BGI Sequencing Data Report

2023/10/27

Table of Contents

1 Project Information

Project code: F23A430000924_RIPzryuR
Sample number: 17

2 Data Statistics

Raw reads produced from sequencer contain adapters, unknown or low quality bases. The statistics of raw data is shown below.

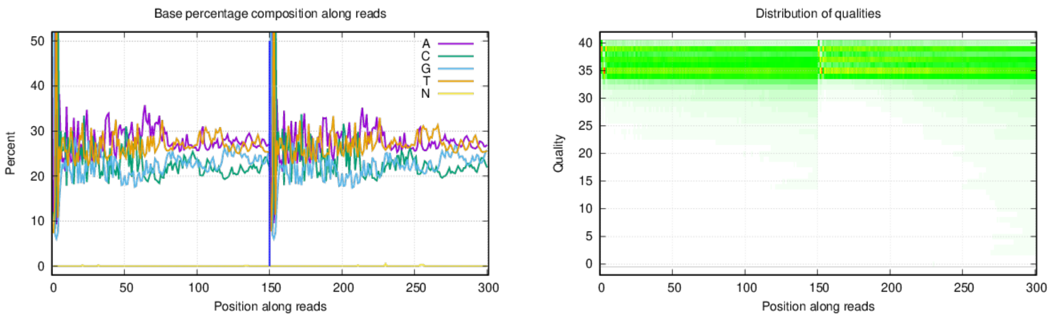
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
718	150;150	96.85;96.28	90.55;91.08	45.32;45.34	7,630,295	2,289,088,500
718_a	150;150	96.80;95.77	90.32;89.42	44.10;44.12	4,617,634	1,385,290,200
Ch	150;150	97.28;95.74	91.62;89.37	43.42;43.48	9,926,576	2,977,972,800
Ch_a	150;150	97.21;95.66	91.42;89.01	43.23;43.27	10,893,890	3,268,167,000
ar1	150;150	95.75;97.37	87.97;94.60	54.14;54.14	523,748	157,124,400
ar10	150;150	98.95;97.47	96.68;94.76	53.77;53.93	538,970	161,691,000
ar11	150;150	98.33;97.50	94.80;94.83	53.94;53.91	516,931	155,079,300
ar12	150;150	98.83;97.50	96.31;94.81	53.76;53.93	544,290	163,287,000
ar2	150;150	97.20;97.50	91.71;94.86	54.05;54.12	506,978	152,093,400
ar3	150;150	98.41;97.48	95.10;94.77	54.00;54.09	600,003	180,000,900
ar4	150;150	96.81;97.33	90.74;94.44	53.96;53.91	580,981	174,294,300
ar5	150;150	98.15;97.47	94.32;94.72	53.89;53.89	565,970	169,791,000
ar6	150;150	99.02;97.59	96.88;95.03	54.00;53.96	564,591	169,377,300
ar7	150;150	98.35;97.55	94.89;95.00	53.80;54.07	515,841	154,752,300
ar8	150;150	96.18;97.61	89.05;95.14	54.07;54.02	521,593	156,477,900
ar9	150;150	97.29;97.57	92.00;95.04	53.88;54.07	593,355	178,006,500
ip8M	150;150	97.22;95.59	91.44;88.72	44.61;44.57	406,379,160	121,913,748,000

Table Format:
1. Sample: The name of sample
2. Length: The Length of reads
3. Q20 (%): The proportion of nucleotides with quality value larger than 20
4. Q30 (%): The proportion of nucleotides with quality value larger than 30
4. GC Content(%): The proportion of bases G and C
5. Total Reads: The total number of raw read pairs
6. Total Bases: The total nucleotides number of raw reads

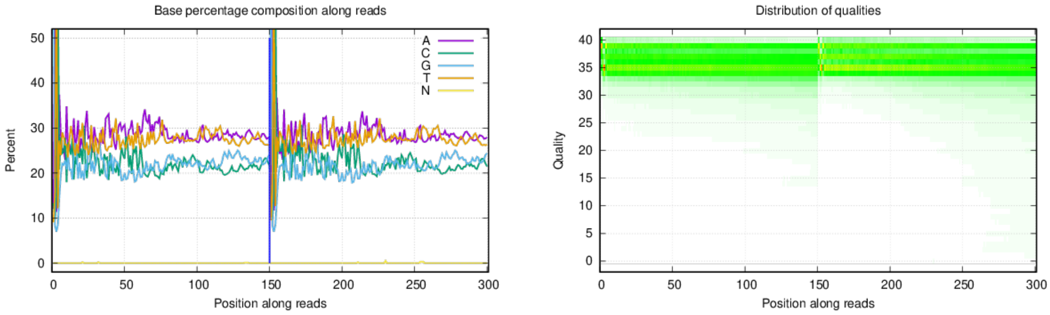
3 Data Quality Control

The distribution of base percentage and qualities along reads in data filtering are shown as following(If a sample has multiple lanes, only one of them will be displayed). The left picture is base percentage distribution along reads the sample, the right picture is distribution of qualities along reads of the sample.

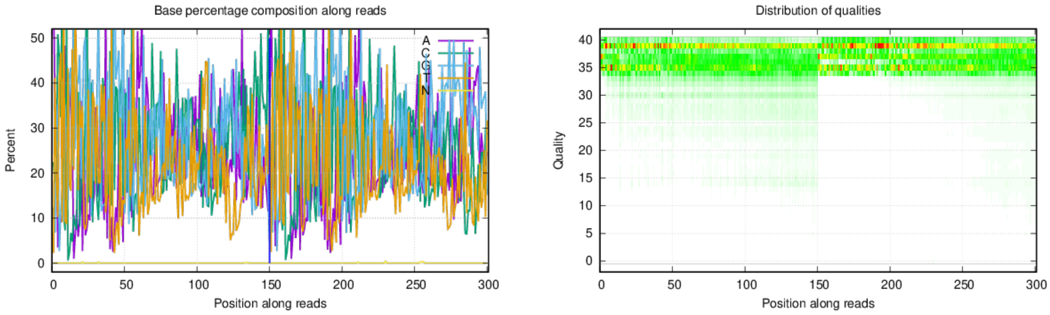
Quality control of sample 718



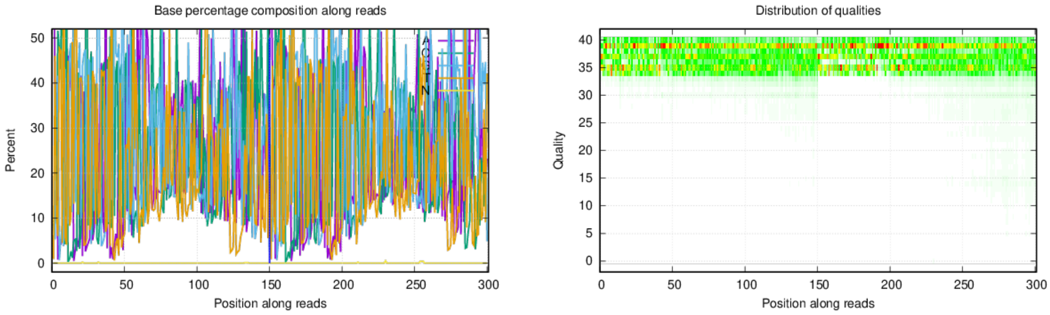
Quality control of sample 718_a



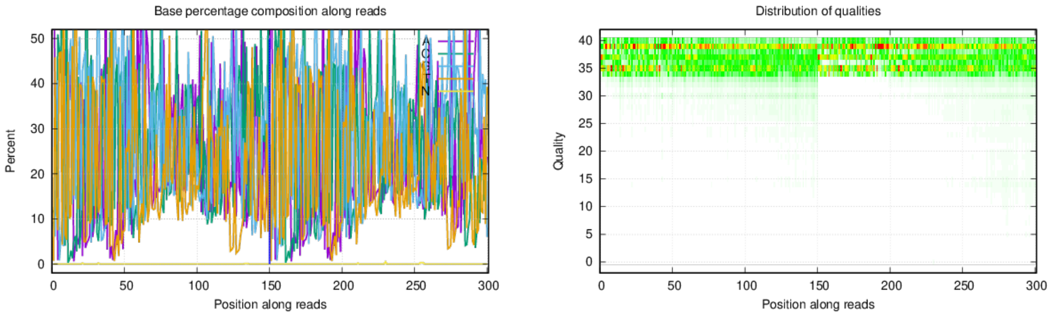
Quality control of sample ar1



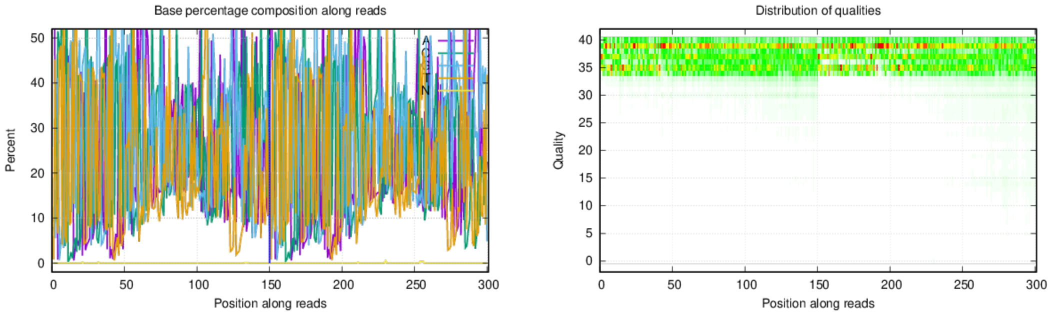
Quality control of sample ar10



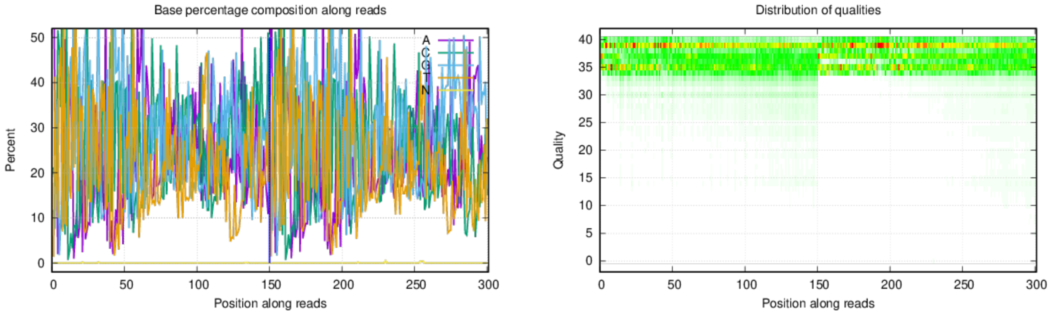
Quality control of sample ar11



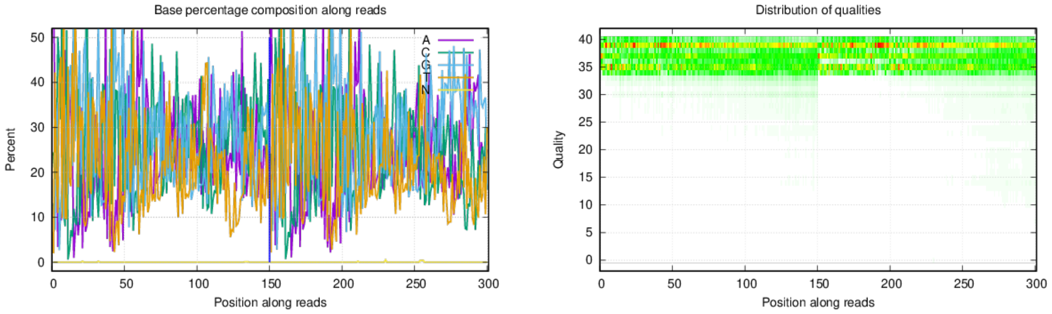
Quality control of sample ar12



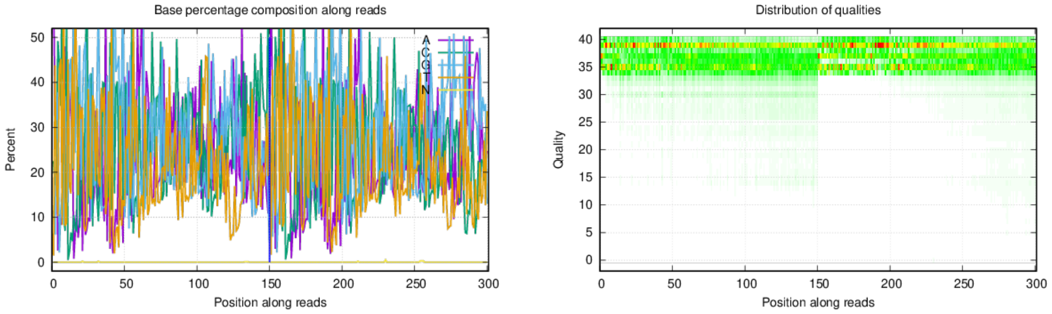
Quality control of sample ar2



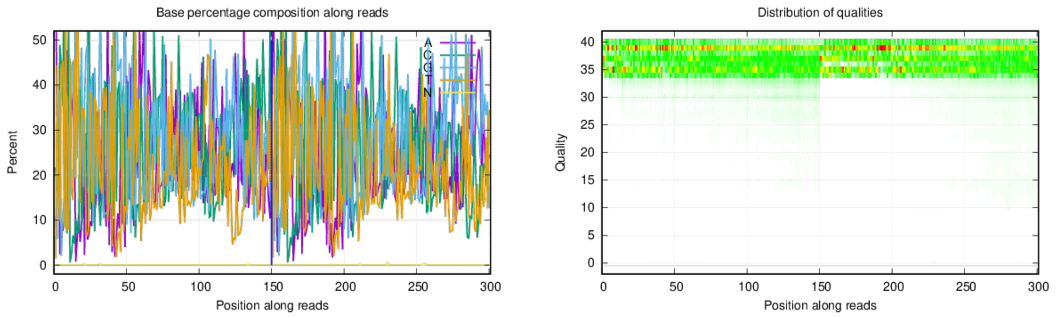
Quality control of sample ar3



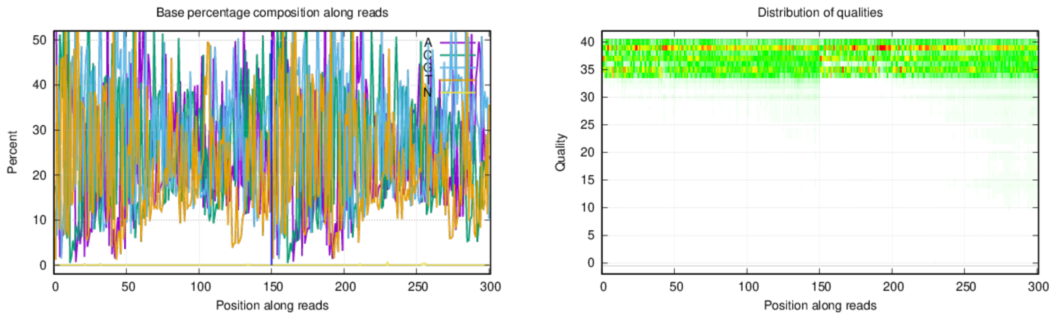
Quality control of sample ar4



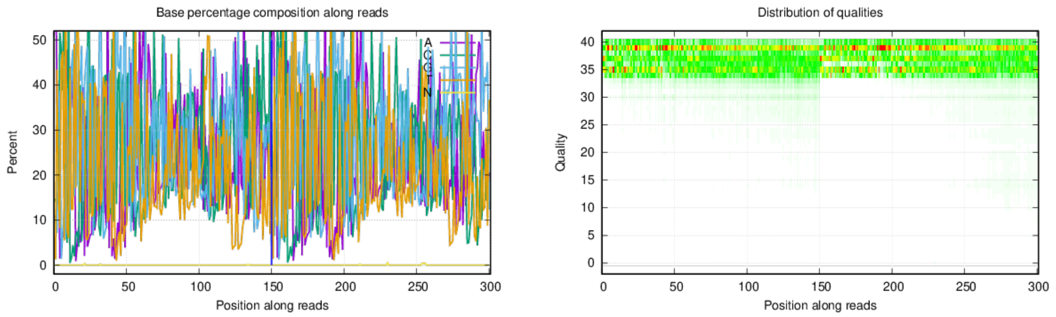
Quality control of sample ar5



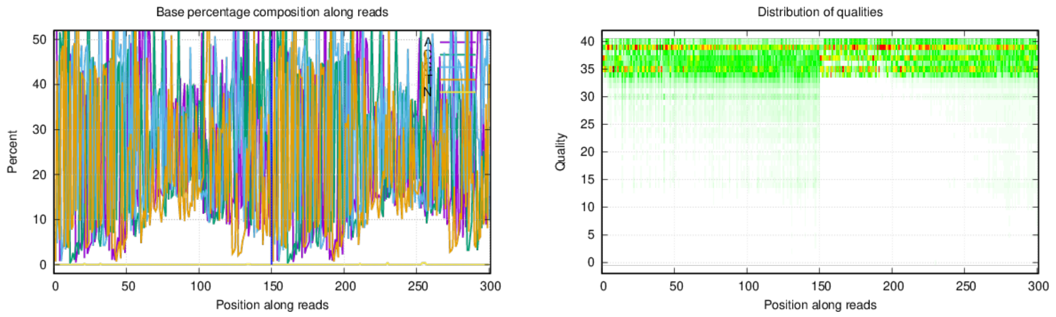
Quality control of sample ar6



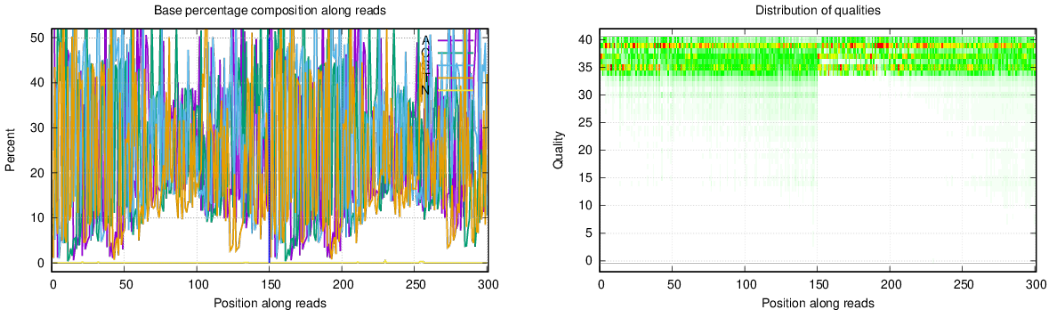
Quality control of sample ar7



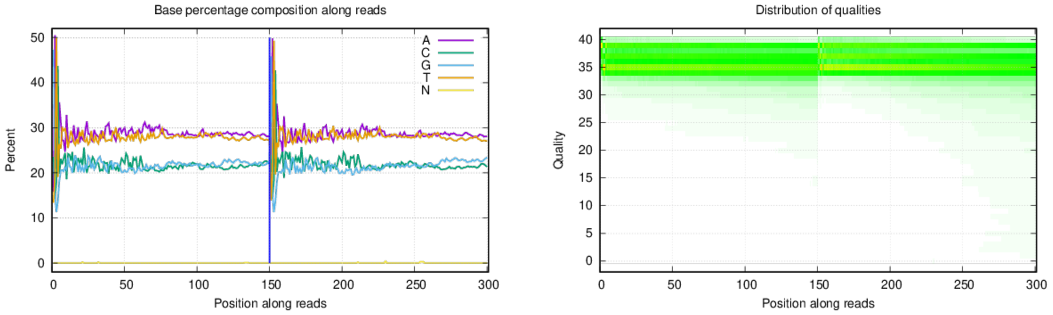
Quality control of sample ar8



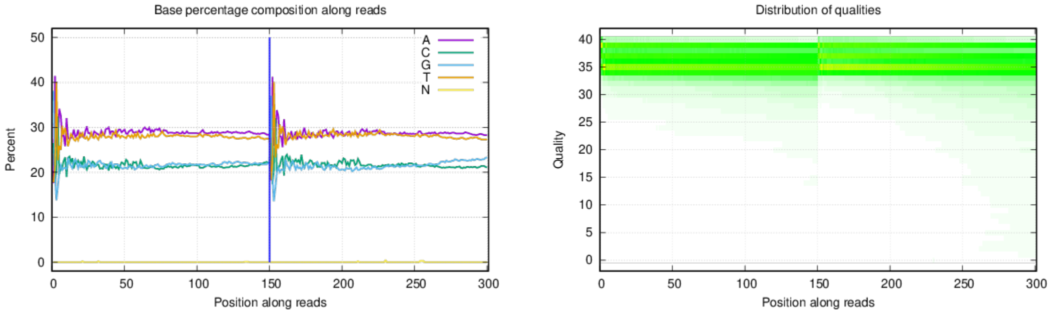
Quality control of sample ar9



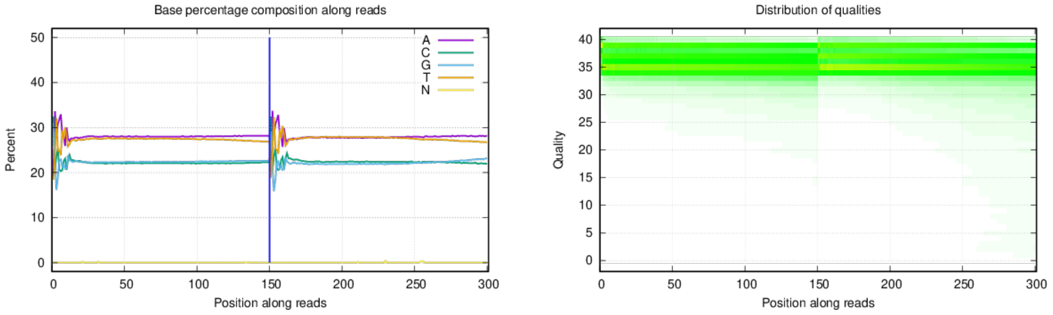
Quality control of sample Ch



Quality control of sample Ch_a



Quality control of sample ip8M



4 Help Document

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. Each entry in a FASTQ files consists of 4 lines:

1. A sequence identifier with information about the sequencing run and the cluster. The exact contents of this line vary by based on the BCL to FASTQ conversion software used.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

Here is an example of a single entry in a FASTQ file:

```
@V300029029L1C001R0010000210/1
GCGACCCCAGGTCAGTCGGGACTACCCGCTGAAGTCGGAGGCCAAGCGGT
+
FFFCFFFFFFFFFDFEFFFFEFEF0FFFFEFFFFFFFEFFFFFECGFFFF
```

The relationship between DNBseq sequencer sequencing error rate and the sequencing quality value is shown in the following formula. Specifically, if the sequencing error rate is denoted as "E", DNBseq sequencer base quality value is denoted as "sQ", the relationship is as follows:

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

Sequencing error rate	Sequencing quality value	Character of Phred +33 quality system
5%	13	.
1%	20	5
0.1%	30	?