filter - DNBseq WGBS

Report ID: F25A430000133_ANIdaxnH

Date: 2025.02.26

BGI Genomics Co.,Ltd.

1. Result

1.1 Project Information

The basic information of the project is shown below:

- Project ID: F25A430000133_ANIdaxnH
- Product name: filter - DNBseq WGBS
- Sample size: 6
- Library type: DNBseq Whole genome bisulfite library
- Sequencing Platform: DNBSEQ
- Sequencing read Length: PE150
- Clean fastq phred quality score encoding: Phred+33

1.2 Data Production

After sequencing, the raw reads were filtered. Data filtering includes removing adapter sequences, contamination and low-quality reads from raw reads.

Table 1 Statistics of clean data

Sample Name	Clean Reads	Clean Base	Read Length	Q20(%)	Q30(%)	GC(%)
MSQ43_R1	15,177,772	4,553,331,600	PE150	96.48	88.15	21.58
MSQ43_R2	14,699,786	4,409,935,800	PE150	96.42	88.14	21.92
MoPh11	360,304,377	108,091,313,100	PE150	96.13	87.96	22.51
MoPh14	360,314,625	108,094,387,500	PE150	96.40	88.89	22.59
MoPh15	360,312,518	108,093,755,400	PE150	96.68	89.25	22.70
MoPh7	360,143,930	108,043,179,000	PE150	96.80	89.45	22.02

- Sample Name: Sample Name;
- Clean Reads: Clean Reads;
- Clean Base: Clean Bases;
- Read Length: Read Length;
- Q20(%): Proportion of Q20;
- Q30(%): Proportion of Q30;
- GC(%): Proportion of GC.
- For better view, datas are formatted.

1.3 Quality Control

The quality of data was examined after filtering.

1.3.1 The distribution of base percentage and qualities along reads after data filtering

In the left figure, x-axis represents base position along reads, y-axis represents base percentage at the position; each color represents a type of nucleotide. Under normal conditions, the sample does not have AT/GC separation. It is normal to see fluctuations in the first several bp positions, which is caused by random primer and the instability of enzyme-substrate binding at the beginning of the sequencing reaction. In the right figure, x-axis represents base position along reads, y-axis represents base quality; each dot represents the base quality of the corresponding position along reads, color intensity reflects the number of nucleotides, a more intense color along a quality value indicates a higher proportion of this quality in the sequencing data.

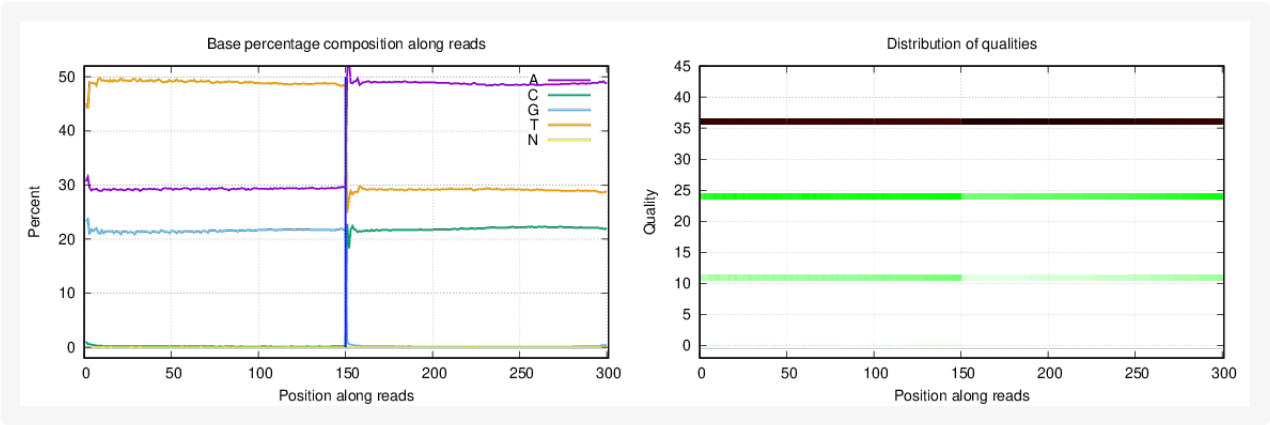

Figure 1-1 Distribution of base percentage and qualities of sample MSQ43_R2.

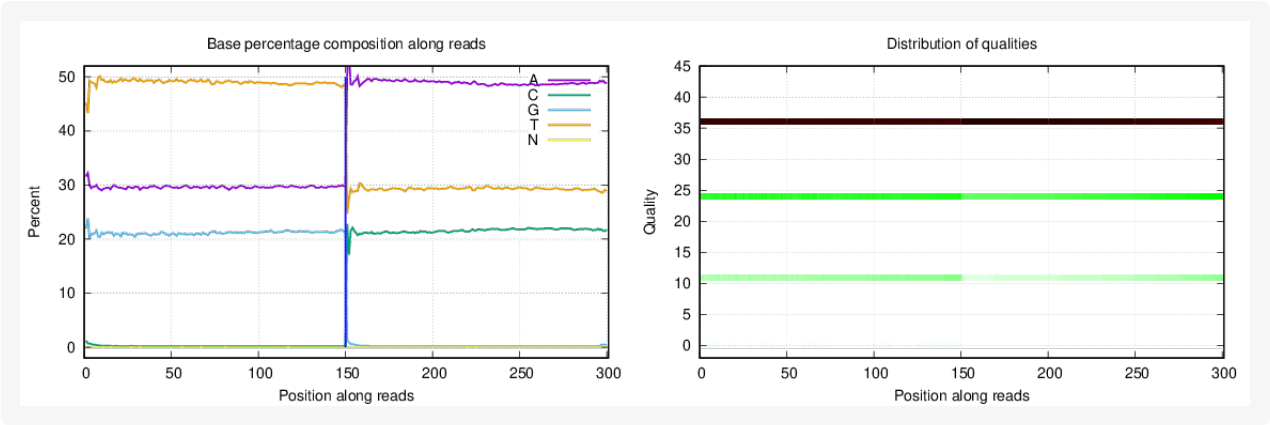

Figure 1-2 Distribution of base percentage and qualities of sample MSQ43_R1.

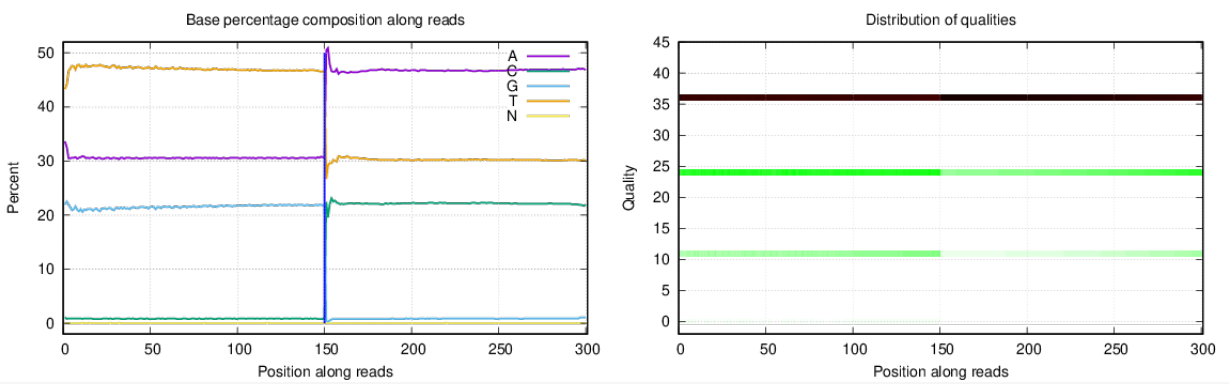

Figure 1-3 Distribution of base percentage and qualities of sample MoPh15.

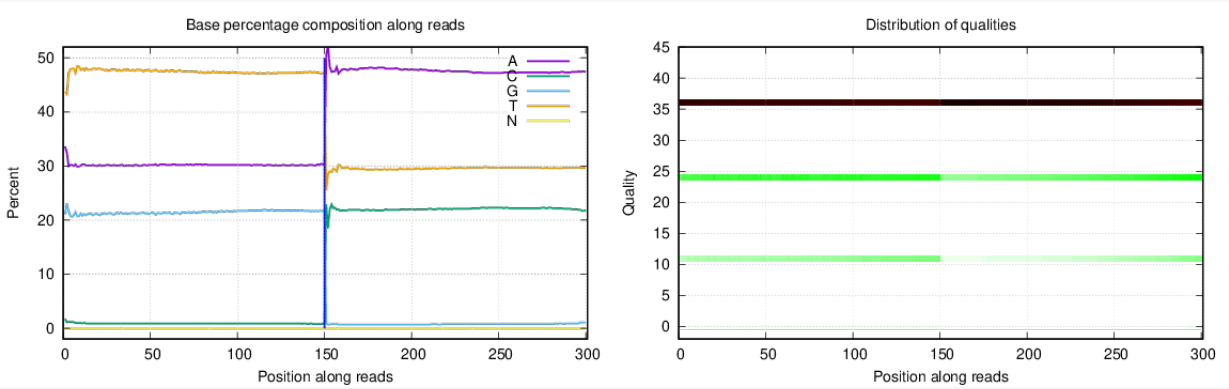

Figure 1-4 Distribution of base percentage and qualities of sample MoPh14.

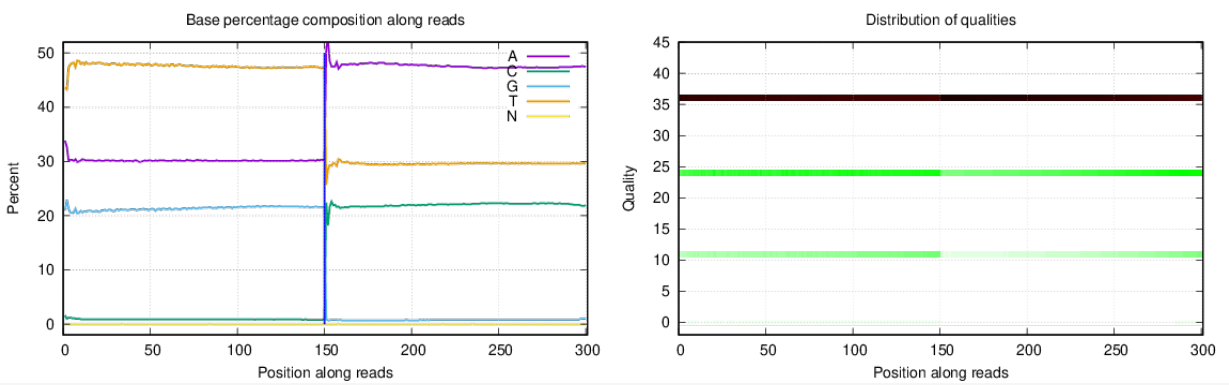

Figure 1-5 Distribution of base percentage and qualities of sample MoPh11.

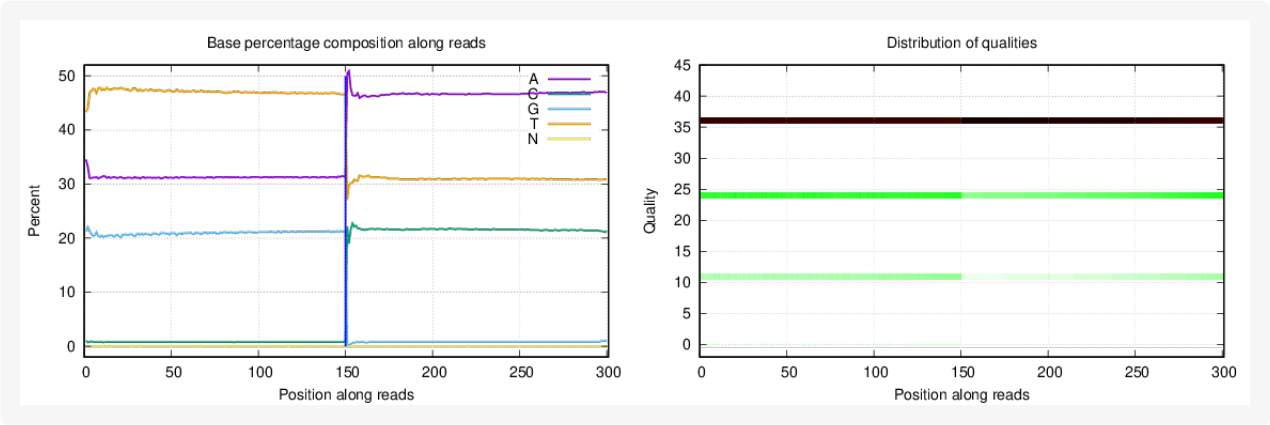

Figure 1-6 Distribution of base percentage and qualities of sample MoPh7.

2. Methods

2.1 Experimental Procedure

The library construction method and sequencing process are carried out according to the following steps:

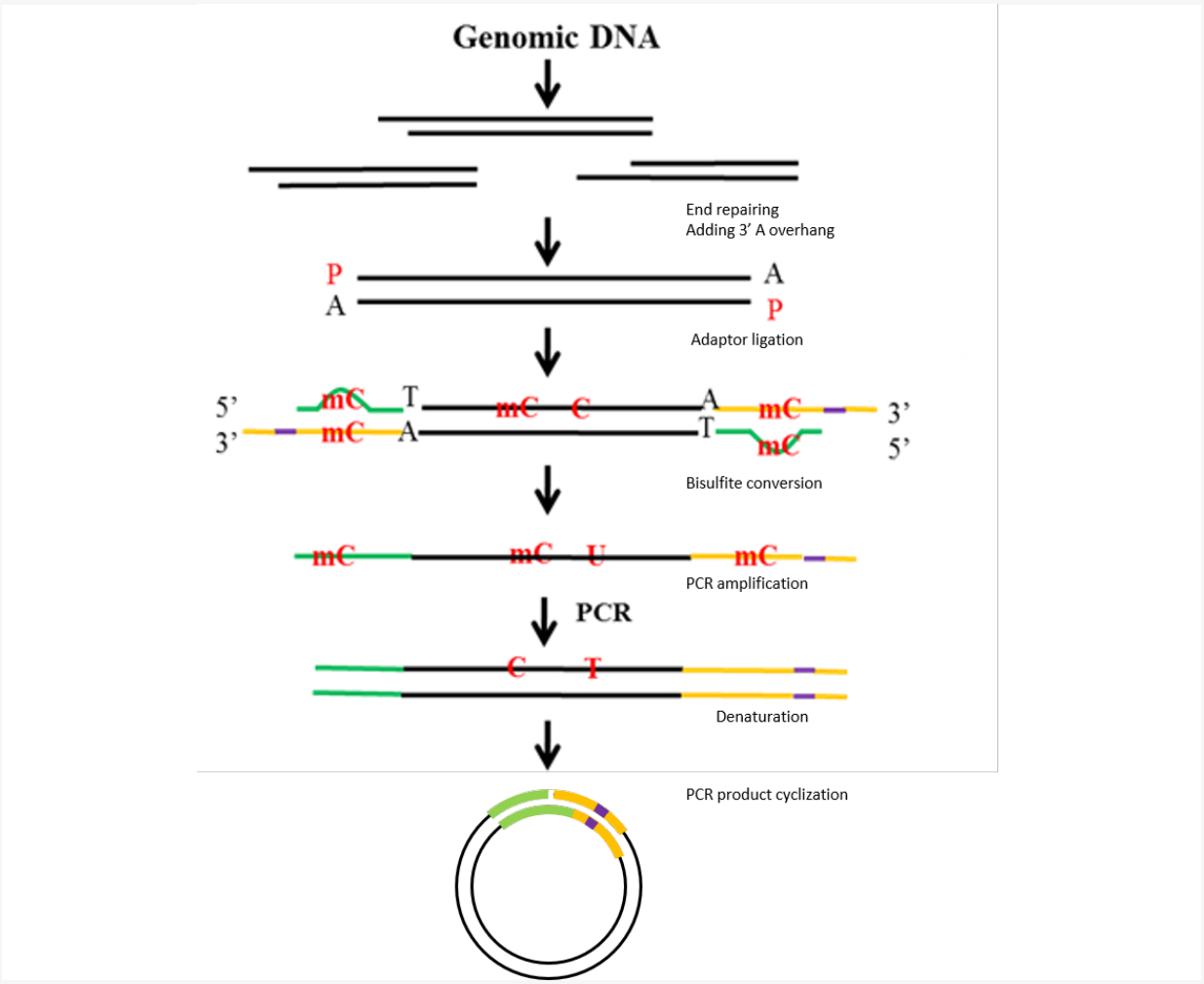

Figure 2 Workflow of experiment.

1. Take genome DNA samples for quality control;
2. Fragmented genome DNA to the mean size of 200-350bp;
3. Blunt-ending and 3'-end dAaddition;
4. Methlated adaptor ligation;
5. Bisulfite conversion with EZ DNA Methylation-Gold kit (ZYMO);
6. PCR amplification;
7. Library quality control;
8. Library circularization;
9. The library was amplified to make DNA nanoball (DNB);
10. Sequencing on DNBSEQ (DNBSEQ Technology) platform.

2.2 Bioinformatic Analysis Workflow

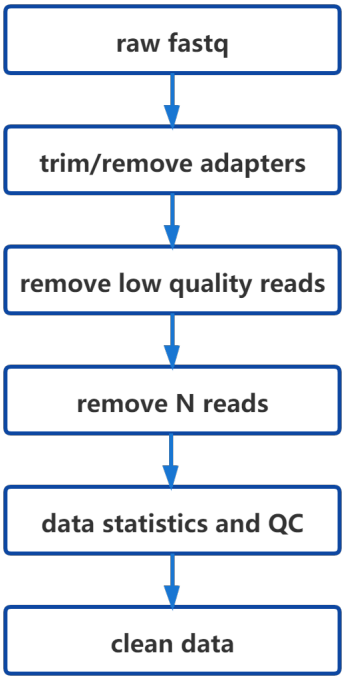

Figure 3 Bioinformatic analysis workflow.

2.3 Parameters for Data Filtering

Raw data with adapter sequences or low-quality sequences was filtered. We first went through a series of data processing to remove contamination and obtain valid data. This step was completed by SOAPnuke [1] developed by BGI.

SOAPnuke software filter parameters: "-n 0.01 -l 20 -q 0.4 --adaMR 0.25 --ada_trim --minReadLen 150", steps of filtering:

1. Filter adapter: if the sequencing read matches 25.0% or more of the adapter sequence (maximum 2 base mismatches are allowed), cut the adapter;
2. Filter read length: if the length of the sequencing read is less than 150 bp, discard the entire read;
3. Remove N: if the N content in the sequencing read accounts for 1.0% more of the entire read, discard the entire read;
4. Filter low-quality data: if the bases with a quality value of less than 20 in the sequencing read account for 40.0% or more of the entire read, discard the entire read;
5. Obtain Clean reads: the output read quality value system is set to Phred+33.

3. Help

3.1 FASTQ format description

Images generated by sequencers are converted by base calling into nucleotide sequences, which are called raw data or raw reads and are stored in FASTQ format. FASTQ files are text files that store both reads sequences and their corresponding quality scores. Each read is described in four lines as follows:

```
@V300029258L2C001R0010000017/1
TTTTTCTGCTCCTTTTGATGCTATTAACAATTGCTTCAAGTTCAAGGGCACCTGCCTCAAAGTCCCTTTCTTCCAGACAAAATCTGAGA
GTTTCATCTTT
+
=,DDE@EFFF=DFDEFCCFDEFEGFEEAFDFFE=FFCFFEEEDFDEEEFDF8FFEFFEFF:FFEDF=EFDGE<1FDCEFFFFFDFEDCF
FFFFFFF9GCF
```

The first line is the sequence identifier and related description information, starting with'@'; the second line is the base sequence information; the third line starts with'+', followed by the sequence identifier, description information, or nothing; The four lines are quality information, which corresponds to the sequence in the second line. Each base has a quality score. Depending on the scoring system, each character represents a different quality value.

The figure below shows the concise correspondence between the sequencing error rate and the sequencing quality value. Specifically, if the sequencing error rate is denoted by E and the base quality value is denoted by SQ, there are the following relationships:

$$SQ = -10 * (log\frac{E}{1-E})/log10$$

> In which:
>
> $$E = \frac{Y}{1+Y}$$
>
> $$Y = \frac{SQ}{e^{-10*log10}}$$

1）For the quality system data with a sequencing quality value of 33: the sequencing quality value of the base = the ASCII value corresponding to the quality information character -33, for example, the ASCII value corresponding to A is 65, then the corresponding base quality value is 65-33 = 32. The base quality value of the DNBSEQ sequencing platform ranges from 2 to 42.

2）For quality system data with a sequencing quality value of 64: the sequencing quality value of the base = the ASCII value corresponding to the quality information character -64, for example, the corresponding ASCII value of c is 99, then the corresponding base quality value is 99-64 =35. The base quality value of the DNBSEQ sequencing platform ranges from 2 to 43.

Table 2 A summary table of the relationship between sequencing error rate and sequencing quality

Sequencing error rate	Sequencing quality value	Character(Phred64)	Character
5%	13	M	.

Sequencing error rate	Sequencing quality value	Character(Phred64)	Character
1%	20	T	5
0.1%	30	^	?

- Sequencing error rate： Sequencing error rate
- Sequencing quality value：Sequencing quality value
- Character(Phred64)：ASCII code under Phred+64
- Character(Phred33)：ASCII code under Phred+33

4. References

1. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience, 7(1), 1-6. https://doi.org/10.1093/gigascience/gix120 ↩

Omics For All

Contact us

Website: www.bgi.com
E-mail: info@bgi.com