ENCODE, ChIA-PET
Data processing pipeline description for Ruan Lab

ChIA-PIPE is a fully automated data processing pipeline to process ChIA-PET data for both in-situ ChIA-PET and long-read ChIA-PET (Lee et al., 2020 Genome Biology. doi:10.1126/sciadv.aay2078). The complete code is available in github site (https://github.com/TheJacksonLaboratory/ChIA-PIPE).

1. Pipeline install and launch the process

After users download and install ChIA-PIPE from the github site, they can fill in pipeline parameters in a config file, and then launch ChIA-PIPE on HPC. Key pipeline parameters in a config file follow, using a GM10248 CTCF ChIA-PET experiment as an example (ENCODE accession: ENCSR627XQX):

```
# The name of the sequencing run
run="LHG0047H"

# The type of sequencing run:
#    "miseq" - around 30 million reads
#    "hiseq" - around 300 million reads
#    "pooled" - around 1 billion reads
# It adjusts a proper computing resource (processors, memory, and walltime).
run_type="hiseq"

# The factor for which the IP was performed
ip_factor="CTCF"

# Cell type
cell_type="GM10248"

# The directory containing the input FASTQ files
data_dir= "/projects/ruan-lab/processing/fastq/${run}"

# The name of the primary genome
# For example: "hg19", "hg38", "dm3", "mm9", "mm10"
genome="hg38"

# The reference genome FASTA file for aligning the reads
# (The same directory must also contain the BWA index files)
# Users download and pass the fasta file.
fasta="/projects/ruan-lab/processing/genomes/hg38/hg38.fa"

# The chrom.sizes file from UCSC Genome browser
# for the relevant genome build
chrom_sizes="/projects/ruan-lab/processing/genomes/hg38/hg38.chrom.sizes"

# The BAM file for the ChIP-seq input control
# (Required for spp; not required for macs2)
# If not available, set to "none"
```

input_control="none"

# The peak-calling algorithm ("macs2" or "spp")
peak_caller="macs2"

Then, users can launch ChIA-PIPE using the following command:
qsub -F "--conf config_file.sh" chia_pipe_installed_folder/0.chia_pipe_hpc.pbs

After the pipeline finish to process ChIA-PET sequencing reads (fastq files), users can check HPC output file (.o) and log file (.log) to check whether the pipeline works properly. The pipeline generates a summary report file (.tsv), and users can check the following quality metric value.

2. Quality metric

The pipeline generates a summary report file (.tsv), and users can check the following quality metric value in the report file. First, deep sequencing ChIA-PET read count is mostly more than 200 million reads, and “fraction read pairs with linker” is more than 0.5. SPP called peaks with input control are less noisy than MACS2 called peaks without input control file. Peak number from SPP or MACS2 will be more than 5,000, if experiment and pipeline are properly processed. Inter-chromosomal PET is more random event while intra-chromosomal PET is meaningful biological contacts. Ratio of intra-chromosomal PET over inter-chromosomal PET and also ratio of intra-chromosomal PET cluster (ChIA-PET loop) over inter-chromosomal PET cluster are indication, more than 1 for in-situ ChIA-PET and less than 1 for long-read ChIA-PET. Long-read ChIA-PET contains more background noise, while in-situ ChIA-PET contains less background noise. The following quality assessment table shows high quality in-situ ChIA-PET data sets.

Key parameters for in-situ ChIA-PET library quality assessment

1. **Total read pair**: for the ENCODE4 project, we routinely generate 180-550 million paired end reads (2X150 bp) for each library dataset. In average, 300 million paired end reads are generated.
2. **Read pair with bridge linker**: we use a bridge linker to connect two chromatin fragments in ChIA-PET protocol. The higher fraction of the read pairs containing the linker, the higher numbers of chromatin interactions will be captured. Usually, the fraction value is around 0.7-0.9.
3. **Non-redundant PET**: The most useful data for chromatin interaction is the uniquely mapped non-redundant PET. The more the better. Usually >10 million non-redundant PET per dataset will provide robust interactions data.
4. **PET redundancy**: The PET redundancy reflects the complexity of a library. Lower PET redundancy means higher library complexity. Therefore, further sequencing depth can go for additional data. Most of the ChIA-PET libraries would reach plateau after total reads beyond 500 million.
5. **Intra-chr PET**: The intra-chromosome PETs are the most useful data for discovery of cis-acting elements involved in chromatin interactions. The number of PETs varies in the range of 10 to 100 millions depending on library sequencing depth and protein factors.
6. **Ratio of intra/inter PET**: This is a very useful value to assess the noise level of proximity ligation. Most of the inter-chromosome PET were derived from random non-specific ligation, except possible translocation events in cancer cells and likely rare specific inter-chr contacts, which will be analyzed separately. In our experience, most of the high quality ChIA-PET library datasets have the ratio value higher than 2, meaning that more intra-chromosome PETs than inter-chromosome PETs.

7. **Intra-chr PET cluster**: The cluster of PETs are a group of PETs that are overlapped in mapping alignment to the reference genome, indicating that these PETs are derived from chromatin interactions of same loci and repeatedly captured from different cells. The minimal PET count of a cluster is 2 and could number up to 1000s. The number of a PET cluster measure the relative frequency of chromatin interactions between the two loci in a cell population. Higher PET count in a cluster suggests higher contact frequency and higher confident of the interaction data.
8. **Ratio of intra/inter PET cluster**: The PET clusters derived from inter-chr PET data can be used as random background noise, and the ratio of intra/inter PET cluster provides the best indicator of the quality of a ChIA-PET library.

Below is an example of statistics of the CTCF and RNAPII ChIA-PET libraries. The key parameters are in BOLD. As shown, the quality of the CTCF ChIA-PET library data is usually higher than that of the RNAPII ChIA-PET library data. The RNAPII ChIA-PET library is capturing RNAPII associated chromatin interactions (promoter to enhancer), related to transcriptional activities of cells.

Library ID	LHG0047H	LHG0118V	Note
Seq method (2X150)	Hiseq	Novaseq	
Reference genome	hg38	hg38	
Cell type	GM10248	GM10248	
Factor	CTCF	RNAPII	
Total read pair	**340,840,469**	**316,510,593**	Library seq depth & quality
Read pair with bridge linker	**320,834,185**	**286,944,905**	Higher fraction indicates higher quality of the library
Fraction of read pairs with linker	**0.94**	**0.91**	
Quality non-redundant tag	441,205,365	74,642,122	Tags used for peak calling # of peaks varies depending on factors
Protein factor binding peak	929,221	222,016	
Paired-end-tag (PET)	218,869,947	170,109,710	These PETs are used for interaction analysis
Uniquely mapped PET	162,534,265	127,491,848	
Non-redundant PET	**142,662,939**	**16,107,717**	
PET redundancy	**0.12**	**0.87**	Library complexity
Self-ligation PET (< 8 kb)	27,534,605	10,575,144	Chromatin fragment
Inter-ligation PET (> 8 kb)	115,128,334	5,532,573	Long range interaction
Intra-chr PET	**83,450,947**	**3,997,431**	Most useful data are intra-chr. Higher ratio (>1) indicates higher quality of Interaction data.
Inter-chr PET	31,677,387	1,535,142	
Ratio of intra/inter-chr PET	**2.63**	**2.6**	
Singleton	101,710,672	4,135,031	
Intra-chr singleton	**70,457,789**	**2,909,661**	
Inter-chr singleton	31,252,883	1,225,370	
PET cluster	4,453,925	498,026	
Intra-chr PET cluster (≥2)	4,245,707	379,818	Mostly useful data
Inter-chr PET cluster (≥2)	208,218	118,208	Mostly noise

Ratio of intra/inter-chr PET cluster (≥2)	**20.39**	**3.21**	Frequently recurrent PET cluster Higher ratio means higher quality of the data
Ratio of intra/inter-chr PET cluster (≥5)	**1,914.56**	**3.82**	
Ratio of intra/inter-chr PET cluster (≥10)	**13,248.57**	**6.97**	
Intra-chr PET cluster			
PET number_2	3,097,854	263,388	
PET number_3	606,301	54,569	
PET number_4	214,162	24,180	
PET number_5	101,160	13,720	
PET number_6	56,565	8,343	
PET number_7	35,610	5,186	
PET number_8	23,968	3,290	
PET number_9	17,347	2,180	
PET number_10	12,873	1,298	
PET number >10	79,867	3,664	
Inter-chr PET cluster			
PET number_2	201,273	86,018	
PET number_3	6,234	15,427	
PET number_4	540	6,887	
PET number_5	99	3,942	
PET number_6	45	2,354	
PET number_7	11	1,483	
PET number_8	7	871	
PET number_9	2	514	
PET number_10	0	308	
PET number >10	7	404	

Processed files submitted to ENCODE **(**pipeline schematic and details can be viewed at https://www.encodeproject.org/pipelines/ENCPL169TBL/):

1. Mapped file
Input: raw fastq file for R1 and R2 (fastq.gz)
Processing module: CPU (https://github.com/cheehongsg/CPU) for filter linker; map Paired-End Tags; Uniquely mapped PETs; De-duplicated PETs
CPU (ChIA-PET Utilities) version: 0.0.1a-r2
Output: bam file

2. ChIA-PET loop file
Input: bam file
Processing module: CPU (cluster commend), version: 0.0.1a-r2
Output: bedpe file (7 columns: chr1, start, end, chr2, start, end, score)

3. ChIA-PET peak file

Input: bam file
Processing module: SPP version 1.13 or MACS2 version 2.1.0
When input control file is available, ChIA-PIPE used SPP. If not, ChIA-PIPE used MACS2.
Output: bed file (SPP: broadPeak file, 9 columns or MACS2: narrowPeak file, 10 columns)

4. ChIA-PET binding coverage file
Input: bam file
Processing module: bedtools (genomecov commend), version 2.26.0
Output: bedgraph file (4 columns: chr, start, end, value)

5. ChIA-PET 2D contact map file
Input: bam file
Processing module 1: bam2pairs
No version assigned for bam2pairs util, https://github.com/4dn-dcic/pairix/blob/master/util/bam2pairs/README.md
Processing module 1 output: pairs format v1.0, (pairs.gz), input file for processing module 2
Processing module 2: Juicer tools (pre commend), version 1.7.5
Output: hic file (Juicebox readable 2D contact map file)

6. Converted loop file to visualize using UCSC genome browser
Input: ChIA-PET loop file (bedpe)
Processing module: bedToBigBed version 2.7
Output: bigInteract file

7. Converted peak file to visualize using UCSC genome browser
Input: ChIA-PET peak file (broadPeak or narrowPeak)
Processing module: bedToBigBed version 2.7
Output: bigBed file

8. Converted binding coverage file to visualize using UCSC genome browser
Input: ChIA-PET binding coverage file (bedgraph)
Processing module: bedGraphToBigWig version 4
Output: bigwig file

Brief summary of processing steps in ChIA-PIPE automated process:

1. Identify bridge linker sequence from sequencing reads, and classify each read as no-linker detected ('none'), single linker detected from either read1 or read2 ('singlelinker-single'), and single linker detected from both read1 and read2 ('singlelinker-paired'). We used 19bp bridge linker sequence with the beginning A, coming from A tailing step. ChIA-PIPE searches for the 20bp sequence (ACGCGATATCTTATCTGACT) by allowing 2bp mismatches. Detected linker sequences are filtered out before proceeding to the mapping step.

2. Map sequencing reads to the reference genome using BWA-MEM and BWA-ALN mapping tools that are integrated into ChIA-PIPE. Reads with low mapping quality (MAPQ <30) are filtered out, and non-redundant PETs are retained in a bam file format. Bam file from singlelinker-paired reads are used to produce 2D contact map file and ChIA-PET loop file. Bam files from no-linker detected, singlelinker-single, and singlelinker-paired reads are

merged to produce ChIA-PET protein binding coverage file and peak file in subsequent steps.

3. Produce 2D contact map (hic) file using singlelinker-paired bam file and Juicer tool. The file can be visualized via Juicebox.

4. Produce loop (bedpe) file. First, ChIA-PIPE filters out inter-chromosomal contacts and self-ligation PETs with genomic span less than 8 kbp. Then, the 5' end of each inter-ligation PET is extended by 500 bp along the reference genome, which makes chromatin fragments closer to the actual protein binding position in DNA. Finally, ChIA-PIPE merges overlapping extended inter-ligation PETs into a cluster with the number of overlapped inter-ligation PET as a PET count of the loop. Loops with PET count =1 are called singletons and we filter them out to retain only highly confident frequent loops. We submitted ChIA-PET loop file with PET count ≥ 3 to ENCODE. We also convert the loop file to bigInteract file for users to visualize it using UCSC genome browser.

5. Produce binding coverage (bedgraph) file using the merged bam file in the step 2 and bedtools genomecov, and then we convert bedgraph file to bigwig file using bedGraphToBigWig tool to be visualized by UCSC genome browser.

6. Produce peak file (bed) using the merged bam file in the step 2 and peak-calling software SPP or MACS2. We convert peak file to bigBed file to be visualized using UCSC genome browser.