Manual of FitHiChIP (Version 2.0)

Dr. Sourya Bhattacharyya
Supervisor: Dr. Ferhat Ay
Vijay-Ay lab
La Jolla Institute for Allergy and Immunology
San Diego, CA 92037, USA

December 1, 2017

FitHiChIP computes statistically significant interactions between fixed size bins (user specified) of a given HiChIP [3] or PLAC-seq [2] data, by applying the method FitHiC [1]. FitHiChIP produces statistical significant *cis* interactions from a given HiChIP or PLAC-seq alignment or fastq files. The code supports multiprocessing, and can be executed either in a laptop or in a computational cluster, in Linux environment.

1 Changes made in Release 2.0

FitHiChIP 2.0 is a major upgraded package from its earlier version 1.0. Below are the list of changes found in this version:

1. Added processing of the valid pairs (txt or txt.gz) file generated by HiC-Pro [4] pipeline.

2. Processing input parameters through a configuration file. Sample examples of configuration file and corresponding script is also included in the package.

3. FitHiC module is now incorporated with bias correction (normalization) technique. User can either employ raw (earlier) FitHiC or use bias correction through specific input parameters (described subsequently).

4. Minor modification in the contact count determination module.

5. Merging significant interactions originated and terminated in neighboring (fixed size) bins, to form a single interaction. This is an optional feature.

6. Mandatory requirement of a peak detection file, either generated by MACS2 [5] or related packages, or downloaded from ENCODE.

7. Computation of coverage for individual bins is modified to support BAM, PAIRIX or HiC-Pro generated Validpairs files together.
8. Detection of contacts (among bins) is first performed without any distance threshold. Later, contacts are filtered according to distance thresholds.

2 Installation

FitHiChIP requires a Linux environment (with bash) for execution. It is developed using Python version 2.7.13, R version 3.3.3, and perl (included in any standard linux distribution). It also requires the following packages (values within brackets denote the package versions we have used):

1. Python packages:
   - pypairix (0.1.7) to support the pairix data format (https://github.com/4dn-dcic/pairix)
   - pysam (0.10.0) to process the bam file (https://github.com/pysam-developers/pysam)
   - networkx (https://networkx.github.io/)
2. R libraries:
   - splines (FitHiC implementation)
   - fdrtool (FitHiC implementation)
   - ggplot2 (various plotting routines)
   - optparse (parsing command line arguments)
   - GenomicRanges
   - Parallel (for multiprocessing)
3. Package MACS2 for peak calling (https://github.com/taoliu/MACS)
4. HiC-Pro (https://github.com/nservant/HiC-Pro) if HiC-Pro validpairs file is provided as input.
5. bedtools (http://bedtools.readthedocs.io/en/latest/) version 2.26

Creation of alignment files from a given pair of fastq reads requires the following packages to be installed:

- samtools (version 1.6) (http://www.htslib.org/doc/samtools.html)
- BWA (http://bio-bwa.sourceforge.net/)
- Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- picard tool (http://broadinstitute.github.io/picard/)

- pairix (https://github.com/4dn-dcic/pairix)
- bam2pairs (https://github.com/4dn-dcic/pairix)

Paths of the respective packages are to be included in the system path.

3 Types of interactions supported

FitHiChIP produces four different types of interactions (contacts) among fixed size genomic segments (bins):

- **ALL to ALL**: Interactions between every pairs of bins, subject to the distance thresholds employed.
- **Peak to Peak**: Interactions between two peak regions, where peaks are provided as a separate input file. Peaks can be either computed using the packages like MACS2 [5], or can be downloaded from ENCODE.
- **Peak to Non Peak**: Interaction between a peak segment and a non peak segment (fixed size bin which does not overlap with any peaks).
- **Peak to ALL**: Encapsulates both peak to peak and peak to non peak interactions. Such an interaction contains a peak segment in one or both ends.

4 Command line options

FitHiChIP requires all input files and parameters to be specified via a configuration file. A sample configuration file is provided along with this package.

The shell script **FitHiChIP.sh** is the main executable. It must be invoked with the following command (assuming the executable and the configuration file is in present working directory):

./FitHiChIP.sh -C *ConfFile*

where, *ConfFile* denotes the name of configuration file. A sample script for executing FitHiChIP is also provided along with this package. Input parameters of FitHiChIP (to be mentioned in the configuration file) are described below. Each option has the format **param=value**.

4.1 Generic command line options

1. InpFile=*InputAlignmentFile*
    - Input alignment file, in either BAM, PAIRIX, or HiC-Pro pipeline generated validpairs format.
    - BAM file, if provided, should be sorted and indexed.

- Pairix alignment file *pairix.gz* should be accompanied with an index file *pairix.px2*.
- HiC-Pro generated validpairs file can be either in .txt or .txt.gz format.

2. PeakFile=*PeakFile*

   - BED formatted file with the detected peak regions. Computed either by MACS2 (or equivalent packages), or downloaded from reference such as ENCODE.
   - If the input alignment file is BAM formatted, and peak file is not provided, peaks are computed using MACS2. For PAIRIX or Validpairs formatted input files, however, peak file should be provided separately.

3. OutDir=*OutDir*

   - Base directory storing all the output results. Default is present working directory.

4. BINSIZE=*BINSIZE*

   - Size of a bin in bp (default = 5000, means 5 Kb bins).

5. LowDistThr=*LowDistThr*

   - Lower distance threshold (in terms of bp) for considering an interaction between two fixed size bins. Default 20000 (20 Kb).

6. UppDistThr=*UppDistThr*

   - Upper distance threshold of interaction between two bins, in terms of bp. Default 2000000 (2 Mb).

7. QVALUE=*QVALUE*

   - Minimum FDR (q-value) cutoff for significant interaction (default = 0.01).

8. NBins=*NBins*

   - Max no of bins (equal occupancy bins) employed in FitHiC. Default 200.

9. BeginBiasFilter=*BeginBiasFilter*

   - Value of 1 or 0 (default).
   - If specified 1, interactions whose both intervals have bias values within the interval *(biaslowthr, biashighthr)* (described below) are only retained for statistical significance estimation.

10. EndBiasFilter=*EndBiasFilter*

    - Value of 1 or 0 (default).
    - If specified 1, P value of any interaction depends on the bias values. Probability of any interaction is multiplied by the bias values of both ends, before computing the P values.

11. biaslowthr=*biaslowthr*

    - if ”BeginBiasFilter” parameter is set as 1, interactions whose one or both ends have bias value lower than this threshold, are discarded from subsequent significance estimation process.
    - Default value of this parameter is 0.2.

12. biashighthr=*biashighthr*

    - if ”BeginBiasFilter” parameter is set as 1, interactions whose one or both ends have bias value higher than this threshold, are discarded from subsequent significance estimation process.
    - Default value of this parameter is 5.

13. MergeInt=*MergeInt*

    - Value of 1 (default) or 0.
    - If specified 1, significant interactions (detected by FitHiC) originating and terminating in nearby bins are merged to represent one single interaction.
    - The merged interactions are written in separate text file.

14. PREFIX=*PREFIX*

    - Prefix string of output files (Default = ‘FitHiChIP’).

15. Draw=*Draw*

    - Specified as 1 or 0 (default). If 1, various analysis plots regarding the performance of FitHiChIP are generated.

16. TimeProf=*TimeProf*

    - Specified as 1 or 0 (default). If 1, a log file showing the time for individual execution steps is generated.

4.2 Command line options specific to processing BAM or PAIRIX alignment files

These options are applicable when input file is either in BAM or PAIRIX format.

1. IntType=*TypeOfInteraction*
   - Type of interactions to be derived.
   - Can have values 1 (Peak to Peak), 2 (Peak to Non Peak), 3 (ALL to ALL), or 0 (Peak to ALL - default). For details, please refer here.
2. Threads=*NumThreads*
   - Number of threads for multiprocessing. Default 1.
3. GSIZE=*GenomeSizeMACS2*
   - Genome size parameter for MACS2
   - Applicable when input alignment file is in BAM format, and no peak detection file is provided separately. In such a case, MACS2 is used to derive the peaks from the alignment file.
   - Default value = 'hs'

4.3 Command line options specific to HiC-pro pipeline output

These options are used when HiC-pro pipeline generated validpairs file is used as the input.

1. Interval=*IntervalFile*
   - If user has computed the distance matrix from the valid pairs file, he / she can provide the path of the interval file (containing fixed size bin intervals) via this option. The file name, by default, ends with the suffix '_abs.bed'.
   - Optional parameter. If not provided, computed by FitHiChIP.
2. Matrix=*DistanceMatrixFile*
   - If user has computed the distance matrix from the valid pairs file, he / she can provide the path of the distance matrix file (containing contact counts among individual pairs of bin intervals). The file name, by default, ends with the suffix '.matrix'.
   - Optional parameter. If not provided, computed by FitHiChIP.
3. ChrSizeFile=*RefChrSizeFile*
   - File containing the information of reference chromosome size

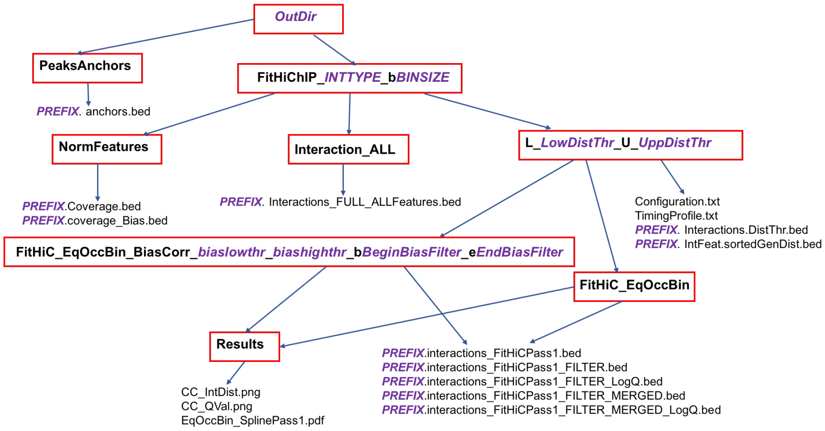
OutDir
PeaksAnchors
FitHiChIP_INTTYPE_bBINSIZE
PREFIX. anchors.bed
NormFeatures
Interaction_ALL
L_LowDistThr_U_UppDistThr
PREFIX.Coverage.bed
PREFIX.coverage_Bias.bed
PREFIX. Interactions_FULL_ALLFeatures.bed
Configuration.txt
TimingProfile.txt
PREFIX. Interactions.DistThr.bed
PREFIX. IntFeat.sortedGenDist.bed
FitHiC_EqOccBin_BiasCorr_biaslowthr_biashighthr_bBeginBiasFilter_eEndBiasFilter
FitHiC_EqOccBin
Results
PREFIX.interactions_FitHiCPass1.bed
PREFIX.interactions_FitHiCPass1_FILTER.bed
PREFIX.interactions_FitHiCPass1_FILTER_LogQ.bed
PREFIX.interactions_FitHiCPass1_FILTER_MERGED.bed
PREFIX.interactions_FitHiCPass1_FILTER_MERGED_LogQ.bed
CC_IntDist.png
CC_QVal.png
EqOccBin_SplinePass1.pdf

Figure 1: Directory structure produced by processing BAM or PAIRIX alignment files.

- Mandatory parameter.

4. HiCProBasedir=*HiCProBasedir*

   - Base directory of HiC-pro package (to be already installed in the system).
   - Required if the interval or matrix files are not provided as inputs and are to be computed from the input valid pairs file.

4.4 Support for PLAC-seq pipeline

PLAC-seq pipeline [2] maps the peaks from the short range segment (read length < 1 Kb) with the alignment consisting of the long range reads (read length $>$ 10 Kb). User may refer to this section for creating such alignment files, and obtaining peaks from the short reads. These alignment and peak files are then to be provided via the above mentioned configuration file.

5 Description of output files and directories

5.1 BAM or PAIRIX alignment file

Fig. 1 shows the directories and files generated by FitHiChIP pipeline when executed on BAM or PAIRIX input files. Entries within rectangles represent folders, while those not within rectangles represent files generated. Arrows represent subdirectories. Texts in violet color and italics represent input parameters (as mentioned in the configuration file) whose values will be substituted in actual folder or file names. The topmost directory is *OutDir* (mentioned in the configuration file).

Below we mention individual files and directories:

1. **PeaksAnchors**: Folder containing the peaks (anchors). The corresponding file (bed formatted) is *PREFIX*.anchors.bed

2. **FitHiChIP_*INTTYPE*_b*BINSIZE***: Depending on the interaction type, this folder contains all results of FitHiChIP. Value of *INTTYPE* is a string. Its values are 'ALL2ALL', 'Peak2ALL', 'Peak2NonPeak', 'Peak2Peak' (depending on the type of interaction specified as input).

3. **NormFeatures**: Folder storing the normalization related features. The files *PREFIX*.Coverage.bed and *PREFIX*.coverage_Bias.bed within this directory stores coverage and bias for individual bins. The second file has six columns:

   (a) chrName, chrStart, and chrEnd: three columns depicting the genomic segment (bin) considered.

   (b) coverage: The coverage (integer value) of a bin.

   (c) isPeak: a boolean variable representing whether the corresponding bin overlaps with a peak segment (provided as the input).

   (d) Bias: bias value of this bin, computed with respect to the coverage distribution of peaks and non-peaks.

4. **Interaction_ALL**: This folder has a file *PREFIX*.Interactions_FULL_ALLFeatures.bed which lists all the interactions (according to the specified interaction type) between every possible pairs of bins. The file has 7 columns:

   (a) chr1Name: name of the chromosome for first interacting segment.

   (b) chr1Start: starting coordinate of the first segment. Always a multiple of *BINSIZE*.

   (c) chr1End: end coordinate of the first segment. Also, a multiple of *BINSIZE*. Further, (chr1End - chr1Start) = *BINSIZE*.

   (d) chr2Name, chr2Start, chr2End: Fields for the 2nd chromosome.

   (e) ContactCount: Number of (nonzero) contacts (interactions) between this pair of bins.

   (f) Columns 8, 9, and 10 list the 'coverage', 'isPeak', and 'Bias' values for the first interacting segment.

   (g) Columns 11 to 13 indicate the same for the second interacting segment.

5. **L_*LowDistThr*_U_*UppDistThr***: FitHiChIP detected interactions and associated results, with respect to the distance thresholds specified as input.

   (a) The file **Configuration.txt** summarizes input parameters.

(b) File **TimingProfile.txt** logs the time elapsed during various stages of execution, provided the option 'TimeProf' in the configuration file is 1.

(c) ***PREFIX*.Interactions.DistThr.bed** stores the interactions and associated features with respect to the input distance thresholds.

(d) ***PREFIX*.IntFeat.sortedGeneDist.bed** contains interactions sorted with respect to the increasing genomic distance between the interacting segments (difference between the columns 2 and 5). For two entries (rows) with equal genomic distance, decreasing order of contact count (column 7) is employed for sorting.

6. **FitHiC_EqOccBin** (directory): Stores results of FitHiC, employed with equal occupancy binning.

   - ***PREFIX*.interactions_FitHiCPass1.bed**: Lists the interactions with four additional columns: 1) prior probability, 2) probability of binomial distribution for the observed contact count, 3) P-value, and 4) Q-value for individual interactions. The Q-value is computed from the P-values using the BH correction.
   - ***PREFIX*.interactions_FitHiCPass1_FILTER.bed**: Lists only those interactions having Q value less than the specified Q value threshold (default 0.01).
   - ***PREFIX*.interactions_FitHiCPass1_FILTER_LogQ.bed**: Significant interactions are listed for displaying in *WashU epigenome browser*.
   - **Bin_Info.log**: A log file describing individual bins in FitHiC. Each bin (individual rows) is characterized by the average genomic distance, average contact count, and the prior probabilities.
   - The folder **Results** contains following files:
     - **EqOccBin_SplinePass1.pdf**: Displays the spline fit between the interaction distance and the contact count.
     - **CC_Qval.png**: Box plot of contact count distributions for the significant (Q $< QVALUE$) and insignificant (Q $\geq QVALUE$) interactions.
     - **CC_IntDist.png**: Plot showing the interaction distance vs contact count for individual interactions belonging to either significant (Q $< QVALUE$) or insignificant (Q $\geq QVALUE$) categories.

7. **FitHiC_EqOccBin_BiasCorr_*biaslowthr_biashighthr* _b*BeginBiasFilter*_e*EndBiasFilter***: This folder is created only if one of the parameters *BeginBiasFilter* or *EndBiasFilter* is 1. Contains interactions from FitHiC with bias correction method enabled. Here the probability, P and Q values of individual interactions are computed by taking into account of the bias factor (see [1] for the definition of bias factor and its computation). The directory structure is similar to **FitHiC_EqOccBin**.

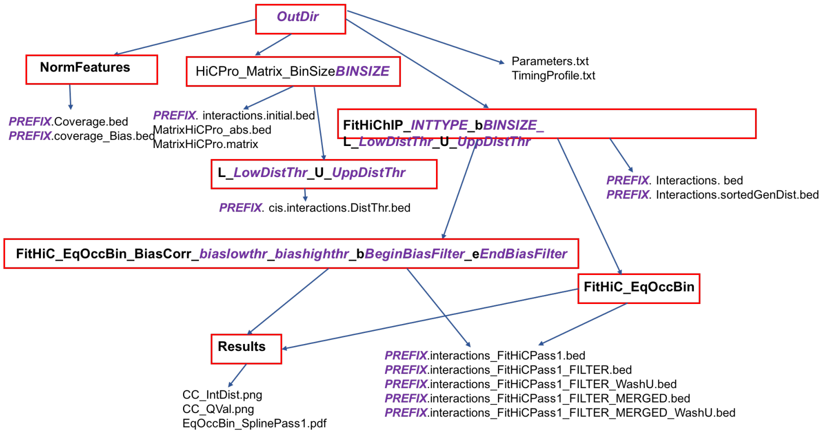
OutDir
NormFeatures
HiCPro_Matrix_BinSizeBINSIZE
Parameters.txt
TimingProfile.txt
PREFIX.Coverage.bed
PREFIX.coverage_Bias.bed
PREFIX. interactions.initial.bed
MatrixHiCPro_abs.bed
MatrixHiCPro.matrix
FitHiChIP_INTTYPE_bBINSIZE_
L_LowDistThr_U_UppDistThr
L_LowDistThr_U_UppDistThr
PREFIX. cis.interactions.DistThr.bed
PREFIX. Interactions. bed
PREFIX. Interactions.sortedGenDist.bed
FitHiC_EqOccBin_BiasCorr_biaslowthr_biashighthr_bBeginBiasFilter_eEndBiasFilter
FitHiC_EqOccBin
Results
CC_IntDist.png
CC_QVal.png
EqOccBin_SplinePass1.pdf
PREFIX.interactions_FitHiCPass1.bed
PREFIX.interactions_FitHiCPass1_FILTER.bed
PREFIX.interactions_FitHiCPass1_FILTER_WashU.bed
PREFIX.interactions_FitHiCPass1_FILTER_MERGED.bed
PREFIX.interactions_FitHiCPass1_FILTER_MERGED_WashU.bed

Figure 2: Directory structure produced by processing HiC-pro pipeline.

5.2 Processing HiC-Pro validpairs file

When HiC-pro pipeline generated validpairs file is provided in FitHiChIP, a few additional files and directories are created. Fig. 2 depicts such directory structure.

1. The folder **HiCPro_Matrix_BinSize*BINSIZE*** contains within it the distance matrices generated from the valid pairs file. These matrices are generated using the utility function 'build_matrix' of the HiC-pro package.

    - The file ***PREFIX*.interactions.initial.bed** converts the HiC-pro generated distance matrix into .bed format.
    - Another folder **L_*LowDistThr*_U_*UppDistThr*** stores the bed formatted interactions which satisfy the input distance thresholds.

2. These interactions are further used in all four different categories of interactions. The folder **FitHiChIP_*INTTYPE*_b*BINSIZE*_L_*LowDistThr*_U_*UppDistThr*** (where *INTTYPE* varies in all four categories of interactions) stores interactions and FitHiC results for individual categories.

3. Folders of individual categories of interactions follow the same structure as described in BAM or PAIRIX formatted input.

6 Creating paired end alignment files

FitHiChIP includes three utility scripts to produce paired end alignment files (in either BAM or PAIRIX formats) from a pair of .fastq reads, or from a given alignment (BAM) file. These scripts are placed within the folder **Preprocess**.

Table 1: List of "required / essential" parameters for Prep_PLAC.sh

Option	Value	Type	Details
-f	*fastq1*	String	Read 1 (fastq).
-f	*fastq2*	String	Read 2 (fastq).
-X	*align*	String	Input alignment (BAM format) if user has already computed the alignment.
-G	*RefGenome*	String	For Fastq reads, this is the reference genome with respect to BWA aligner.
-p	*PicardExec*	String	Path of the Picard tool executable. Used for removing duplicates of the given (or derived) alignment.
-d	*OutDir*	String	Directory storing the output results.
-n	*PREFIX*	String	Prefix string of output files (Default = empty string).

Table 2: List of "optional" parameters for Prep_PLAC.sh

Option	Value	Type	Details
-c	*REFile*	String	File containing the RE cut sites (bed formatted). User can use RE cut sites used in HiC-pro pipeline.
-t	*Threads*	Integer	Set number of threads. Default 1.
-q	*MAPQThr*	Integer	Quality threshold applied on the alignment file. Default 30.
-i	*InsertSize*	Integer	For PLAC seq experiment, minimum distance between the read ends to define a long segment. Default 10000 (corresponds to 10 Kb). This parameter should be greater than 1000 (corresponds to 1 Kb).
-m	*MaxMem*	String	Memory specification for applying the Picard tool jar file. Default "1G".
-g	*GSIZE*	String	If MACS2 is called for the peak calling, genome size parameter. Default 'hs'.

6.1 Utility 1 - Prep_PLAC.sh

As FitHiChIP supports PLAC-seq data [2], this script is included to produce alignment files used in the PLAC seq pipeline. Given a pair of fastq reads, in addition of creating a paired end alignment file (sorted and duplicate removed), the script produces two separate alignments of short (< 1 Kb) and long (default > 10 Kb) reads.

Changes in version 2.0: **This script is modified to create alignments in both BAM and PAIRIX formats.**

Details of the output files are described below. Command line options for this script are provided in the Tables 1 and 2.

Requirements:

- Two fastq files (consisting of single reads), or a pre-computed alignment file (in BAM format).
- If fastq files are provided, user needs to provide the reference genome file with respect to the BWA aligner (recommended in the PLAC seq pipeline).
- Path of picard tool executable (after installation in the system), to be provided as a command line option. This tool is used to remove PCR duplicates from the alignment.

Processing restriction enzyme cuts

The option '-c' (Table 2) is used to specify a file (bed formatted) containing restriction enzyme cuts. For such an input file, only the reads within a specific distance threshold from the nearest RE cuts are used for the subsequent analysis. User can follow the RE site file used and described in HiC-pro pipeline.

Output files and directories:

Within the specified output directory *OutDir*, following files and folders are created:

- **Alignment_MAPQ*MAPQThr*** (directory)
  1. Sorted, duplicate removed, and indexed alignment file ***PREFIX*.paired.cis.RE.filtered.sorted.nodup.bam**.
  2. Its PAIRIX version named ***PREFIX*.paired.cis.RE.bsorted.pairs.gz** and its index file ***PREFIX*.paired.cis.RE.bsorted.pairs.gz.px2**
- **Segments** (directory)
  1. ***PREFIX*.long.bam** and ***PREFIX*.short.bam**.
  2. These bam files are sorted and indexed.
  3. Denote alignments with long reads (default > 10 Kb) and short reads (< 1 Kb), respectively.
  4. Their PAIRIX conversions ***PREFIX*.long.bsorted.pairs.gz** and ***PREFIX*.short.bsorted.pairs.gz**, which are also indexed.
- **PeaksAnchors_ShortSegment** (directory)
  1. Peak segments detected (using MACS2) from the short read alignment *PREFIX*.short.bam).
  2. Applied on FitHiChIP pipeline.
- **PeaksAnchors_ALL** (directory)
  1. Peaks detected from the complete alignment *PREFIX*.paired.cis.RE.filtered.sorted.nodup.bam
  2. Can be used as reference for comparison.

6.2 Utility 2 - Preprocess_HiChIP_fastq.sh

We have included a separate script for creating alignments of HiChIP data. Similar to the script for PLAC-seq, this script uses two fastq reads and creates alignments of short and long reads (in both BAM and PAIRIX formats). In addition, it generates peaks (using MACS2) of the alignments of short reads. Main objective of this new script is to apply strand specific filtering of HiChIP data, to filter out the dangling reads of short distance.

6.3 Utility 3 - Preprocess_HiChIP_BAM.sh

Another script for processing BAM formatted HiChIP input alignments. Its command line options are similar to the Fastq version.

7 Contact

For any queries, please e-mail:

- Sourya Bhattacharyya (sourya@lji.org)
- Ferhat Ay (ferhatay@lji.org)

References

[1] Ferhat Ay, Timothy L. Bailey, and William S. Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*, 24(6):999–1011, 2014.

[2] R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, and B. Ren. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.*, 26(12):1345–1348, Dec 2016.

[3] M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, 13(11):919–922, Nov 2016.

[4] N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker, and E. Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, 16:259, Dec 2015.

[5] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.