SINBAD: A pipeline for processing SINgle cell Bisulfite sequencing samples and Analysis of Data

USER MANUAL

1. Introduction

SINBAD is a high-throughput analysis tool for single cell DNA methylation data. It can be used to process both raw and processed data generated with single cell bisulfite sequencing experiments. It is designed in a modular architecture and modules can be used independently. As an example, a user can use SINBAD starting from raw reads (fastq) or methylation calls (bed). As the output, SINBAD generates quality and summary statistics in addition to bed and matrix files, which can be used as input for other analysis tools. In addition, certain downstream analyses can be performed by SINBAD, such as dimensionality reduction, clustering and differential methylation analysis. It can be used in both command line mode and graphical user interface.

2. Installation

R 3.6.0 or later version is required for SINBAD installation. To install SINBAD, type the following commands in R command prompt:

```
devtools::install_github("yasin-uzun/SINBAD")
```

Once you have installed the SINBAD: A pipeline for processing SINgle cell Bisulfite sequencing samples and Analysis of Data package, verify that it is installed correctly as follows:

```
SINBAD::test()
```

If SINBAD is installed without any problems, you should see the following message:

```
SINBAD installation is ok.
```

Details and dependencies are available at the project portal:
https://github.com/yasin-uzun/SINBAD

4. Configuration

SINBAD uses three configuration files:

1) config.general.R : Sets the program paths to be used by SINBAD. This file only needs to be edited once.

2) config.genome.R : Sets the genomic information and paths to be used by SINBAD. You need to generate one for each organism. We provide the built-in configuration by hg38.

3) config.project.R : You need to configure this file for your project.

The templates for the configuration files can be downloaded from here and customized according to the needs.

5. Running SINBAD

SINBAD can be run in either of the two modes: 1) By executing R functions. 2) Using the graphical user interface.

5.1. R function mode

SINBAD is a software package implemented in R. The researchers who are familiar with R programming language can find command-line execution mode practical, in which they can simple R functions to run SINBAD.

In the command SINBAD is run in two steps:

1. Read configuration files:

```
read_configs(config_dir)
```

The input parameter config_dir is a character vector pointing to the configuration file directory (mentioned above).

2. Process data:

```
process_sample_wrapper(raw_fastq_dir, demux_index_file, working_dir, sample_name)
```

The input arguments are as follows:

- raw_fastq_dir: a character vector pointing to the directory containing fastq files as the input.
- demux_index_file: a character vector pointing to the demultiplexing index file for the fastq files.

- working_dir: a character vector pointing to the directory where all the outputs will be placed into.
- sample_name: (optional) a character vector pointing the name for the sample or project.

This function reads fastq files, demultiplexes them into single cells, performs filtering, mapping (alignment), DNA methylation calling and quantification, dimensionality reduction, clustering and differential methylation analysis for the given input. All the outputs are placed into related directories in working_dir.

5.2. Graphical user interface mode

SINBAD also supports a graphical user interface (GUI) for the users who don't have any experience in running R scripts. Graphical interface can be directly launched through RStudio. The input parameters are handled through the "Settings" panel and the processing is run through the "Execute" panel.

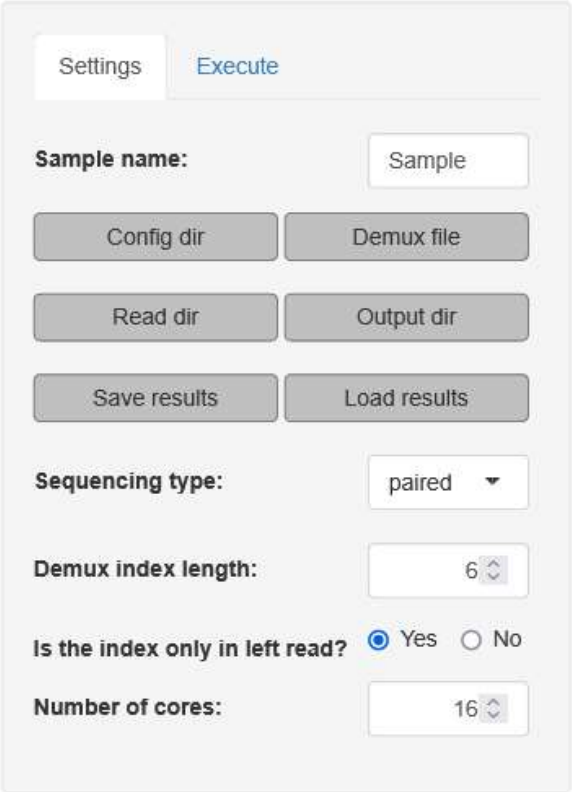

Figure 1. Settings panel

In the settings panel, the user can enter the sample name, which is a character string and will be used for selecting the raw read files and the naming the outputs. Clicking onto the "Config dir" button opens a browser window for the user to select the directory, where the configuration files

are stored for execution. If demultiplexing needs to be done for the raw data, the demultiplexing index file is provided by a browser view that opens via clicking the “Demux file” button. In this case, the index length should be provided in the “Demultiplexing index length” numeric input box. If there are no separate indices for right reads and the left reads indices are used for both types of reads as in the case of snmC-Seq data, the radiobutton for the “Is right index embedded in left reads?” should be selected as “Yes”.

If the processing will be done starting from the raw reads, the directory where the fastq files are present is given by a file browser that opens clicking onto “Read dir” button. Similarly, independent of the input type (raw reads, methylation calls), the output directory should be provided by clicking onto “Output dir” button.

The intermediate results for SINBAD during data processing can be saved into a file on disk. For this purpose, the user can provide the file path by clicking the “Save results” button. When it is desired to continue processing a specific execution, the saved results can be reloaded by clicking onto the “Load results” button which launches a file browser. SINBAD also supports parallel processing using the R doSNOW package. The user can specify how many threads to be used by using the input field “Number of cores”.

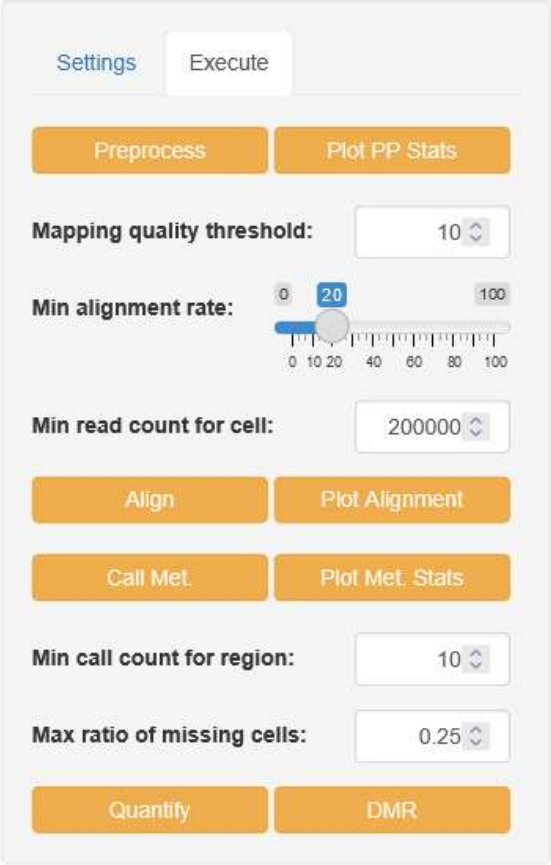

Figure 2. Execution panel

Once the running parameters are set, SINBAD is ready for processing the data using the "Execute" panel. When the processing is starting from raw reads, preprocessing steps need to be run by clicking onto the "Preprocess" button. "Plot PP Stats" generates and plots preprocessing statistics as shown below:

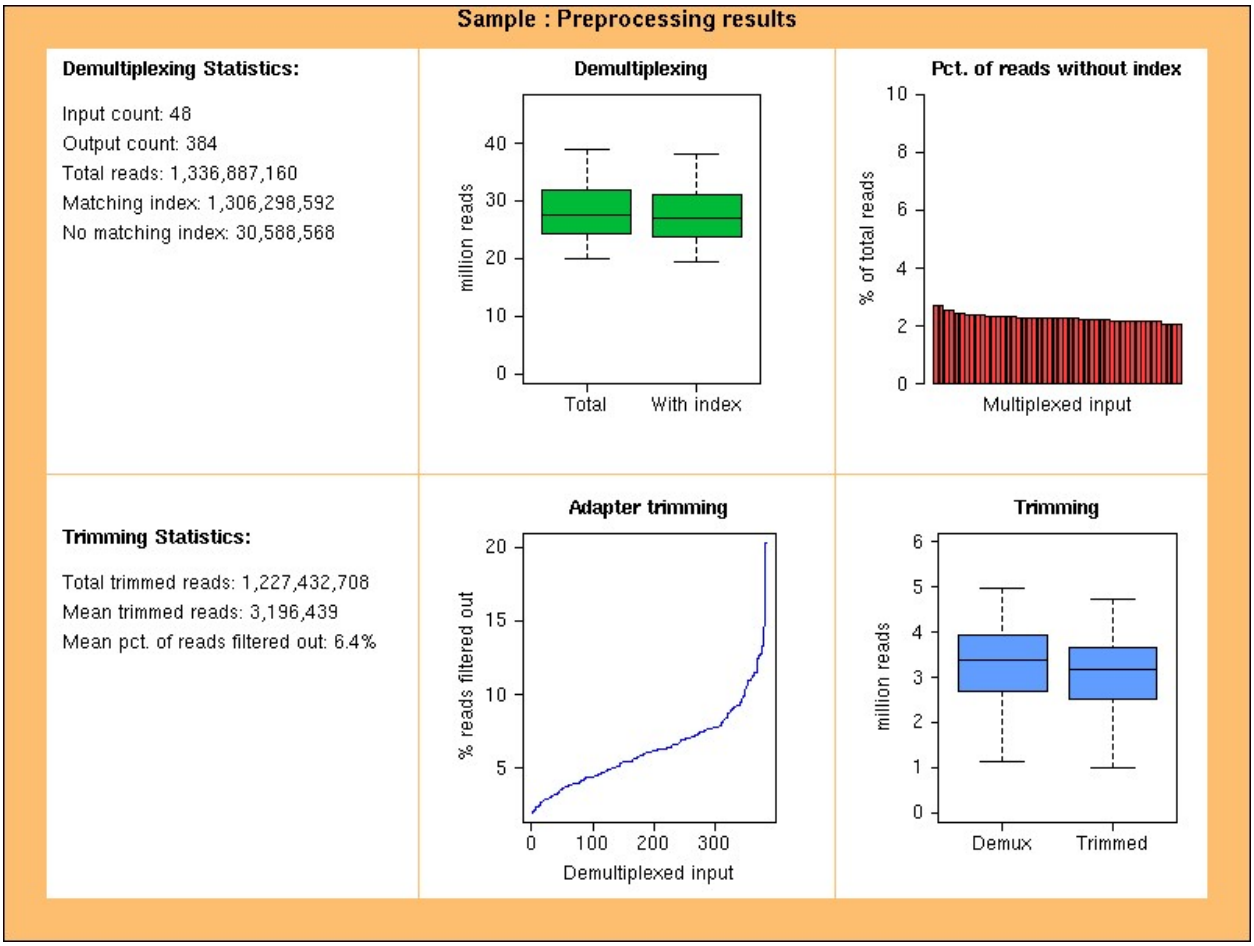

Figure 3. Preprocessing statistics

There are 3 special settings for alignment. 1) Mapping quality threshold: Minimum MAPQ value for filtering aligned reads. 2) Minimum alignment rate: The minimum overall alignment rate of a cell in order to include it for downstream analysis. 3) Minimum read count for cell: Minimum number of aligned reads that a cell must have for inclusion. Clicking the "Align" button starts alignment and filtering procedures. The alignment statistics can be visualized by clicking "Plot Alignment", which displays a window as shown below:

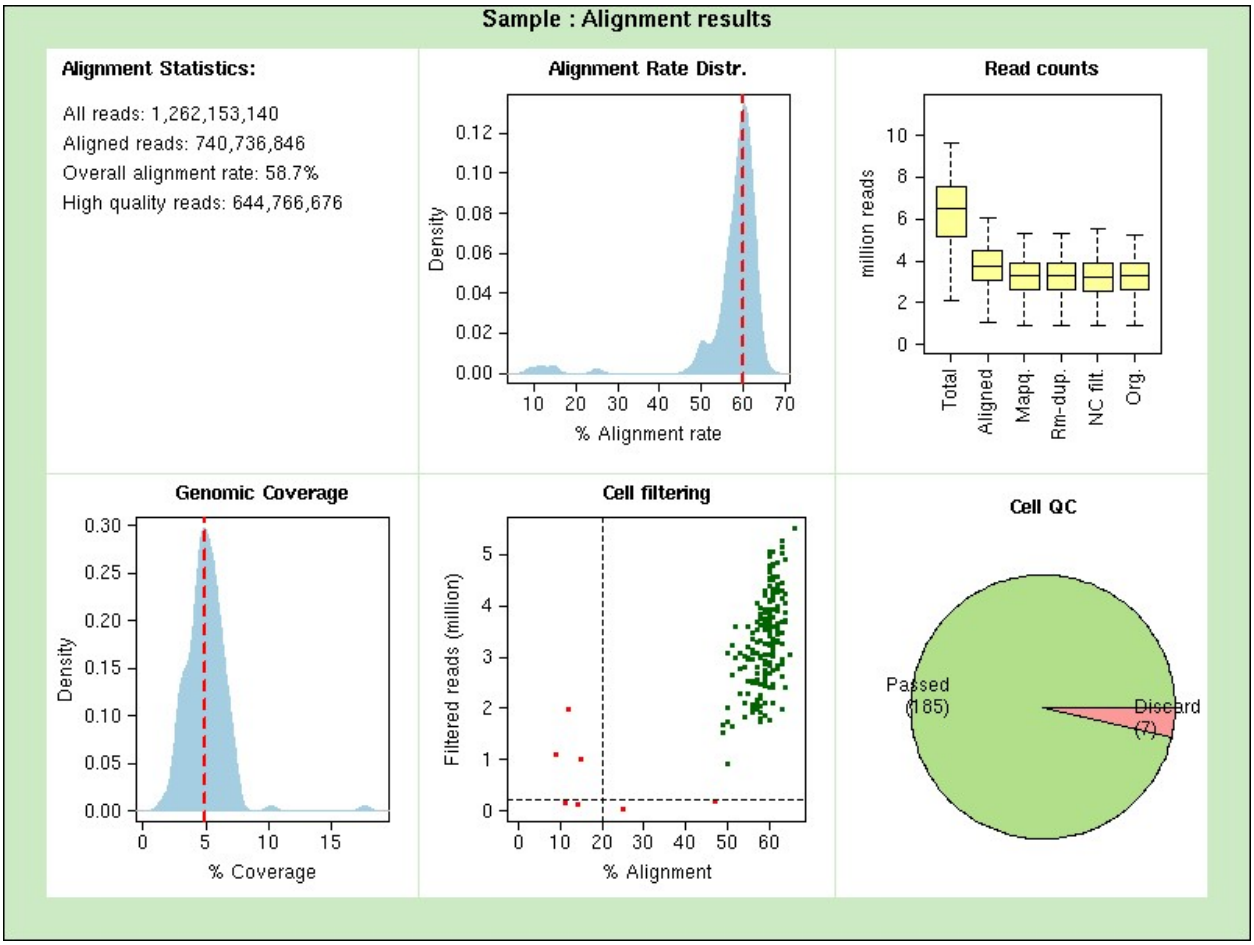

Figure 4. Alignment statistics

Once the alignment is completed, clicking onto the “Call Met.” button calls methylation sites and clicking the “Plot Met. Stats” shows the statistics as shown below:

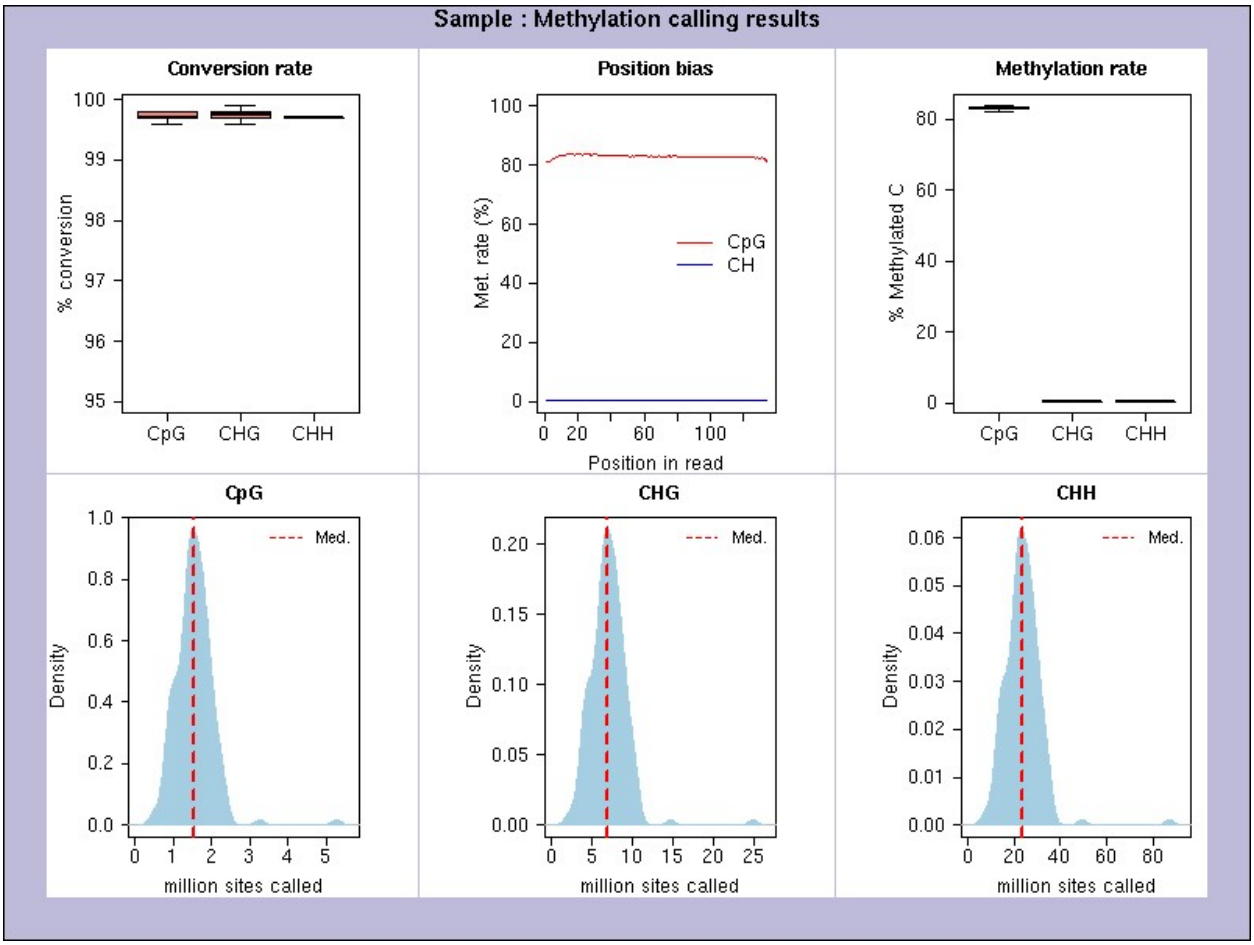

Figure 5. Methylation calling statistics

For generating methylation matrices, there are two parameter settings: 1) Minimum call count for region: The number of minimum cytosine calls for a region-cell pair to make an estimate for the methylation level. 2) Maximum ratio of missing cells for region: The maximum allowable ratio of missing values for a region to be included in the quantification matrix. Once these parameters are set, quantification can be run by clicking the “Quantify” button. As a result, the matrices are generated, dimensionality reduction is performed and clusters are identified as shown below:

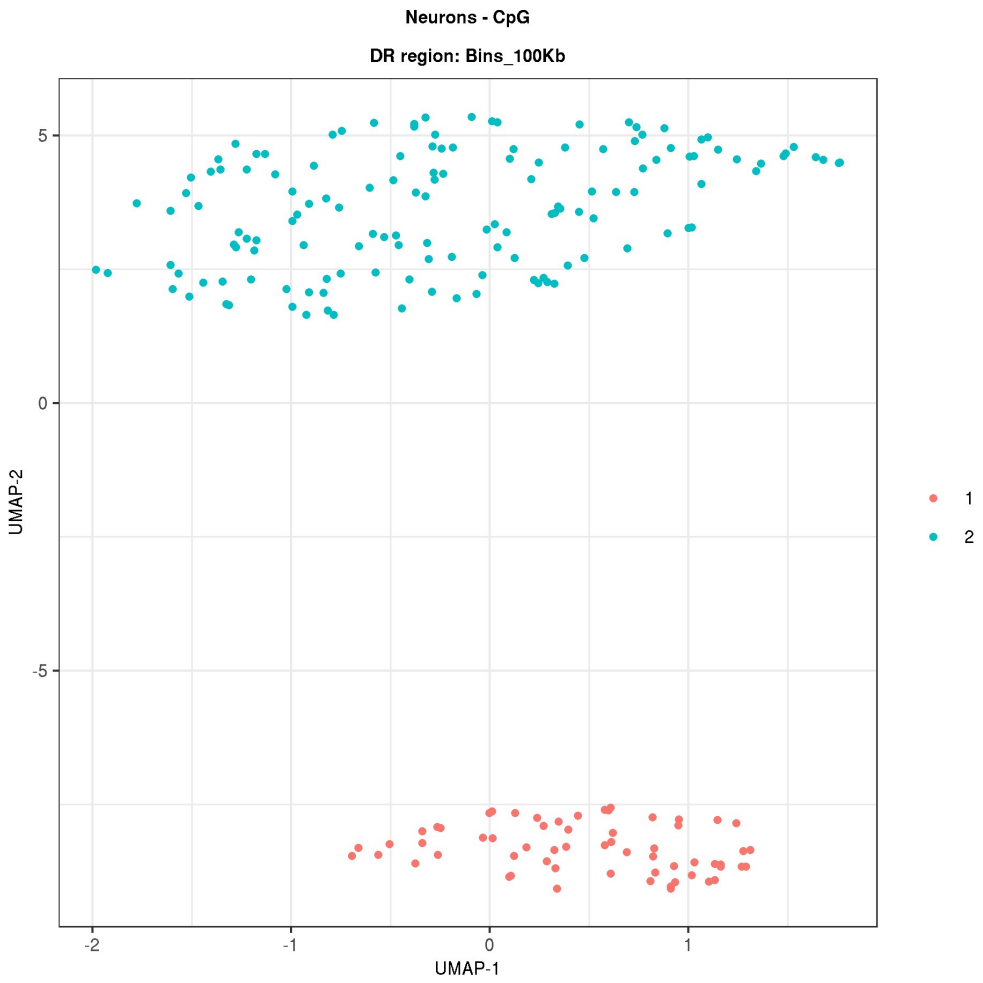

Figure 6. Dimensionality reduction and clustering

Once the clusters are generated, it is possible to perform differential methylation analysis by clicking the “DMR” button. As a result, differentially methylated regions are computed and a heatmap is generated as follows:

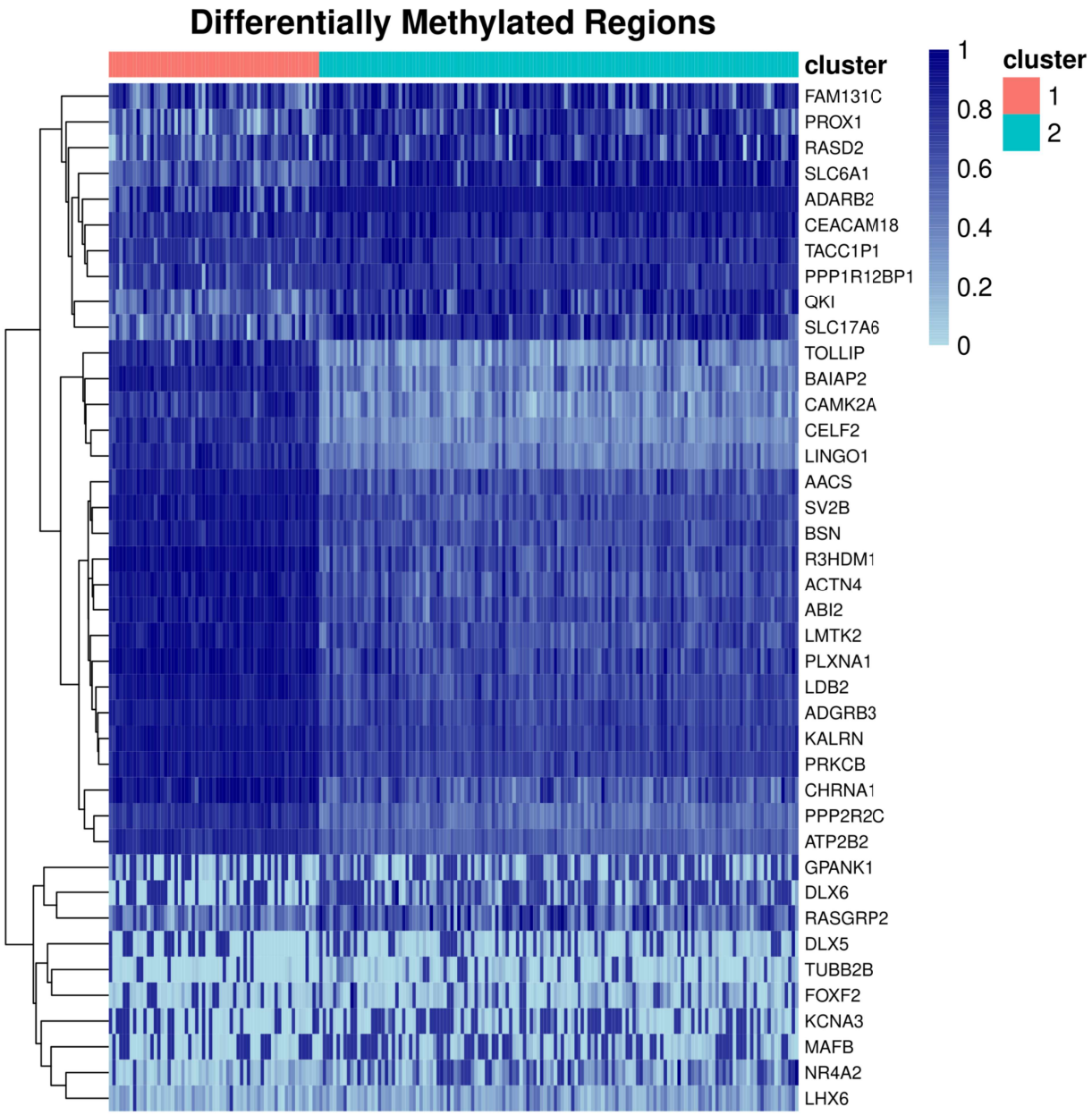

Figure 7. Hetmap showing differentially methylated regions

5. Citing SINBAD

If you use SINBAD in your analysis, please cite it as follows:

Uzun Y, Yu W, Chen C, Tan K. SINBAD: a flexible tool for single cell DNA methylation data. bioRxiv, doi: https://doi.org/10.1101/2021.10.23.465577.