ChIA-PET Tool User Manual

Contents

Introduction....1
Download....1
Installation....1
Prerequisites before execution....2
Required supporting software....2
Required supporting data....2
Example ChIA-PET data....3
Compile the Java package....3
Set up the running information and configurations....3
Execution....6
Outputs....14
Abbreviations....21
References....21
Contact us....21

ChIA-PET Tool User Manual

Introduction

Understanding chromatin interactions is important since chromatin interactions create chromosome conformation and link the *cis*- and *trans*- regulatory elements to their target genes for transcription regulation. **Ch**romatin **I**nteraction **A**nalysis with **P**aired-**E**nd **T**ag (**ChIA-PET**) sequencing is a genome-wide high-throughput technology that detects chromatin interactions associated with a specific protein of interest (Fullwood *et al.,* 2009). **ChIA-PET Tool** (Li *et al.,* 2010) is a computational package to process the next-generation sequence data generated from ChIA-PET wet-lab experiments, which contains 7 steps:

1) linker filtering
2) mapping the paired-end reads to a reference genome
3) purifying the mapped reads
4) dividing the reads into different categories
5) peak calling
6) interaction calling
7) visualizing the results

ChIA-PET Tool was originally published in the journal *Genome Biology* in 2010(Li *et al.,* 2010). After that, the package and its modifications were used in many research projects for publications in high-profile journals (Dowen et al., 2014, Kieffer-Kwon et al., 2013, Li et al., 2012, Zhang et al., 2013). The modifications include revising the linker filtering scripts, adopting the state-of-the-art mapping tools (such as BWA and Bowtie), generating the statistics report of the data, and evaluating the quality of the data.

The current ChIA-PET Tool is a command-line program whose execution requires a terminal program. ChIA-PET Tool is mainly coded in Java. Shell scripts are used to glue the different steps in ChIA-PET Tool as a single pipeline. R scripts are used to calculate p-values and generate figures.

Download

The package can be downloaded from https://github.com/GuoliangLi-HZAU/ChIA-PET_Tool/archive/master.zip, which includes ChIA-PET Tool source code in Java, a precompiled JAR file, shell scripts, R scripts and some example files.

Installation

Unpack the package using the following command in your selected directory:

```
$ unzip master.zip
```

Change the working directory to ChIA-PET_Tool-master

```
$ cd ChIA-PET_Tool-master
```

Prerequisites before execution

Required supporting software

ChIA-PET Tool is a pipeline primarily coded with Java language (http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html), with the combination of R scripts and shell scripts. In order to run ChIA-PET Tool properly, it requires the following software.
BWA (http://bio-bwa.sourceforge.net/) is used to map ChIA-PET sequence reads to a reference genome, which could be replaced by other mapping tools, such as Bowtie (http://sourceforge.net/projects/bowtie-bio/files/bowtie).
SAMtools (http://samtools.sourceforge.net/) is used to convert the mapping output from SAM format to BAM format.
BEDtools (https://bedtools.googlecode.com/files/BEDTools.v2.17.0.tar.gz) is used to convert the files from BAM format to bedpe format (a genomic data format defined in BEDtools).
R (http://www.r-project.org/) environment is used to compute the p-values in peak calling and interaction calling and R packages xtable (http://cran.r-project.org/web/packages/xtable/index.html) and RCircos (http://cran.r-project.org/web/packages/RCircos/index.html) are used to generate the graphs for visualization.
Install each software package according to the corresponding instructions and test each software to be run properly.

Required supporting data

To run ChIA-PET Tool, the genome sequence and its mapping index, lengths of individual chromosomes, and cytoband data of the interested genome are required. The genome index for mapping is required in ChIA-PET Tool and needs to be built with BWA in advance.

In our test, human hg19 reference genome (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz), chromosome lengths (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes) and cytoband data (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cytoBandIdeo.txt.gz) were downloaded from UCSC. The random sequences in the genome were removed before further processing.

Mouse mm10 reference genome (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz), chromosome lengths

(ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/mm10.chrom.sizes), and cytoband data (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/database/cytoBandIdeo.txt.gz) were downloaded from UCSC too.

Example ChIA-PET data

In our test, we used published ChIA-PET data associated with RNA polymerase II (RNAPII) from human breast cell line MCF7 and leukemia cell line K562 (Li *et al.,* 2012), which could be downloaded from NCBI Gene Expression Omnibus (GEO) with accession numbers GSM832458 and GSM832459.

You can prepare your own ChIA-PET data for the application of ChIA-PET Tool.

Compile the Java package

A precompiled JAR file is included in the ChIA-PET Tool package. If you need to modify the Java source code, change your working directory to: ChIA-PET_Tool-master/program/LGL/src/, and use the following commands to compile and pack the files.

```
mkdir ../classes

javac -d ../classes @path.txt

cd ../classes/

jar -cvf LGL.jar LGL

rm ../../LGL.jar

cp LGL.jar ../../
```

Set up the running information and configurations

1. Edit the variables in the shell script to set up the information about the supporting tools, data and parameters required for ChIA-PET Tool. Especially, the directories of the programs and data should be set properly to make sure that the programs could run smoothly. We assume that Java and R are available as system commands, and there is no need to set the shell variables for them. If BWA, SAMtools and BEDtools are available as system commands, there is no need to specify the absolute paths for these tools. The following shell variables will be illustrated further in the corresponding steps.

```
PROGRAM_DIRECTORY='/home/local/ChIA-PET/program'

INPUT_INFO_FILE='MCF7.input.information.txt'

INPUT_ANCHOR_FILE='null' ### The path and file name of the given anchors for clustering. If you don't have this file, please specify its value as 'null'.
```


OUTPUT_DIRECTORY='/home/data/results'

OUTPUT_PREFIX='MCF7'

SPECIES='1' ##1:human ; 2:mouse

CYTOBAND_DATA=' hg19_cytoBandIdeo.txt '

CHROM_SIZE_INFO='hg19.chromSize.txt'

GENOME_LENGTH='3E9'

GENOME_COVERAGE_RATIO='0.8' ### the proportion of the genome covered by the reads

GENOME_INDEX='/home/data/genome/hg19.genome.fa'

BWA='/home/local/bwa-0.7.10-r789/bwa'

NTHREADS='6' ### number of threads used in mapping reads to a reference genome

SAMTOOLS='/home/local/samtools-1.1/samtools'

BAM2BEDPE='/home/local/bedtools-2.17.0/bin/bamToBed -bedpe'

MAPPING_CUTOFF='20' ### cutoff of mapping quality score for filtering out low-quality or multiply-mapped reads

MERGE_DISTANCE='2'

SELF_LIGATION_CUFOFF='8000'

EXTENSION_LENGTH='500'

PVALUE_CUTOFF_PEAK='0.00001'

PVALUE_CUTOFF_INTERACTION='0.05'

PEAK_MODE='2' ### value: 1: the peak is an enriched region; 2: the peak is a local summit

MIN_DISTANCE_BETWEEN_PEAK='500' ### minimum distance between two different peaks

MIN_COVERAGE_FOR_PEAK='5' ### minimum coverage for peaks by extended reads

2. Edit the configuration files

There are two configure files: one is the linker file and another one is the sample input information file. In the package, we include two linker files we have used, and one of

them was chosen as an example for illustration.
The linker file has three columns as follows. The first column contains the half-linker sequences. The second column is the start position of the barcodes in the half-linkers. The third column is the length of the barcodes.

GTTGGATAAGATATCGCGG 7 4

GTTGGAATGTATATCGCGG 7 4

The sample input information file is used to specify the input sequence files, linker file, and parameters. In this protocol, we use ChIA-PET data associated with RNA polymerase II (RNAPII) from human breast cancer cell line MCF7 as an illustration. The file "MCF7.input.information.txt"contains the following parameters:

fastq_file_1	The path and file name of read1fastq file in paired-end sequencing. (Mandatory)
fastq_file_2	The path and file name of read2fastq file in paired-end sequencing. (Mandatory)
linker_file	A file specifies linker information, including linker sequences, the start positions and length of barcodes, as described above. (Mandatory)
minimum_linker_alignment_score	Specifies the allowed minimum alignment score.Default:14 (Optional)
minimum_tag_length	Specifies the minimum tag length. Tag is the sequence after linker removal. This parameter is better to be set above 18bp.Default:20 (Optional)
maximum_tag_length	Specifies the maximum tag length. Default:1000 (Optional)
minSecondBestScoreDiff	Specifies the minimum difference of alignment scores between the best-aligned and the second-best aligned linkers. Default:3 (Optional)
output_data_with_ambiguous_linker_info	Determines whether to output the PETs with ambiguous linkers. Value 1 means to output the PETs with ambiguous linkers to specific files, while other values mean not to output the PETs with ambiguous linkers. Default:1 (Optional)

In the current design, the length of the half-linkers is 19 base pairs (bps), and the sequencing mode is 2*36bp. So, the min_linker_alignment_score is set to 14 as the default value, and min_tag_length is set to 20 as the default value. There is a distribution of score differences between best-aligned linker and second-best-aligned linker, which can be used as the reference to set the minSecondBestScoreDiff. If the linker sequences and lengths are changed, the parameters should be changed accordingly.

Execution

ChIA-PET Tool is an easy-to-use pipeline and you can simply run it with one command line after you setup all the required tools, data and parameters. The details of parameters and their meanings will be illustrated in the following steps. In this user manual, we take ChIA-PET data associated with RNA polymerase II (RNAPII) from human breast cancer cell line MCF7 as an example for illustration.

Run ChIA-PET Tool:

```
$ sh run.MCF7.sh
```

The pipeline includes multiple steps, which are explained as below.

Step 1. Linker filtering

```
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar
LGL.chiapet.LinkerFiltering_FastQ_PET ${INPUT_INFO_FILE}
${OUTPUT_DIRECTORY} ${OUTPUT_PREFIX}
```

This command tests all the designed linker sequences with those in the real reads. If the linker sequences in the real reads match one of the designed linker sequences with the specified criteria, the PETs will be output to specific fastq files with the linker sequences excluded. The shell variables in the command line are:

PROGRAM_DIRECTORY: specifies the directory of Java and R scripts in ChIA-PET Tool.

INPUT_INFO_FILE: specifies the file contains input information and parameters for linker filtering. Some details can be seen in "MCF7.input.information.txt" above.

OUTPUT_DIRECTORY: specifies the directory to store the output data from ChIA-PET Tool

OUTPUT_PREFIX: specifies the prefix of all the output files

Output files from linker filtering:

MCF7.runningInformation.LinkerFiltering_FastQ_PET.txt:for the running information

MCF7.1_1.R1.fastq

MCF7.1_1.R2.fastq

MCF7.1_2.R1.fastq

MCF7.1_2.R2.fastq

MCF7.2_1.R1.fastq

MCF7.2_1.R2.fastq

MCF7.2_2.R1.fastq

MCF7.2_2.R2.fastq

MCF7.ambiguous.R1.fastq

MCF7.ambiguous.R2.fastq

MCF7.linker_alignment_score_distribution.txt

MCF7.tag_length_distribution.txt

MCF7.linker_alignment_score_difference_distribution.txt

MCF7.linker_composition_distribution.txt

The main output files from linker filtering program are named as *.m_n.R1.fastq and *.m_n.R2.fastq. For example, the files *.2_1.R1.fastq and *.2_1.R2.fastq are a pair of files with linker B from read1 input file and linker A from read2 input file. The PETs in the files with labels 1_1 or 2_2 are defined as **same-linker PETs**, and the PETs in the files with labels 1_2 or 2_1 are defined as **different-linker PETs**.

Step 2. Mapping to genome

After linker filtering, mapping tool BWA is used to map reads to a reference genome. We assume that the genome index is already built with BWA and the path and the prefix of the index files are specified with shell variable GENOME_INDEX. Since the DNA sequences after linker filtering are around 20-21bp, option "aln" in BWA is used for reads mapping and option "sampe" is used to convert the paired-end mapping results into SAM format. In the following script, the fastq files with linker composition A_A (with label "1_1" in the file name) is used for illustration. All files with other linker compositions should be mapped and processed similarly.

```
> ${BWA} aln -n 2 -t ${NTHREADS} ${GENOME_INDEX}
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R1.fastq 1>
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R1.sai
2>${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R1.sai.output.info.txt
> ${BWA} aln -n 2 -t ${NTHREADS} ${GENOME_INDEX}
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R2.fastq 1>
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R2.sai
2>${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R2.sai.output.info.txt
> ${BWA} sampe -o 1 ${GENOME_INDEX}
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R1.sai
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R2.sai
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R1.fastq
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.R2.fastq
1>${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.sam
```

```
2>${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.sam.output.info.txt
```

In these commands, the shell variables are:

BWA: specifies the path of executable BWA program.

NTHREADS: the number of threads used in mapping.

GENOME_INDEX: specifies the path and prefix of the genome index files built by BWA.

After mapping, Java program is used to extract the uniquely mapped reads with high mapping quality score. The cutoff for mapping quality score is 20. The uniquely mapped reads from the SAM file are defined as the reads with annotation tags: XT:A:U, X0:i:1 and X1:i:0, which means that the mapping is unique, there is only one position with the best alignment score and the number of suboptimal hits found by BWA is 0. SAMtools is used to convert the files from SAM format to BAM format, and BEDtools is used to convert the files from BAM format to bedpe format (bedpe format is a genomic data format defined in BEDtools).

```
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.util.MappingStatistics
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.sam
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1 ${MAPPING_CUTOFF}
> ${SAMTOOLS} view -Sb
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.temp.sam >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bam
> ${BAM2BEDPE} -i ${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bam >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe
```

In these commands, the shell variables are:

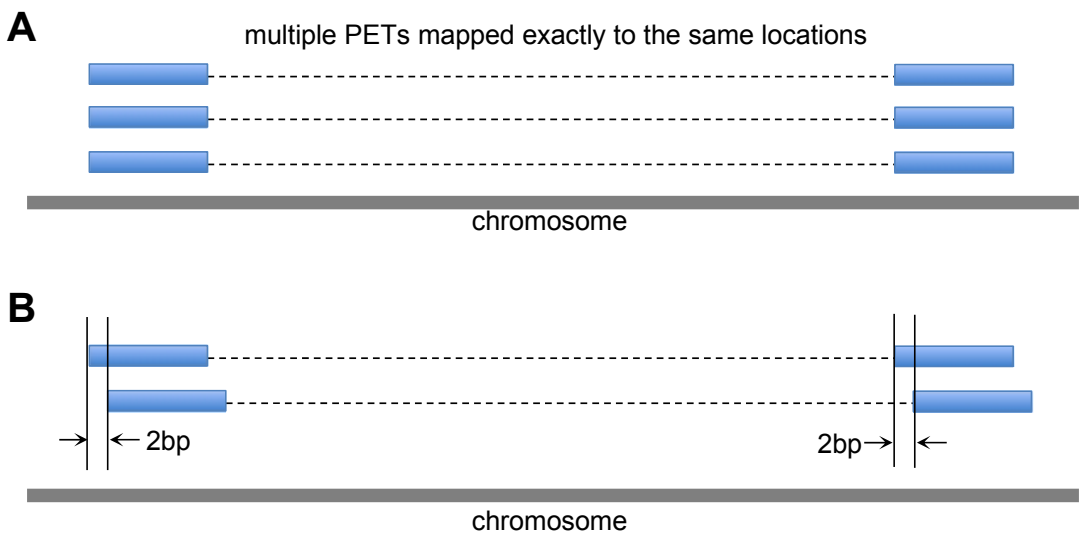
MAPPING_CUTOFF: The cutoff to remove the PETs with low mapping quality score.

SAMTOOLS: specifies the path of SAMtools.

BAM2BEDPE: specifies the path of program bamToBed in the BEDtools package with option -bedpe.

Step 3. Cleaning the mapped PETs

There are two main operations in this steps: 1) Merge all the same PETs (as shown in Figure 1A), 2) Merge all the similar PETs (within ±2bp at the both ends of different PETs, as shown in Figure 1B) from the selected PETs into one unique PET to remove duplicate PETs from PCR amplification and other noises



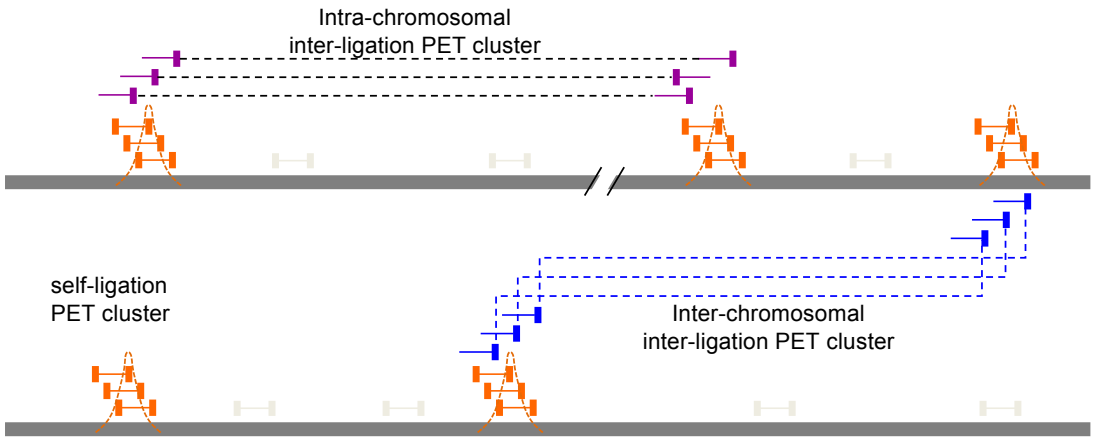

```
> awk -v mapping_cutoff=${MAPPING_CUTOFF} '{if(($1!=".") &&
($4!=".")){print
$1"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t.\t"mapping_cutoff"\t"$9"\t"$10}}'
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.txt
> LANG=C sort <
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.txt >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.sorted.txt
> uniq <
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.sorted.txt >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.unique.txt
> awk '{if($9=="+"){printf $1"\t"$3"\t"$9"\t"}else{printf
$1"\t"$2"\t"$9"\t"}; if($10=="+"){print $4"\t"$6"\t"$10}else{print
$4"\t"$5"\t"$10}}'< ${OUTPUT_DIRECTORY}/${OUTPUT_
PREFIX}.1_1.bedpe.selected.unique.txt | LANG=C sort -k1,1-k4,4-k3,3-k6,6 >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.unique.pet.txt
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.chiapet.MergeSimilarPETs2
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.unique.pet.txt
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.1_1.bedpe.selected.merged.pet.txt
${MERGE_DISTANCE}
```

In these commands, the shell variables are:

MERGE_DISTANCE: specifies the distance limit to merge the PETs with similar mapping locations

Step 4. Categorization of the PETs

The same-linker PETs are classified into three categories based on the coordinates and strand compositions of mapping results: self-ligation PETs, inter-ligation PETs (including intra-chromosomal and inter-chromosomal inter-ligation PETs) and other PETs with short distance as defined in Figure 2.




```
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.chiapet.PetClassification
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.bedpe.selected.pet.txt
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.ipet
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.spet
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.opet ${SELF_LIGATION_CUFOFF}
```

In these commands, the shell variables are:

SELF_LIGATION_CUFOFF specifies the distance threshold between self-ligation PETs and intra-chromosomal inter-ligation PETs.

Step 5. Peak calling

Overlapping regions of self-ligation PETs are used to define transcription factor binding sites and Poisson distribution is used to calculate the p-values for the peaks.

```
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.chiapet.BindingSitesFromPETs
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.spet
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.peak  ${EXTENSION_LENGTH}
```

```
${SELF_LIGATION_CUFOFF} ${MIN_COVERAGE_FOR_PEAK}  ${PEAK_MODE}
${MIN_DISTANCE_BETWEEN_PEAK}
> R --no-save --no-readline --args genomeLengthStr=${GENOME_LENGTH}
genomeCoverageRatioStr=${GENOME_COVERAGE_RATIO}
extensionLengthStr=${EXTENSION_LENGTH}< ${PROGRAM_DIRECTORY}/pois.r
> paste ${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.peak.5K_5K
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.pvalue.pois.txt |awk -v
pvalue_cutoff=${PVALUE_CUTOFF_PEAK} '{if($8<pvalue_cutoff+0.0){print
$1"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8}}' >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.peak.FDRfiltered.txt
```

The meanings of the parameters are as follows：

EXTENSION_LENGTH: specifies the extension length from the location of each tag, which is determined by the median span of the self-ligation PETs.

MIN_COVERAGE_FOR_PEAK: specifies the minimum coverage to define peak regions.

PEAK_MODE: There are two modes for peak calling. One is "peak region" mode, which takes all the overlapping PET regions above the coverage threshold as peak regions, and another one is "peak summit" mode, which takes the highest coverage of overlapping regions as peak regions. The “peak summit” mode is used as the default mode in the package.

MIN_DISTANCE_BETWEEN_PEAK: specifies the minimum distance between two peaks. If the distance of two peak regions is below the threshold, only the one with higher coverage will be kept.

GENOME_LENGTH: specifies the number of base pairs in the whole genome.

GENOME_COVERAGE_RATIO: specifies the estimated proportion of the genome covered by the reads.

PVALUE_CUTOFF_PEAK: specifies p-value to filter peaks that are not statistically significant.

Step 6. Interaction calling

Overlapping regions of inter-ligation PETs are used to define interacting regions and hyper-geometric distribution is used to calculate p-value for the interactions.

```
> java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.file.Pet2Cluster1
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.ipet
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.pre_cluster.txt ${EXTENSION_LENGTH}
${PROGRAM_DIRECTORY}/${CHROM_SIZE_INFO}
> if [ ${INPUT_ANCHOR_FILE} != 'null' ];then
>  java -Xmx10G -cp ${PROGRAM_DIRECTORY}/LGL.jar
LGL.chiapet.PetClusterWithGivenAnchors
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.pre_cluster.sorted
${INPUT_ANCHOR_FILE} ${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.cluster 1
> else
>  java -cp ${PROGRAM_DIRECTORY}/LGL.jar LGL.chiapet.PetCluster2
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.pre_cluster.sorted
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.cluster
> fi
> R --vanilla < ${PROGRAM_DIRECTORY}/hypergeometric.r
> awk -v
pvalue_cutoff=${PVALUE_CUTOFF_INTERACTION}'{if($13<pvalue_cutoff+0.0)pr
int}'< ${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.cluster.withpvalue.txt >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.cluster.FDRfiltered.txt
```

The parameters are as follows：

CHROM_SIZE_INFO: specifies the file that contains the length of each chromosome.

INPUT_ANCHOR_FILE: a file which contains user-specified anchors for interaction calling. It is a 3-column bed format file. If you don't have this file, please specify the value of this variable as “null” instead.

PVALUE_CUTOFF_INTERACTION: specifies p-value to filter interactions which are not statistically significant.

Step 7. Visualization Report

During the execution of ChIA-PET Tool, the statistics of the library is generated and summarized in a html report file. Another script is used to convert the mapping results, peaks and chromatin interactions to BedGraph (peaks) and GFF(clusters) format for visualization in UCSC browser.

```
> Rscript --verbose
${PROGRAM_FOLDER}/Rscript_and_genome_data/ChIA-PET_Tool_Report.r
${PROGRAM_FOLDER} ${OUTPUT_FOLDER}/files_for_report ${OUTPUT_PREFIX}
${CYTOBAND_DATA} ${SPECIES}
> awk -v OUTPUT_PREFIX=${OUTPUT_PREFIX} 'BEGIN{print "browser position
chr1:9997500-10922500\nbrowser hide all\ntrack name=ChIAPET
description=\"ChIA-PET "OUTPUT_PREFIX" human\""}$1==$4{print
$1"\tChIAPET\t"OUTPUT_PREFIX"\t"$2"\t"$3"\t.\t.\t.\ttouch"NR"\n"$4"\tCh
IAPET\t"OUTPUT_PREFIX"\t"$5"\t"$6"\t.\t.\t.\ttouch"NR}'
${OUTPUT_FOLDER}/${OUTPUT_PREFIX}.cluster.FDRfiltered.txt >
${OUTPUT_FOLDER}/${OUTPUT_PREFIX}.toUCSC.gff
> awk -v OUTPUT_PREFIX=${OUTPUT_PREFIX} 'BEGIN{print "browser position
chr17:73626165-73912870\nbrowser hide all\nbrowser pack refGene
encodeRegions\nbrowser full altGraph\n#300 base wide bar graph, autoScale
is on by default == graphing\n#limits will dynamically change to always show
full range of data\n#in viewing window, priority = 20 positions this as the
second graph\n#Note, zero-relative, half-open coordinate system in use for
bedGraph format\ntrack type=bedGraph name=\""OUTPUT_PREFIX" ChIAPET peak\"
description=\""OUTPUT_PREFIX" ChIA-PET peaks\" visibility=full
color=32,178,170 altColor=0,255,127 priority=20"}{print
$1"\t"$2"\t"$3"\t"$4}'
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.peak.FDRfiltered.txt >
${OUTPUT_DIRECTORY}/${OUTPUT_PREFIX}.peak.toUCSC.bedGraph
```

CYTOBAND_DATA: specifies the ideogram data used to plot intra-chromosomal peaks and interactions.

SPECIES: specifies the genome used to plot inter-chromosomal interactions, 1 for human and 2 for mouse. Currently only two species are supported in R package RCircos.

Outputs

ChIA-PET Tool can process the next-generation sequence data from ChIA-PET experiment with 7 steps to generate enriched binding peaks of the protein of interest and the related chromatin interactions. We demonstrated the application of ChIA-PET Tool with RNAPII-associated ChIA-PET data from human breast cancer cell line MCF7 as an example. The summary statistics were listed in the following table.

Summary statistics of RNAPII-associated ChIA-PET data

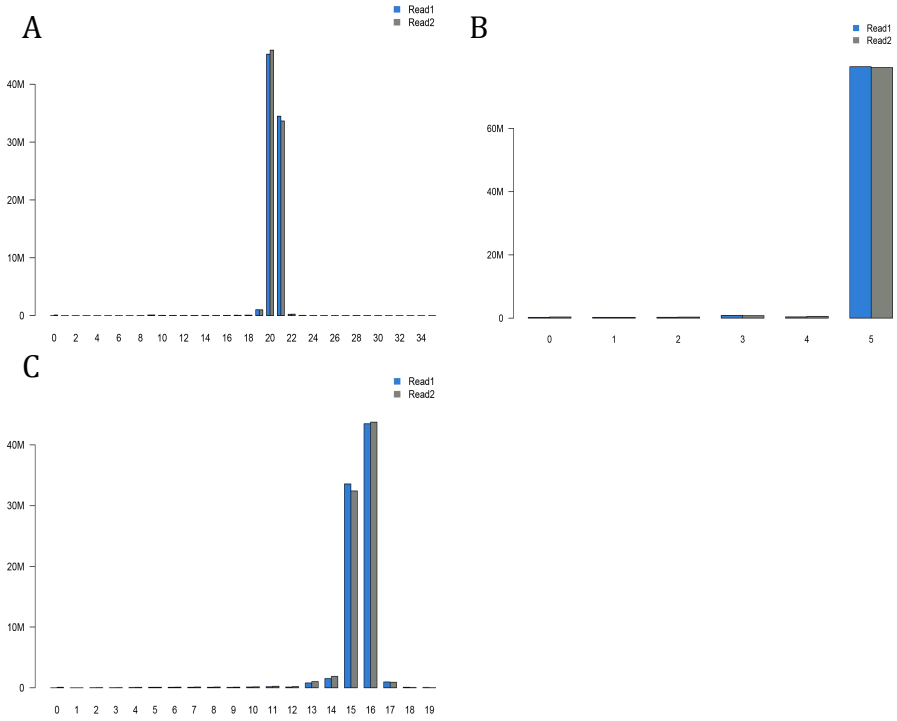
Category	Number	Percentage	Percentage of total PETs	Order
Total PETs	81,557,570	NA	NA	(1)
Same-linker PETs after linker filtering	66,821,587	81.93% of (1)	81.93% of (1)	(2)
Uniquely mapped same-linker PETs	9,371,948	14.03% of (2)	11.49% of (1)	(3)
Merging same same-linker PETs	7,894,875	84.24% of (3)	9.68% of (1)	(4)
Merging similar same-linker PETs	7,884,966	99.87% of (4)	9.67% of (1)	(5)
Self-ligation PETs	2,406,617	30.52% of (5)	2.95% of (1)	(6)
Inter-ligation PETs	5,118,291	64.91% of (5)	6.28% of (1)	(7)
Other PETs with short distance	360,058	4.57% of (5)	0.44% of (1)	(8)
Peaks from self-ligation	8,526	NA	NA	(9)
Interacting pairs	13,627	NA	NA	(10)

Note: NA – Not applicable

The results from linker filtering (Refer to the report HTML file for more details):
(1) Distribution of linker alignment scores. Figure 3A reflects the distribution of best alignment scores from the designed linker sequences to the reads. We can see that most of the alignment scores are 15 or 16, which means that most of the linker sequences in the reads are 15 or 16 base pairs.

(2) Distribution of linker alignment score differences. Figure 3B reflects the distribution of alignment score difference distribution between the best-aligned linker and the second-best aligned linker. This distribution is used to check how different it is from the best-aligned linker to the second-best aligned linker, in case there are sequence errors or ambiguities in the reads.

(3) Distribution of tag lengths. Figure 3C reflects the distribution of tag lengths after trimming the best-aligned linker sequences from the reads. We can see that most of the tag lengths are 20 or 21bp, which is consistent with enzyme *Mme*I's digestion property.



(4) Linker composition statistics. This file reports the proportion of each linker combination in the reads. The results show that most (81.93%) of the PETs are composed of same-linker (A_A or B_B). This indicates proper proximity-ligation within individual molecules.

	A_A	A_B	B_A	B_B	Ambiguous	Total
Numbers	33,822,895	4,744,854	4,748,192	32,998,692	5,242,937	81,557,570
Percentage	41.47%	5.82%	5.82%	40.46%	6.43%	100%

A_A: refers to PETs that both reads optimally aligned to linker A.

A_B: refers to PETs that read1 and read2 optimally aligned to linker A and linker B respectively.

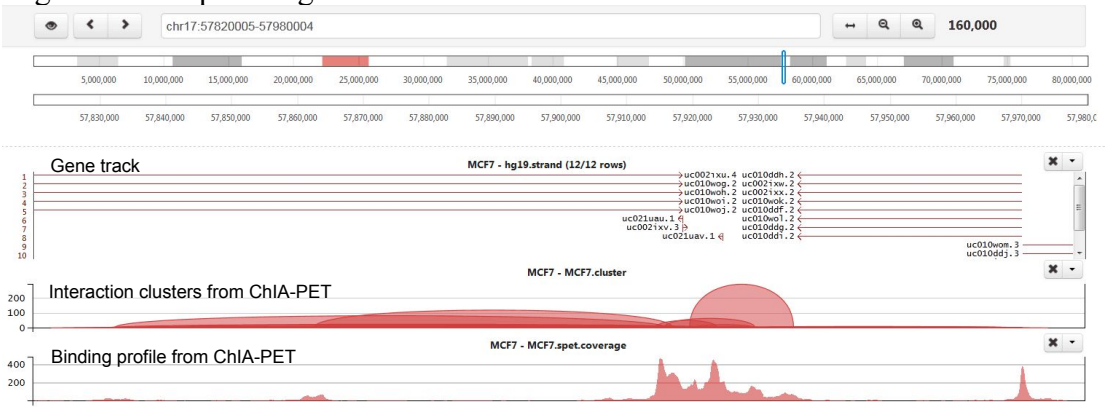
B_A: refers to PETs that read1 and read2 optimally aligned to linker B and linker A respectively.

B_B: refers to PETs that both reads optimally aligned to linker B.

Ambiguous: Ambiguous refers to PETs that not satisfy either one of the criteria below: 1) best linker alignment score is more than the cutoff, 2) the difference between second best and best score exceeds score cutoff, 3) tag length should conform to the specified range, and 4) the barcodes must be completely matched with reads.

Total: Total number of PETs.

Figure 4 shows a screenshot of binding profile and chromatin interactions from the example data set. We can see that the peaks from self-ligation PETs of RNAPII-associated ChIA-PET are mainly enriched around gene promoter regions. The interactions are mainly between gene promoter regions and other regulatory elements, which indicate that RNAPII-associated chromatin interactions are involved in gene transcription regulation.



The first 6 lines of peak output file from Step 5-peak calling are in the following table.

chrom	summit start	summit end	peak coverage	p-value	p.adjust
chr1	853737	853882	56	2.64e-12	4.54e-11
chr1	854798	854877	49	4.89e-08	5.17e-07
chr1	901960	902162	27	1.59e-09	1.97e-08
chr1	935402	935433	44	2.18e-14	4.74e-13
chr1	955915	955938	24	3.67e-10	4.91e-09
chr1	1005210	1005211	49	2.72e-08	2.95e-07

chrom: chromosome name

summit start: The start coordinate of peak summit.

summit end: The end coordinate of peak summit.

peak coverage: Highest coverage by tags in a peak.

p-value:This value represent the statistical significance of a peak, which is calculated by Poisson distribution.

p.adjust: P.adjust means p-value adjusted with Benjamini-Hockberg method for multiple hypothesis testing.

The first 6 lines of interaction output file from Step 6 interaction calling are in the following table.

chrom1	start1	end1	chrom2	start1	end2	ipet counts	type	distance
chr1	856188	856759	chr1	876341	876951	2	1	21857

chr1	858046	858611	chr1	867685	868224	2	1	11791
chr1	858822	859887	chr1	876936	878041	3	1	155161
chr1	859010	859668	chr1	872863	873824	2	1	47494
chr1	860536	861482	chr1	870654	871570	2	1	146003
chr1	895260	896033	chr1	908502	909403	2	1	144478

tag count within anchor 1	tag count within anchor 2	p-value	p.adjust	-log10(p-value)	-log10(p.adjust)
34	22	2.47E-09	4.94E-09	8.61	8.31
13	8	4.17E-11	1.77E-10	10.38	9.75
71	64	1.33E-11	6.68E-11	10.88	10.17
47	13	1.61E-09	3.49E-09	8.79	8.46
84	18	1.02E-08	1.62E-08	7.99	7.79
13	13	1.16E-10	4.12E-10	9.94	9.38

chrom1: The name of the chromosome on which the cluster anchor 1 exists.

start1: The start coordinate 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 coordinate of cluster anchor 2.

end2: The end coordinate of cluster anchor 2.

ipet count: Number of PETs between cluster anchors 1 and 2.

type: Interactions type. Value 1 represents intra-chromosomal interaction, and 0 represents inter-chromosomal interaction.

distance: Distance between anchors of an intra-chromosomal interaction cluster. If two anchors are located on different chromosomes, the value is set to 2,147,483,647.

tag count within anchor 1: Number of tags that fall in the cluster anchor 1.

tag count within anchor 2: Number of tags that fall in the cluster anchor 2.

p-value: This value represents the statistical significance of a chromatin interaction, which is calculated by hyper-geometric distribution.

p.adjust: P.adjust means p-value adjusted with Benjamini-Hockberg method for multiple hypothesis testing.

-log10(p-value): The negative logarithm of p-value.

-log10(p.adjust): The negative logarithm of adjusted p-value.

The statistics of the clusters are summarized in the following two tables.

Statistics of clusters with different PET counts and intra-chromosomal information

PET counts	No. of clusters	No.intra chrom clusters	No.inter chrom clusters	Percent of intra chrom clusters
2	10216	9953	263	97.42%
3	1709	1693	16	99.06%
4	625	618	7	98.88%
5	326	319	7	97.85%
6	198	196	2	98.99%
7	133	130	3	97.74%
8	73	73	0	100%
9	59	57	2	96.61%
>=10	288	277	11	96.18%
Total	13627	13316	311	97.72%

We can see that majority of the clusters are intra-chromosomal interactions.

Span distribution of the interactions.

Distance	Frequency	Interaction type
< 100Kb	10851	Intra-chromosomal
[100Kb, 1Mb)	2345	Intra-chromosomal
[1Mb, 10Mb)	103	Intra-chromosomal
> 10Mb	17	Intra-chromosomal
Different chromosomes	311	Inter-chromosomal

Majority of the interactions are with spans within 1Mb.

Figure 5 shows the chromosome view of binding peaks and intra-chromosomal chromatin interactions. Blue bars above the chromosomes are transcription factor binding sites, and red curves under chromosomes are intra-chromosomal chromatin interactions. The binding peaks are distributed allover the whole genome. Most of the intra-chromosomal chromatin interactions are within a short span (within 1Mb) and a small portion of chromatin interactions can span a very long distance.

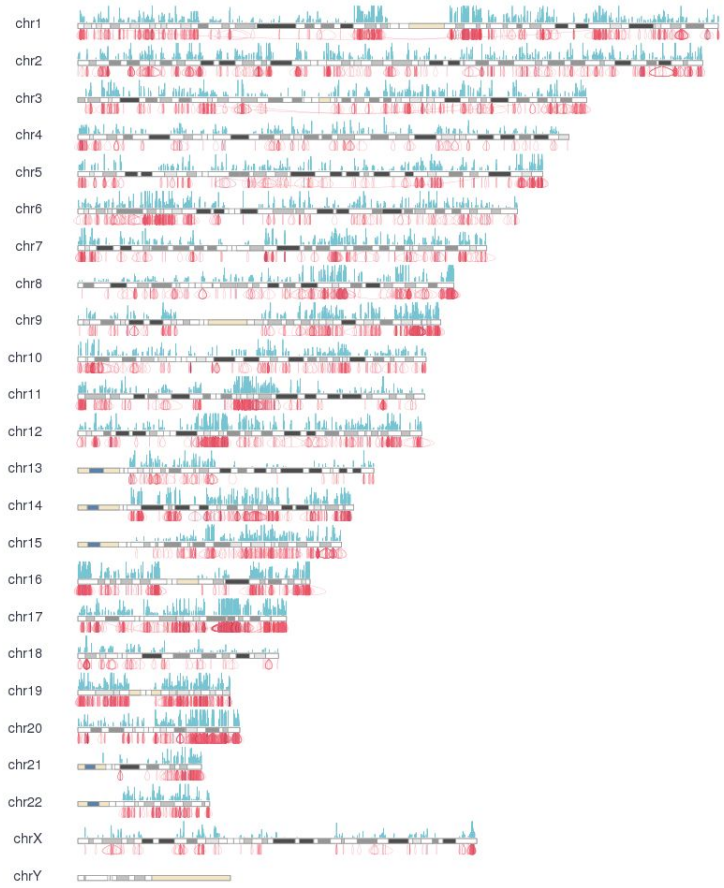
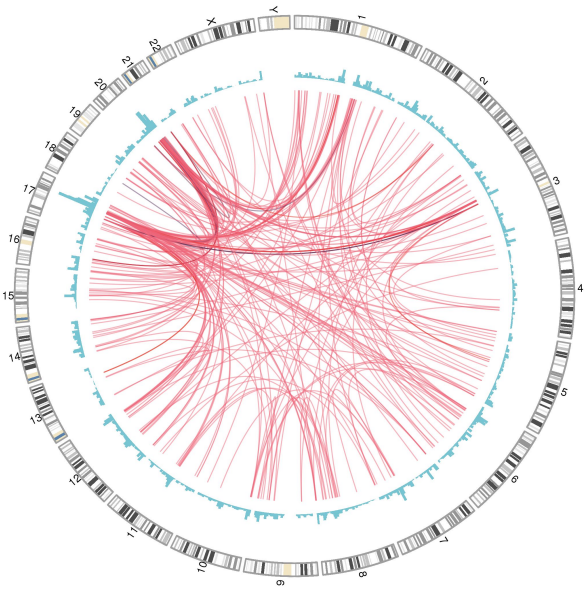


Figure 6 shows the circular view of the inter-chromosomal chromatin interactions. Compared to intra-chromosomal chromatin interactions, there are fewer inter-chromosomal chromatin interactions.



Evaluation of the data quality

One concern from the users is the quality of the ChIA-PET libraries they generate or use. In this pipeline, we summarize the statistics of ChIA-PET libraries in Table 1 for quality evaluation:

1) Percentage of the same-linker PETs over the total PETs. By the nature of the ChIA-PET design, there will be more same-linker PETs than the different-linker PETs. In general, we saw the percentage of the same-linker PETs varied from 60% to ~99%. If such percentage is above 75%, the library is good in the linker composition level;

2) Percentage of the uniquely mappable PETs over the total PETs, which varies with the libraries;

3) Percentage of the PETs after merging those mapped to exactly the same positions over the uniquely mappable PETs. This is more about reducing the redundancy from PCR amplification. If this percentage is too low, see less than 30%, it means that there are more PETs from PCR amplification and the data is already near saturation. Then it does not worth to re-sequence this library for more distinct PETs. If the PETs after merging those mapped exactly to the same positions are 70% or more of the uniquely mappable PETs, deeper sequencing can be applied to get more data for this library.

4) Percentage of inter-ligation PETs over all the PETs after purification. This is about the efficiency catching the interacting PETs. If this percentage is low, it means that the library has too few inter-ligation PETs and it is not good enough for chromatin interaction detection.

5) Percentage of intra-chromosomal inter-ligation PETs over the total inter-ligation PETs. From the current understanding, most of the chromatin interactions are within the individual chromosomes. Therefore, there should be more intra-chromosomal PETs than inter-chromosomal PETs. If there are more inter-chromosomal PETs, it means that the proximity ligation introduces many random ligations.

6) Peak number. This depends on the transcription factor used, and should be compared with the background knowledge or ChIP-Seq data available. For RNAPII and CTCF, there are tens of thousands of peaks in a good ChIA-PET library from human and mouse.

7) Interaction number. This depends on the transcription factors used. For RNAPII and CTCF, there are tens of thousands of interactions in a good ChIA-PET library from human and mouse.

Table 1. Some statistics of a ChIA-PET library

Category	From the test data	Note
Percent of same-linker PETs over total PETs	81.93%	About the linker ligation
Percent of uniquely mappable PETs over total PETs	11.49%	About the mapping of the PETs
Percent of PETs after merging those mapped to the exact same positions over uniquely mappable PETs	84.24%	About redundancy from PCR amplification
Percent of inter-ligation PETs over PETs	64.91%	Efficiency in catching the

after purification		interaction PETs
Percent of intra-chromosomal inter-ligation PETs over inter-ligation PETs	97.72%	
Number of peaks	8,526	
Number of interactions	13,627	

Abbreviations

PET – Paired-End Tag sequences;
ChIA-PET – Chromatin Interaction Analysis with Paired-End Tag sequencing

References

- Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in Mammalian chromosomes. Cell 159, 374-387, doi:10.1016/j.cell.2014.09.030 (2014).
- Fullwood, M. J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58-64 (2009)
- Kieffer-Kwon, K. R. et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell 155, 1507-1520, doi:10.1016/j.cell.2013.11.039 (2013).
- Li, G. et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biology 11(2):R22 (2010)
- Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).
- Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504, 306-310, doi:10.1038/nature12716 (2013).

Contact us

If you have any problems or suggestions, you could email to Dr. Guoliang Li (guoliang.li@mail.hzau.edu.cn).