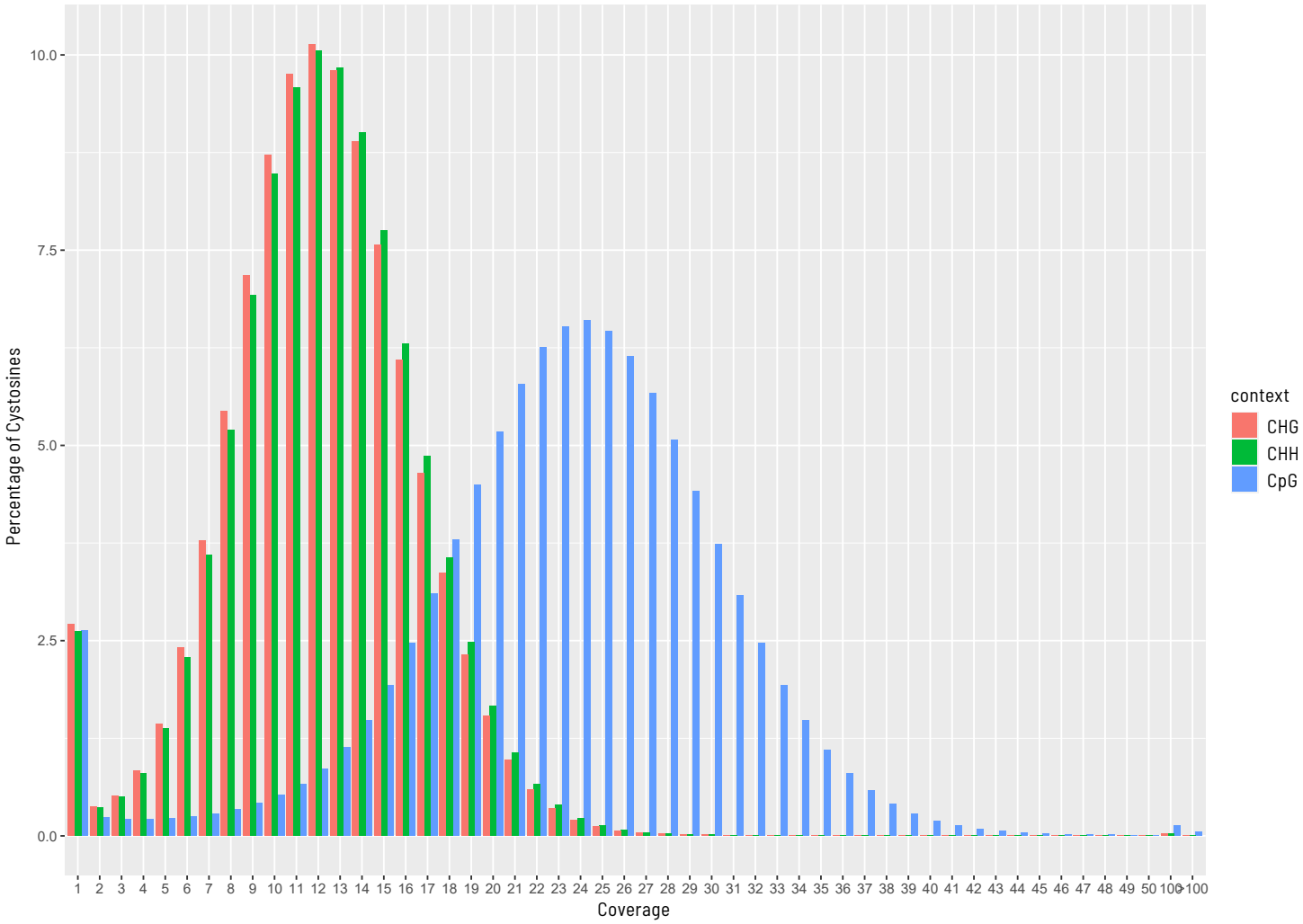


Figure 1 | Cytosine coverage plot.

Methylation profiles of C's in these contexts differ significantly between, e.g., mammals and plants but also between different tissues and between healthy and cancer samples (Law *et al.* 2010). As most samples contain a mix of different cells and maybe even tissues, methylation values follow a bimodal distribution where most C's have either low (0%) or high (100%) methylation.

The methylation profile in the different contexts is provided in a TSV file, which is also shown in table 6. In the project report, the percentages of each context are provided in a separate table. Additionally, the methylation profile is visualized in a bar plot in PDF format, as shown in figure 2.

Table 6 | Methylation percentages in different contexts

percent_methylated	CpG	CHH	CHG
5	24.4546284367836	99.724706196277	99.6914907584202
15	6.16238986649709	0.213521874059844	0.24162303490143
25	5.01059299410357	0.0205325856409485	0.020704689960124
35	5.49987431609235	0.0135573091385415	0.013392664591491
45	6.2652400151512	0.0100815141623037	0.0105525586350432
55	4.76470622691816	0.00416517963841758	0.00461914139947316
65	5.63600368614583	0.0032458822453126	0.00369622036992036
75	7.43097144728667	0.00248343459329501	0.00312465265456728
85	10.5354349123445	0.00222547510175459	0.0031790876750771
95	24.240158098677	0.00548054914256292	0.00761719139270322

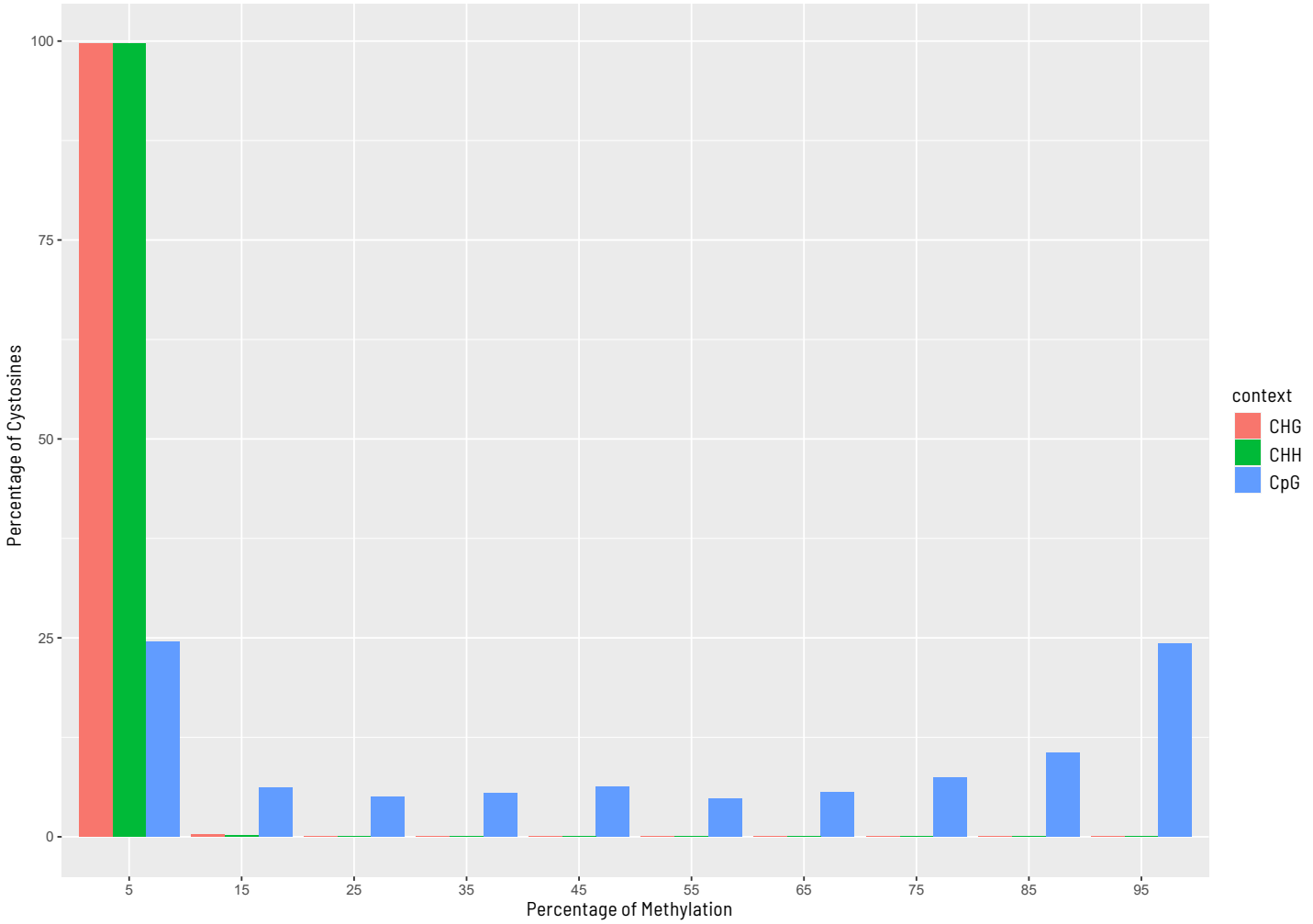


Figure 2 | Methylation profile in different contexts.

Methylation statistics, such as the mapping efficiency, or the total number of C's analyzed, are provided in TSV format. Every entry in this file consists of a section description, its value as count or ratio, and, where applicable, the percentage.

The methylation proportion across each possible position in the read is provided as an M-bias report in TXT format. This report contains a table for each C-context. Each table consists of a series of records, where each record is a read base position. An excerpt of the CpG context M-bias file is shown in table 7. The counts of the methylated C's and unmethylated C's are given for every position. These counts are restricted to those reads in which the first base is aligned to the specific context, in our example, to a CpG location in the genome. The percentage of methylated C bases is given, as well as the sum of the methylated and unmethylated C counts, which is indicated as coverage in the file.

Table 7 | Excerpt of the CpG context M-bias file.

position	count methylated	count unmethylated	% methylation	coverage
1	1658817	1275081	56.54	2933898
2	1681334	1076955	60.96	2758289
3	1707806	1421905	54.57	3129711
4	1828572	1526794	54.50	3355366
5	1809420	1503156	54.62	3312576

References

Krueger, Felix; Andrews, Simon R (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics (Oxford, England), 27(11), 1571–1572.

Law, Julie A; Jacobsen, Steven E (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. 11(3), pp. 204–220.

Morrison Jacob; Koeman Julie M; Johnson Benjamin K *et al.* (2021): Evaluation of whole-genome DNA methylation sequencing library preparation protocols. Epigenetics Chromatin. 14(1): p. 28.

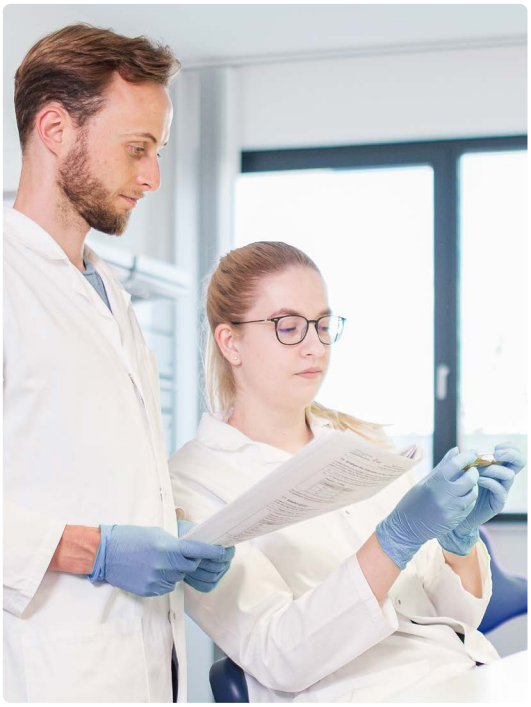
Vaisvila, Romualdas, *et al.* (2021): Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Research 31 (7), pp. 1280-1289.

About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00
Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

For more details please visit
www.cegat.com/rps