The MiClip package

Tao Wang
The University of Texas Southwestern Medical Center,
5323 Harry Hines Boulevard Dallas, Texas, 75390
tao.wang@utsouthwestern.edu

December 12, 2012

Abstract

There has been increasing interest in the role of RNA-binding proteins in biological processes. Crosslinking and immunoprecipitation (CLIP) experiments have made it possible to identify binding sites of RNA-binding proteins in various cell culture and tissue types. There are now mainly three generations of CLIP-Seq experiments, hist-Clip, par-Clip and iCLIP. Here we present MiClip, an R package for identification of binding sites in CLIP-Seq experiments. The MiClip package employs two rounds of Hidden Markov Model to identify enriched regions and further high-confidence binding sites from raw sequencing data.

Contents

1 Installation

R (http://www.r-project.org/) needs to be installed first for *MiClip* and the installation of the *MiClip* package follows the regular method for R package installation.

However, *MiClip* also requires Perl to be installed. Perl should ship along with any standard UNIX distribution. But Windows and MacOS users probably need to install Perl themselves (`http://www.perl.com/`). The users can use the following command in the command console to check if Perl has been installed properly.

```
perl -v
```

2 Raw sequencing file

2.1 Trimming adaptor

During CLIP-Seq experiments, RNAs are usually digested to short fragments. So it is quite often for sequenced reads to have adaptor contamination at the 3' end, while it is relatively rare for the 5' end of short reads to have adaptor contamination. Thus, it is necessary to trim contaminating adaptor sequences before running alignment. We provide a helper function to remove adaptor sequence at the 3' end of short reads. The helper function can handle single-end fastq or fasta file in basespace. Users are welcome to use other available softwares to trim adaptors. Here we used a small portion of the sequencing data provided in [1] for demonstration purpose.

```
> library("MiClip")
> MiClip.adaptor(file="test.fastq",
+                  adaptor="TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC")
```

After trimming, ”ACCGGGTGCTGTAGGCTTTTGGAATTCTCGGGTGCCAAGGAACTCCAGT” will be cut down to ”ACCGGGTGCTGTAGGCTTT”. A new file with ”.removed” suffix will be generated in the same folder of the original fastq file.

2.2 Alignment

The raw sequencing file may be single-end or paired-end in basespace or colorspace. Any mainstream alignment software can be used to align the short reads. The output format must be SAM format and in basespace. *MiClip* can work on both single-end and paired-end alignment files. In the case where the user wishes to pool the alignment files from several experiments, the user can just concatenate the SAM files simply by typing the following in the command console.

```
cat example1.sam example2.sam > example.sam
```

Importantly, the SAM file needs to have MD field present, please read the SAM formation specification (`http://samtools.sourceforge.net/SAM1.pdf`). If not, the user should install and use samtools to populate the MD field, please see the instructions by typing the following command in command console.

```
samtools fillmd
```

However, this command runs very slowly. So it will be much better if the user can choose an alignment software which will give MD fields to all sequencing tags in the very beginning.

2.3 Paired-end reads

For paired-end reads, the users must look at the sequencing files and provide the suffix for the forward strand and the negative strand. For example, the mate in the sequencing dataset may be named like ”694_122_1972-F3” and ”694_122_1972-F5-RNA”, where ”694_122_1972” is the id number of the mates, ”F3” means forward strand and ”F5-RNA” means backward strand. Then the suffix should be ”F3” and ”F5-RNA” or ”F3” and ”F5-RNA” or ”3” and ”5-RNA”.

3 Running *MiClip*

3.1 Construct a *MiClip* class object for following analysis

The analysis of *MiClip* starts by constructing a MiClip object. Here we used a small portion of the HITS-CLIP data provided in [2] for demonstration purposes.

```
> library("MiClip")
> test=MiClip(file="test.sam")
>
> # for paired-end data
> # test=MiClip(file="test.sam",paired=TRUE,suffix=c("F3","F5-RNA"))
```

This command returns a MiClip object for further analysis. Following sections will explain some of the available parameters in constructing this object. For detailed descriptions of all parameters, please refer to the *MiClip* manual.

3.2 Read raw sequencing data and mutation data

The MiClip.read function calls some embeded perl scripts to form clusters by overlapping reads and collect tag pile-up as well as mutation information from the provided SAM file. This process will usually take a few minutes depending on the size of the file.

```
> test=MiClip.read(test) # read raw data
```

3.3 Identify enriched bins

The MiClip.enriched function first collects tag pile-up information on a step bp basis (bins) and estimates the paramters for a two-poisson mixture model for the count values. Because we are running a truncated part of the real data for demonstration, so the model estimation will not be accurate. Then the first Hidden Markov Model will try to identify the enriched bins vs. non-enriched bins in CLIP clusters.

```
> test=MiClip.enriched(test,quiet=FALSE) # identify enriched regions
```

```
Initialization of the first HMM finished!
>>>>>
Iterations of the first HMM finished!
Viterbi algorithm of the first HMM finished!
```

The empirical parameter is devised to adjust model estimation in this step. The default for empirical is ”auto”, which lets the algorithm decides its value. The user can adjust the estimation to be more conservative or less conservative by setting the empirical value. User can set this value to roughly the minimal number of overlapping tags for a ”true” cluster according to user’s experience and experimental design.

3.4 Identify binding sites

The `MiClip.binding` function first concatenate neighboring enriched bins and then expand each chain of adjacent bins into single base pairs. Then `MiClip.binding` collects the tag pile-up and mutant pile-up information on each base for estimation of a mixture model of one zero-inflated binomial distribution and a binomial distribution. Then the second Hidden Markov Model is run to identify significant binding sites.

```
> test=MiClip.binding(test,quiet=FALSE) # identify binding sites
```

```
Initialization of the second HMM finished!
>>>>
Iterations of the second HMM finished!
Viterbi algorithm of the second HMM finished!
```

The `model.cut` parameter is devised to adjust model estimation in this step. The default for `model.cut` is 0.2. User can set this value to roughly the minimal proportion of mutant tag vs. total tag on true binding sites according to user's experience and experimental design. Larger values will lead to more conservative predictions.

4 Output of *MiClip*

4.1 Output format

`MiClip.binding` returns a `MiClip` object which comprises of three data frames.

```
> enriched=test$enriched # test is a list of 3 components
> sites=test$sites
> clusters=test$clusters
> head(enriched)
```

```
  region_id   chr   start     end strand tag enriched probability
1         1 chr19 3182690 3182694      -   2    FALSE           1
2         1 chr19 3182695 3182699      -   2    FALSE           1
3         1 chr19 3182700 3182704      -   2    FALSE           1
4         1 chr19 3182705 3182709      -   2    FALSE           1
5         1 chr19 3182710 3182714      -   2    FALSE           1
6         1 chr19 3182715 3182719      -   2    FALSE           1
```

```
> head(sites)
```

```
  region_id sub_region_id strand   chr     pos tag mutant sites probability
1        12             1      - chr19 3456025   6      0 FALSE           1
2        12             1      - chr19 3456026   5      0 FALSE           1
3        12             1      - chr19 3456027   6      0 FALSE           1
4        12             1      - chr19 3456028   6      0 FALSE           1
5        12             1      - chr19 3456029   6      0 FALSE           1
6        12             1      - chr19 3456030   7      0 FALSE           1
```

```
> head(clusters)
```

```
  region_id   chr strand   start     end enriched sites
1         1 chr19      - 3182690 3182729    FALSE FALSE
2         2 chr19      - 3271010 3271044    FALSE FALSE
3         3 chr19      - 3278390 3278429    FALSE FALSE
4         4 chr19      - 3448825 3448864    FALSE FALSE
5         5 chr19      - 3454480 3454519    FALSE FALSE
6         6 chr19      - 3454970 3455039    FALSE FALSE
```

`enriched` is the output of the first Hidden Markov Model. `region_id` is the id number for each cluster. `chr`, `strand`, `start` and `end` specify the genomic location of each bin. `tag` is the rounded average tag count in each bin. `enriched` and `probability` are the inference results.

`sites` is the output of the second Hidden Markov Model. `region_id` is the id number for the cluster which each base resides in. `sub_region_id` is the id number of the concatenated segment. Sometimes one enriched cluster has multiple modes, so it may be cut into two or more segments. `chr`, `strand` and `pos` specify the genomic location of each base. `tag` is the tag count and `mutant` is the mutant count on each base. `sites` and `probability` are the inference results.

`clusters` is the summary of results for all clusters. `region_id` is the id number for each cluster. `chr`, `strand`, `start` and `end` specify the genomic range. `enriched` specifies whether a cluster is found to have at least one enriched bin, and `sites` specifies whether a cluster is found to have at least one significant binding site.

4.2 summary

MiClip provides method dispatch for the generic `summary`. Because we are using a very small toy sample data, the results presented are not realistic.

```
> summary(test)

For identifying enriched regions
# of clusters: 619
# of identified enriched clusters: 100
# of bins: 5932
# of bins in each state:
FALSE  TRUE
 5194   738
Statistics of probability
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2300  1.0000  1.0000  0.9806  1.0000  1.0000
Average tag count of enriched bin: 8
Average tag count of not enriched bin: 2


For identifying binding sites
# of enriched clusters: 100
# of sub enriched clusters: 102
# of enriched clusters with identified binding sites: 10
# of bases: 3644
# of bases in each state:
```

```
FALSE  TRUE
 3634    10
Statistics of probability
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7400  1.0000  1.0000  0.9989  1.0000  1.0000
Average tag count of binding site: 8
Average mutant count of binding site: 3
Average tag count of not binding site: 8
Average mutant count of not binding site: 0
```

summary gives the basic statistics on the results of the two rounds of Hidden Markov Model.

5 Session Info

```
> sessionInfo()

R version 2.15.0 (2012-03-30)
Platform: x86_64-redhat-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    splines   stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
[1] MiClip_1.0   VGAM_0.8-7   moments_0.13

loaded via a namespace (and not attached):
[1] tools_2.15.0
```

References

[1] Macias, S., et al. (2012) DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nat Struct Mol Biol* 19(8): p. 760-6.

[2] Chi SW, Zang JB, Mele A, Darnell RB. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 2009 Jul 23;460(7254):479-86. Epub 2009 Jun 17.