Post-Sequencing Quality Control Process Of Dovetail™ HiChIP Libraries

Introduction

A key component of working with any NGS-based assay is the processing and quality control of the data that come off the sequencer. The Dovetail™ HiChIP *MNase* assay combines the benefits of ChIP-seq with Dovetail™ Micro-C, an MNase-based proximity ligation method. Therefore, the main QC goals for the HiChIP libraries are (I) to classify and assess the distance information captured by ligation events from high-quality read pairs and (II) to assess the extent of the chromatin immunoprecipitation (ChIP) enrichment. To make the data processing and QC of the libraries easier, Dovetail Genomics has designed a workflow that incorporates 4D Nucleome best practices to help you accurately assess the quality of libraries generated with Dovetail™ HiChIP *MNase* kit. A detailed breakdown of the tools can be found here: https://hichip.readthedocs.io/en/latest/

We recommend shallow sequencing of your library to 20 million read pairs to get an initial assessment of library quality. This document walks you through the consecutive post-sequencing QC process while clarifying what the different QC metrics indicate.

How Is A Valid Proximity-Ligation Read Pair Defined?

Before we can discuss the QC process, we must first define a valid read pair as not all read pairs produced in a proximity ligation library are of equal interest. Read pairs result from one of three ligation events:

1. Self-ligation
2. Re-ligation

} Invalid read pair

3. Valid ligation — Valid read pair

The first two ligation events are of low interest while the third ligation event - the desired class - yields a valid read pair. Figure 1 provides a detailed schematic defining each class and how it is generated for both restriction enzyme (RE) and DNase/MNase-based approaches. The percentage of read pairs that fall into the valid ligation class is, therefore, an important QC

metric. It should be noted that self-ligation products are not a concern when working with Dovetail™ HiChIP *MNase* Kit as the workflow does not require sonication, and thus, these products cannot physically be converted into sequenceable molecules.

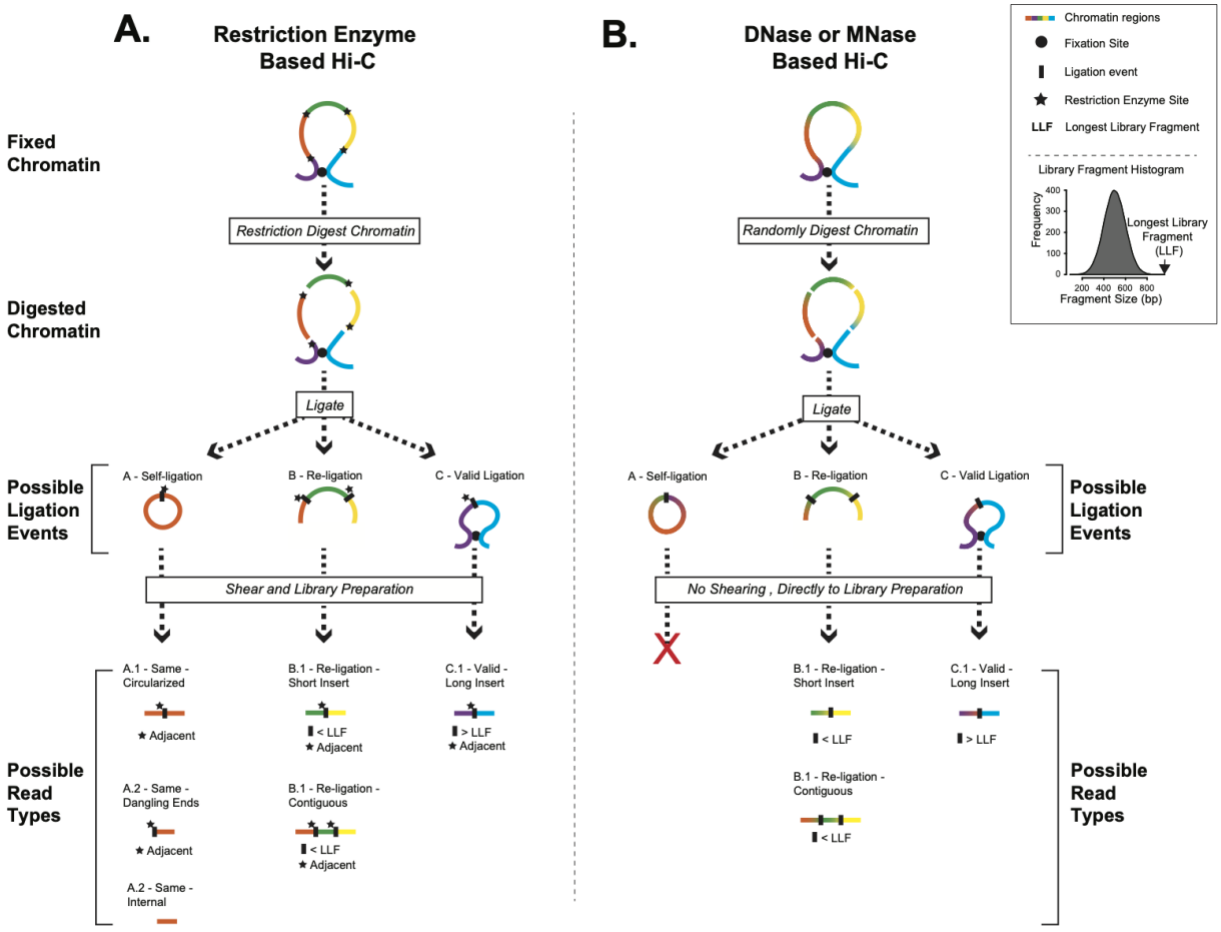


Figure 1. Classification schematic of reads generated from restriction enzyme (RE)-based and RE-free proximity ligation assays. Possible ligation events and resulting read types are depicted. The RE-based (A) and DNase/MNase (B) proximity ligation workflows are shown in parallel for direct comparison. Chromatin regions are denoted by different colors, the change in color either abrupt at a RE site (star) or blended to represent a sequence independent view of chromatin digestion. Ligation events are shown as a vertical black bar. The longest ligation fragment (LLF) is defined as the upper limit of the library size distribution as shown in the library fragment size histogram depicted in the inset. *Trans* read pairs, where each read from a pair maps to two different chromosomes, are considered valid read pairs but not pictured.

Post-Sequencing QC Analysis Workflow Overview

After sequencing the library to 20 M read pairs, the QC analysis workflow is completed in two parts: Part I is Proximity Ligation Assessment and Part II is ChIP Enrichment Assessment. The workflow is outlined in detail in the readthedocs pages for HiChIP *MNase* Kit. In this document, we will discuss the QC metrics according to the step in which they are computed. To clarify each group of metrics, graphical representations of the data are occasionally included, however, these graphs are not part of the QC analysis output file.

Part I of The QC Analysis: Proximity Ligation Assessment

Part I of the QC analysis consists of the following 2 steps:

- Step 1. Aligning raw reads and filtering for unmapped, low mapping quality and PCR duplicate read pairs.
- Step 2. Classifying filtered read pairs as *cis* or *trans* and characterizing insert distance to identify valid read pairs.

Step 1. Aligning raw reads and filtering for unmapped, low mapping quality and PCR duplicate read pairs.

After sequencing, the read pairs are aligned using **BWA MEM** to the appropriate reference genome. The alignment step results in:

a) Unmapped read pairs
b) Mapped read pairs with a mapping quality (MAPQ) value < 40
c) Mapped read pairs with a mapping quality (MAPQ) value ≥ 40

Unmapped and low MAPQ read pairs are removed from the subsequent steps. (Note: Low MAPQ read pairs are not reported in the QC table output by the script.)

Mapped read pairs with MAPQ ≥ 40 are processed by **pairtools** to flag and remove PCR duplicates. Only non-duplicate mapped read pairs with MAPQ ≥ 40 (referred to as No-Dup Read Pairs) progress into step 2.

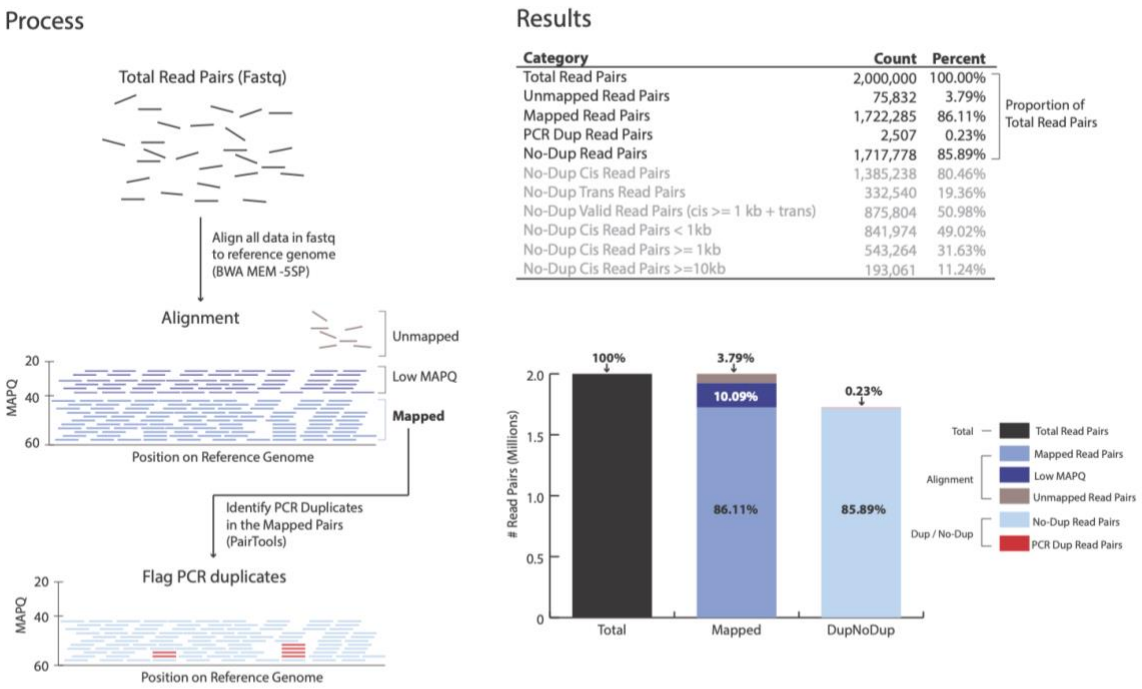


Figure 2. Aligning raw reads and filtering for unmapped, low mapping quality and PCR duplicate read pairs.
Process – Total reads are aligned to a reference genome. The reads are then characterized as unmapped, low MAPQ (< 40), or mapped read pairs (≥ 40). PCR duplicates are then flagged and filtered from the mapped read pairs using **pairtools**.
Results – The results of this step are captured in the first 5 rows of the QC table.
Graphical Representation – The three bars represent each step in the alignment and filtering process with number of read pairs on the y-axis. Total Read Pairs represents the denominator used to calculate percentages.

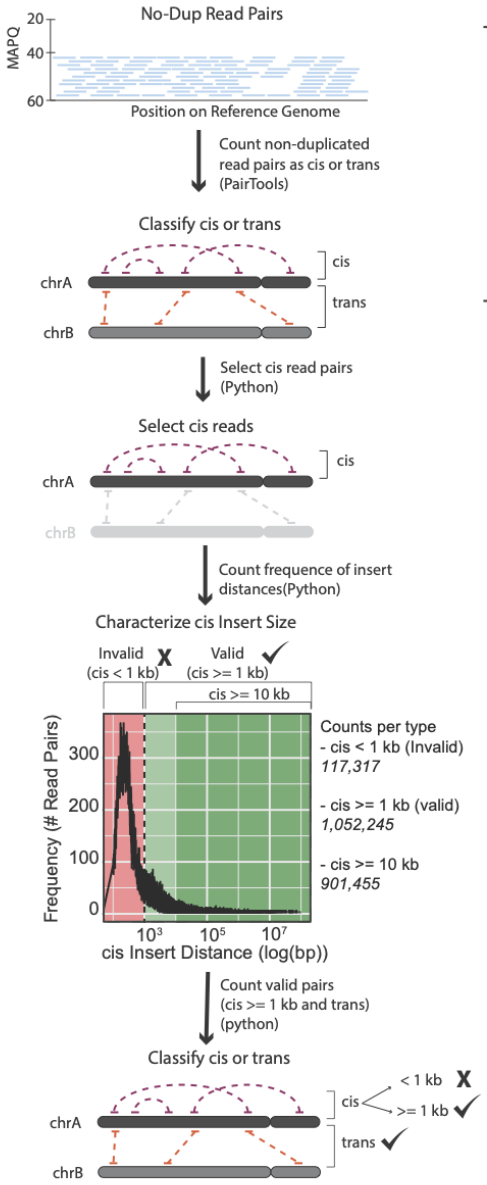
Step 2. Classifying filtered read pairs as *cis* or *trans* and characterizing insert distance to identify valid read pairs.

The non-duplicate mapped read pairs with MAPQ ≥ 40 (No-Dup Read Pairs) from step 1 are categorized by **pairtools** as valid if they meet one of the following criteria:

a) the pair maps to different chromosomes (*trans*).
b) the pair maps to the same chromosome (*cis*) and the distance between the interacting points is > 1 kb.

In addition to looking at the percentage of valid read pairs as a QC metric, another consideration is how these valid read pairs are partitioned across the two valid categories of *trans* and *cis* > 1 kb.

Process



Results

Category	Count	Percent	
Total Read Pairs	2,000,000	100.00%	
Unmapped Read Pairs	75,832	3.79%	
Mapped Read Pairs	1,722,285	86.11%	
PCR Dup Read Pairs	2,507	0.23%	
No-Dup Read Pairs	1,717,778	85.89%	
No-Dup Cis Read Pairs	1,385,238	80.46%	Proportion of No-Dup Read Pairs
No-Dup Trans Read Pairs	332,540	19.36%	
No-Dup Valid Read Pairs (cis >= 1 kb + trans)	875,804	50.98%	
No-Dup Cis Read Pairs < 1kb	841,974	49.02%	
No-Dup Cis Read Pairs >= 1kb	543,264	31.63%	
No-Dup Cis Read Pairs >=10kb	193,061	11.24%	

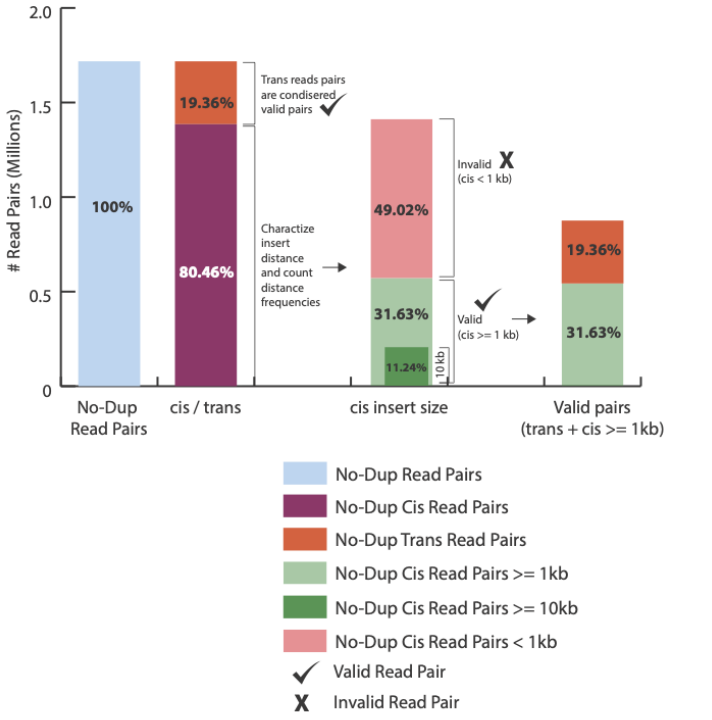


Figure 3. Classifying filtered read pairs as *cis* or *trans* and characterizing insert distance to identify valid read pairs.
Process – No-Dup read pairs are classified as *cis* or *trans* using **pairtools**. All *trans* read pairs are considered valid. By contrast, valid *cis* reads must have an insert size greater than the Longest Library Fragment (LLF). 1 kbp is used as the LLF cut-off for the HiChIP libraries. Since the libraries are size selected, physical insert sizes range from 350 bp to 1 kbp. Therefore, mapped insert size < 1 kbp represent re-ligation events (and are invalid).
Results – The results of this step are captured in rows 6 – 11 of the QC table.
Graphical Representation – On the left is a plot of *cis* read pair insert size (frequency); color changes mark the 1 kb, 10 kb and >10kb insert size bins. The bar chart on the right plots the reads classified in the QC table. The No-Dup read pairs count is the denominator for the percentages calculated.

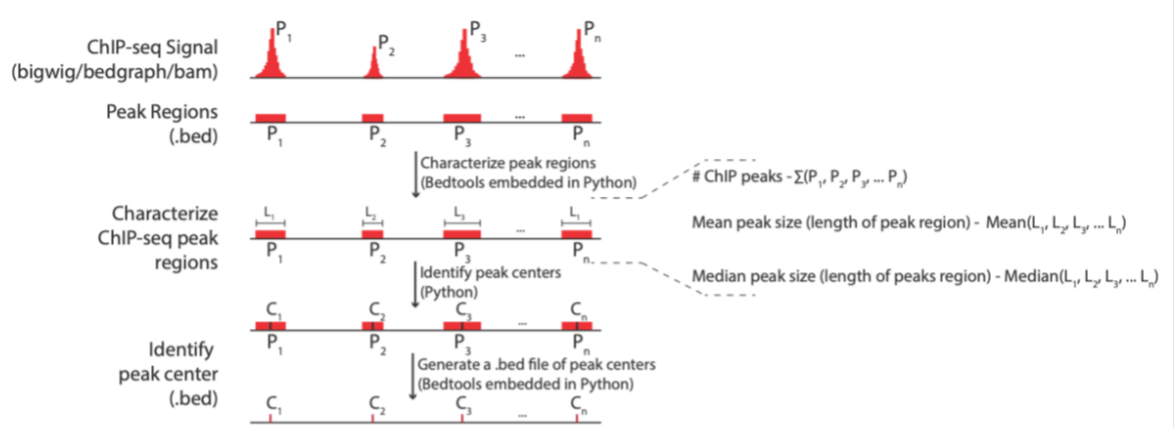
Part II of The QC Analysis: ChIP Enrichment Assessment

The second half of assessing a HiChIP library is determining the extent of the chromatin immunoprecipitation (ChIP) enrichment. It is worth noting that HiChIP data is composed of primary peaks, reflecting direct protein binding, and secondary peaks, resulting from interactions occurring in 3D chromatin space. For the purpose of assessing our ChIP success, we will be focusing on primary peaks in our HiChIP data. Therefore, to fully complete the analysis, in addition to our HiChIP data, we require a ***.bed** or a ***.narrowpeaks** file generated by MACS2 from a previous ChIP-seq experiment or derived from the ENCODE database.

To assess ChIP enrichment, we characterize the user provided ChIP-seq peaks, calculate observed/expected score evaluating actual HiChIP read coverage at primary peaks versus expected uniform coverage over the fraction of the genome containing peaks, finally we compute and plot an average per base coverage profile of the HiChIP data over the primary ChIP-seq peaks. The process consists of three steps and uses the ***.bam** file generated from the previous steps and a ***.bed** file containing the location of primary peaks as input files. We will now break this process down into its three steps.

Step 1. Counting and characterizing user-provided ChIP-seq peaks

To get some basic information about the primary peaks, we start by using **bedtools** (embedded in a short Python script) to characterize the primary peaks ***.bed** file reporting the total number of peaks, the mean, and the median ChIP peak size. Additionally, the centers of each peak are identified and captured in a temporary ***.bed** defining peak centers.

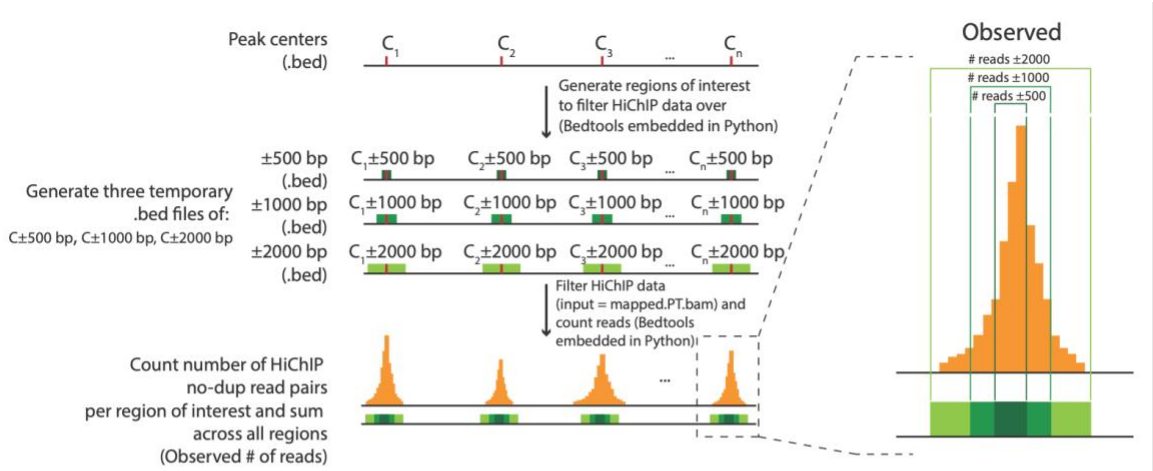


Category	Value	Percent
Total ChIP peaks	41,017	NA
Mean ChIP peak size	309 bp	NA
Median ChIP peak size	309 bp	NA
Total reads in 500 bp around center of peaks	393,163	9.46%
Total reads in 1000 bp around center of peaks	519,272	12.49%
Total reads in 2000 bp around center of peaks	692,305	16.66%
Observed/Expected ratio of reads in 500 bp around center of peaks	14.25	NA
Observed/Expected ratio of reads in 1000 bp around center of peaks	9.41	NA
Observed/Expected ratio of reads in 2000 bp around center of peaks	6.27	NA

Figure 4. Characterizing ChIP-seq peak signal and identifying peak centers.
Process - Using **bedtools** (as part of a Python script), the number of ChIP-seq peaks and the mean and median peak sizes are calculated. The centers of each peak are identified and a new peak center ***.bed** file is produced.
Results - Reported are the number of peaks and their respective mean and median size.

Step 2.A. Counting the number of observed HiChIP read pairs in regions surrounding the peak centers

Following primary peak characterization, we then assess HiChIP coverage around those peaks. The number of read-pairs that occur in a window over a peak center constitutes our observed signal. To evaluate observed signal, we look to see how the number of read-pairs behave within three windows centered on the peak center: ±500 bp, ±1 kbp, and ±2 kbp. We expect the observed read-pair accumulation rate to be highest at the peak center and then fall as we move farther from the peak center. As output from the analysis, an observed signal metric is reported as total read-pair count and percent of non-duplicate (no-dup) read pairs.



Category	Value	Percent
Total ChIP peaks	41,017	NA
Mean ChIP peak size	309 bp	NA
Median ChIP peak size	309 bp	NA
Total reads in 500 bp around center of peaks	393,163	9.46%
Total reads in 1000 bp around center of peaks	519,272	12.49%
Total reads in 2000 bp around center of peaks	692,305	16.66%
Observed/Expected ratio of reads in 500 bp around center of peaks	14.25	NA
Observed/Expected ratio of reads in 1000 bp around center of peaks	9.41	NA
Observed/Expected ratio of reads in 2000 bp around center of peaks	6.27	NA

Figure 5. Calculating the observed HiChIP coverage signal.
Process - Using **bedtools** (in a Python script), new ***.bed** files are generated at ±500 bp, ±1 kbp, and ±2 kbp from the peak centers. For each ***.bed** file, **bedtools** is used to count the number of read-pairs per ***.bed** entry and sum the number of read-pairs across all entries.
Results - The total number of read pairs in the windows of interest are reported as a value and as a percentage of total no-dup read pairs.

Step 2.B. Calculating the number of expected HiChIP read-pairs in regions around the peak centers

Next, we want to compare the observed signal at primary peaks against an expected value which assumes that coverage is evenly distributed over the fraction of the genome that consists of primary peak sites. That is, the expected value assumes that the total number of HiChIP read-pairs are evenly distributed across the genome resulting in uniform coverage independent of peak position.

This value is calculated by taking the fraction of the genome length (bp) contained in each window (±500 bp, ±1 kbp, ±2 kbp centered on each peak) and multiplying by the total number of no-dup read pairs. The expected coverage value for each window should decrease as window length increases and the same number of no-dup reads are distributed across a larger fraction of

the genome. The computed coverage value for each window gives us an expected signal for our following analysis.

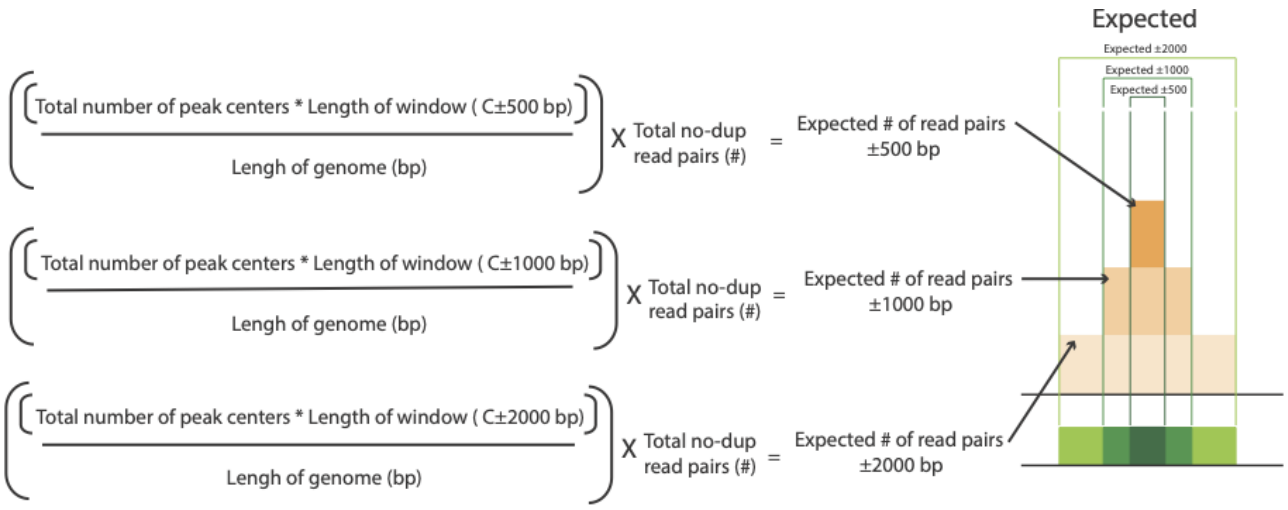
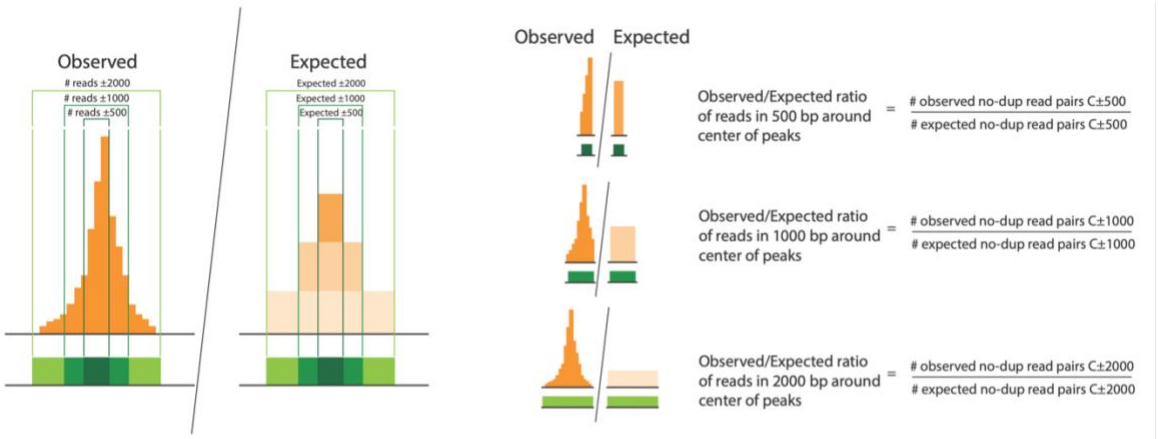


Figure 6. Calculating the expected signal.
Process - For each ***.bed** file (±500 bp, ±1 kbp, ±2 kbp from the peak centers), the total number of peak centers was multiplied by the length of the window, then divided by the genome's length to get the fraction of the genome enclosed in each window. Next, that fraction was multiplied by the number of total no-dup read pairs to evenly distribute the read pairs over the fraction of the genome contained in each window.
Result - A single coverage value is generated for each window (not reported).

Step 2.C. Calculating the observed to expected ratio

With the observed and expected signal values determined for the three windows around the primary peak centers, we can now calculate our observed:expected ratio. For a protein factor that binds a discrete recognition site, we expect a lower ratio value at larger window sizes indicating a successful ChIP.



Category	Value	Percent
Total ChIP peaks	41,017	NA
Mean ChIP peak size	309 bp	NA
Median ChIP peak size	309 bp	NA
Total reads in 500 bp around center of peaks	393,163	9.46%
Total reads in 1000 bp around center of peaks	519,272	12.49%
Total reads in 2000 bp around center of peaks	692,305	16.66%
Observed/Expected ratio of reads in 500 bp around center of peaks	14.25	NA
Observed/Expected ratio of reads in 1000 bp around center of peaks	9.41	NA
Observed/Expected ratio of reads in 2000 bp around center of peaks	6.27	NA

Figure 7. Calculating the observed to expected ratio.
Process - The observed value (the number of read pairs in the three windows around peak centers) is divided by the expected signal.
Result - The ratio value is reported for each window. As the window size expands, the ratio should be lower.

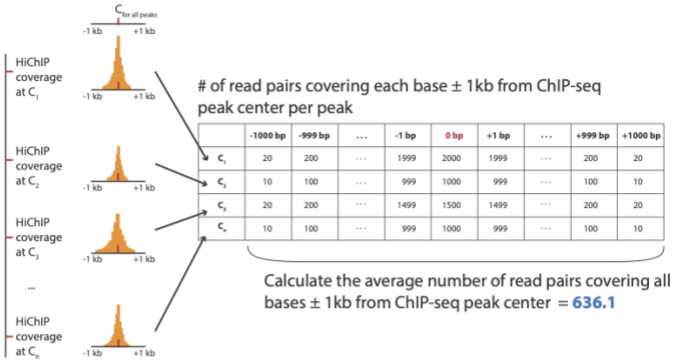
Step 3. Read density enrichment of HiChIP data over all primary peaks.

Now that we have assessed HiChIP coverage over primary peaks, it is useful to evaluate the HiChIP coverage averaged across all peaks. The result will be a plot that depicts the average coverage from the mean at each base ±1 kbp surrounding the primary peak centers. To achieve this, we use **samtools -mpileup** to build a data table with the read-pair coverage across all primary peaks. The columns in the data table contain the base position referenced to the primary peak center (… -3, -2, -1, 0, +1, +2, +3…) and there is a row entry per primary peak center. We then average the number of read-pairs at every position for every primary peak to obtain a mean coverage score.

Next, we normalize across all primary peak centers by dividing every entry in the table by the mean coverage score. Finally, to visualize the coverage across all primary peaks, the average coverage is calculated and plotted (Y-axis) against each base position (X-axis).

Step 3 - Read density enrichment of HiChIP data over ChIP-seq peaks

3.1) Average HiChIP per base coverage at each base position ± 1kb from ChIP-seq peak center to generate a mean HiChIP coverage across all bases ± 1kb from ChIP-seq peak center (samtools -mpileup embedded in Python)



	-1000 bp	-999 bp	...	-1 bp	0 bp	+1 bp	...	+999 bp	+1000 bp
C_1	20	200	...	1999	2000	1999	...	200	20
C_2	10	100	...	999	1000	999	...	100	10
C_3	20	200	...	1499	1500	1499	...	200	20
C_n	10	100	...	999	1000	999	...	100	10

3.2) For each base position ± 1kb from ChIP-seq peak center, calculate the coverage fold change from the mean coverage value in step 1 (python)

Fold change is calculated by dividing # of read pairs covering each base ± 1kb from ChIP-seq peak center per peak by the mean # of read pairs in step 3.1

	-1000 bp	-999 bp	...
C_1	20 / 636.14	200 / 636.14	...
C_2	10 / 636.14	...	...
...	...	...	...

Fold change from mean read pair coverage for each base ± 1kb from ChIP-seq peak center per peak

	-1000 bp	-999 bp	...	-1 bp	0 bp	+1 bp	...	+999 bp	+1000 bp
C_1	0.03	0.31	...	3.14	3.14	3.14	...	0.31	0.03
C_2	0.02	0.16	...	1.57	1.57	1.57	...	0.16	0.02
C_3	0.03	0.31	...	2.36	2.63	2.63	...	0.31	0.03
C_n	0.02	0.16	...	1.57	1.57	1.57	...	0.16	0.02

3.3) For each base position ± 1kb from ChIP-seq peak center, calculate and plot the mean of the coverage fold change (calculated in step 3.2) across all ChIP-seq peak centers (Python)

Average the fold change from mean read pair coverage for each base ± 1kb from ChIP-seq peak center across all peaks

	-1000 bp	-999 bp	...	-1 bp	0 bp	+1 bp	...	+999 bp	+1000 bp
C_1	0.03	0.31	...	3.14	3.14	3.14	...	0.31	0.03
C_2	0.02	0.16	...	1.57	1.57	1.57	...	0.16	0.02
C_3	0.03	0.31	...	2.36	2.63	2.63	...	0.31	0.03
C_n	0.02	0.16	...	1.57	1.57	1.57	...	0.16	0.02

Avg Avg Avg Avg Avg Avg Avg

	-1000 bp	-999 bp	...	-1 bp	0 bp	+1 bp	...	+999 bp	+1000 bp
$C_{mean\ fc}$	0.02	0.24	...	2.16	2.16	2.16	...	0.24	0.02

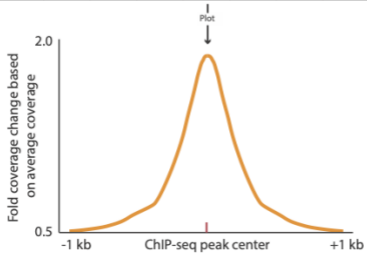


Figure 8. Read density enrichment of HiChIP data over all primary peaks.
Process - The number of read-pairs covering each base ±1 kbp from peak centers for each peak entry is calculated and organized into a table with **samtools -mpileups**. Then, an average coverage score is calculated across all cells in the table. Next, the coverage across all primary peaks are normalized by dividing each cell's coverage value by the mean coverage value. Finally, a mean coverage is calculated for each base position (columns) across all primary peaks.
Result - The resulting foldchange from the mean is plotted with the base position on the x-axis and the average fold change from the mean on the y-axis.