filter - DNBSEQ Eukaryotic Strand-specific Transcriptome Resequencing

Report ID: F23A430001415_HOMyebiR

Date: 2023.11.30

BGI Genomics Co.,Ltd.

1. Result

1.1 Project Information

The basic information of the project is shown below:

- Project ID: F23A430001415_HOMyebiR
- Product name: filter - DNBSEQ Eukaryotic Strand-specific Transcriptome Resequencing
- Sample size: 31
- Library type: DNBSEQ Eukaryotic Strand-specific mRNA library
- Sequencing Platform: DNBSEQ
- Sequencing read Length: PE100
- 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(%)
1-1	24077344	4815468800	PE100	97.42	92.01	51.53
1-2	24064708	4812941600	PE100	97.50	92.23	51.35
2-1	24182507	4836501400	PE100	97.48	92.17	51.05
2-2	24244058	4848811600	PE100	97.41	91.94	51.14
3-1	24169085	4833817000	PE100	97.51	92.27	50.92
3-2	24126196	4825239200	PE100	97.61	92.56	51.34
4-1	24159101	4831820200	PE100	97.51	92.28	51.45
4-2	24119521	4823904200	PE100	97.51	92.25	51.05
5-1	24172906	4834581200	PE100	97.48	92.18	51.25
5-2	24130161	4826032200	PE100	97.47	92.13	51.33
6-1	24237112	4847422400	PE100	97.31	91.63	51.16
6-2	24128920	4825784000	PE100	97.46	92.11	51.30
APV3	24200438	4840087600	PE100	98.25	94.24	52.34
BLA3	24316636	4863327200	PE100	98.24	94.22	51.82

Sample Name	Clean Reads	Clean Base	Read Length	Q20(%)	Q30(%)	GC(%)
EVV2	24162168	4832433600	PE100	97.62	92.62	52.57
GAS2	23094671	4618934200	PE100	97.45	92.19	52.11
GGA1	24116396	4823279200	PE100	97.59	92.52	52.38
GLG1	24140364	4828072800	PE100	98.27	94.38	51.49
IAS1	24185617	4837123400	PE100	97.76	93.03	52.27
IAV1	24243798	4848759600	PE100	97.47	92.25	51.88
KAG2	21985428	4397085600	PE100	97.72	93.00	52.17
LIV2	22516953	4503390600	PE100	97.56	92.53	55.09
LKN1	24330280	4866056000	PE100	97.99	93.41	52.97
MAV2	24163742	4832748400	PE100	98.14	93.91	52.06
SAU2	24230555	4846111000	PE100	97.56	92.38	53.20
SHV1	24168472	4833694400	PE100	98.05	93.60	52.35
Shua2	24246993	4849398600	PE100	97.46	92.19	51.71
TSV2	24070947	4814189400	PE100	97.67	92.76	52.69
Yu1	22229737	4445947400	PE100	97.71	92.90	52.16
ZUA2	24402344	4880468800	PE100	98.13	93.86	52.74
ZYAM3	24191929	4838385800	PE100	97.70	92.83	52.22

- 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.

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.

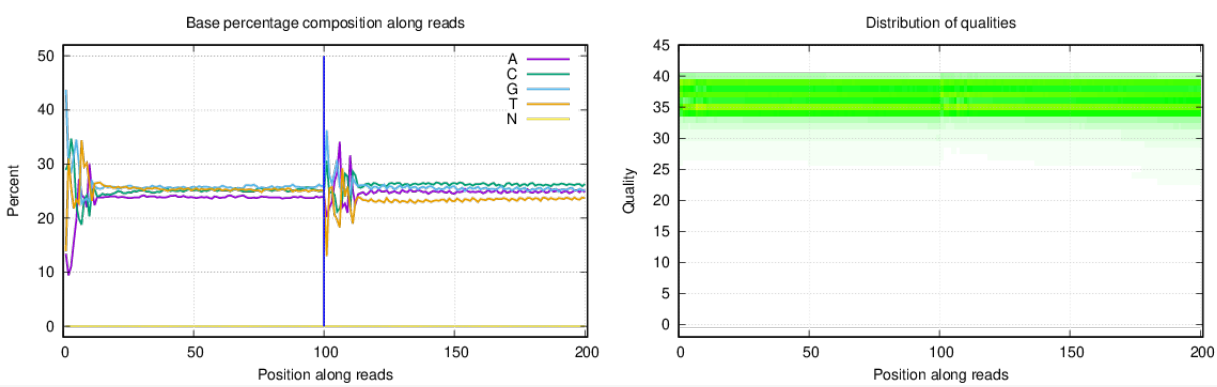

Figure 1-1 Distribution of base percentage and qualities of sample GLG1

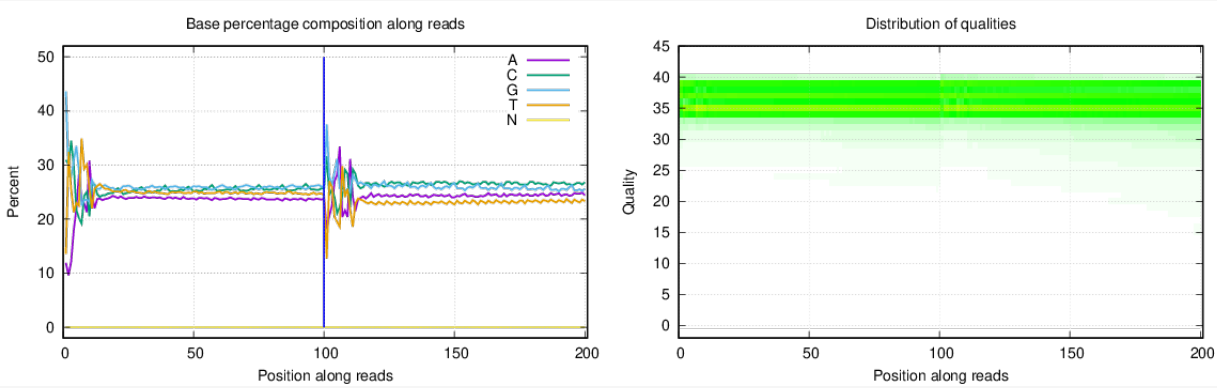

Figure 1-2 Distribution of base percentage and qualities of sample GAS2

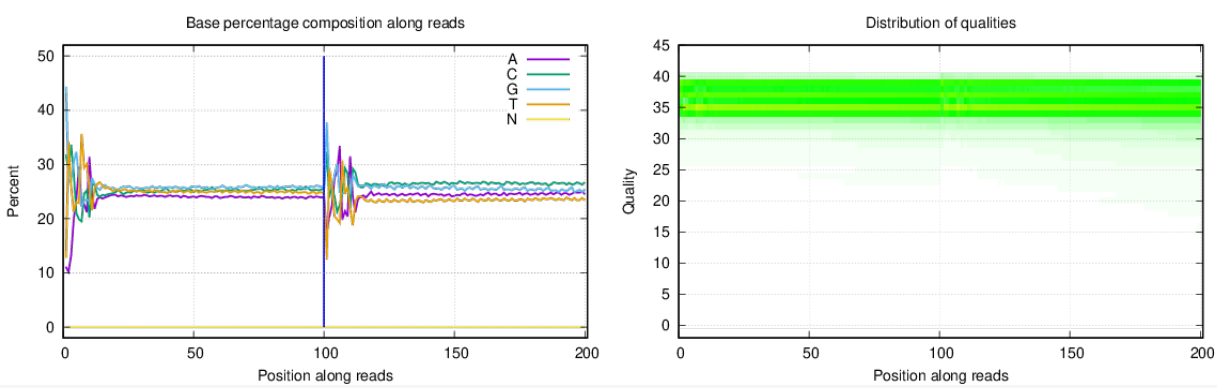

Figure 1-3 Distribution of base percentage and qualities of sample Shua2

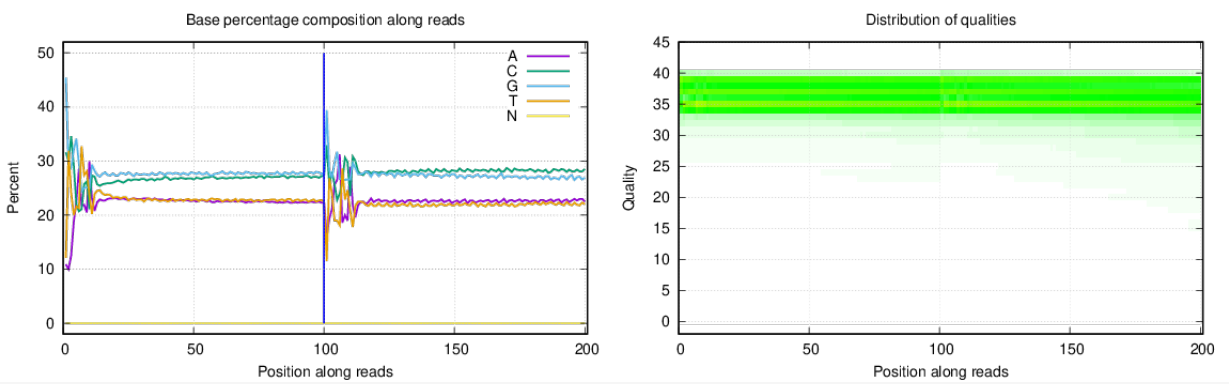

Figure 1-4 Distribution of base percentage and qualities of sample LIV2

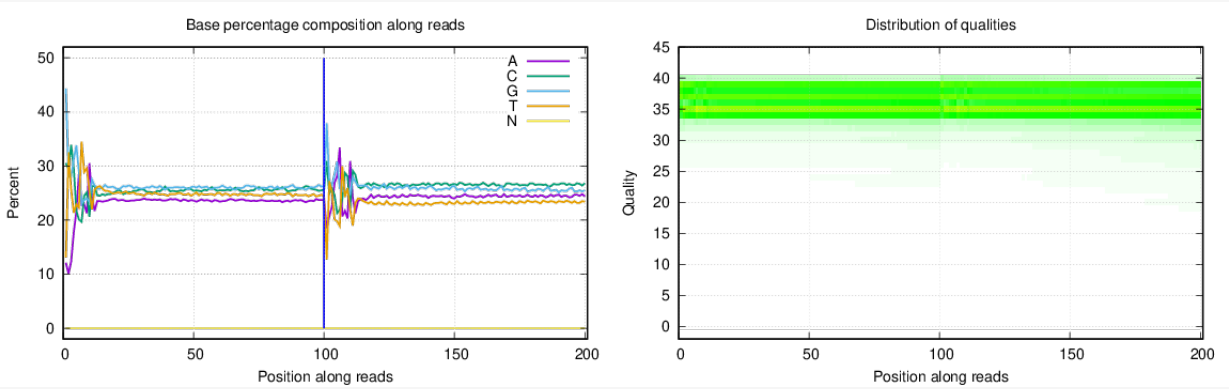

Figure 1-5 Distribution of base percentage and qualities of sample KAG2

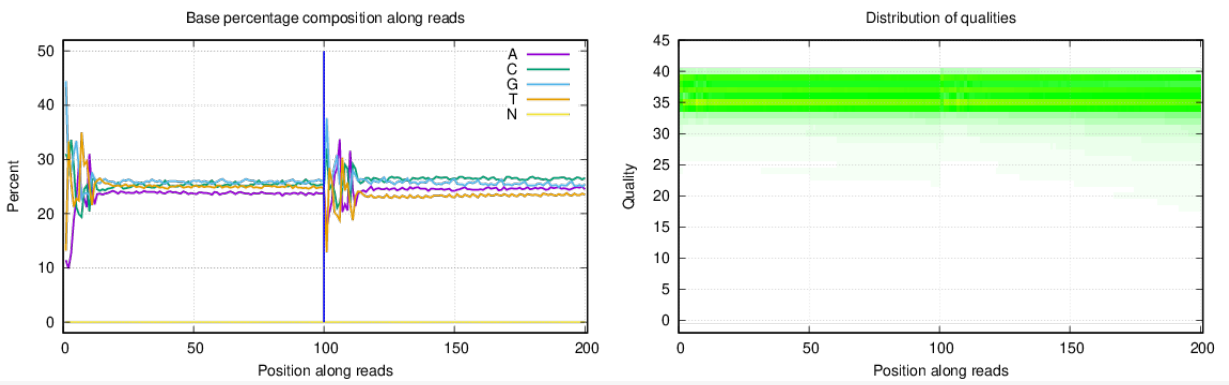

Figure 1-6 Distribution of base percentage and qualities of sample IAV1

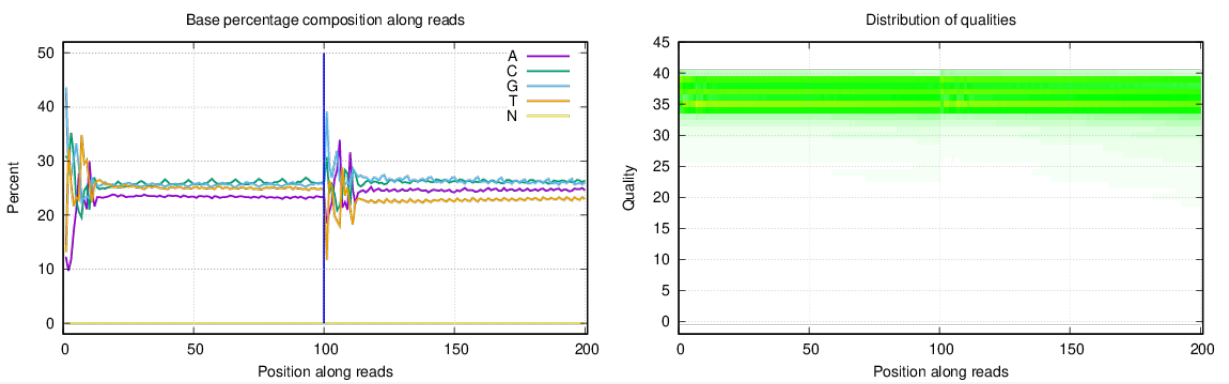

Figure 1-7 Distribution of base percentage and qualities of sample IAS1

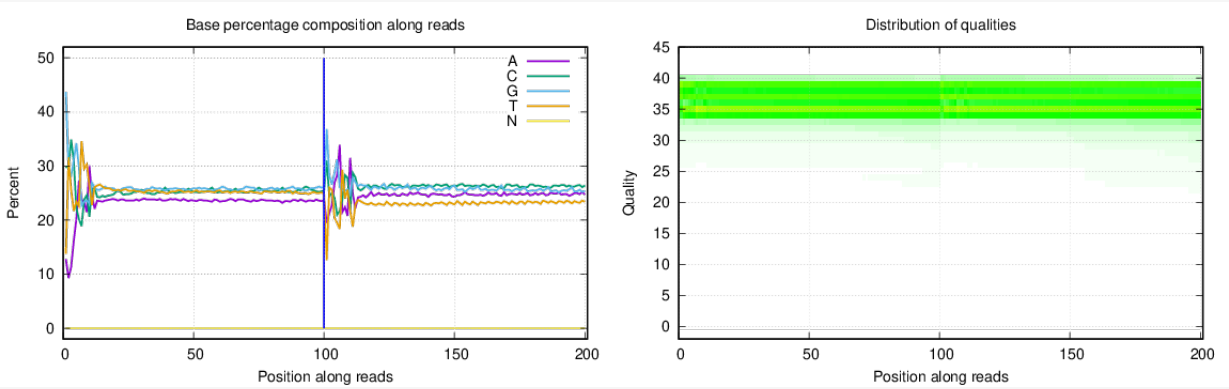

Figure 1-8 Distribution of base percentage and qualities of sample BLA3

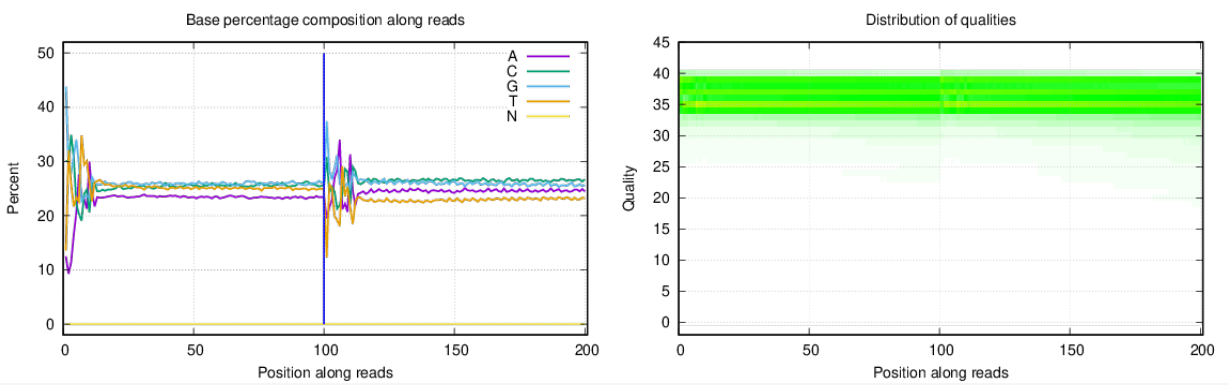

Figure 1-9 Distribution of base percentage and qualities of sample Yu1

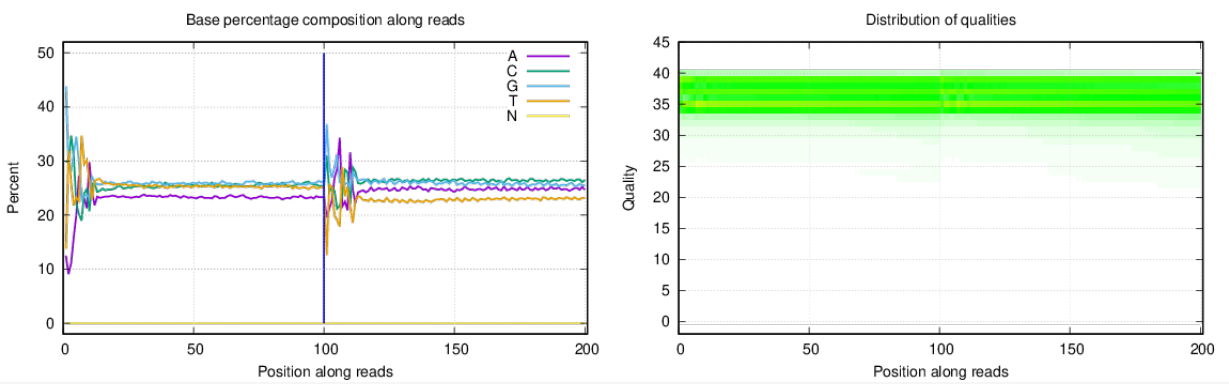

Figure 1-10 Distribution of base percentage and qualities of sample MAV2

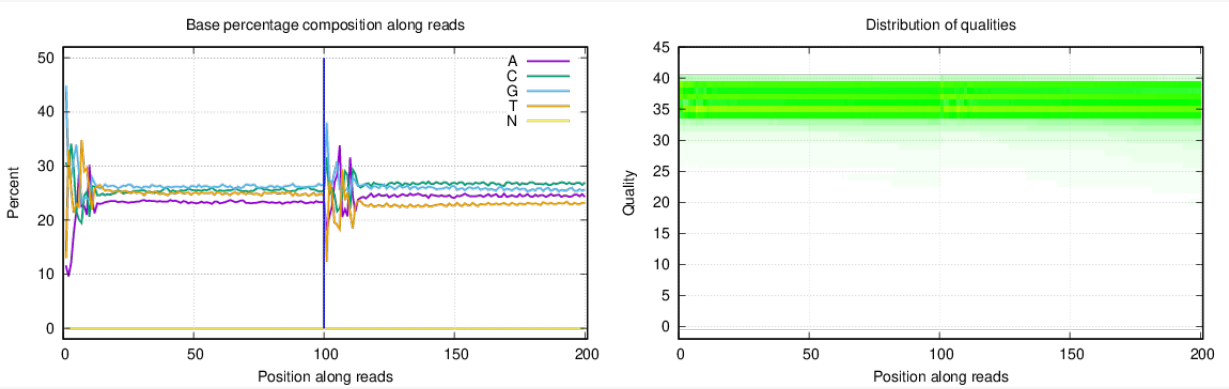

Figure 1-11 Distribution of base percentage and qualities of sample SHV1

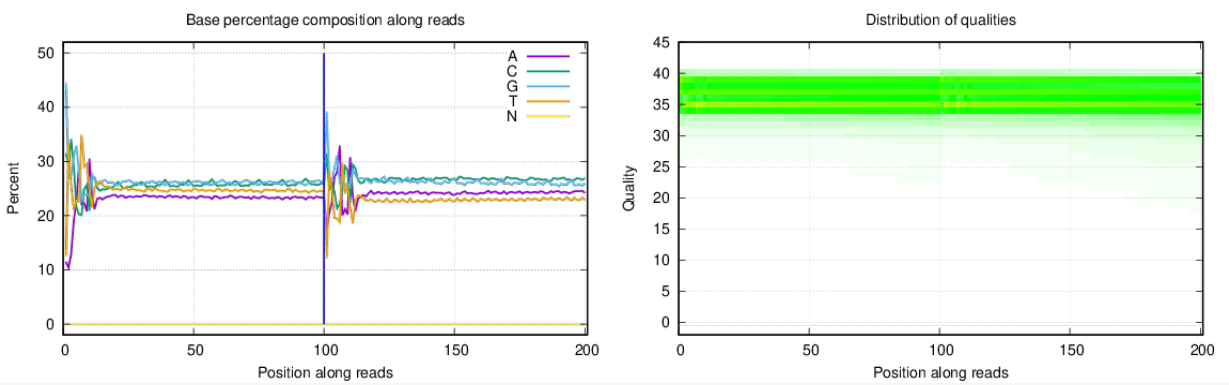

Figure 1-12 Distribution of base percentage and qualities of sample EVV2

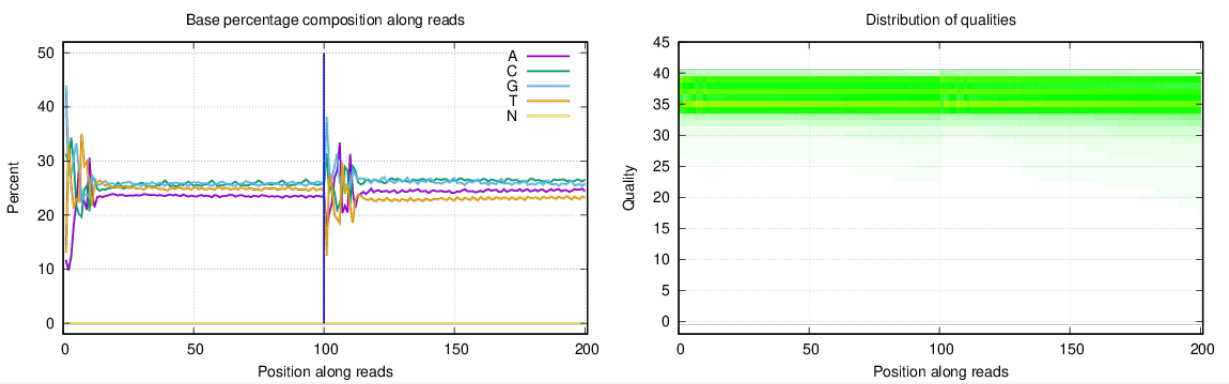

Figure 1-13 Distribution of base percentage and qualities of sample ZYAM3

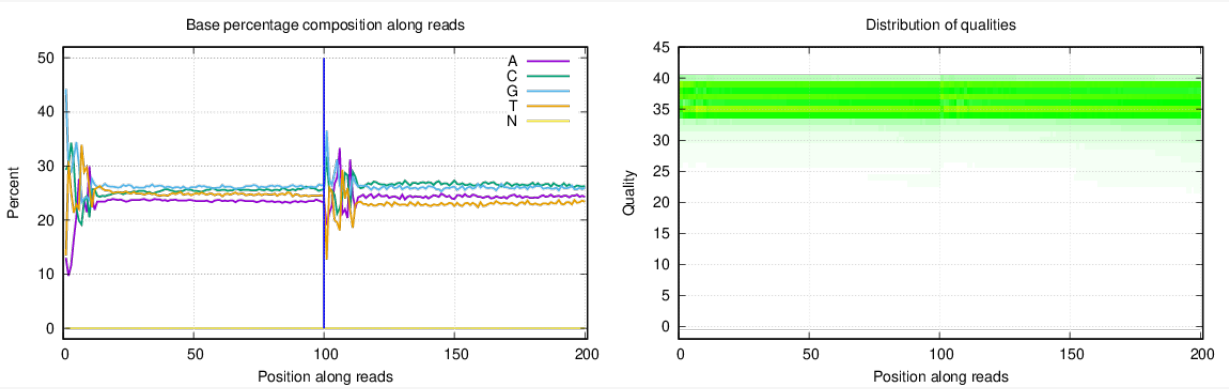

Figure 1-14 Distribution of base percentage and qualities of sample APV3

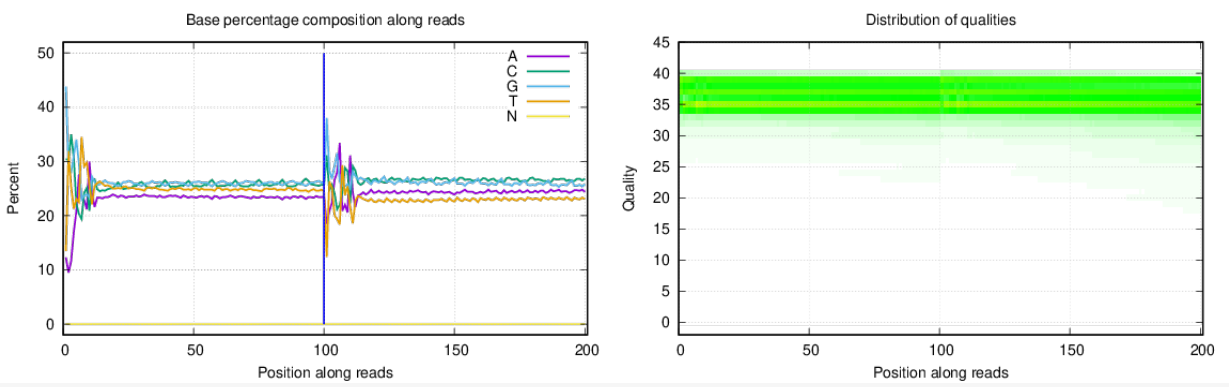

Figure 1-15 Distribution of base percentage and qualities of sample GGA1

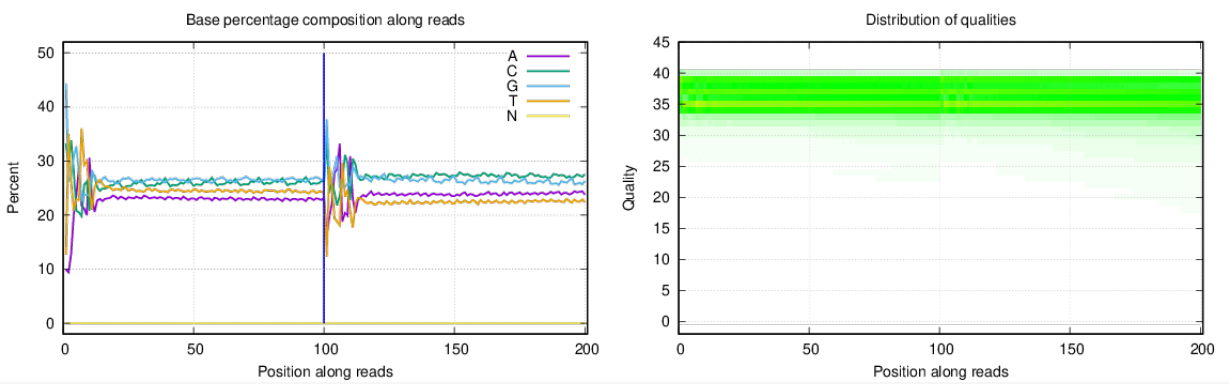

Figure 1-16 Distribution of base percentage and qualities of sample SAU2

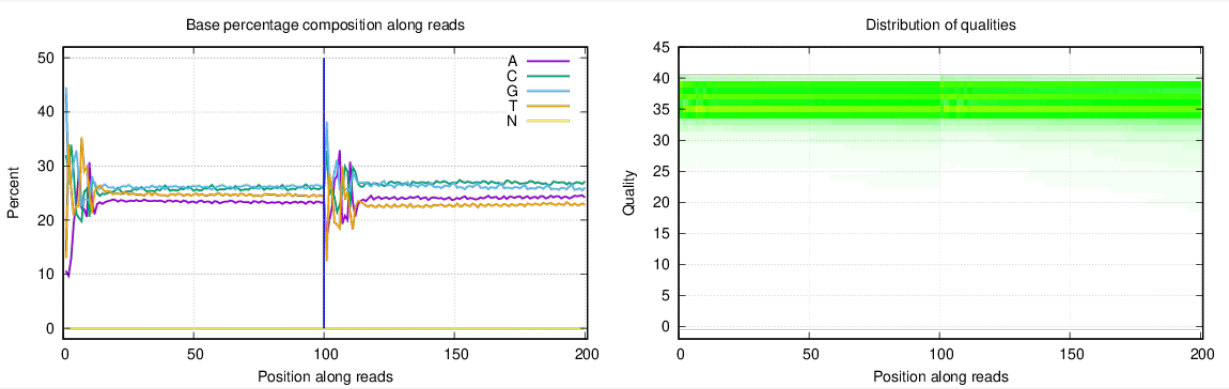

Figure 1-17 Distribution of base percentage and qualities of sample TSV2

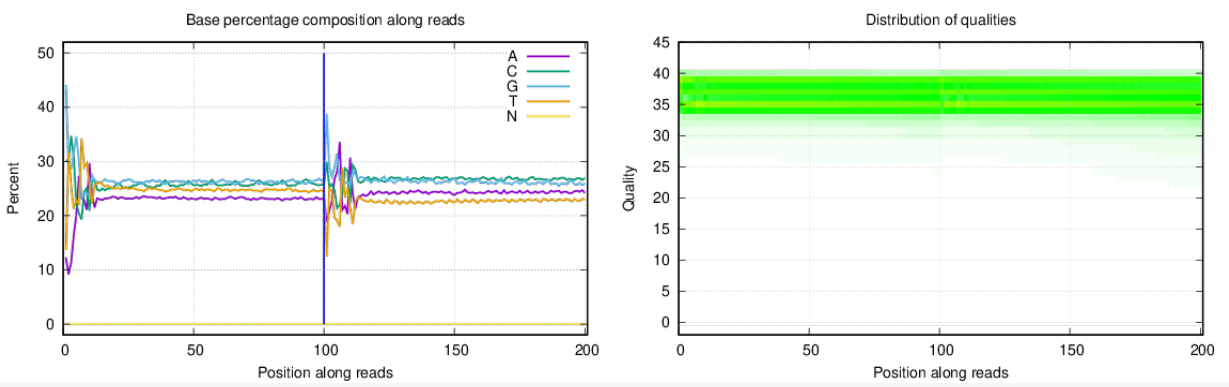

Figure 1-18 Distribution of base percentage and qualities of sample ZUA2

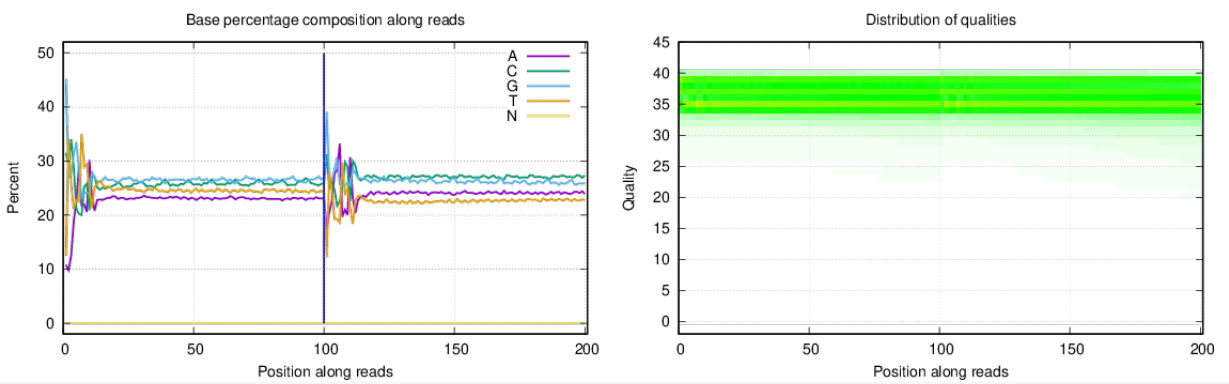

Figure 1-19 Distribution of base percentage and qualities of sample LKN1

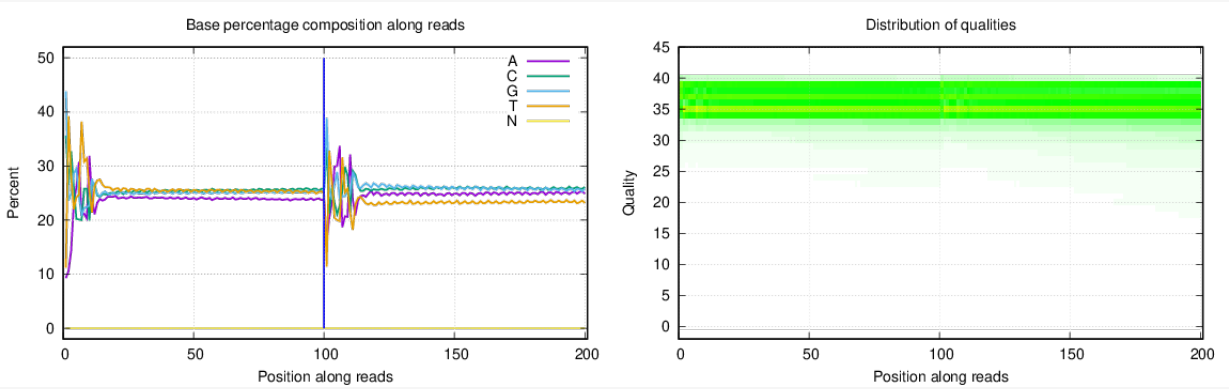

Figure 1-20 Distribution of base percentage and qualities of sample 6-2

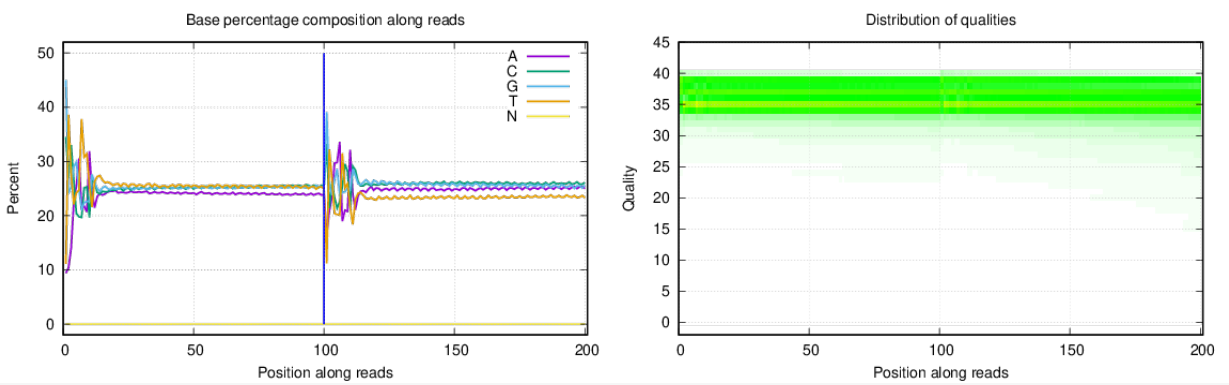

Figure 1-21 Distribution of base percentage and qualities of sample 6-1

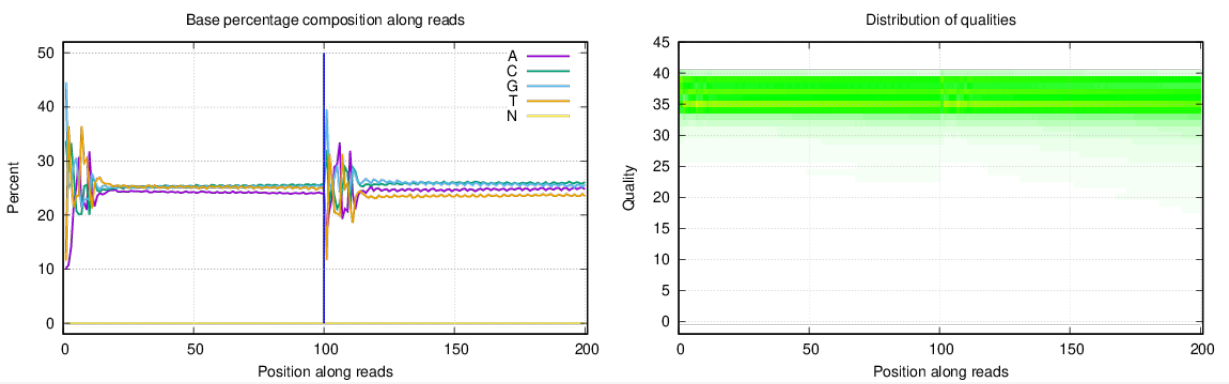

Figure 1-22 Distribution of base percentage and qualities of sample 5-2

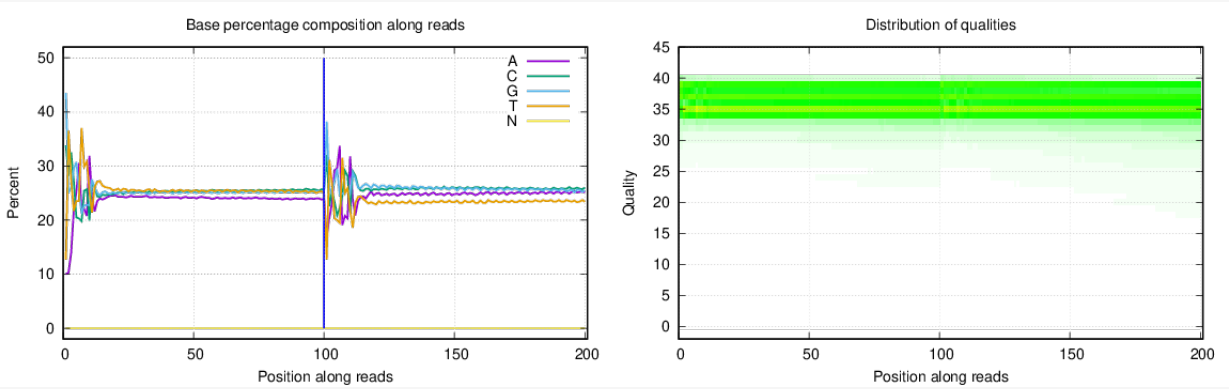

Figure 1-23 Distribution of base percentage and qualities of sample 5-1

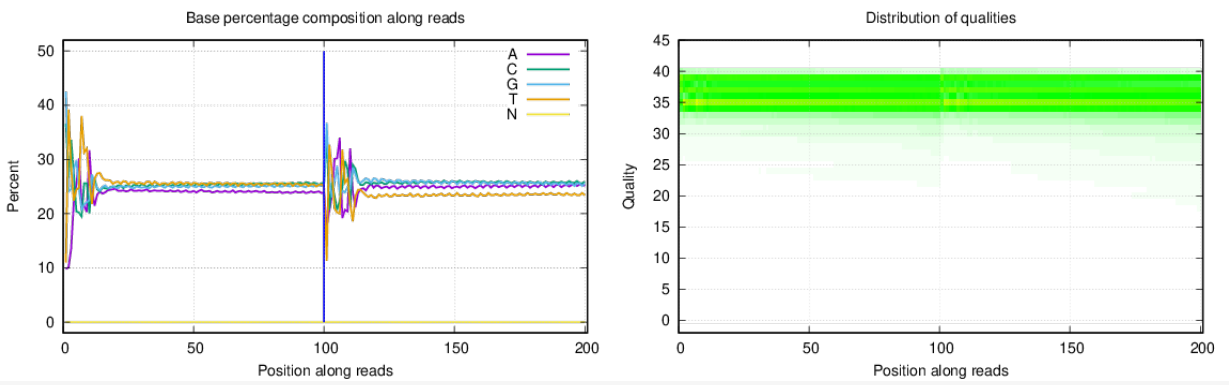

Figure 1-24 Distribution of base percentage and qualities of sample 4-2

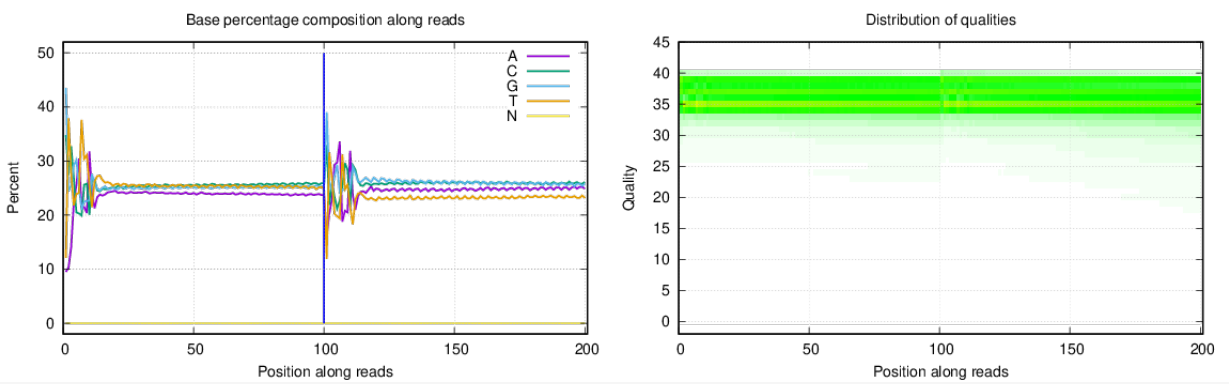

Figure 1-25 Distribution of base percentage and qualities of sample 4-1

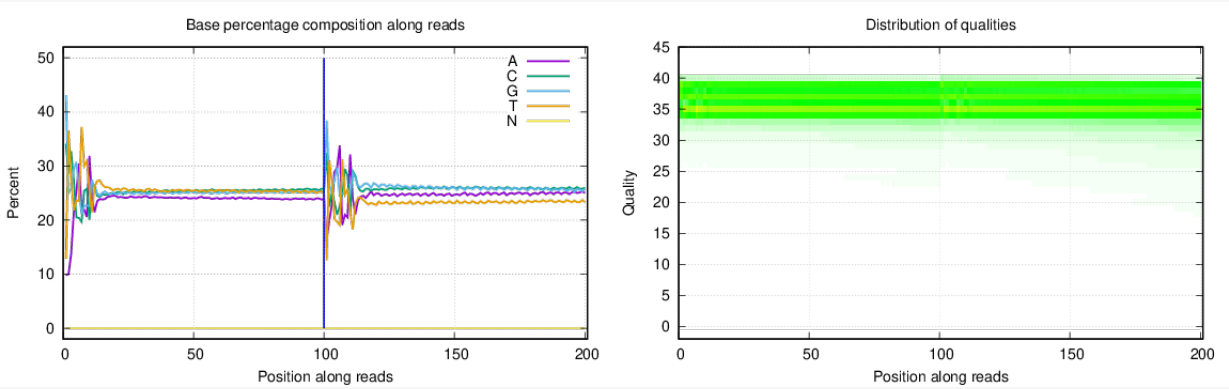

Figure 1-26 Distribution of base percentage and qualities of sample 3-2

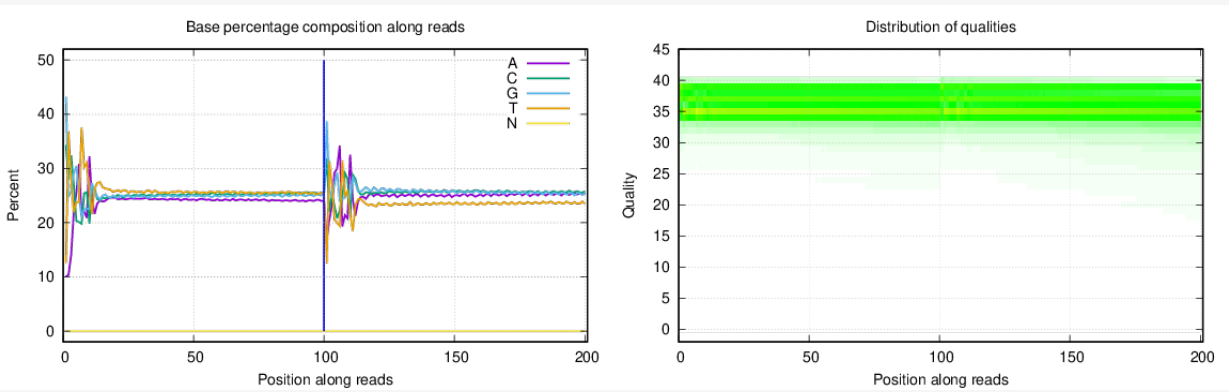

Figure 1-27 Distribution of base percentage and qualities of sample 3-1

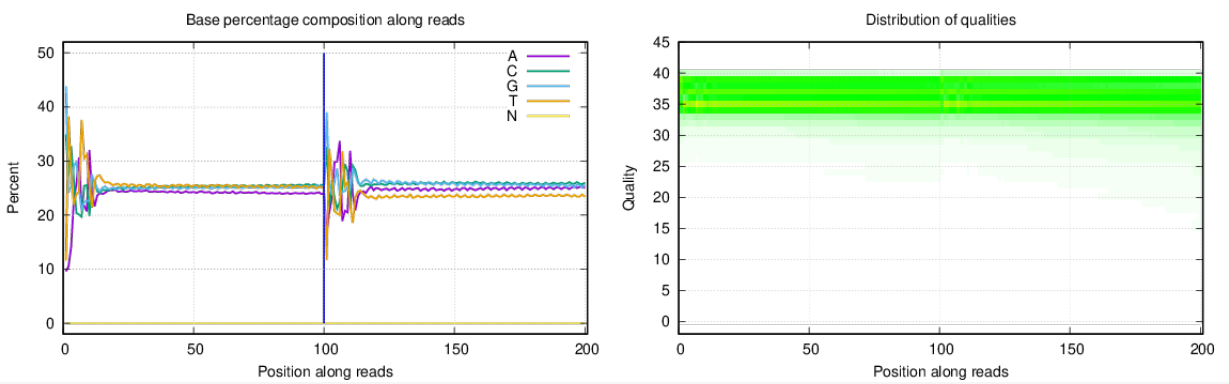

Figure 1-28 Distribution of base percentage and qualities of sample 2-2

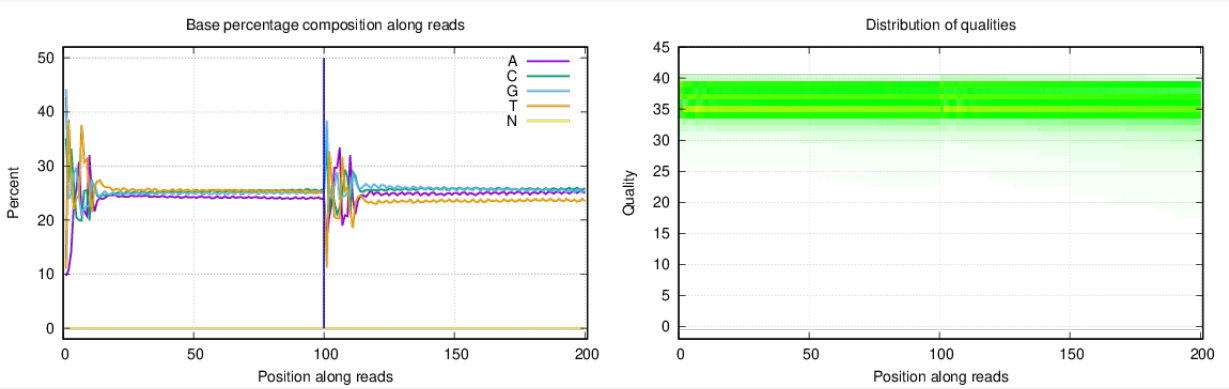

Figure 1-29 Distribution of base percentage and qualities of sample 2-1

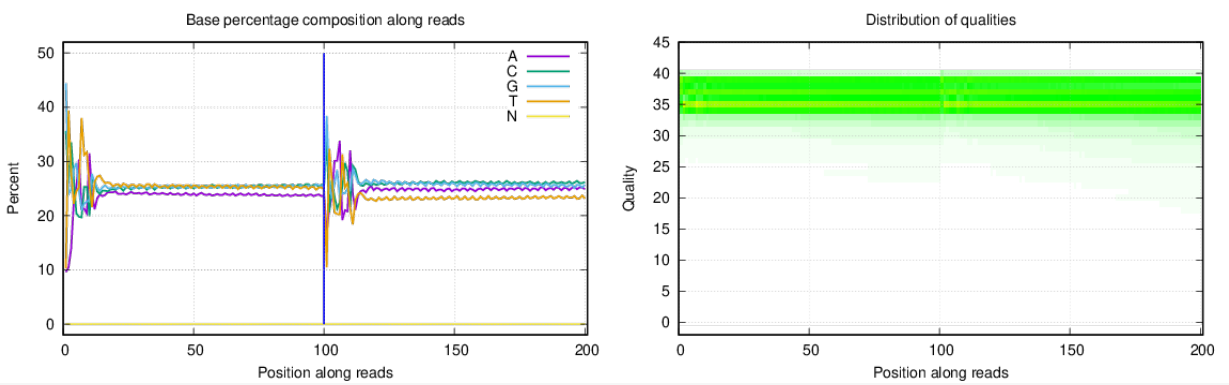

Figure 1-30 Distribution of base percentage and qualities of sample 1-2

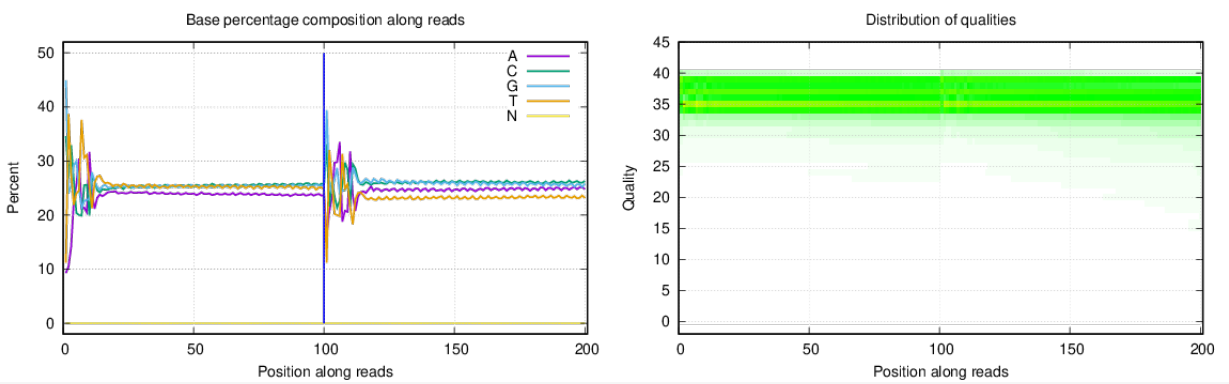

Figure 1-31 Distribution of base percentage and qualities of sample 1-1

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 a certain amount of total RNA samples, and use oligo dT beads to enrich mRNA with poly A tail; (2) mRNA molecules were fragmented into small pieces; (3) The fragmented mRNA was synthesized into first strand cDNA using random primers; (4) The second strand cDNA was synthesized with dUTP instead of dTTP; (5) The synthesized cDNA was subjected to end-repair and 3' adenylated. Adaptors were ligated to the ends of these 3' adenylated cDNA fragments; (6) Digest the U-labeled second-strand template with Uracil-DNA-Glycosylase (UDG) and perform 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 0.001 -l 20 -q 0.4 --adaMR 0.25 --polyX 50 --minReadLen 100", 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), remove the entire read;
2. Filter read length: if the length of the sequencing read is less than 100 bp, discard the entire read;
3. Remove N: if the N content in the sequencing read accounts for 0.1% or more of the entire read, discard the entire read;
4. Remove polyX: if the length of polyX (X can be A, T, G or C) in the sequencing read exceeds 50bp, discard the entire read;
5. 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;
6. 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	.
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