The BMix Toolbox
v 1.0
User Manual

Monica Golumbeanu[1,2], Pejman Mohammadi[1,2], Niko Beerenwinkel[1,2]

[1]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[2]Swiss Institute of Bioinformatics, Basel

Contact: monica.golumbeanu@bsse.ethz.ch

Contents

1 Introduction

BMix is a novel probabilistic method based on a constrained three-component mixture, which identifies high confidence T-to-C substitutions in PAR-CLIP data, and, based on these, reports putative RNA-protein cross-link sites. Starting from observed substitution counts throughout the genome, BMix classifies all the loci with observed T-to-C alterations in three groups: (i) background, (ii) sequence variants, and (iii) cross-link loci.
The BMix toolbox is modular and comprises shell, awk and Matlab programs destined to pre-process PAR-CLIP data, identify the high confidence substitutions and report RNA-protein binding sites. A main program performing all these operations is provided.

2 Requirements

In order to successfully run the programs of the BMix toolbox, the following requirements need to be assured:

- Unix shell command terminal
- Matlab (R2013 or later)
- samtools (`http://www.htslib.org/`)
- awk (`http://www.gnu.org/software/gawk/manual/gawk.html`)
- bedtools (`http://bedtools.readthedocs.org/en/latest/`)

3 Analyze PAR-CLIP data with the BMix Toolbox

The BMix toolbox gives the user the possibility to run the entire pipeline in one go.

3.1 Perform the entire analysis with *BMix*

Once the user has clipped and aligned the PAR-CLIP sequencing reads and has produced a sorted `.bam` file (steps not performed by BMix), they can employ the *BMix* program from the toolbox to analyze the data and retrieve a list of candidate binding sites in `.bed` format. In order to do so, a configuration file is needed. The file contains the following fields which need to be specified:

- `BAM_FILE` - path to the input `.bam` file
- `REF_FILE` - path to the `fasta` file containing the reference genome (the same as the one used for alignment)
- `SAMPLE_NAME` - chosen name for the experiment (the produced files will contain this name)
- `WORK_FOLDER` - path of the folder where the output will be saved (a new folder will be created if it does not exist already)
- `COV_MIN` - minimum coverage to consider (default is 5)
- `REFINE_COV` - the tails of the binding sites with coverage lower than this value will be trimmed (default is 1)
- `CONFIDENCE_PER` - threshold for the posterior probability used to classify substitutions (default is 0.95)

Once the configuration file has been created, the pipeline can be ran with the following shell command executed in the folder containing the BMix toolbox programs:

```
./BMix path_to_config_file
```

3.1.1 Example of Use

On the BMix Git repository, under the folder `test/`, a sample dataset is provided in the folder `data/`, as well as a sample configuration file `CONFIG.txt`. The dataset consists of PAR-CLIP reads aligned to Chromosome 21 extracted from a published AGO2 dataset [1], as well as the reference genome `fasta` file for chromosome 21. The `CONFIG.txt` file is filled accordingly and indicates BMix to create the folder `BMix_output` where the results are stored:

```
#!/bin/bash

BAM_FILE="../test/data/AGO2_reads_chr21.bam"
REF_FILE="../test/data/hg19_chr21.fa"
SAMPLE_NAME="test"
WORK_FOLDER="../test/BMix_output/"
COV_MIN=5
REFINE_COV=1
CONFIDENCE_PER=0.95
```

By downloading the contents of the BMix repository and keeping the same folder hierarchy, the user can go to the command terminal, change (`cd`) to the `source/` folder where the BMix program is stored, and run:

```
./BMix ../test/CONFIG.txt
```

The folder `BMix_output/` is created under the folder `test/` and contains the results of the pipeline. The constructed binding sites are stored in the file `Sites_sorted.bed` located in folder `BindingSites/`, under the `BMix_output/` directory.

3.1.2 Description of the pipeline output

Under the folder `test/`, on the BMix Git repository, a sample BMix result is provided in the `BMix_sample_output/` folder. It contains several files and folders produced during the BMix execution on the provided sample data:

- File `Log.txt` - contains the execution time in seconds of the whole pipeline
- File `test.mpileup` - contains the alignment summary produced by the `samtools mpileup` command. This summary is further employed by BMix to construct substitution summaries.
- Folder `MismSummaries` - contains the substitution summaries produced by BMix from the previously mentioned `.mpileup` file. A mismatch summary file contains, for each position on the genome where the coverage is larger than COV_MIN, the number of times a specific substitution was observed, as well as the coverage at the respective position. For example, the T-to-C mismatch profile file will contain the number of T-to-C substitutions and the coverage observed at each position on the genome. The folder contains the T-to-C, A-to-C and G-to-C mismatch profile files for the forward strand of the genome, and the A-to-G, T-to-G and C-to-G mismatch profile files for the reverse strand of the genome.
- Folder `TC_Results` - contains the results of the classification of T-to-C and A-to-G loci for the forward and reverse strand, respectively (files `TC_f.results` and `AG_r.results`)

in (i) background, (ii) sequence variants, and (iii) cross-link loci. The inferred parameters of the statistical model are reported in the file `parameters.txt` for the forward and reverse strand. BMix extracts the T-to-C (on the forward strand) and A-to-G (on the reverse strand) loci classified as a cross-link with posterior probability larger than CONFIDENCE_PER and stores them in the files `TC_f.parclip.bed` and `AG_r.parclip.bed`. The reads covering these selected substitutions are stored in the files `TC_f.parclip.reads.bed` and `AG_r.parclip.reads.bed`. Additionally, two figures depicting the classified loci on the forward and reverse strand are created and stored in the folder `Figures`. For each T locus, the x-axis of the figures corresponds to $log10$ of the coverage, while the y-axis represents the observed substitution frequency.

- Folder `BindingSites` - contains the constructed binding sites from the previously mentioned cross-link loci (file `Sites_Sorted.bed`).

References

[1] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.," *Nature methods*, vol. 8, pp. 559–64, July 2011.