ChIAMM: a Mixture Model for Statistical Analysis of Long-Range Chromatin Interactions from ChIA-PET Experiments

Yibeltal Arega[1], Hao Jiang[1], Shuangqi Wang[2], Jingwen Zhang[2], Xiaohui Niu[1], Guoliang Li[1,2*]

[1] Agricultural Bioinformatics Key Laboratory of Hubei Province, Hubei Engineering Technology Research Center of Agricultural Big Data, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China.

[2] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China.

*** Correspondence:**
Guoliang Li
E-mail: guoliang.li@mail.hzau.edu.cn

Introduction

Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing is a technology to study genome-wide long-range chromatin interactions bound by protein factors. ChIAMM is a statistical technique for processing and analyzing ChIA-PET sequence data using the Mixture model in the Bayesian framework.

The ChIAMM requires the following dependencies:

- R (≥ 3.4.0)
- StanHeaders (≥ 2.18.1)
- rstan (version ≥ 2.18.2)
- ggplot2 (≥ 2.0.0)
- tidyverse (≥ 1.2.1)
- bayesplot (≥1.7.1)
- bedtools (≥ 2.25.0)
- ngs-tools (≥1.2)

Data preparation

Before executing the ChIAMM, you need to analyze the raw ChIA-PET data using ChIA-PET Tool V3 without any FDR cutoff value. From the ChIA-PET Tool output files, the file *out.cluster.withpvalue.txt* will be used for downstream analysis in ChIAMM. Specifically, the first 11 columns in the file *out.cluster.withpvalue.txt* will be used: chrom1, start1, end1, chrom2, start2, end2, ipet count, type, distance, tag count within anchor 1 and tag count within anchor 2. We call chrom1, start1, end1 as **Anchor1**, and chrom2, start2, end2 as **Anchor2**.

Computing the Systematic Biases

Average tag counts

Call the variables tag count within anchor 1 and tag count within anchor 2 as tagcou1 and tagcou2 respectively, and compute the average of it call the variable name "*tagcouAvg*".

Self-ligation PETS

Using the *out.spet* file in the ChIA-PET Tool V3 output, we can compute the self-ligation PETs of anchors using the commands as follows:

- awk '{if($2<$5){print $1"\t"$2"\t"$5}else{print $1"\t"$5"\t"$2}}' out.spet > out.spet.bed3
- bedtools coverage -a Anchor1.bed -b out.spet.bed3 > self1.bed
- bedtools coverage -a Anchor2.bed -b out.spet.bed3 > self2.bed
- then compute the average of it and call the variable name "*selfAvg*"

Mappability

Download the mappability score for human genome version hg19 using http://genome-asia.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability. The file is in bigWig format, and we need to convert it to .bed format using ***bigWigToWig*** and BEDOPS ***wig2bed***. Then, find the overlap region between .bed file and Anchor regions using ***Bedtools map.***

- bedtools map -a Anchor1.bed -b mappability.bed -c 5 -o mean > mappability1.bed
- bedtools map -a Anchor2.bed -b mappability.bed -c 5 -o mean > mappability2.bed
- Compute the average of 5th column in mappability1.bed and mappability2.bed, and call the variable name "***mappaAvg***"
- If you don't find the mappability in the above link, you can prepare by yourself using ***ngs-tools***.
    - Example for rice RS1 reference genome we can find like
    - sh mappability.sh -i mhRS63.fa -l 35 -p mappability > mappability.bed

GC content

The GC content of anchors are computed using ***bedtools nuc***

- bedtools nuc -fi hg19.fa -bed Anchor1.bed > gc1.bed
- bedtools nuc -fi hg19.fa -bed Anchor2.bed > gc2.bed
- Compute the average of 5th column in the gc1.bed and gc2.bed and call variable name "***gcAv***"

Input files

We need to organize the input file for inter- and intra-chromosomal interaction data separately. As an example, the input file of intra-chromosomal interactions should have such kind of variables arrangement.

chrom1	start1	end1	chrom2	start2	end2	ipet
chr1	840068	840732	chr1	855579	856565	2
chr1	913753	914300	chr1	1199808	1200630	2
chr1	919077	919880	chr1	998944	999920	4
chr1	919648	920151	chr1	1219377	1220316	2
chr1	968089	969098	chr1	998105	999837	6
chr1	994350	995077	chr1	1003901	1004606	2
distance	**tagcou1**	**tagcou2**	**tagcouAvg**	**selfAvg**	**gcAv**	**mappaAvg**
15672	3	6	4.5	9.0	0.66	0.69
286192	46	33	39.5	3.0	0.60	0.51
79953	30	13	21.5	35.0	0.66	0.51
299947	56	10	33.0	22.5	0.62	0.32
30377	27	6	16.5	34.5	0.70	0.63
9540	26	25	25.5	21.0	0.74	0.78

Test data sets

- RNAPII ChIA-PET data from human MCF7:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM832458 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM832459
- RNAPII ChIA-PET data from human K562:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM832464 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM832465
- CTCF ChIA-PET data from human MCF7:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM970215
- CTCF ChIA-PET data from human K562:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM970216
- RNAPII ChIA-PET data from rice MH63:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3767548 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3767549
- H3K9me2 ChIA-PET data from rice MH63:
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3767546 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3767547

Usage

Rscript ChIAMM.R -h

```
Usage: ChIAMM-jh.R [-[-input|i] <character>] [-[-prefix|p] [<character>]] [-[-inter|e]] [-[-iter|r]
[<integer>]] [-[-warmup|w] [<integer>]] [-[-help|h]]
      -i| --input          input file
      -p| --prefix         output prefix (default "out")
      -e| --inter          the input is inter-chromosomal, if not, intra-chromosomal interactions data
```

-r| --iter iteration number (default 5000)
-w| --warmup warmup value (default 750)
-h |--help print help

Example

For intra- and inter-chromosomal interactions, we need to run the ChIAMM.R separately.

- For intra-chromosomal interaction data
    - ChIAMM.R -i intra_input_file.txt
- For inter-chromosomal interaction data
    - ChIAMM.R -i inter_input_file.txt -e

Result file

We will get the result file names out_significant_interaction.txt.

chrom1	start1	end1	chrom2	start2	end2	ipet	W_{1i}
chr1	919077	919880	chr1	998944	999920	4	0.59
chr1	919648	920151	chr1	1219377	1220316	2	0.51
chr1	968089	969098	chr1	998105	999837	6	0.73
chr1	994350	995077	chr1	1003901	1004606	2	0.53

- **chrom1**: The name of the chromosome on which the cluster anchor 1 exists
- **start1**: The start coordinates of cluster anchor 1
- **end1**: The end coordinate of cluster anchor 1
- **chrom2**: The name of the chromosome on which the cluster anchor 2 exists
- **start2**: The start coordinates of cluster anchor 2
- **end2**: The end coordinate of cluster anchor 2
- **ipet**: Number of PETs between cluster anchor 1 and cluster anchor 2
- W_{1i} : The probability of pair i being a true pair