BGI Bioinformatics Report for SmallRNA Pure-sequencing Project

1 Data Production

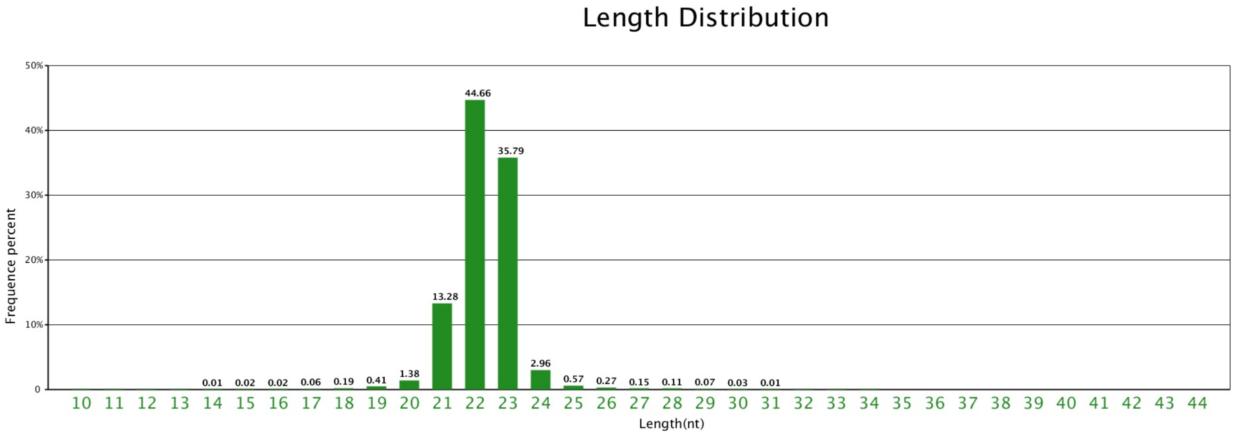
After sequencing ,the raw reads were filtered.Data filtering includes removing adaptor sequences, contamination and low-quality reads from raw reads.Next, we get the statistics of data production.Table 1-1 shows statistical results after data treatment.

Table 1-1 Reads statistics results

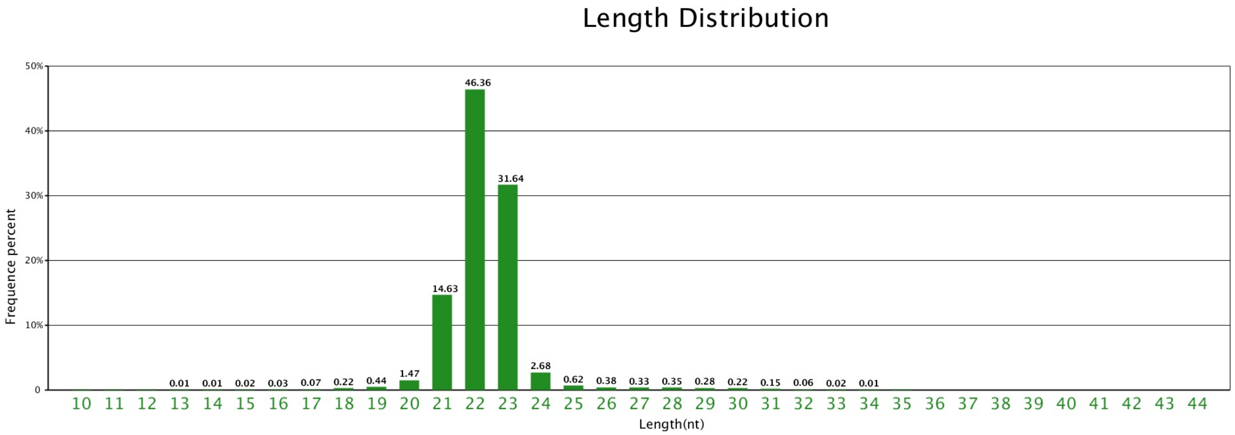
Sample Name	Clean Reads	Clean bases	Read length (bp)	Insert Size (bp)	Q20(%)	GC(%)
Sample 10	9,551,423	212,883,532	49	108.954	-	43.42%
Sample 2	9,749,880	217,461,388	49	113	-	43.58%
Sample 5A	9,491,692	210,151,099	49	108.518	-	43.39%
Sample 6A	8,494,120	188,579,007	49	113	-	43.46%
Sample 7A	8,746,016	194,421,428	49	112	-	43.33%
Sample 4A	10,434,197	230,603,170	49	108.676	-	43.80%
Sample 8	9,849,533	217,214,557	49	113	-	43.27%
Sample 3A	8,603,165	191,226,469	49	109.524	-	43.10%
Sample 9	10,644,560	236,661,838	49	112	-	43.03%
Sample 1	10,230,623	227,698,246	49	112	-	43.40%

2 Data Quality Control

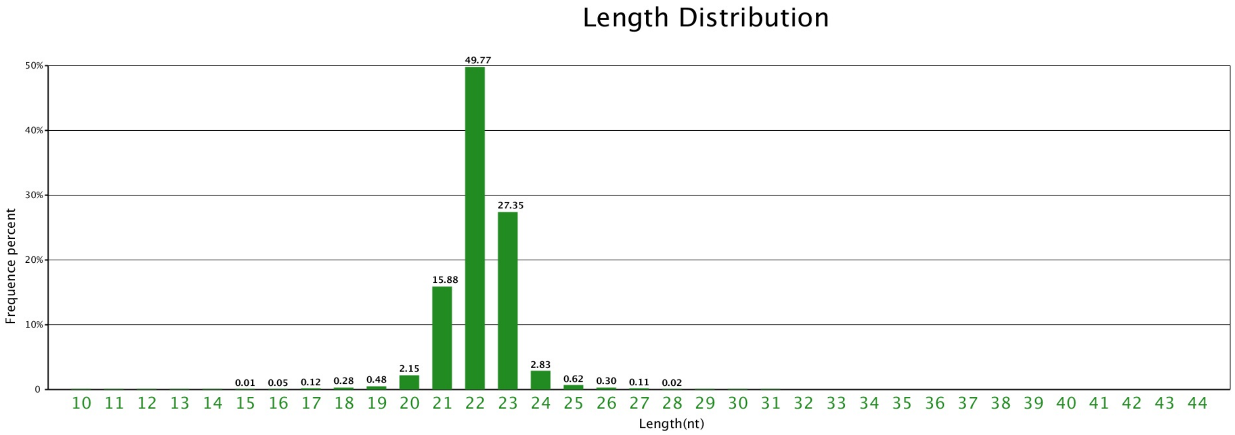
2.1 Length distribution



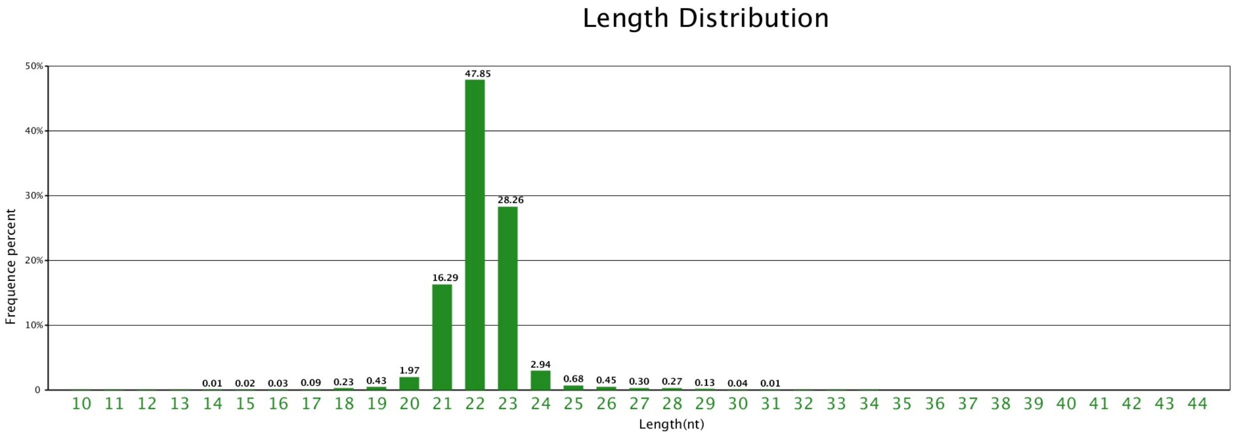
Length distribution of sample Sample 10



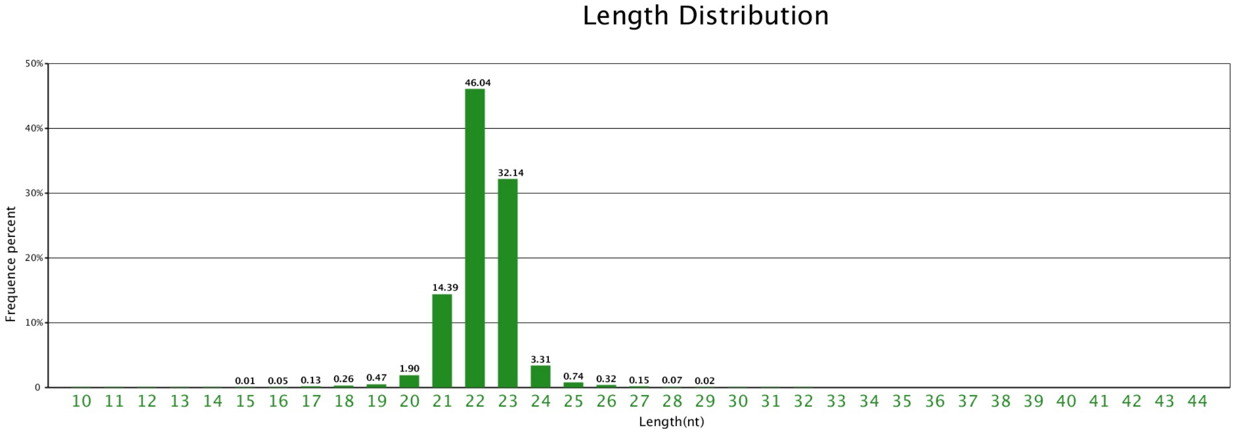
Length distribution of sample Sample 2



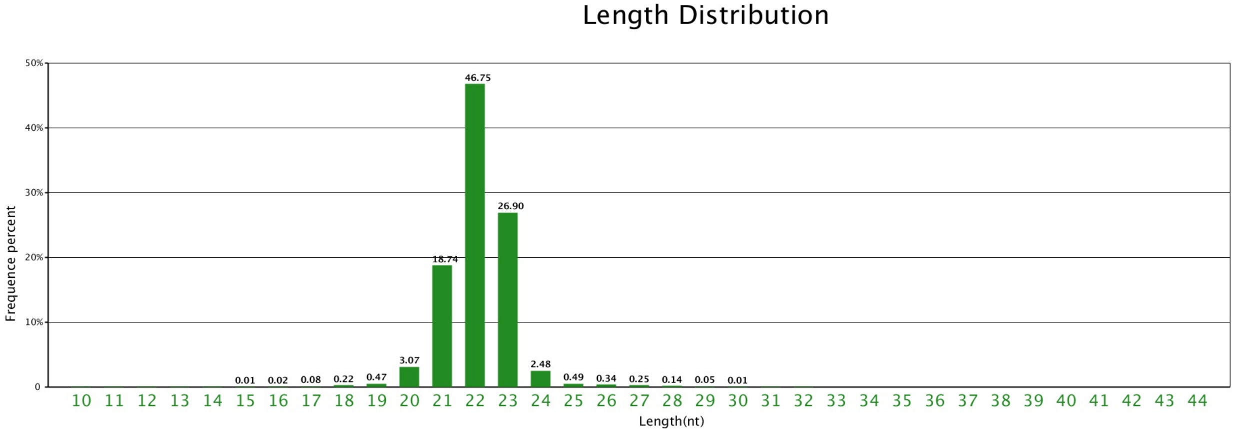
Length distribution of sample Sample 5A



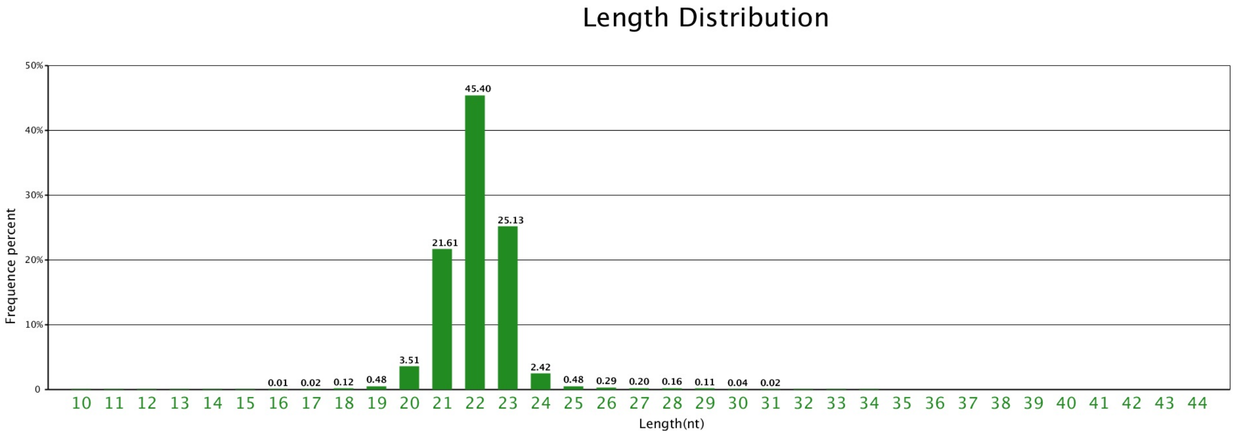
Length distribution of sample Sample 6A



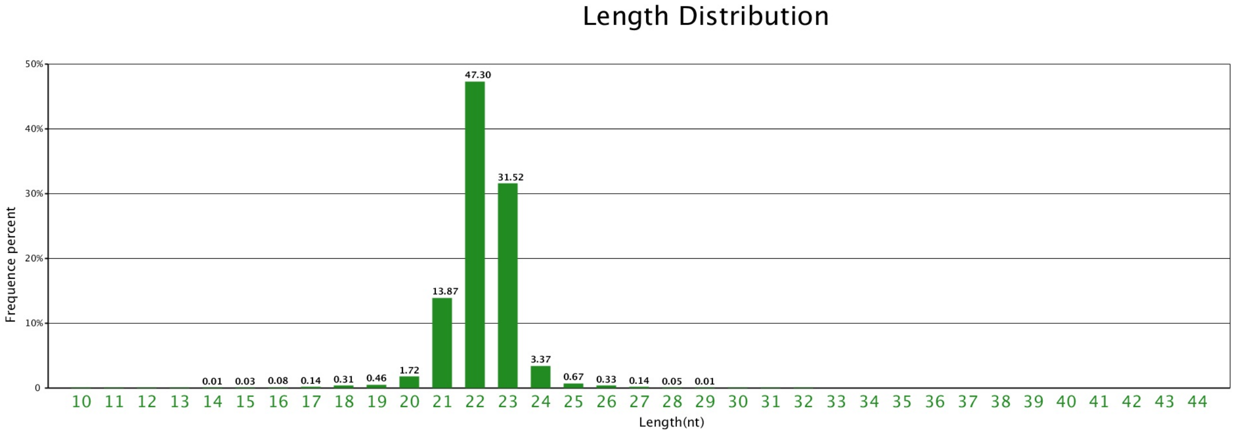
Length distribution of sample Sample 7A



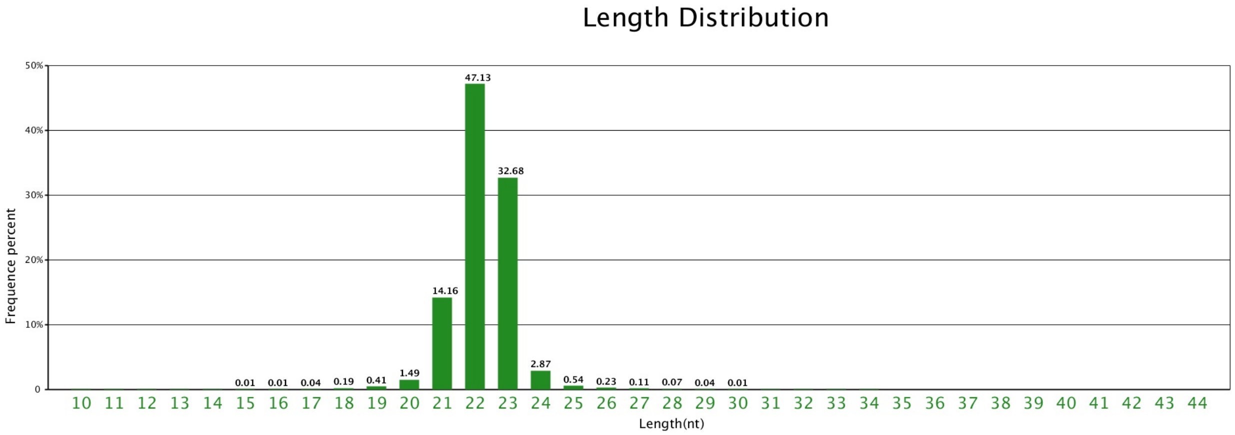
Length distribution of sample Sample 4A



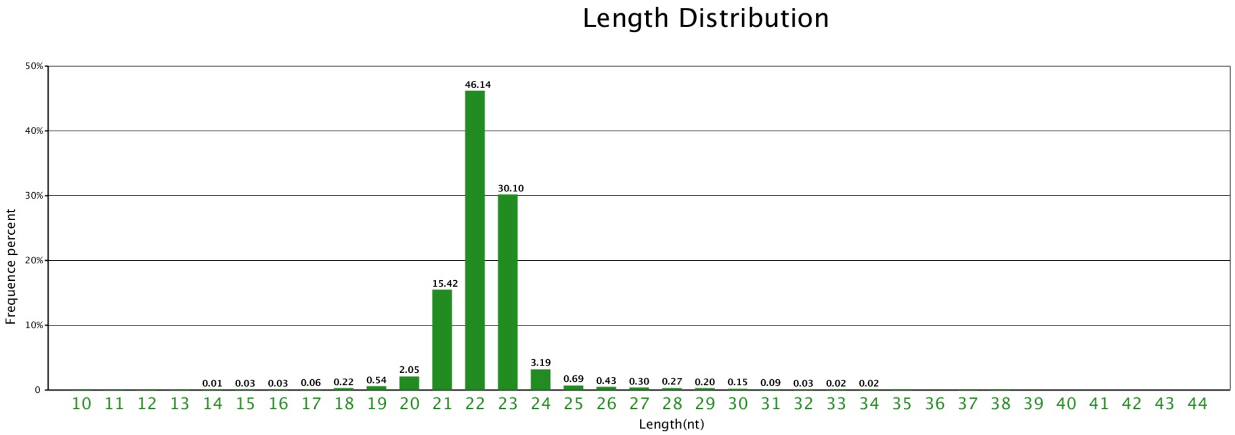
Length distribution of sample Sample 8



Length distribution of sample Sample 3A



Length distribution of sample Sample 9



Length distribution of sample Sample 1

3 HELP DOCUMENT

3.1 Experiment process

Small RNA is an special kind of molecules in organisms which induces the gene silence and plays an important role in the regulation of cell growth, gene transcription and translation. The small RNA digitalization analysis based on HiSeq high-throughput sequencing takes the SBS-sequencing by synthesis, which can decrease the loss of nucleotides caused by the secondary structure. It is also strong for its small requirement of sample quantity, high through-put, high accuracy with simply operated automatic platform. Such analysis can obtain millions of small RNA sequence tags in one shot, identify small RNA of certain species in certain condition comprehensively, predict novel miRNA and construct the small RNA differential expression profile between samples, which could be used as a powerful tool on small RNA function research. The experiment process1 of small RNA sequencing is shown in Figure3-1-1 and Figure 3-1-2:

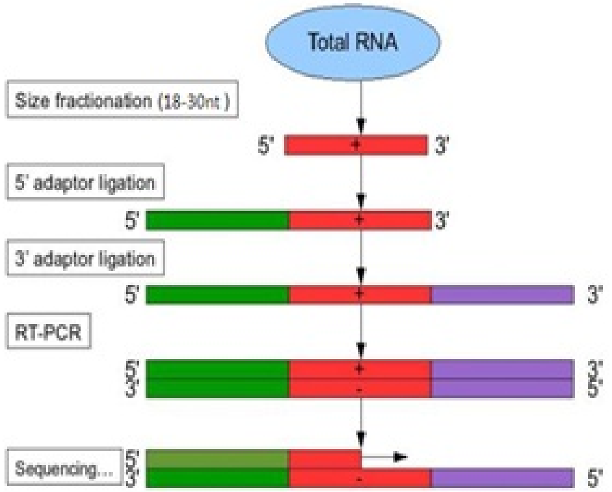


Figure 3-1-1 RNA sequencing experiment process for plant and blood serum

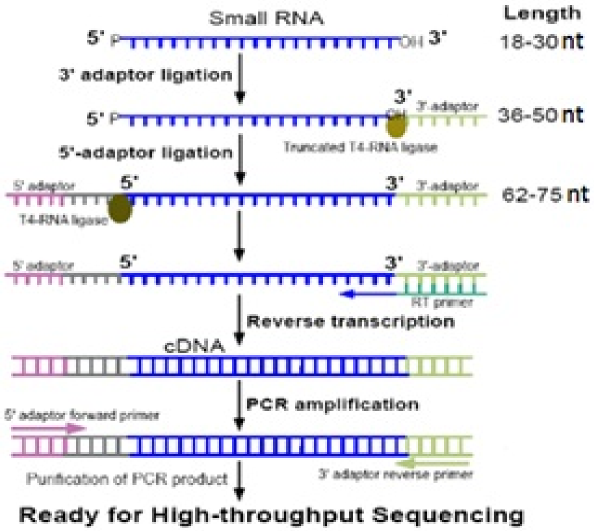


Figure 3-1-2 RNA sequencing experiment process of animals

3.2 Raw sequence data

The original image data is transferred into sequence data via base calling, which is defined as raw data or raw reads and saved as FASTQ file. Those FASTQ files are the original data provided for users, and they include the detailed read sequences and the read quality information. In each FASTQ file, every read is described by four lines, listed as follows:

```
@A80GVTABXX:4:1:2587:1979#ACAGTGAT/1
NTTTGATATGTGTGAGGACGTCTGCAGCGTCACCTTTATCGGCCATGGT
+
BTTMKZXUUUdddddddddddddddddddddddddddaddddddd^WYYU
```

The first and third lines are sequences names generated by the sequence analyzer; the second line is sequence; the fourth line is sequencing quality value, in which each letter corresponds to the base in line 2; the base's quality is equal to ASCII value of the character in line 4 minus 64, e.g. the ASCII value of c is 99, then its base quality value is 35. Starting from the Illumina GA Pipeline v1.5, the range of base quality values is from 2 to 41. Table 3-2 demonstrates the relationship between Illumina HiSeq^{TM} 2000 sequencing error rate and the sequencing quality value. Specifically, if the sequencing error rate is denoted as E, Illunima HiSeq^{TM} 2000 base quality value is denoted as sQ, the relationship is as follows:

$$sQ = -10\log_{10} E$$

Table 3-2 Relationship between Illumina HiSeq^{TM} 2000 error rate and sequencing quality

Sequencing error rate	Sequencing quality value	Character
5%	13	M
1%	20	T
0.1%	30	^

3.3 Column description for data production

Table 3-3 Column description for data production

Header	Description
Sample Name	Sample name that identify each sample
Insert Size (bp)	The length of sequencing fragment (bp)
Read length (bp)	The length of the total reads (bp)
Clean Reads	Total clean reads number
Clean bases	Total clean bases number
Q20 Percentage	The number of nucleotide with quality higher than 20/nucleotide(clean read1,read2)
GC Percentage	GC number / nucleotide(clean read1,read2)