Manual of FitHiChIP (Version 1.0)

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

September 10, 2017

FitHiChIP computes statistically significant interactions between different pairs of genomic intervals of a given HiChIP [4] or PLAC-seq [3] data. The code applies the FitHiC method [1] on a given HiChIP or PLAC-seq data. FitHiChIP offers a complete pipeline for finding the statistical significant cis interactions from a given HiChIP or PLAC-seq alignment or fastq files. Current version of FitHiChiP supports processing only 'cis' interactions. The code supports multiprocessing, and can be executed either in a laptop or in a computational cluster. The code is executed and tested in Linux environment.

1 Installation

FitHiChIP requires a Linux environment (with bash) with Python and R installed. FitHiChIP is developed using Python version 2.7.13, R version 3.3.3, and perl (included in any standard linux distribution). Following packages are also required (values within the 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)
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)

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

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

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

2 Pipeline of FitHiChIP

- Fastq files are serially processed for alignment of single end reads.
- Aligned reads are merged to form the paired end reads.
- Reads are applied quality based thresholding, and duplicates are removed from the given paired end alignment.
- For processing PLAC seq data, the alignment file is further processed to separate the short reads (< 1 Kb) and long reads (> 10 Kb) in two distinct alignment files.
- If interactions involving peak segments are sought, and no peak information is provided, the pipeline computes the peaks from given .bam formatted alignment file. Peaks are detected using the package MACS2.
- According to the bin size provided by the user, individual genome segments are divided in fixed size genomic intervals (or bins).
- Depending on the type of interactions to be computed, contacts between different pairs of bins are computed.
- These pairs of bins and the associated non-zero contact counts are applied on FitHiC module to derive the statistical significant contacts and the associated bin pairs.

3 Input files required

FitHiChIP mainly requires following two input files:

3.1 Paired end alignment file (BAM or PAIRIX format)

FitHiChIP requires a paired end alignment file in either BAM or PAIRIX formats (see creation of alignment files from a given pair of fastq reads).

3.2 Peak detection file (BED format)

To find the interactions associated with individual peak segments, a file containing the peak regions is required. User needs to provide pre-computed peak detection output file (bed formatted).

- If the alignment file is BAM formatted, user may skip pre-computing the peaks. In such a case, FitHiChIP computes the peaks from the given BAM file, using the package MACS2.
  - Command for the MACS2 based peak calling: macs2 callpeak -f AUTO -g *GSIZE* –keep-dup all –outdir *macs2dir* -n *PREFIX* –nomodel –extsize 147 –call-summits -t *INPFILE*
  - Details of this command is provided in the MACS2 webpage: https://github.com/taoliu/MACS
- **For PAIRIX formatted alignment, however, user must separately provide the peak detection file.**
- For PLAC-seq experiments, first user needs to create alignment files of short and long reads. Alignment consisting of the short reads are used to compute the peaks, which are used to find interactions involving the short read peaks.

4 Types of interactions supported

FitHiChIP can derive four types of interactions between individual pairs of genomic segments (represented as fixed size binned intervals):

- **ALL to ALL**: Interactions between every pairs of bins are computed, subject to the distance thresholds between these segments. Here, user does not need to provide any peak detection file.
- **Peak to Peak**: Possible when user provides a separate peak detection output file (using the -P option). Otherwise, for a BAM alignment file, FitHiChIP computes the peaks using MACS2. Here, fixed size bins which overlap with the input (or derived) peak regions are marked as the *peak*

segments. Otherwise, they are denoted as the *non-peak segments*. Interactions between any pairs of such fixed size peak segments are computed, provided that distance between these segments are within the specified distance thresholds.

- **Peak to Non Peak**: Interaction between a peak segment (fixed size bin overlapping with a peak region) and a non-peak segment (fixed size bin which does not overlap with any peaks) is computed, subject to the distance thresholds.
- **Peak to ALL**: Encapsulates both peak to peak and peak to non peak interactions. One side of such an interaction is a fixed sized peak segment, while the other end (fixed size bin) can either be a peak or a non-peak segment.

5 Command line options

The shell script **FitHiChIP.sh** is the main executable. User should invoke this script with the following command line parameters. Some of these parameters are marked as 'required', while others are 'optional'. Table 1 shows the list of required (or essential) parameters, while Table 2 lists the optional parameters (along with their default values) for this package.

5.1 Support for PLAC-seq pipeline

PLAC-seq pipeline [3] 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.

Suppose, *ShortPeakFile* denotes the peak file generated from the short reads. Further, let *LongAlign* be the alignment consisting of the long range reads. FitHiChIP command (recommended) for obtaining the statistical significant interactions from *ShortPeakFile* to *LongAlign* is as follows:

FitHiChIP.sh -I *LongAlign* -P *ShortPeakFile* -M 0 -o *OutDir* -n *PREFIX* -b *BinSize* -t *Threads* -L *LowDistThres* -U *UppDistThres*

Where, -M 0 indicates that interactions from the peaks to all segments (peaks or non-peaks) of the file *LongAlign* is computed.

5.2 Notes on the optional command line parameters

- Parallel processing is supported by specifying the number of threads (using the option -t) a value greater than 1. We have used -t 8 during the implementation.

Table 1: List of "required" parameters

Option	**Value**	**Type**	**Details**
-I	*InpFile*	String	Input alignment file, in either BAM or PAIRIX formats.
-M	*Method*	Integer (0 to 3)	**3 (ALL)**: Default value. Computes statistical significant interactions among all binned segments. Bin size is determined by the option -b (mentioned in the list of optional parameters). In this case, peak detection file is not required. **2 (Peak to Non Peak)**: Statistical significant interactions from the peaks to the non peak segments. A peak file is to be provided using the option -P. Otherwise, for BAM formatted input alignment, peaks are computed using MACS2. **1 (Peak to Peak)**: Statistical significant interactions between the fixed size binned peak segments. A peak detection file (or output from MACS2) is to be provided by the option -P. **0 (Peak to ALL)**: Statistical significant interactions between a peak and any fixed size binned segment (peak or non-peak) are derived. A peak detection file (or output from MACS2) is to be provided by the option -P.
-P	*PeakFile*	String	BED formatted file with the detected peak regions (computed by MACS2 or similar packages). Should be provided for the methods (option -M) 0 to 2. If not provided, and the input alignment file is in BAM format, peaks are computed using MACS2.
-o	*OutDir*	String	Base directory storing all the output results.
-n	*PREFIX*	String	Prefix string of output files (Default = 'FitHiChIP').
-b	*BinSize*	Integer	Size of a bin in bp (default = 5000, means 5 Kb bins). Partitioning of chromosomes into fixed size intervals occur according to this parameter. Interaction between pairs of fixed size bins are computed.

Table 2: List of "optional" parameters

Option	**Value**	**Type**	**Details**
-t	*Threads*	Integer	Set number of threads for peak calling and finding the interactions. Default 1.
-L	*LowDistThres*	Integer	Lower distance threshold of interaction between two segments in terms of bp. Default 20000 (20 Kb).
-U	*UppDistThres*	Integer	Upper distance threshold of interaction between two segments in terms of bp. Default 2000000 (2 Mb).
-f	*FitHiCBinMethod*	Integer	1 (default) or 0. If 1, equal occupancy (contact count) bins are employed for FitHiC. Else equal length bins are used. Recommended value is 1.
-N	*NBins*	Integer	Max no of bins (equal occupancy or equal length) employed in FitHiC. Default 200.
-B	*BinomDistr*	Integer	0 (default) or 1. If 1, binomial distribution model between the observed genomic distance and the contact count is also computed [2] and compared with the FitHiC output.
-q	*QVALUE*	Float	Minimum FDR (q-value) cutoff for interaction detection [default = 0.01].
-v	*verbose*	Integer	Specified as 1 or 0 (default); if 1, a log file showing the time for individual execution steps is generated.
-D	*DrawFig*	Integer	Specified as 1 or 0 (default); if 1, various analysis plots regarding the performance of FitHiChIP are generated.
-g	*GSIZE*	String	If MACS2 is used for peak detection, its genome size parameter. Default 'hs'.

- The -L and -U options specify the distance thresholds within which the interactions between any pair of bins are considered. Default values of these thresholds are 20 Kb and 2 Mb, respectively. Most of our experiments, however, used L = 10 Kb and U = 3 Mb.
- The option -f is by default 1 (equal occupancy binning). It is the recommended value since it produces superior performance compared to equal length binning (f = 0).
- The option -N specifies the number of bins considered for FitHiC (default = 200). We recommend users to keep this setting.
- The option -B, if set as 1, additionally models the genomic distance vs contact count as a binomial distribution [2]. If set, outputs of both binomial model and FitHiC will be generated. This option is thus useful for comparative analysis with FitHiC. By default, the option is 0.

5.3 Examples of a few commands

Following commands assume that user has a paired end, sorted, and indexed alignment file (in either BAM or PAIRIX formats). User may refer to this section for creating a paired end alignment file from two separate fastq reads.

Example 1:

FitHiChIP.sh -I *inp.bam* -M 3 -o */home/sourya/FitHiChIP/* -n *'tempFitHiChIP'* -b 5000 -t 8

- BAM formatted Input alignment file.
- Bin size = 5 Kb
- Number of threads = 8
- Interactions among all possible fixed size bins are sought.

Example 2:

FitHiChIP.sh -I *inp.pairix* -P *sample.peak.bed* -M 2 -o */home/sourya/FitHiChIP/* -n *'tempFitHiChIP'* -b 5000 -t 8 -B 1

- PAIRIX formatted alignment file.
- Peak detection file is also provided.
- Objective: find the interactions between peaks and non-peak segments.
- Both binomial and FitHiC based spline distribution models are checked (modeling of the contact count and interaction distance).

Example 3:

FitHiChIP.sh -I *inp.bam* -M 1 -o */home/sourya/FitHiChIP/* -n *'tempFitHiChIP'* -b 5000 -t 8 -L 10000 -U 3000000

- Interactions between peak segments.
- Peaks are computed by FitHiChIP using MACS2.
- Distance thresholds for valid interactions are set as 10 Kb and 3 Mb, respectively.

5.4 Examples of a few erroneous commands

Example 1:

FitHiChIP.sh -I *inp.pairix* -M 2 -o */home/sourya/FitHiChIP/* -n *'tempFitHiChIP'* -b 5000 -t 8 -B 1

- PAIRIX formatted alignment file.
- Interactions between peaks and non-peak segments are sought.
- No peak detection file is provided - **error**.

6 Description of output files and directories

6.1 Example 1: ALL to ALL interactions

A general command:

FitHiChIP.sh -I *AlignFile* -M 3 -o *OutDir* -n *PrefixStr* -b *BinSize* -t *NThread* -v 1 -L *LowDist* -U *HighDist* -f 1 -B 1 -D 1

- Variables are used to denote the parameters.
  - *AlignFile*: Given alignment file.
  - *BinSize*: Bin size employed.
  - *LowDist* and *HighDist*: distance thresholds.
  - *OutDir*: Stores all output results.
  - Individual output file names start with the string *PrefixStr*.
- '-v 1' indicates that a time log will be generated.
- '-f 1' means using equal occupancy bins in FitHiC.
- '-B 1' indicates that a separate binomial distribution based model of the interaction distance vs contact count is to be computed.

- '-D 1' enables the plotting of various figures for subsequent analysis.

Example of a specific command:
An instance of the above mentioned generalized command is as follows:

FitHiChIP.sh -I *inp.bam* -M 3 -o */home/sourya/FitHiChIP/* -n *'tempFitHiChIP'* -b 5000 -t 8 -v 1 -L 10000 -U 3000000 -f 1 -B 1 -D 1

- *inp.bam*: given alignment file.
- Bin size = 5 Kb.
- Base output directory */home/sourya/FitHiChIP/*.
- *'tempFitHiChIP'*: prefix string of individual output files.
- Lower and upper distance threshold values are 10 Kb and 3 Mb, respectively.

6.1.1 Description of the output files with respect to the generalized command:

Considering the previously mentioned (generalized) command, the following folder will be created under the directory *OutDir*:

FitHiChIP_ALL2ALL_b*BinSize*_L*LowDist*_U*HighDist*

Under this folder, following files and directories will be created:

A: Configuration.txt: Text file summarizing the parameters employed for this execution.

B: TimingProfile.txt: If the 'verbose' option (-v) is set as 1, this text file depicts the time elapsed during various stages of execution.

C: *PrefixStr*.anchors.interactions.initial.bed: Initial list of interactions among individual pairs of genomic segments (of size = *BinSize*). The file has 7 columns:

1. chr1Name: name of the 1st chromosome.
2. chr1Start: start of the first segment. This coordinate is always a multiple of *BinSize*.
3. chr1End: end of the first segment. Also, a multiple of *BinSize*. Further, (chr1End - chr1Start) = *BinSize*.
4. chr2Name: name of the 2nd chromosome.
5. chr2Start: start of the second segment. This coordinate is always a multiple of *BinSize*.

6. chr2End: end of the second segment. Also, a multiple of *BinSize*. Further, (chr2End - chr2Start) = *BinSize*.
7. ContactCount: Number of (nonzero) contacts (interactions) between this pair of genomic segments (bins).

D: ***PrefixStr*.anchors.interactions.bed**: File (C) may contain repeated entries since all pairs of genomic regions are scanned for finding the interactions. This file removes the repeated entries (same pairs of genomic intervals). Interactions reported in this file are used for the subsequent analysis.

E: ***PrefixStr*.Coverage.bed**: Text file with the coverage of individual genomic segments (having a size of *BinSize*) with respect to the given alignment file. This file has five columns.

1. chrName, chrStart, and chrEnd: three columns depicting the genomic segment (bin) considered.
2. coverage: The coverage (integer value) of this segment with respect to the input alignment.
3. isPeak: a boolean variable representing whether the current bin overlaps with a peak segment. For method (-M) 0, 1, and 2, peak segments are provided (or computed using MACS2). Otherwise, for M = 3, no peak information is available. Here, this field is 0 for all the binned segments.

F: ***PrefixStr*.Complete_InteractionFeatures.bed**: This file has 11 columns, and has a header row indicating the names of individual columns. First 7 columns of this file are identical to the file (D).

- Columns 8 and 9 list the 'coverage' and 'isPeak' values (obtained from the file (E)) for the first interacting segment.
- Columns 10 and 11 list the corresponding information for the second interacting segment.

G: ***PrefixStr*.Compl_IntFeat.sortedGenDist.bed**: Sorts the file (F) with respect to 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. This file is used for modeling binomial and spline distribution between the contact count and the distance between interacting segments.

H: BinomDistr (directory): If '-B 1' option is provided, this directory stores the binomial distribution model (and significant interactions) computed using the contact count and the distance between interacting segments, according to the method in [2]. Following files exist within this folder:

- ***PrefixStr*.interactions_BinomDistr.bed**: Has the contents of file (G) with four additional columns, representing the prior probability, probability of binomial distribution for the observed contact count, P-value and Q-value for individual interactions. The Q-value is computed from the P-values using the BH correction.
- ***PrefixStr*.interactions_BinomDistr_FILTER.bed**: Contains the **significant interactions** whose Q-value (last column) is lower than a specific threshold (t). Default value of t is 0.01. Value of t can be altered by the option -q (see Table 2).

I: FitHiC_EqOccBin (directory): For equal occupancy binning in FitHiC (indicated by the parameter -f 1), stores the spline model for the observed contact count and the interaction distance. Files within this directory are:

- ***PrefixStr*.interactions_FitHiCPass1.bed**: Same structure like the file *PrefixStr*.interactions_BinomDistr.bed. The only difference is that, FitHiC (Ay et. al. 2014) is used to compute the probability, P and Q values for individual interactions.
- ***PrefixStr*.interactions_FitHiCPass1_FILTER.bed**: Significant interactions according to the Q value - default threshold of 0.01.
- ***PrefixStr*.interactions_FitHiCPass1_FILTER_LogQ.bed**: For all significant interactions, stores the interacting segments and the logarithm (base 10) of Q value. Useful for plotting the significant interactions 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.
- If the option -D 1 is specified, a directory named 'Results' also exists within this folder. Following files are placed within this directory:
  - **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 ($Q < 0.01$) and insignificant ($Q \geq 0.01$) interactions.
  - **CC_IntDist.png**: Plot showing the interaction distance vs contact count for individual interactions belonging to either significant ($Q < 0.01$) or insignificant ($Q \geq 0.01$) categories.
  - **Interaction_BinomDistr_SplineEqOcc.pdf**: If the option -B 1 is provided, this plot displays the common and unique interactions for both binomial and spline fitted model.

J: FitHiC_EqOccBin_BiasCorr (directory): Same directory structure as (K). The only difference is that, here the probability, P and Q values of individual interactions are computed by taking into account of the bias factor (see Ay et. al. 2014 for the definition of bias factor and its computation).

K: FitHiC_EqLenBin (directory): Similar structure to the folder FitHiC_EqOccBin. The directory is generated when equal length binning (instead of equal occupancy binning) is employed in FitHiC.

K: FitHiC_EqLenBin_BiasCorr (directory): Folder storing the results of FitHiC when equal length binning and bias correction are employed together.

6.2 Example 2: Peak to Peak interactions

Generic command:

FitHiChIP.sh -I *AlignFile* -P *PeakFile* -M 1 -o *OutDir* -n *PrefixStr* -b *BinSize* -t *NThread* -v 1 -L *LowDist* -U *HighDist* -f 1 -B 1 -D 1

- *PeakFile*: bed formatted file, storing the peaks of the given alignment.
- Within the directory *OutDir*, following folder is created: FitHiChIP_Peak2Peak_b*BinSize*_L*LowDist*_U*HighDist*
- Files and folders under this directory are similar to those mentioned for the ALL to ALL interactions. There are a few differences, however:
  - For the file (E), the column 'isPeak' has an entry 1 if the corresponding segment overlaps with a given peak region.
  - The columns 9 and 11 of the file (F) are all 1, since interactions between peak segments are considered.
  - ***PrefixStr*.interactions_FitHiCPass1_FILTER.Peakcount.bed**: Two new files of this name are created under the directories (K) and (L). This file contains four columns. First three columns denote individual peak segments (fixed size bins overlapping with the given peak regions). The fourth column denotes the number of interactions each peak is associated with.

6.3 Example 3: Peak to Non Peak interactions

Computed for the option -M 2. Requires peak detection file as input. The base output directory is:

FitHiChIP_Peak2NonPeak_b*BinSize*_L*LowDist*_U*HighDist*

For individual interactions listed in the output files, first segment is a fixed size bin overlapping with a peak region (the entry 'isPeak' is 1 in the file E). Second segment is also a fixed size bin which does not overlap with a peak region (the entry 'isPeak' is 0 in the file E). Folder and file structures for this option follow earlier descriptions.

Table 3: 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).

6.4 Example 3: Peak to ALL interactions

Computed for the option -M 0. Requires peak detection file as the input. The base output directory is:

FitHiChIP_Peak2ALL_b*BinSize*_L*LowDist*_U*HighDist*

The interactions list peak regions (fixed size bins overlapping with a peak segment) as the first segment (chr1) and either peak or non peak regions at the right side (chr2). Rest of the descriptions are similar as before.

7 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'.

7.1 Utility 1 - Prep_PLAC.sh

As FitHiChIP supports PLAC-seq data [3], 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, the script produces two separate alignments of short ($<$ 1 Kb) and long (default $>$ 10 Kb) reads. Command line options for this script are provided in the Tables 3 and 4.

- User can provide either 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).
- User also needs to install the picard tool.

Table 4: List of "optional" parameters for Prep_PLAC.sh

Option	**Value**	**Type**	**Details**
-c	*REFile*	String	File containing the RE cut sites (bed formatted).
-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'.

Table 5: List of parameters for Align.sh

Option	**Value**	**Type**	**Details**
-f	*fastq1*	String	Read 1 (fastq).
-f	*fastq2*	String	Read 2 (fastq).
-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).
-t	*Threads*	Integer	Set number of threads. Default 1.
-q	*MAPQThr*	Integer	Quality threshold applied on the alignment file. Default 30.
-m	*MaxMem*	String	Memory specification for applying the Picard tool jar file. Default "1G".
-a	*Multimap*	Integer	No of multiple maps supported in Bowtie2 (default = 4).

– Needs to provide the path of picard executable as a command line option.

– Used to remove the duplicates from the alignment.

Processing restriction enzyme cuts

The option '-c' (Table 4) 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.

Output files and directories:

A directory named **Alignment_MAPQ*MAPQThr*** under the directory *OutDir* contains a sorted and duplicate removed alignment file '*.sorted.rmdup.bam'. This bam file is also indexed.

In addition, a separate directory **Segments** under the directory *OutDir* contains two indexed bam files: *.long.bam and *.short.bam. The former denotes the alignment with long reads (default $>$ 10 Kb) while the later denotes the alignment with short reads ($<$ 1 Kb).

A directory **PeaksAnchors** contains the peak segments detected from the short read alignment (the file *.short.bam). Peaks are detected using MACS2.

Generated alignment files are applied on the PLAC seq pipeline, according to the command specified here.

7.2 Utility 2 - Prep.sh

This utility has the same command line parameters as Prep_PLAC.sh, except the absence of GSIZE argument. The function creates a sorted and duplicate removed alignment file and places in the directory **Alignment_MAPQ*MAPQThr***. This program, however, does not generate any alignments of short or long range reads. It does not perform any peak calling either.

7.3 Utility 3 - Align.sh

This utility is useful to create a paired end alignment file in both BAM and PAIRIX formats. **If user requires alignment file of PAIRIX format, this script is to be used.**

Here, input fastq reads are applied on Bowtie2 to produce a paired end alignment. User needs to provide the reference genome for Bowtie2 alignment. User may also change the option *Multimap* to specify the number of multiple mappings allowed. Rest of the options are similar to the previous descriptions.

8 Contact

For any queries, please e-mail:

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

References

[1] F. Ay, T. L. Bailey, and W. S. Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome Research*, 24:999–1011, 2014.

[2] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, May 2010.

[3] 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, 2016.

[4] 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, 2016.