MLML User Guide

Jenny Qu Meng Zhou Qiang Song Elizabeth Hong Andrew Smith

July 15, 2013

`mlml` aims to simultaneously make consistent estimation of 5mC and 5hmC levels. It is developed to analyze DNA methylation level data. The input can be any pair of BS-seq, oxBS-seq and Tab-seq data or all of them.

1 System requirements

We tested `mlml` on various platforms, including Ubuntu, Unix/Linux system, and Mac OSX. `mlml` is memory efficient. In our test run on a 64-bit Unix platform with Xeon E5420 2.5GHz processor, the runtime and memory usage for computing 5 million CpG sites are 10 minutes and 37MB respectively.

2 Input file format

The input format is BED format file with 6 columns (`https://genome.ucsc.edu/FAQ/FAQformat.html#format1`). Below is an example for a line of input file:

```
chr1    3001345 3001346 CpG:9   0.777777777778  +
```

The first three columns are genomic coordinates of a region. The fourth column indicates that this site is a CpG site, and the number of reads covering this site is 9. The fifth column is the methylation level of the CpG site, ranging from 0 to 1. This level is usually calculated from the number of methylated reads covering one site dividing by the total number of reads included in the fourth column.

We also provide some sample data with the package. The sample data have real methylation values attached to the last column as a reference for the accuracy of estimation.

3 Usage

The usage of `mlml` is simple: specify the input files with corresponding options to their types. `mlml` can take various types of input combinations. Assume user has three input files ready: `meth_BS-seq.bed`, `meth_oxBS-seq.bed` and `meth_Tab-seq.bed`, all in the format described above. The following command will take all the inputs:

```
$ ./mlml -v -u meth_BS-seq.bed -m meth_oxBS-seq.bed \
         -h meth_Tab-seq.bed -o result.bed
```

Note this command should be run in the path where `mlml` is installed. The option `-o` will direct the output to file `result.bed`. The option `-v` will provide some useful statistics information.

If only two types of input are available, e.g. `meth_BS-seq.bed` and `meth_oxBS-seq.bed`, then use the following command:

```
$ ./mlml -u meth_BS-seq.bed -m meth_oxBS-seq.bed \
         -o result.bed
```

In some cases, user might want to specify the convergence tolerance for EM algorithm. This can be done through `-t` option. For example:

```
$ ./mlml -u meth_BS-seq.bed -m meth_oxBS-seq.bed \
         -o result.bed -t 1e-2
```

This command will make the iteration process stop when the difference of estimation between two iterations is less than 10^{-2}. The value format can be scientific notation, e.g. 1e-5, or float number, e.g. 0.00001.

4 Output file format

The output of `mlml` is also in BED format. Here is an example:

```
chr1    3001345 3001346 mC:0.125        0.652778        +
```

The fourth column is estimated 5mC level, and the fifth column is 5hmC level. Note that the output is always in this format, no matter what combination of inputs are used. Therefore user can easily extract 5mC and 5hmC levels from specific columns.

5 Example data

There are three example data files in `example_data` directory. These files have simulated CpG coverage and methylation level. For each CpG, true methylation levels are simulated from one Dirichlet distribution, so that the sum of all three levels is equal to 1. Then sequencing coverage is simulated and methylation level for each site is determined by binomial sampling. The true levels are attached as the last column in the files.

6 FAQ

- Q: I had some errors in compiling/installing the software.
  A: Compiling errors can be very complicated. But from the tests we have done, we strongly recommend you to check your complier and GSL versions (see readme.txt) and make sure they are up to date and correctly installed.

- Q: I had an error like the following, what's that?

  ```
  Segmentation fault.
  ```

  A: Please check if your hardware conforms with the system requirements, and make sure all the options set for `mlml` are correct, e.g. no bad path and all options are complete.

- Q: I got result like this, what happened?

```
chr1    10570   10571   mC:nan  nan     +
```

  A: Because some sites have 0 coverage, there is not enough information to calculate mC and hmC levels. Therefore the output will be nan instead.