**Hi-C**

**DNA Library Preparation and Sequencing**

**Quality Control**

**Demo Report**

**Overseas Department**

**March, 2019**

Content

[1 Introduction - 2 -](#_Toc3898716)

[2 Experimental Procedure - 2 -](#_Toc3898717)

[2.1 Library Preparation for Sequencing - 2 -](#_Toc3898718)

[2.2 Sequencing - 3 -](#_Toc3898719)

[3 Data Quality Control - 4 -](#_Toc3898720)

[3.1 Raw data - 4 -](#_Toc3898721)

[3.2 Sequencing Data Filtration: - 5 -](#_Toc3898722)

[3.3 Sequencing Error Rate Examination: - 5 -](#_Toc3898723)

[3.4 Sequencing Quality Distribution - 6 -](#_Toc3898724)

[3.5 A/T/G/C Content Distribution - 7 -](#_Toc3898725)

[3.6 Data Quality Control Summary - 8 -](#_Toc3898726)

[3.7 Quality Control for Hi-C Data - 9 -](#_Toc3898727)

[3.8 Insert fragment analysis - 9 -](#_Toc3898728)

# 1 Background

Hi-C is short for High-throughput Chromosome Conformation Capture (Hi-C) technology. The study of the three-dimensional spatial structure and function of genomes is referred to as Three-Dimensional Genomics (3D Genomics). In 2003, Job Dekker and his collaborators proposed chromatin capture imaging (Chromatin Conformation Capture, 3C) to determine interchromosomal interactions between specific points. Hi-C was developed based on this in 2009. After the cross-linking of DNA fragments with a long linear distance and close spatial structure, the cross-linked DNA fragments was enriched, followed by high-throughput sequencing. The analysis of sequencing data can reveal the long-distance interaction of chromatin, thus deducing the three-dimensional structure of the genome. What’s more, after rapid development in recent years, it has been applied to the study of spatial regulation mechanisms of gene expression, construction of chromosome-level reference genomes and construction of haplotype maps.

# 2 Experimental Procedure

**Working Flow:**

## 2.1 Library Preparation for Sequencing

1. Nucleus DNA from tissue of the sample individual was cross-linked.
2. Then nucleus DNA was cutted with a restriction enzyme, leaving pairs of distally located but physically interacted DNA molecules attached to one another.
3. The sticky ends of these digested fragments were biotinylated.
4. Then biotinylated DNA fragments was ligated to each other to form chimeric circles.
5. The protein and DNA uncrosslinked after the protein at the junction point was digested. Genomic DNA were extracted and breaken to 350 bp fragments randomly by Covaris crusher.
6. Library construction: DNA fragments with biotin were captured using the avidin magnetic beads. The library was prepared by the steps of terminal repair, A tail addition, adaptor ligation, PCR amplification and purification.

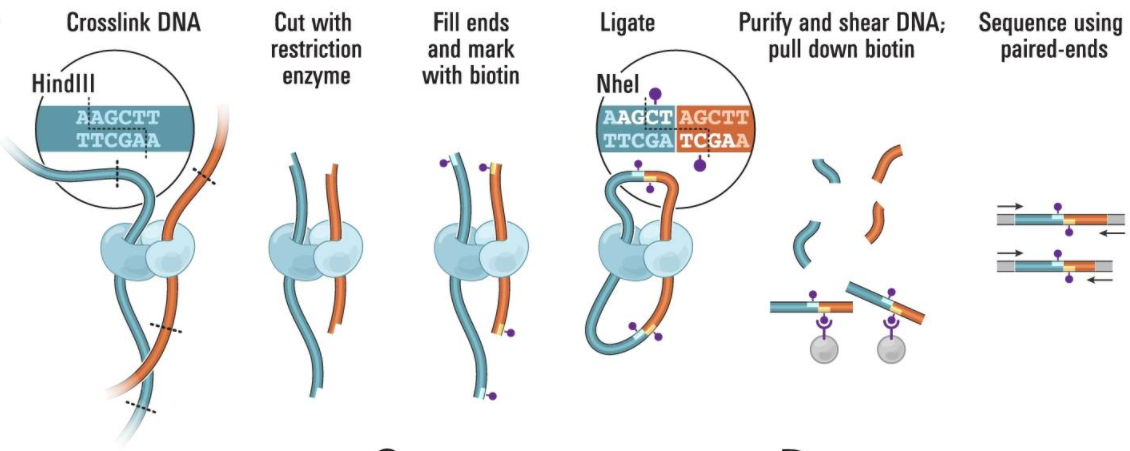


Figure 2.1 Hi-C Genomic DNA extraction and biotin labeling (Lieberman-Aiden E, et al., Science, 2009, 326(5950):289-93)

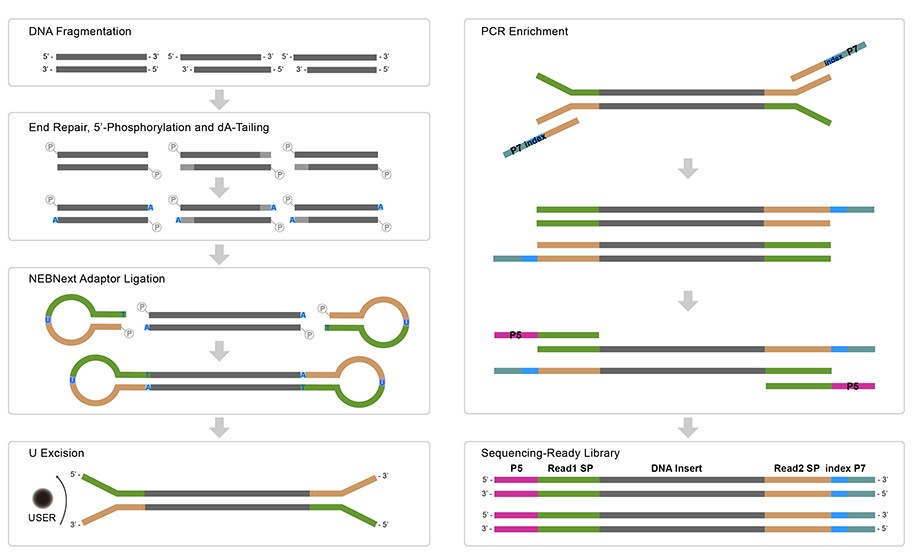


Figure 2.2 Library Construction Workflow

## 2.2 Sequencing

After deproteinization, removal of biotinylated free-ends, and DNA purification, Hi-C libraries were controlled for quality and sequenced on an Illumina sequencer (paired-end sequencing with 150 bp in read length).

# 3 Results and Instructions

## 3.1 Data Quality Control

## 3.1.1 Distribution of Sequencing Quality

The “e” represents the sequence error rate and Qphred represents the base quality value, Qphred=-10log10(e). The relationship between sequencing error rate (e) and sequencing base quality value (Qphred) is as below:

|  |  |  |  |
| --- | --- | --- | --- |
| Phred Score | Error Rate | Correct Rate | Q-score |
| 10 | 1/10 | 90% | Q10 |
| 20 | 1/100 | 99% | Q20 |
| 30 | 1/1000 | 99.9% | Q30 |
| 40 | 1/10000 | 99.99% | Q40 |

## The distribution of quality score is shown in Fig.3.1:

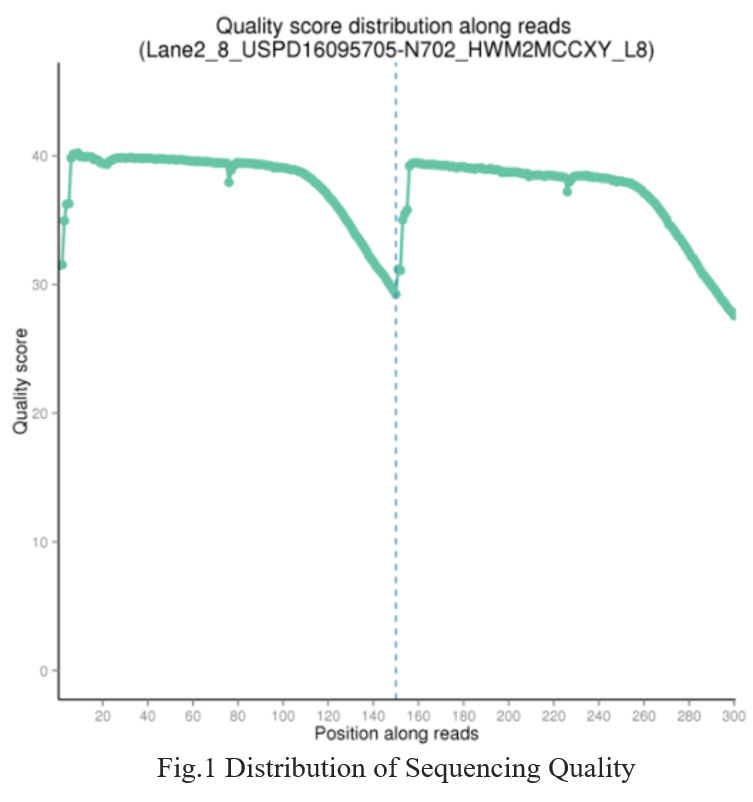


Figure 3.1 Sequencing error rate distribution

The base position is on the horizontal axis and the sequencing quality is on the vertical axis  
The first half part of the distribution is for reads1 and the latter half part is for reads2.

## 3.1.2 Distribution of Sequencing Error Rate

Sequencing error rate is related to the base quality of the obtained sequence. The sequencing platform, chemical reactants, and sample quality can all influence sequencing error rate and herein the base quality. For next-generation sequencing (NGS) with sequencing-by-synthesis strategy, sequencing error rate distribution shows two common features:

(1) Error rate increases with extending of the sequencing reads due to the consumption of chemical reagents, damage of the DNA template by laser irradiation, and possible accumulation of errors during the sequencing cycles. All the Illumina high-throughput sequencing platforms have this feature (Erlich Y. et al. 2008; Jiang et al. 2011).

(2) The sequencing error rate is higher for the first several bases than at other positions, which is likely the result of reading errors during the first few cycles after calibration of the optical instruments.

Generally, single base error rate should be lower than 1%. The error rate of this project is shown in Fig.3.2:

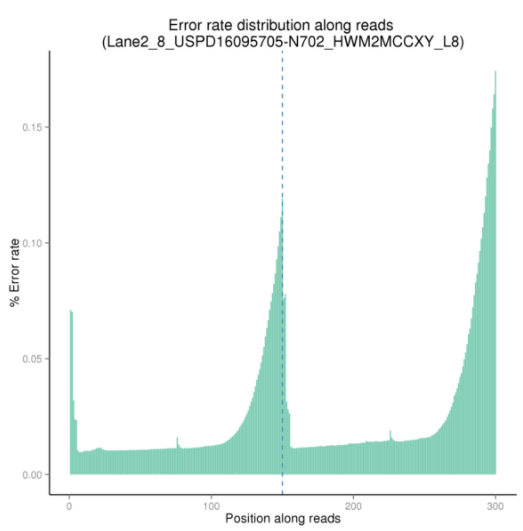


Figure 3.2 Sequencing error rate distribution

The base position is on the horizontal axis and the single base error rate is on the vertical axis  
The first half part of the distribution is for reads1 and the latter half part is for reads2

## 3.1.3 A/T/G/C Content Distribution

GC content distribution is evaluated to detect potential AT/GC separation, which affects subsequent gene expression quantification. Theoretically, G should equal C, and A should equal T throughout the whole sequencing process for non-stranded libraries, whereas AT/GC separation is normally observed in stranded libraries. For DGE (Digital Gene Expression) libraries, a large variation of sequencing error in the first 6-7 bases is allowed due to the use of random primers in library construction.

The distribution of GC content is shown in Fig.3:

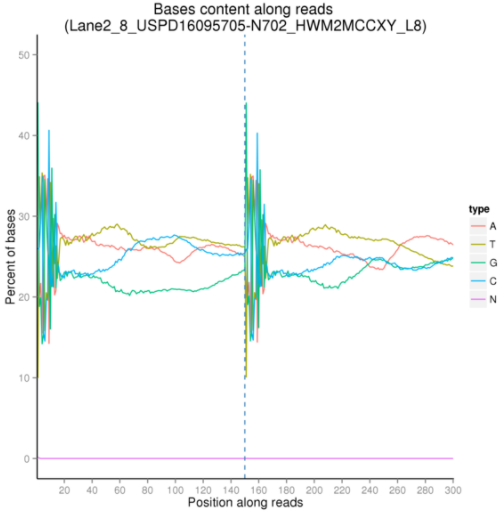


Figure 3.3 GC content distribution

The base position is on the horizontal axis and the single base percentage is on the vertical axis

The first half part of the distribution is for reads1 and the latter half part is for reads2.

## 3.1.4 Sequencing Data Filtration:

Raw sequencing data may contain adapter contaminated and low-quality reads. These sequence artifacts may increase the complexity of downstream analyses, which means that quality control is an essential step. All the downstream analyses will be based on clean reads that pass quality control.

We performed quality control according to the following procedure:

(1) Discard a read pair if either one read contains adapter contamination;

(2) Discard a read pair if more than 10% of bases are uncertain in either one read;

(3) Discard a read pair if the proportion of bases of low quality is over 50% in either one read.

Adapter sequences :

5' Adapter

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

3' Adapter(The underlined 6bp bases is Index)

5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3'

The Sequencing data filtration of this project can be seen in Fig.3.4:

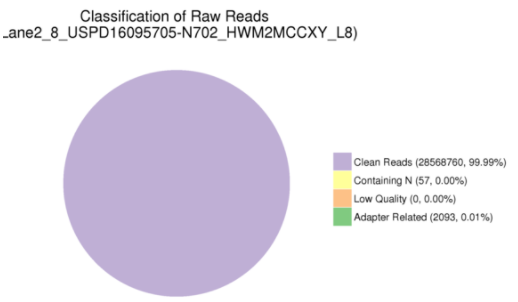


Figure 3.4 Composition of Raw Data

## 3.2 Summary of Sequencing Data Information

Consistent with the Illumina platform sequencing features, the data quality control summary is shown in Table 1.

Table 1 Data Quality Summary

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | Library | Flowcell/Lane | Raw reads | Raw data(G) | Effective  (%) | Error  (%) | Q20(%) | Q30(%) | GC(%) |
| RHC01262\_L4 | USPD16095701-1 | HWM2MCCXY\_L7 | 21621442 | 6.5 | 77.47 | 0.02% | 96.43 | 92.23 | 44.08 |
| RHC01262\_M5 | USPD16095702-5 | HWM2MCCXY\_L7 | 28649626 | 8.6 | 73.46 | 0.02% | 96.49 | 92.34 | 42.86 |

The details for the sequencing data statistics are as follows:

1. Sample name: Sample name.
2. Raw reads: four rows are taken as a unit to calculate the total amount of read1 and read2 in raw data files.
3. Raw bases: (total raw reads) \* (sequence length), calculating in G.
4. Error rate：The average error rate of all bases.
5. Q20：The percentage of bases with Phred score ≥20.
6. Q30：The percentage of bases with Phred score ≥30.
7. GC content：The percentage of G and C in the total bases.

# 4 Appendix

**4.1 Introduction of Sequencing Data Format**

The original raw data from Illumina platform are transformed to Sequenced Reads by base calling. Raw data are recorded in a FASTQ file, which contains sequence information (reads) and corresponding sequencing quality information. Every read in FASTQ format is stored in four lines as follows:

@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT  
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT  
+  
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHFAE0ECFFD/AEHH

Line 1 begins with a '@' character and is followed by the Illumina Sequence Identifiers and an optional description.

Illumina Sequence Identifier details:

|  |  |
| --- | --- |
| Identifier | Meaning |
| HWI-ST1276 | Instrument – unique identifier of the sequencer |
| 71 | run number – Run number on instrument |
| C1162ACXX | FlowCell ID – ID of flowcell |
| 1 | LaneNumber – positive integer |
| 1101 | TileNumber – positive integer |
| 1208 | X – x coordinate of the spot. Integer which can be negative |
| 2458 | Y – y coordinate of the spot. Integer which can be negative |
| 1 | ReadNumber - 1 for single reads; 1 or 2 for paired ends |
| N | whether it is filtered - NB： Y if the read is filtered out, not in the delivered fastq file, N otherwise |
| 0 | control number - 0 when none of the control bits are on, otherwise it is an even number |
| CGATGT | Illumina index sequences |

Line 2 is the raw sequence read.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier and description.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of characters as there are bases in the sequence (Cock et al.).

**4.2 Explanation of Sequencing Data Related**

(1) The data deliverd is a compressed file in format of '.fq.gz'. Before data delivery, we will calculate the md5 value of each compressed file and please check it when you get the

data. There are two ways to check the md5 value. In Linux environment, you can use 'md5sum -c <\*md5.txt>' command under the data directory. In Windows environment, you can

use a calibration tool e.g. hashmyfiles. If the md5 value of compressed file doesn't match with the one we provide in md5 file in data directory, the file may have been damaged during

the transmitting procedure.

(2) For paired-end (PE) sequencing, every sample should have 2 data flies (read1 file and read2 file). These 2 files have the same line number, you could use 'wc -l' command to

check the line number in Linux environment. The line number divide by 4 is the number of reads.

(3) The date size is the space occupied by the data in the hard disk. It's related to the format of disk and compression ratio. And it has no influence on the quantity of sequenced

bases. So the size of read1 file may be unequal to the size of read2 file.

(4) When customer’s samples need large amount of data e.g. whole genome sequencing data, we would use separate-lane sequencing strategy to make sure the quality of data. So

it's possible that one sample has several parts sequencing data. For example, if sample 1 has two read1 files, sample1\_L1\_1.fq.gz and sample1\_L2\_1.fq.gz, that means this sample was

sequenced on different lanes.

(5) About the sequenced reads. The Index is normally in the middle of the adapter during the process of experimenting and sequencing except the special library. We can get the

Read1 sequence and Read2 sequence by Index read. They are all the sequence of samples so that it's no necessary to dispose the beginning and end of reads in the downstream

analysis(e.g. mapping).

(6) Ninety days after the data delivery, we will delete outdated data. So please keep your data properly. If you have any question or doubt, please contact us as soon as possible.

Have a nice day!

**4.3 References**

Cock P.J.A. et al (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.

Erlich Y.et al (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.Nature Methods,5,679-682.

Jiang L.C. et al (2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551.