2025 School of Life Sciences, Peking University
Linux-based Essential Bioinformatics

Epigenetics data analysis

Group 7

Yufei Zhao, Yuchuan Wang,
Yuhua Wang, Shiqi Tang

Peking University

May 18th, 2025

Epigenetics data analysis

- Background & Introduction.
- Data analysis pipeline.
- Process of analysis.
- Results of analysis.
- Downstream analysis: expectations.

Epigenetics data analysis

- Background & Introduction.

Epigenetic modification

- Generality: whole lifespan
- Wide distribution: from virus to mammals
- Heritability: same genome but different offsprings

- Cancer: gene expression and function
- Anti-virus: game between viruses and host

DNA methylation

- DNA methylation: -CH3
- Mostly happens at cytosine
- Gene expression regulation, gene silencing, X chr deactivation, increasing genome stability, environment stress response, etc.
- Dysfunction and disease genesis and development

Datatype & data source

- Bisulfite-seq: 5mC
- Whole-genome Bisulfite-seq (WGBS)
- Reduced Representative Bisulfite-seq (RRBS)
- DRRBS, LHC-BS, oxBS, etc

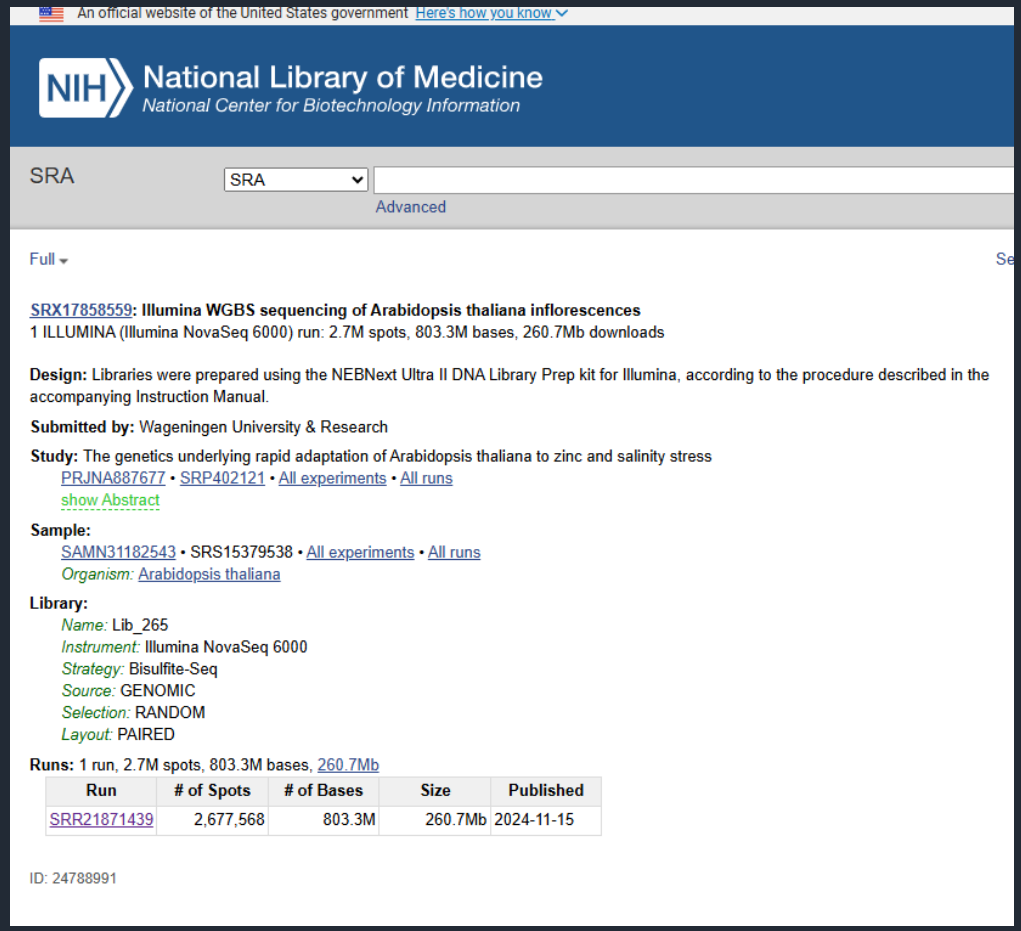
Epigenetics data analysis

- Background & Introduction.
- Data analysis pipeline.

Data acquisition

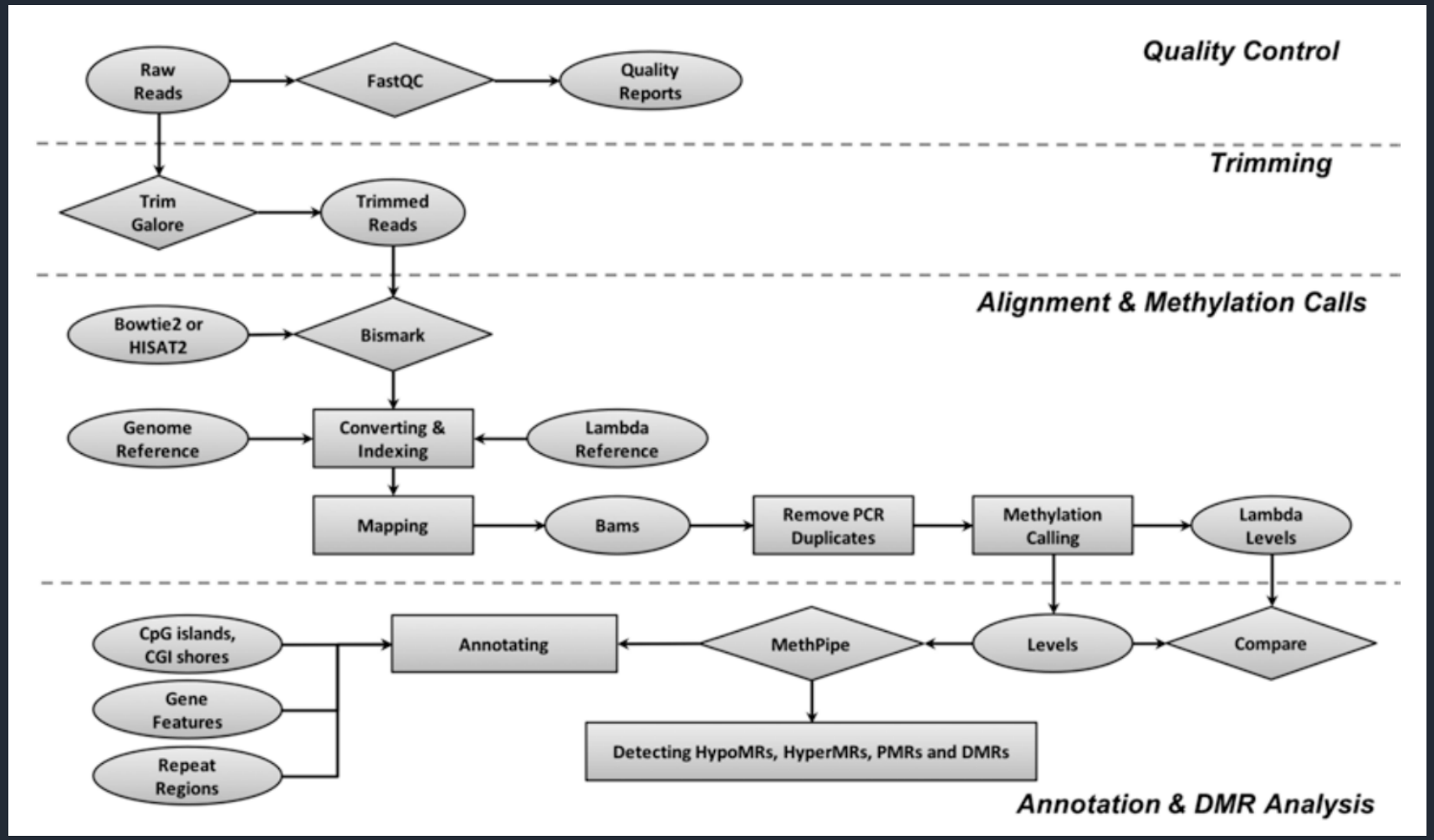
- Data: SRX28746177
- Data type: Bisulfite-Seq
- Species: *Arabidopsis thaliana*
- Sequencer: Illumina NovaSeq 6000
- Download by sra-tools

```
conda install -c bioconda sra-tools
fasterq-dump SRR33511656 --split-files -O ./fastq/
```



https://www.ncbi.nlm.nih.gov/sra/SRX28746177[accn]

Complete analysis pipeline



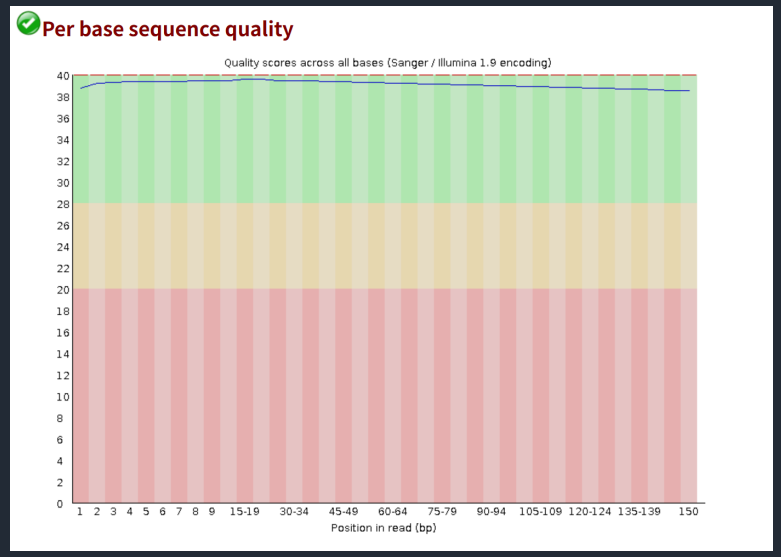
Alexey Ruzov and Martin Gering (eds.), *DNA Modiffcations: Methods and Protocols*, Methods in Molecular Biology, vol. 2198

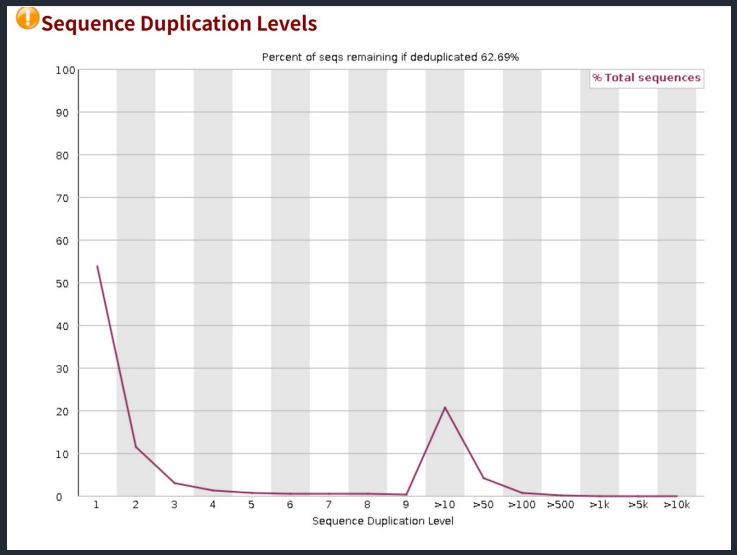
Epigenetics data analysis

- Background & Introduction.
- Data analysis pipeline.
- Process of analysis.

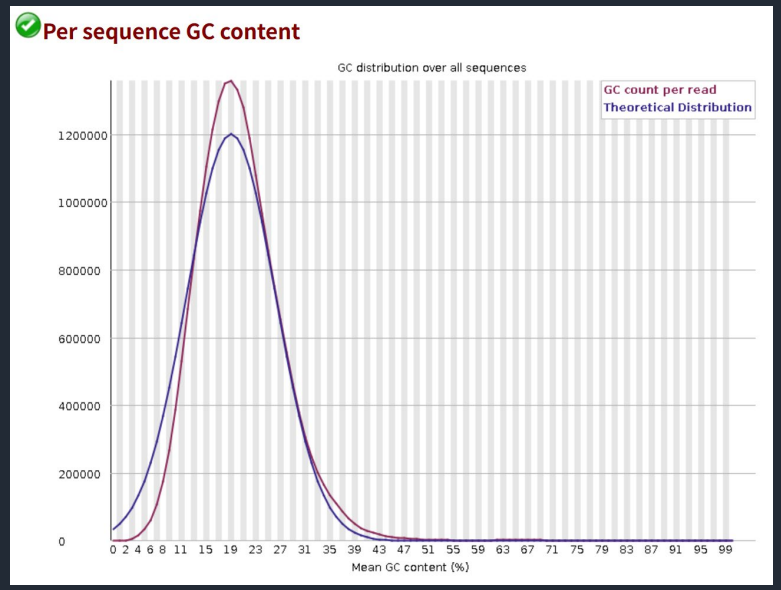
Quality control and trimming

```
$ conda install -c bioconda fastqc
$ fastqc -o qc_results/ fastq/
```





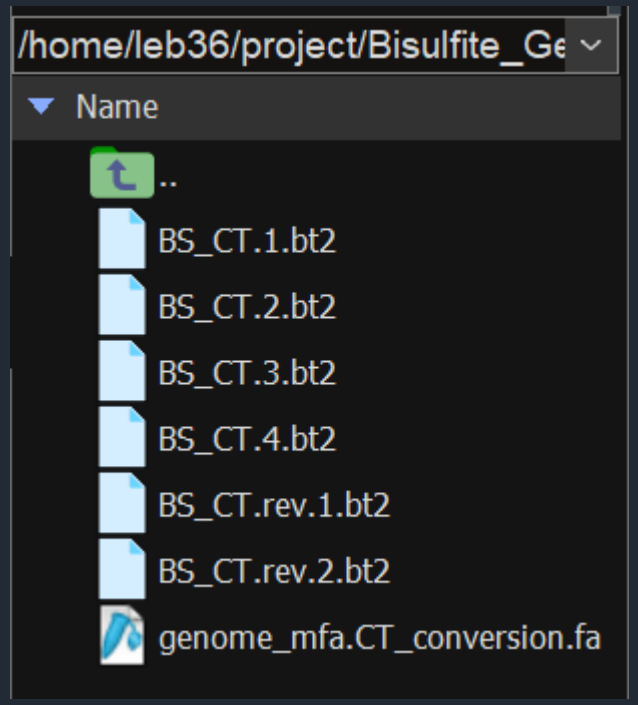

```
$ conda install -c bioconda trim-galore
$ conda install -c bioconda cutadapt  #确保cutadapt已经安
$ trim_galore --paired -j 8 -q 30 -o trimmed/ fastq/SRR3
```

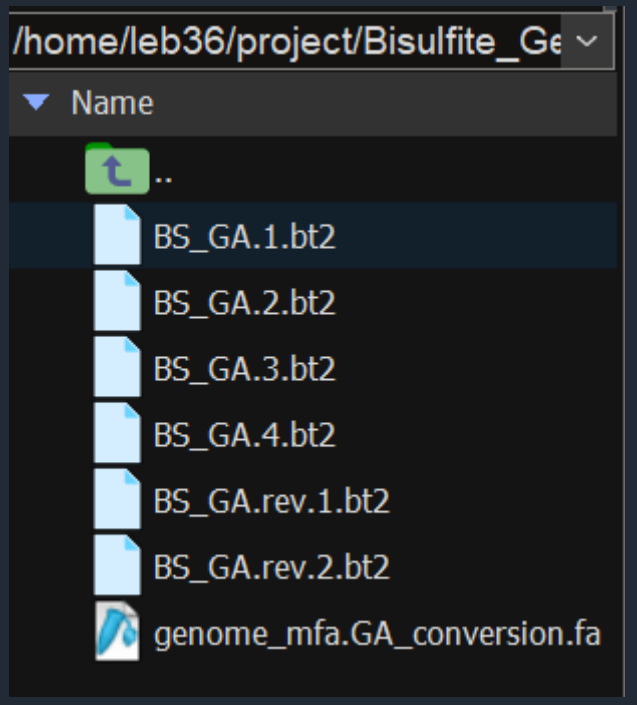



```
=== Summary ===

Total reads processed:              21,527,356
Reads with adapters:                 8,143,192 (37.8%)
Reads written (passing filters):    21,527,356 (100.0%)

Total basepairs processed: 3,229,103,400 bp
Quality-trimmed:              46,435,736 bp (1.4%)
Total written (filtered):  3,164,369,445 bp (98.0%)

=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 8143192 times

Minimum overlap: 1
No. of allowed errors:
1-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 40.2%
  C: 1.5%
  G: 16.3%
  T: 41.5%
  none/other: 0.6%

Overview of removed sequences
length  count   expect  max.err error counts
1       6139380 5381839.0       0       6139380
2       1255517 1345459.8       0       1255517
3       483180  336364.9        0       483180
```

Building reference genome & Bismark index

```
$ aria2c -x 16 -s 16 ftp://ftp.ensemblgenomes.org/pub/release-60/
$ conda install -c bioconda bismark
$ conda install -c bioconda bowtie2
$ bismark_genome_preparation  --bowtie2 --parallel 8 --verbose .
```





Alignment

```
dna_met2 $ bismark -p 16 --genome . \ #(reference基因组 bismark_genome_preparation 创建 index 的地方)
--bowtie2 -1 ./trimmed/SRR33511656_1_val_1.fq -2 ./trimmed/SRR33511656_2_val_2.fq -o ./bismark_output/
```

```
Final Alignment report
======================
Sequence pairs analysed in total:       21331686
Number of paired-end alignments with a unique best hit: 13839275
Mapping efficiency:     64.9%

Sequence pairs with no alignments under any condition:  5219672
Sequence pairs did not map uniquely:    2272739
Sequence pairs which were discarded because genomic sequence could not be extracted:    96

Number of sequence pairs with unique best (first) alignment came from the bowtie output:
CT/GA/CT:       6877703 ((converted) top strand)
GA/CT/CT:       0       (complementary to (converted) top strand)
GA/CT/GA:       0       (complementary to (converted) bottom strand)
CT/GA/GA:       6961476 ((converted) bottom strand)

Number of alignments to (merely theoretical) complementary strands being rejected in total:     0
```

Alignment

```
dna_met2 $ bismark -p 16 --genome . \ #(reference基因组 bismark_genome_preparation 创建 index 的地方)
--bowtie2 -1 ./trimmed/SRR33511656_1_val_1.fq -2 ./trimmed/SRR33511656_2_val_2.fq -o ./bismark_output/
```

```
Final Cytosine Methylation Report
=================================
Total number of C's analysed:   706084350

Total methylated C's in CpG context:    22257770
Total methylated C's in CHG context:    8909639
Total methylated C's in CHH context:    22670228
Total methylated C's in Unknown context:        698559

Total unmethylated C's in CpG context:  75966250
Total unmethylated C's in CHG context:  89144090
Total unmethylated C's in CHH context:  487136373
Total unmethylated C's in Unknown context:      818036

C methylated in CpG context:    22.7%
C methylated in CHG context:    9.1%
C methylated in CHH context:    4.4%
C methylated in Unknown context (CN or CHN):    46.1%
```

```
Bismark completed in 0d 2h 12m 57s
Bismark completed in 0d 2h 12m 57s
Bismark completed in 0d 2h 12m 57s
Bismark completed in 0d 2h 12m 57s
Bismark completed in 0d 2h 12m 57s
```

Deduplication

```
$ deduplicate_bismark --bam ./bismark_output/SRR33511656_1_val_1_bismark_bt2_pe.bam --output_dir ./deduplicated/
```

```
Total number of alignments analysed in ./bismark_output/SRR33511656_1_val_1_bis
mark_bt2_pe.bam:        13839179
Total number duplicated alignments removed:     2991457 (21.62%)
Duplicated alignments were found at:    2264010 different position(s)

Total count of deduplicated leftover sequences: 10847722 (78.38% of total)
```

Extraction of methylation data

```
dna_met2 $ bismark_methylation_extractor \
--paired-end \
--comprehensive \
--gzip \
--bedGraph \
--cytosine_report \
--report \
--genome_folder . \  #(reference基因组 bismark_genome_preparation 创
--output ./methylation_calls/ \
./deduplicated/SRR33511656_1_val_1_bismark_bt2_pe.deduplicated.bam
```

Extraction of methylation data

```
Processed 10847722 lines in total
Total number of methylation call strings processed: 21695444

Final Cytosine Methylation Report
=================================
Total number of C's analysed:   509996034

Total methylated C's in CpG context:    16485131
Total methylated C's in CHG context:    6676026
Total methylated C's in CHH context:    17261468

Total C to T conversions in CpG context:        54288013
Total C to T conversions in CHG context:        64380280
Total C to T conversions in CHH context:        350905116

C methylated in CpG context:    23.3%
C methylated in CHG context:    9.4%
C methylated in CHH context:    4.7%
```

position	count methylated		count unmethylated	% methylation c overage
1	47290	225646	17.33	272936
2	60576	190915	24.09	251491
3	56028	206749	21.32	262777
4	54620	215061	20.25	269681
5	57113	212082	21.22	269195
6	58039	215755	21.20	273794
7	54268	204108	21.00	258376
8	55440	208451	21.01	263891
9	55397	209448	20.92	264845
10	55994	209898	21.06	265892
11	56566	210916	21.15	267482
12	56230	211843	20.98	268073
13	56085	209655	21.11	265740
14	56668	209090	21.32	265758
15	55812	209901	21.00	265713
16	55998	208253	21.19	264251
17	55875	209855	21.03	265730
18	55546	210285	20.90	265831
19	55738	208818	21.07	264556
20	56501	210601	21.15	267102
21	56647	212146	21.07	268793

Epigenetics data analysis

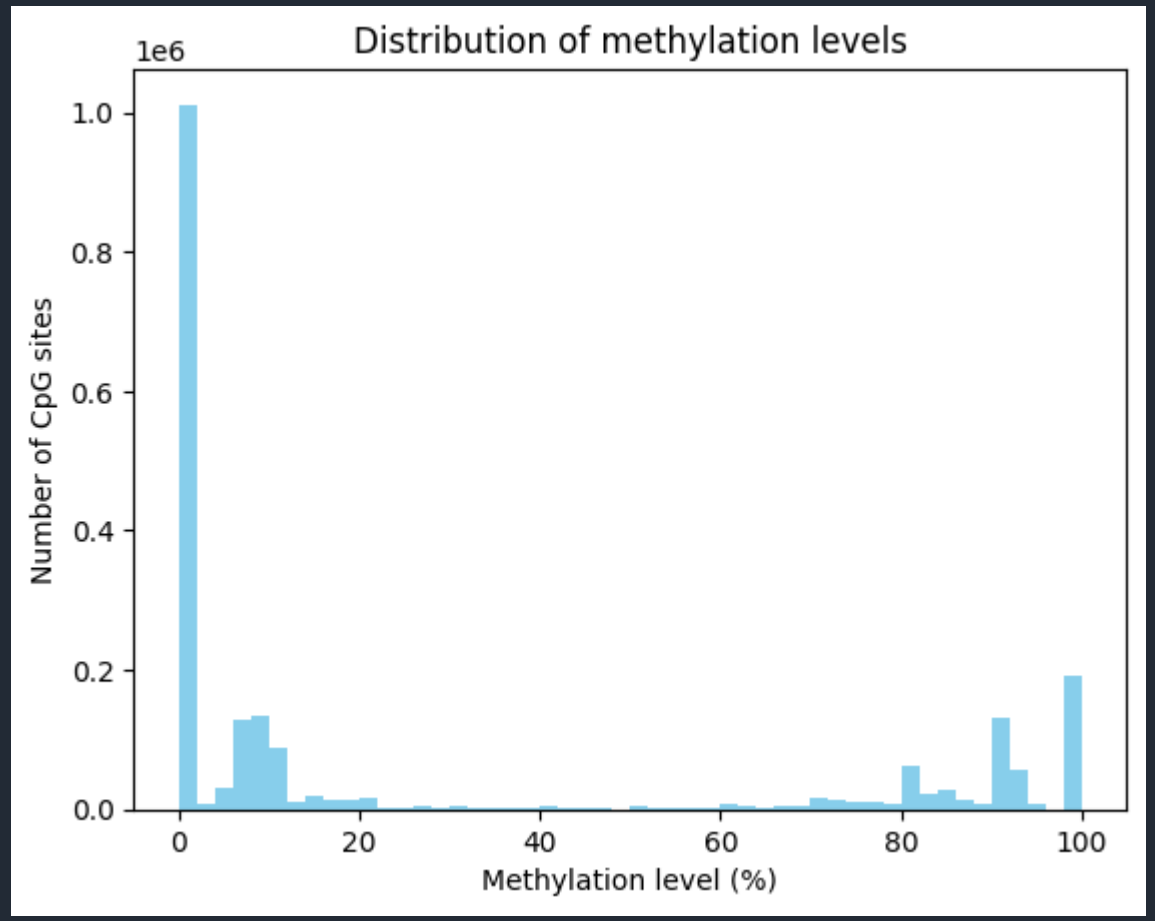
- Background & Introduction.
- Data analysis pipeline.
- Process of analysis.
- Results of analysis.

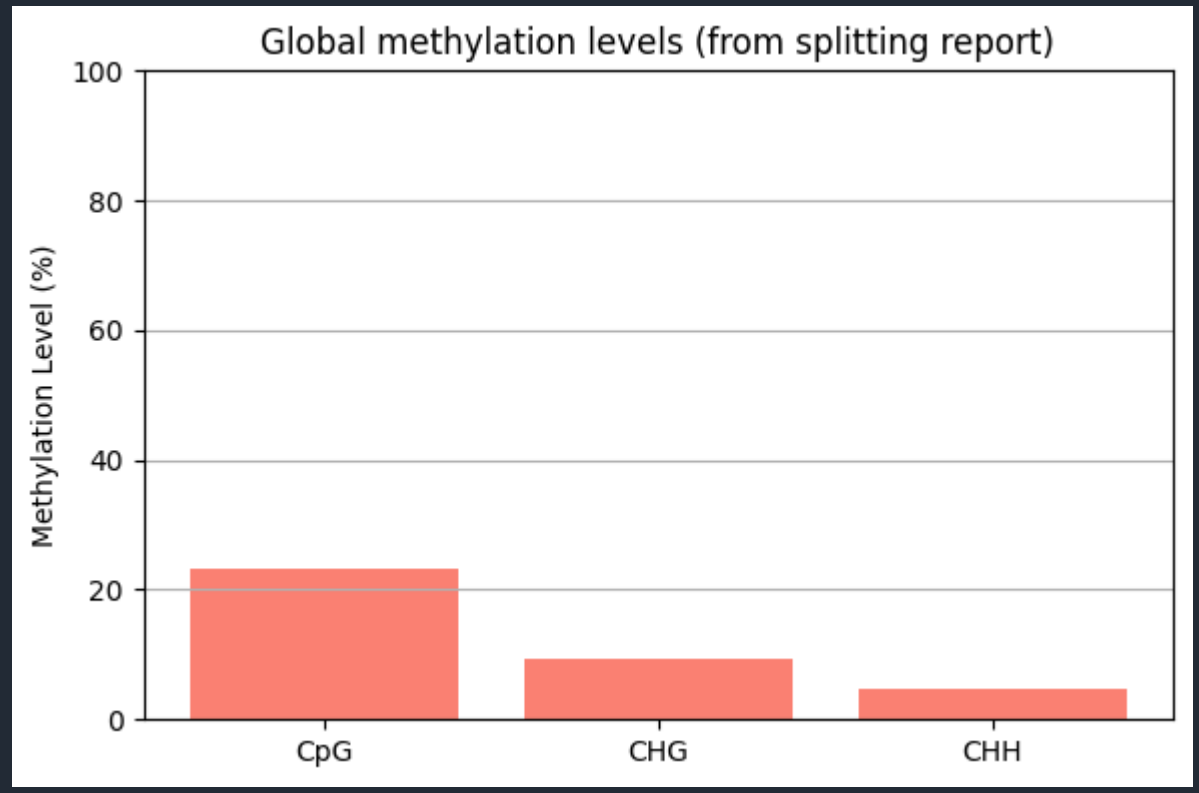
Methylation levels of CpG, CHG & CHH

```python
import pandas as pd #pip/conda install pandas

# 读入cov文件，列名根据bismark cov格式设定
cols = ['chrom', 'pos', 'meth_level', 'meth_count', 'unmeth_count']
file_path = 'methylation_calls/SRR33511656_1_val_1_bismark_bt2_pe.deduplicated.bismark.cov.gz'

#
df = pd.read_csv(file_path, sep='\t', names=cols, compression='gzip')

print(df.head())
```

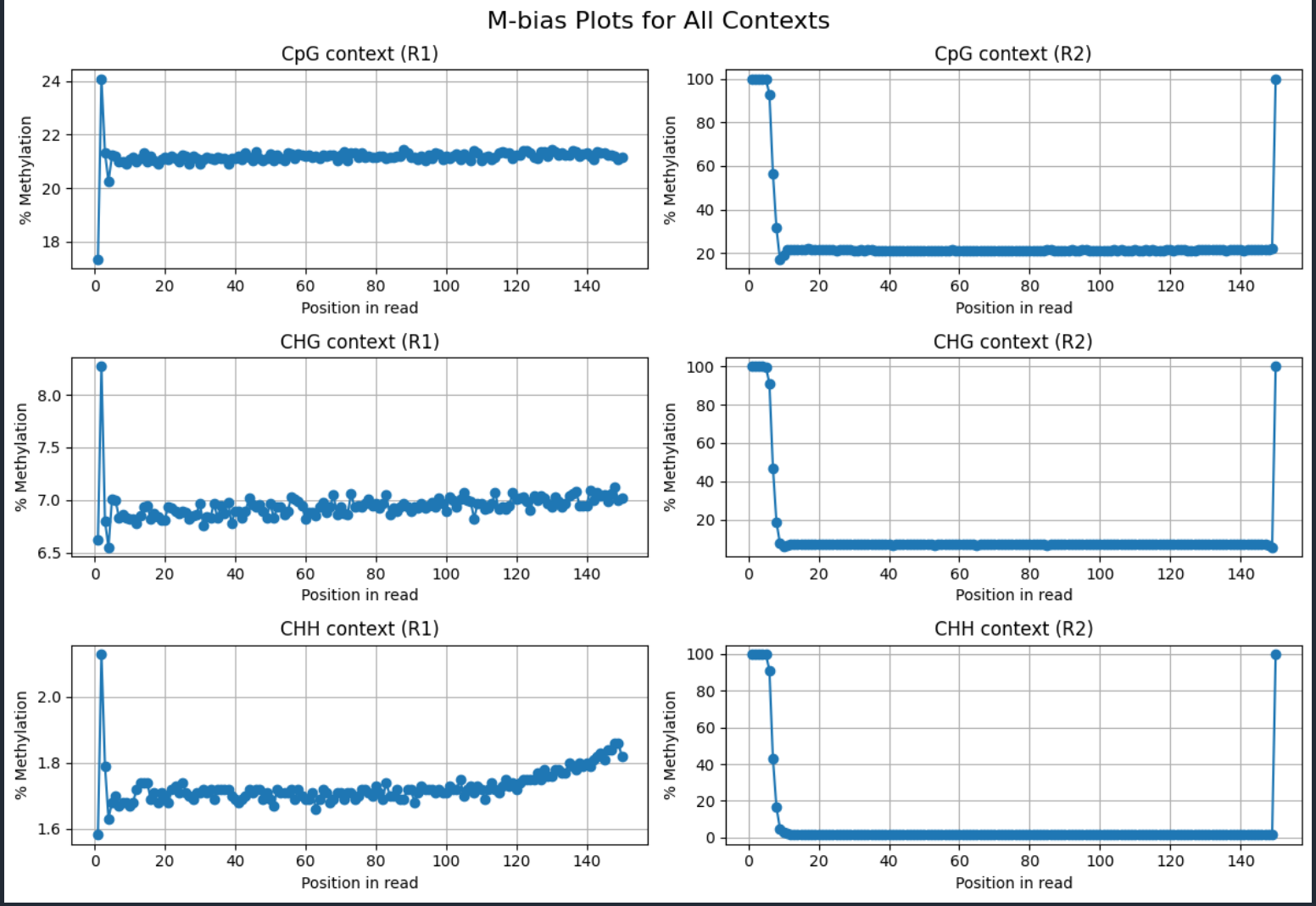
```
   chrom  pos  meth_level  meth_count  unmeth_count
1    109  109       100.0           3             0
1    110  110       100.0           4             0
1    115  115       100.0           3             0
1    116  116       100.0           4             0
1    161  161       100.0           9             0
```

Methylation levels of CpG, CHG & CHH

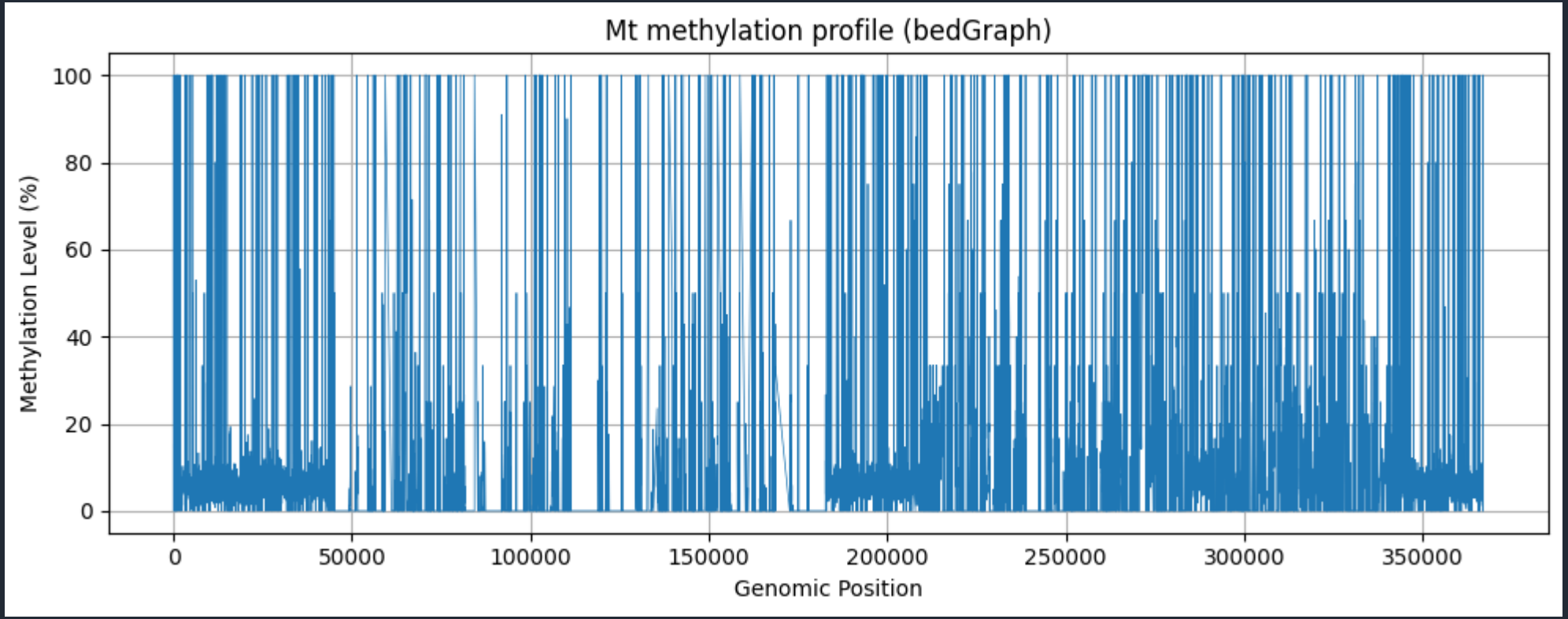




M-bias of different kinds of sites



Methylation levels in certain chr.



Epigenetics data analysis

- Background & Introduction.
- Data analysis pipeline.
- Process of analysis.
- Results of analysis.
- Downstream analysis: expectations.

DNA methylation and further research

- Gene expression regulation
- Embryogenesis and cell differentiation
- Genome structure and stability
- Disease genesis and possible treatment

Thank you for listening!

Questions welcomed!