AutoMethyc

Documentation

AutoMethyc is an integrative pipeline to methylation analysis from raw paired-end sequences obtained from massive parallel bisulfite sequencing.

1 Installation

1.1 docker

We created a docker container with all the necessary dependencies to run the program in order to provide a portable and self-sufficient container. To install it you need to have docker installed and then download the docker image.

Command 1: Download docker container

```
docker pull ambrizbiotech/automethyc
```

1.2 Local installation

For this installation option is necessary to install all the dependencies in the $PATH

Dependencies

- Bowtie2 v2.4.5
- Samtools v1.15.1-12
- Bismark v0.23.0
- python v3.10.6
  - pandas v1.5.2
  - numpy v1.23.1
  - plotly v5.10.0
  - plotly-express v0.4.1
  - scikit-learn v1.1.2
  - tqdm v4.64.1
  - IPython v8.4.0
  - pysam v0.19.1
- fastqc v0.11.9
- TrimGalore v0.6.6
- figlet v2.2.5
- multiqc v1.13
- git v2.34.1
- wget v1.21.2
- curl v7.81.0
- UnZip v6.0
- cutadapt v3.5
- java v11.0.18
- gatk v4.3.0.0
- R v4.1.2
  - gsalib v2.2.1
  - ggplot2 v3.4.2
  - reshape v0.8.9
  - gqplots v3.1.3
  - tidyverse v2.0.0
- revelio

And then move the files from the scr folder to the $PATH

Command 2: Moving the scripts

```
git clone https://github.com/FerAmbriz/AutoMethyc.git && cd AutoMethyc/scr
sudo mv * /usr/bin/
```

2 Usage

We provide a series of default values for simplicity when running with a single command.

Command 3: Running automethyc

```
automethyc -i [fastq_folder] -o [Output_folder] -r [reference genome file] [optional arguments]
```

But you can modify it according to the needs of the project

Command 4: Optional arguments

```
-t --threads  # Number of threads (default=4)
-n --normal   # Folder with fastq of normals (default=False)
-g --genome   # Genome used for request in UCSC (default=hg19)
-b --bed    # File with regions of interest (default=False)
-d --depth    # Minimum depth to consider (default=20)
-q --quality  # Minimum quality (default=30)
--read        # Read type in fastq (default=Paired)
```

In case you are using the version installed with docker, you have to mount the volume (-v) in the corresponding directory and run it interactively (-it).

Command 5: Running automethyc in docker interactively

```
docker run -v [/home]:[/home] -it ambrizbiotech/automethyc
automethyc -i [fastq_folder] -o [Output_folder] -r [reference genome file] [optional arguments]
```

or mount the volume in the corresponding partition and run it in the background (-d) to avoid breaking the process in long execution times

Command 6: Running automethyc in docker interactively

```
docker run -v [/home]:[/home] -d ambrizbiotech/automethyc automethyc -i [fastq_folder] -o
    [Output_folder] -r [reference genome file] [optional arguments]
# The output when executing this command is the "container ID" that will be running in the background.
    To see the execution progress use:
docker logs "container ID"
```

2.1 Format of bed file

The bedGraph file must contain the regions of interest, in order to filter non-specific sequencing products or regions of non-interest. The file format is comma separated values (CSV) with the chromosome, start and end, presenting different formats for greater versatility.

Chr	Start	End
chr10	89619506	89619580
chr11	22647545	22647849

Table 1: In range

Chr	Start	End
chr17	41277106	41277106
chr17	41277115	41277115

Table 2: Specific-site

Chr	Start	End	Gene
chr10	89619506	89619580	KLLN
chr11	22647545	22647849	FANCF

Table 3: With gene

3 Output and interpretation

The output is organized in 4 folders (Bismark, CSV, HTML, VCF).

3.1 Phred score

Base call error probability on logarithmic scale

$$Q = -10log_{10}P \tag{1}$$

3.2 Cgi mapping

- CpG island
- CpG shore
- CpG shelf
- CpG inter

3.3 Normalization

$$Z_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j} \tag{2}$$

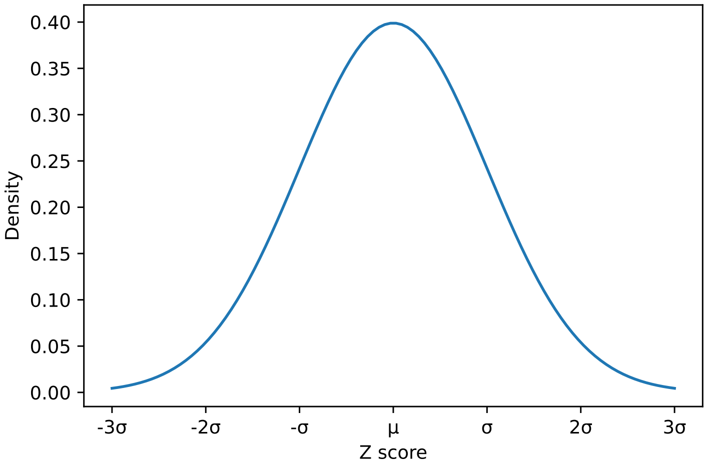


Figure 1: Normal distribution

3.4 PCA

Principal component analysis

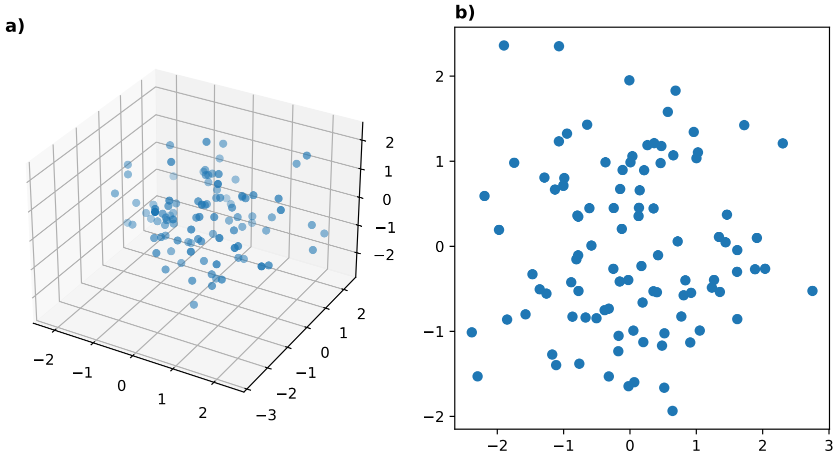
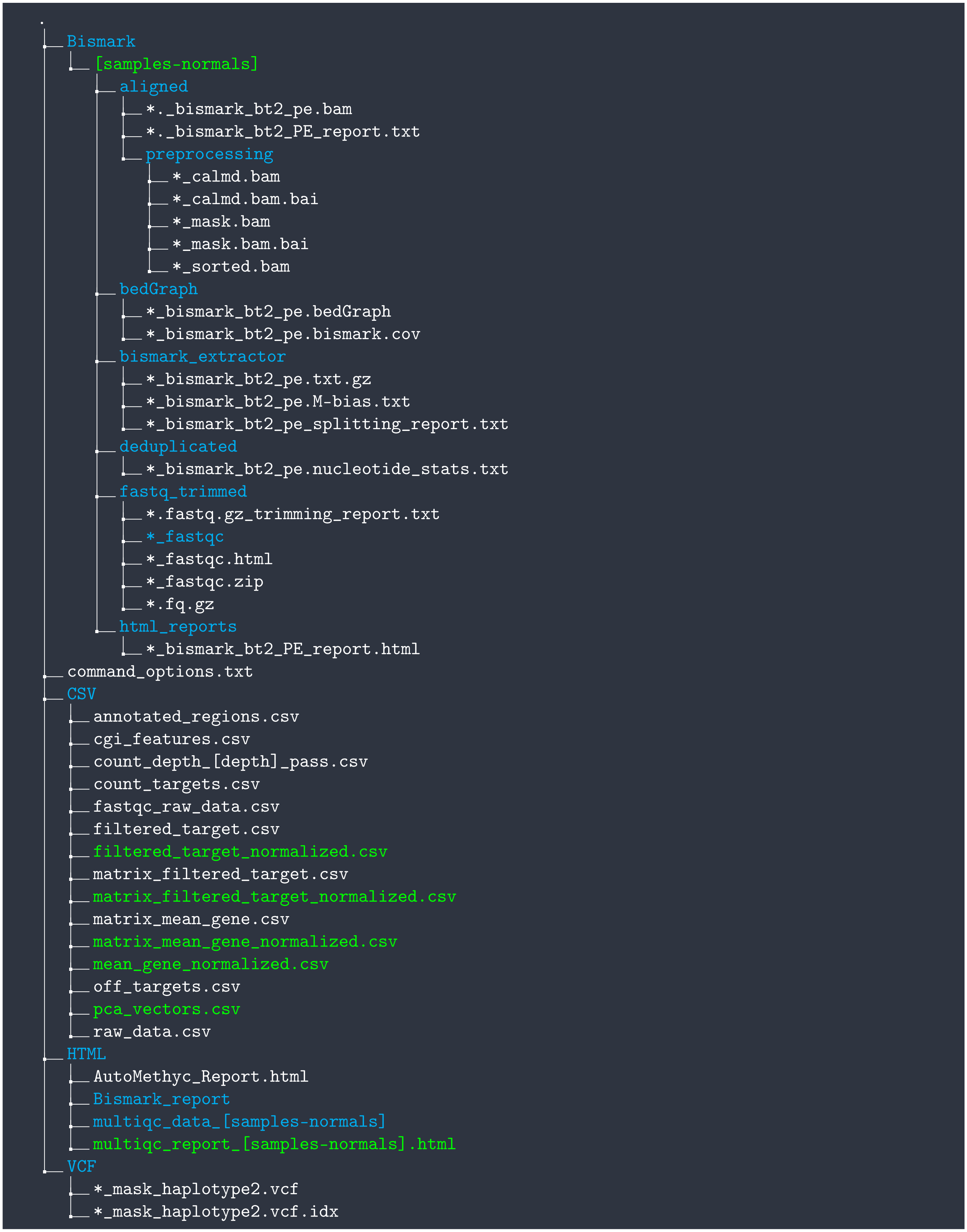


Figure 2: Dimensionality reduction by PCA



```
.
├── Bismark
│   └── [samples-normals]
│       ├── aligned
│       │   ├── *._bismark_bt2_pe.bam
│       │   ├── *._bismark_bt2_PE_report.txt
│       │   └── preprocessing
│       │       ├── *_calmd.bam
│       │       ├── *_calmd.bam.bai
│       │       ├── *_mask.bam
│       │       ├── *_mask.bam.bai
│       │       └── *_sorted.bam
│       ├── bedGraph
│       │   ├── *_bismark_bt2_pe.bedGraph
│       │   └── *_bismark_bt2_pe.bismark.cov
│       ├── bismark_extractor
│       │   ├── *_bismark_bt2_pe.txt.gz
│       │   ├── *_bismark_bt2_pe.M-bias.txt
│       │   └── *_bismark_bt2_pe_splitting_report.txt
│       ├── deduplicated
│       │   └── *_bismark_bt2_pe.nucleotide_stats.txt
│       ├── fastq_trimmed
│       │   ├── *.fastq.gz_trimming_report.txt
│       │   ├── *_fastqc
│       │   ├── *_fastqc.html
│       │   ├── *_fastqc.zip
│       │   └── *.fq.gz
│       └── html_reports
│           └── *_bismark_bt2_PE_report.html
├── command_options.txt
├── CSV
│   ├── annotated_regions.csv
│   ├── cgi_features.csv
│   ├── count_depth_[depth]_pass.csv
│   ├── count_targets.csv
│   ├── fastqc_raw_data.csv
│   ├── filtered_target.csv
│   ├── filtered_target_normalized.csv
│   ├── matrix_filtered_target.csv
│   ├── matrix_filtered_target_normalized.csv
│   ├── matrix_mean_gene.csv
│   ├── matrix_mean_gene_normalized.csv
│   ├── mean_gene_normalized.csv
│   ├── off_targets.csv
│   ├── pca_vectors.csv
│   └── raw_data.csv
├── HTML
│   ├── AutoMethyc_Report.html
│   ├── Bismark_report
│   ├── multiqc_data_[samples-normals]
│   └── multiqc_report_[samples-normals].html
└── VCF
    ├── *_mask_haplotype2.vcf
    └── *_mask_haplotype2.vcf.idx
```

Figure 3: Output directory tree