BGI Sequencing Data Report

2023/5/16

Table of Contents

1 Project Information

Project code: F20FTSEUHT0946-02_MUSbeusR_poolK
Sample number: 27

2 Data Statistics

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

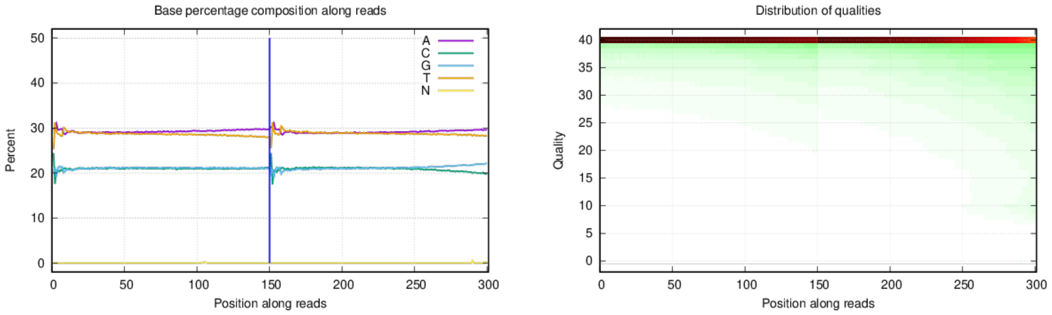
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
1	150;150	98.20;95.73	94.30;88.96	42.09;42.03	2,987,652	896,295,600
2	150;150	98.40;97.26	94.80;92.61	34.26;34.27	136,139,309	40,841,792,700
3	150;150	98.18;96.55	93.97;90.57	34.12;34.13	138,318,931	41,495,679,300
4	150;150	98.43;96.68	94.79;91.15	35.65;35.62	111,415,283	33,424,584,900
Af2-1	150;150	98.24;95.93	95.52;91.32	53.68;53.31	3,889,925	1,166,977,500
Af2-2	150;150	98.31;94.64	95.67;89.16	55.07;54.48	5,654,320	1,696,296,000
Ai4-1	150;150	98.71;95.46	96.43;89.68	60.72;61.33	6,708	2,012,400
BerL-3_2	150;150	98.61;95.59	96.29;89.96	59.03;59.31	13,814	4,144,200
BerL-4	150;150	98.98;96.41	97.31;91.79	57.40;57.54	6,682	2,004,600
Bol1-1	150;150	99.32;97.29	98.01;93.81	49.72;49.95	102,964	30,889,200
Chek1-1	150;150	98.65;95.46	96.42;89.80	60.57;61.18	10,357	3,107,100
Gus1-1	150;150	96.22;90.89	89.44;81.08	56.86;54.75	24	7,200
Gus1-2	150;150	99.03;96.29	97.50;91.47	57.86;56.67	38,490	11,547,000
Gus1-3	150;150	98.90;95.81	97.01;90.59	57.29;57.76	39,748	11,924,400
Gus1-4	150;150	99.00;96.68	97.25;92.25	58.21;58.98	105,524	31,657,200
Gus1-5	150;150	98.89;95.49	96.89;89.98	58.64;58.65	83,466	25,039,800
Gus1-6	150;150	98.35;94.95	95.64;88.63	59.81;59.93	14,067	4,220,100
Im-1	150;150	98.80;95.70	96.91;90.61	56.59;55.90	9,439,442	2,831,832,600
Kam1-1	150;150	99.16;95.97	97.72;90.89	58.41;58.87	2,997	899,100
Ku-1	150;150	98.85;95.97	96.82;90.71	59.76;60.71	73,884	22,165,200
M-1	150;150	96.78;93.62	91.73;85.97	36.45;36.64	67	20,100
Raz1-1	150;150	99.11;96.41	97.58;91.60	61.33;60.78	9,908	2,972,400
Rub-1	150;150	99.47;97.92	98.64;94.74	57.33;57.11	23,069	6,920,700
Rub-2	150;150	99.53;95.80	98.11;90.95	57.26;58.06	130	39,000
Shu-2	150;150	98.22;93.34	95.53;87.23	54.45;52.82	11,505,459	3,451,637,700
Tas-1	150;150	98.77;95.76	96.59;90.28	61.11;60.46	84,466	25,339,800
Ust1-1	150;150	99.23;97.73	97.69;94.66	45.23;45.74	119,098	35,729,400

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

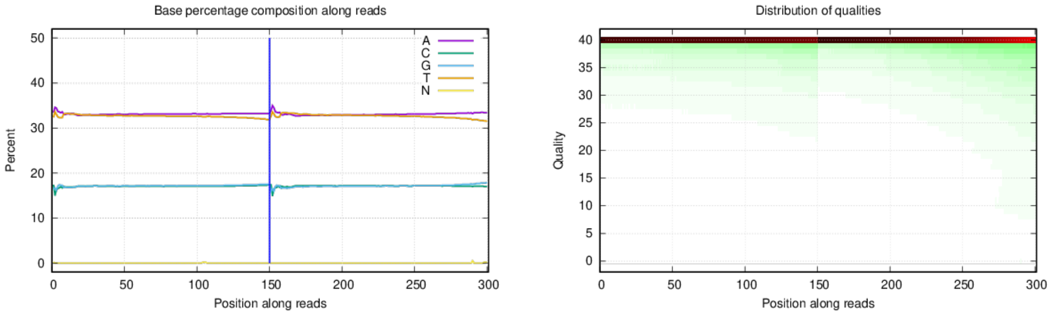
3 Data Quality Control

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.

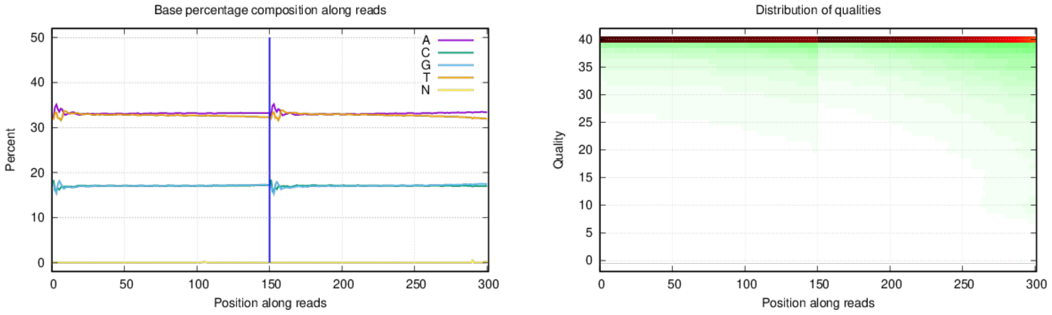
Quality control of sample 1



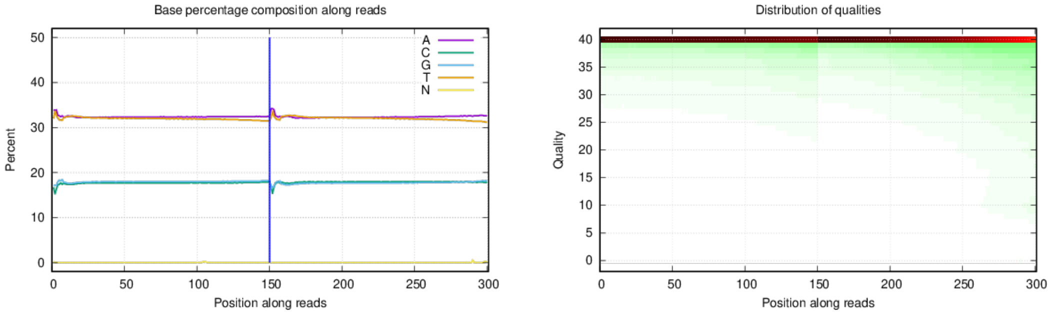
Quality control of sample 2



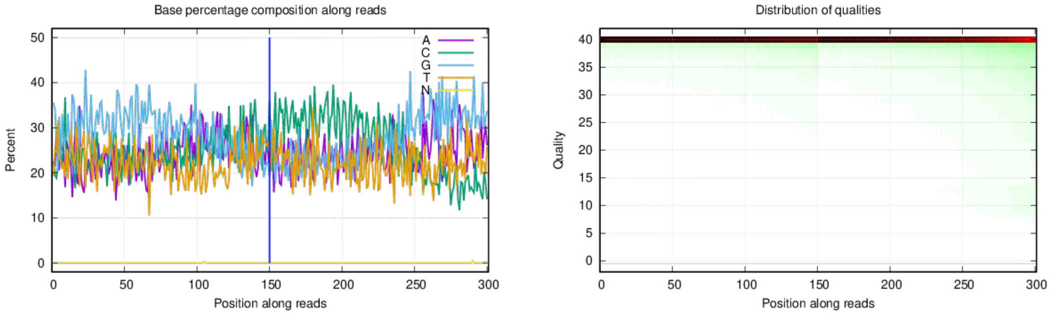
Quality control of sample 3



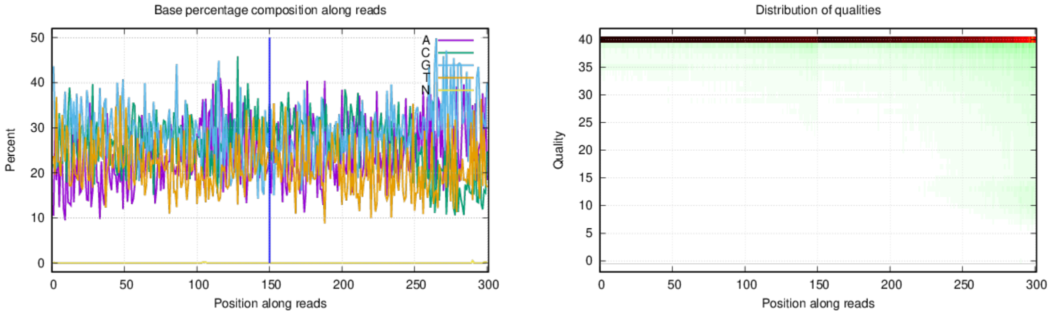
Quality control of sample 4



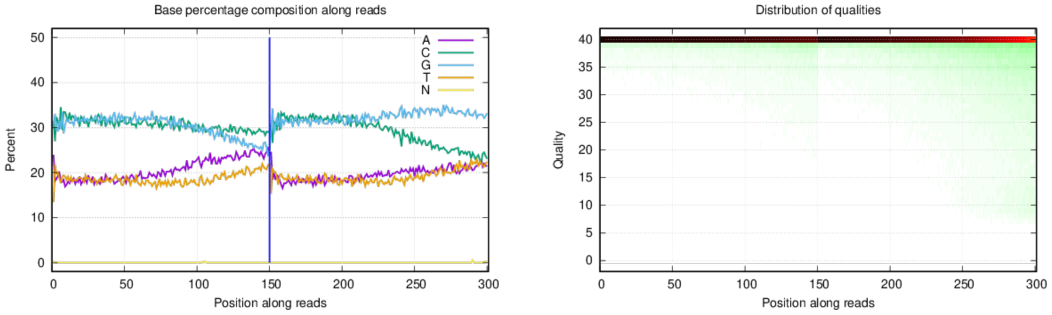
Quality control of sample Af2-1



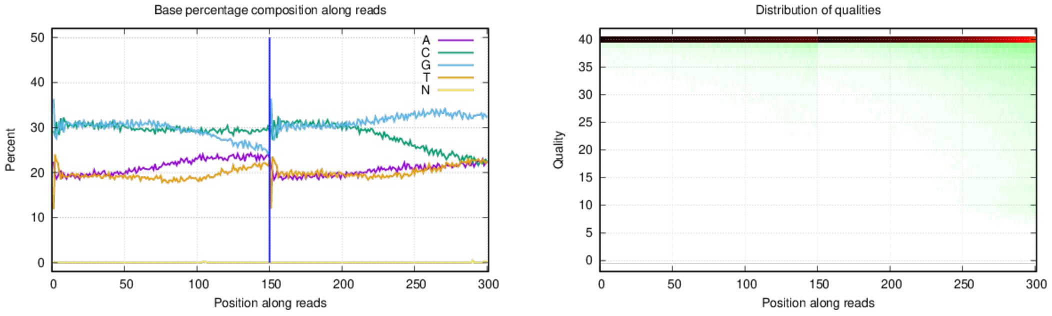
Quality control of sample Af2-2



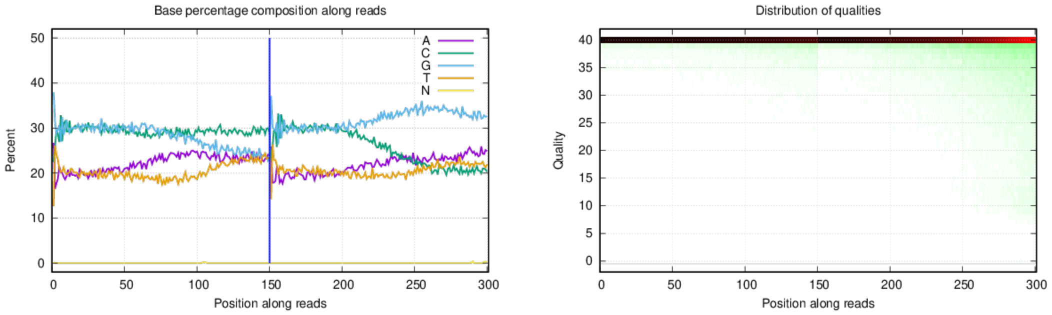
Quality control of sample Ai4-1



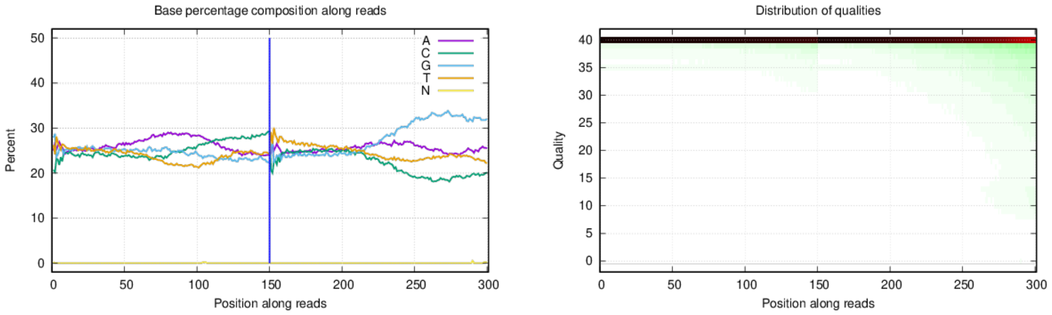
Quality control of sample BerL-3_2



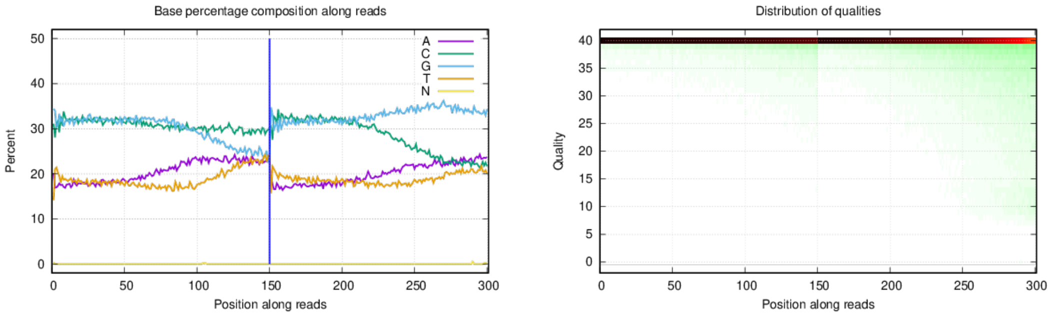
Quality control of sample BerL-4



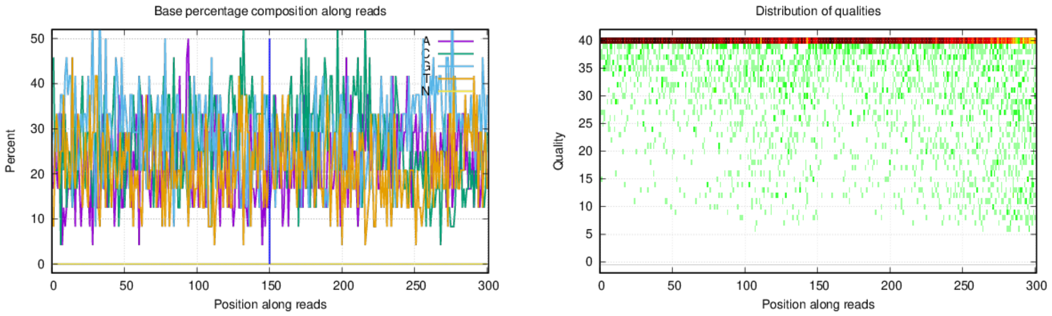
Quality control of sample Bol1-1



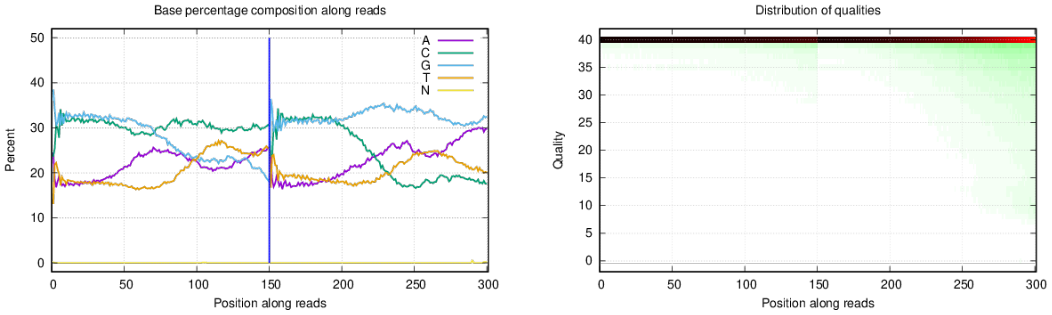
Quality control of sample Chek1-1



Quality control of sample Gus1-1

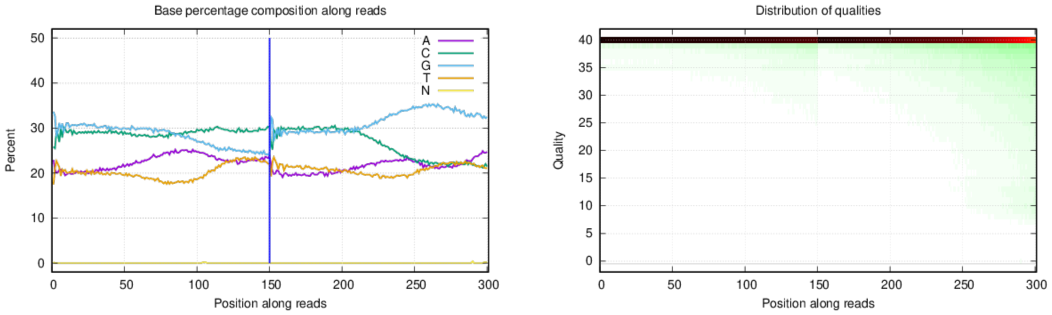


Quality control of sample Gus1-2


Base percentage composition along reads

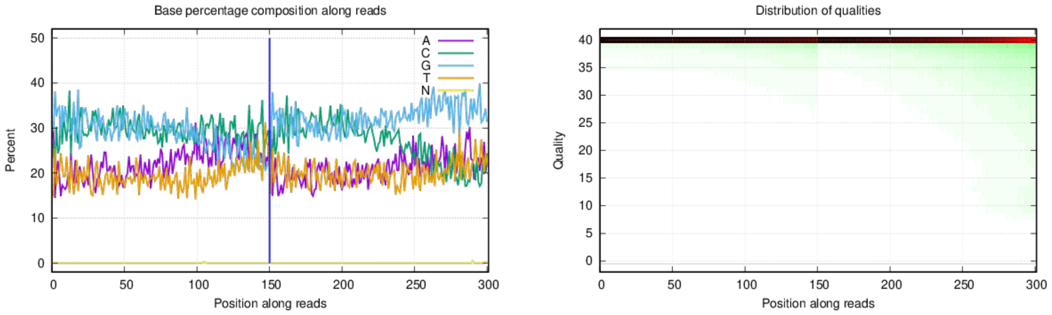
Distribution of qualities

Quality control of sample Gus1-3


Base percentage composition along reads

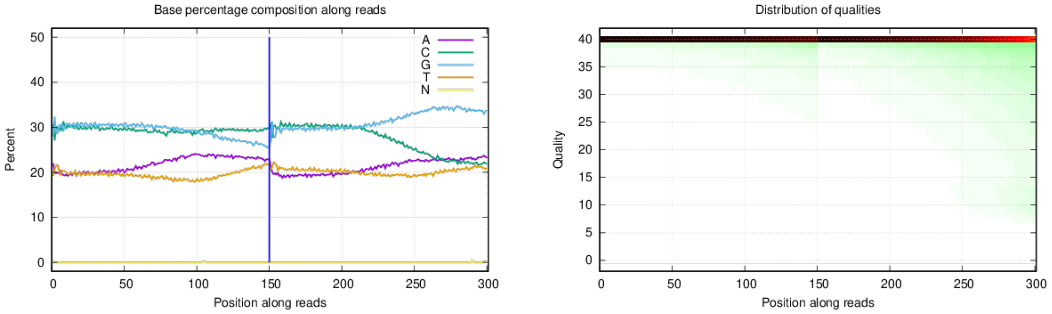
Distribution of qualities

Quality control of sample Gus1-4


Base percentage composition along reads

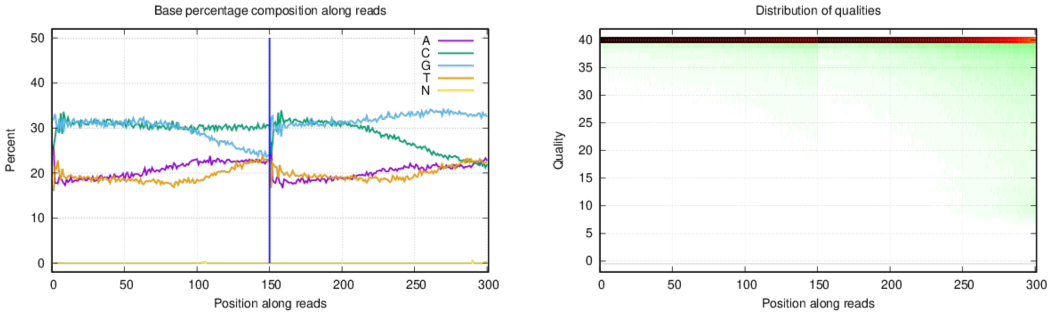
Distribution of qualities

Quality control of sample Gus1-5

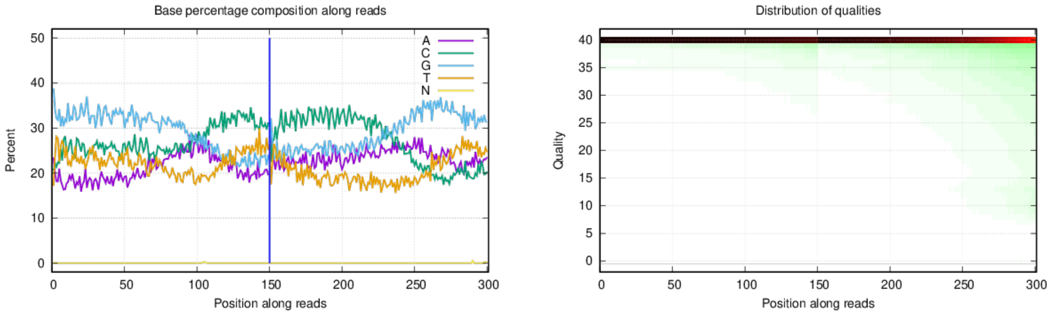

Base percentage composition along reads

Distribution of qualities

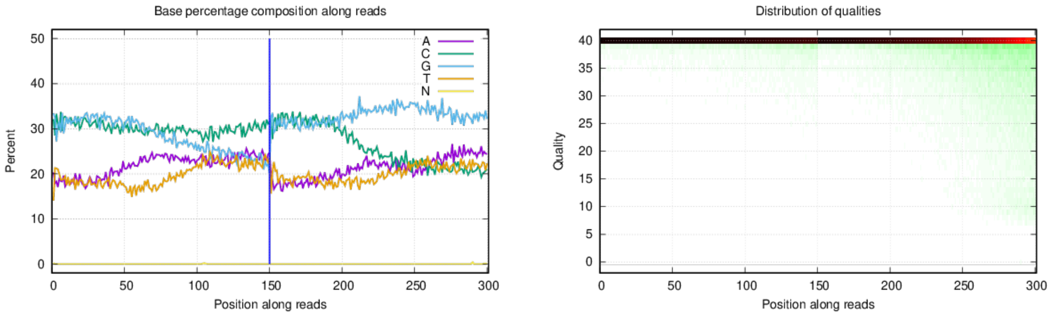
Quality control of sample Gus1-6



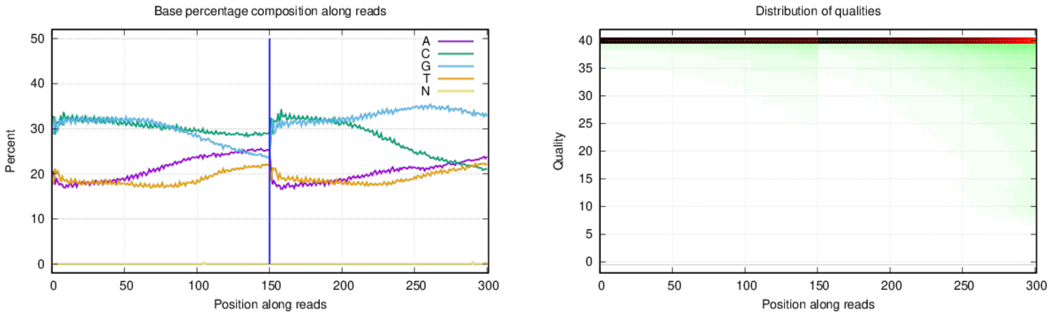
Quality control of sample Im-1



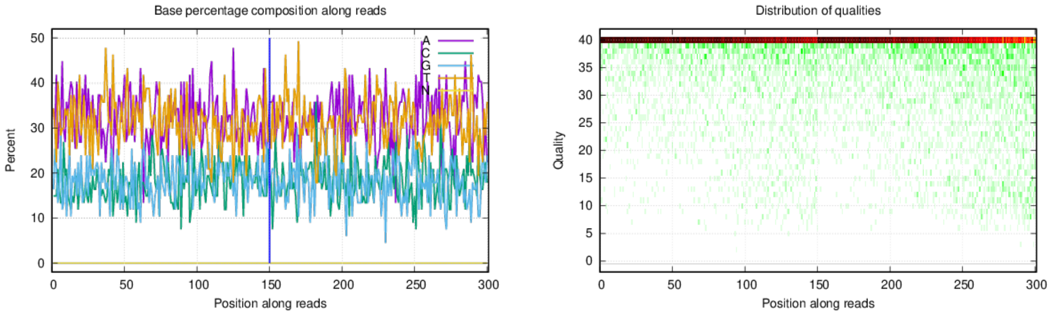
Quality control of sample Kam1-1



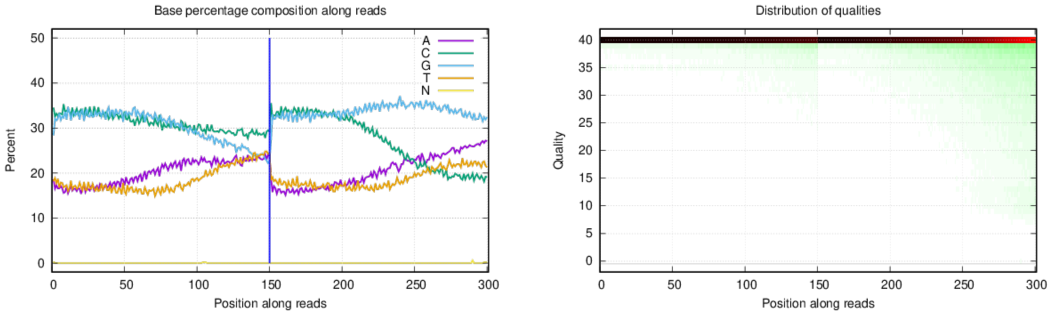
Quality control of sample Ku-1



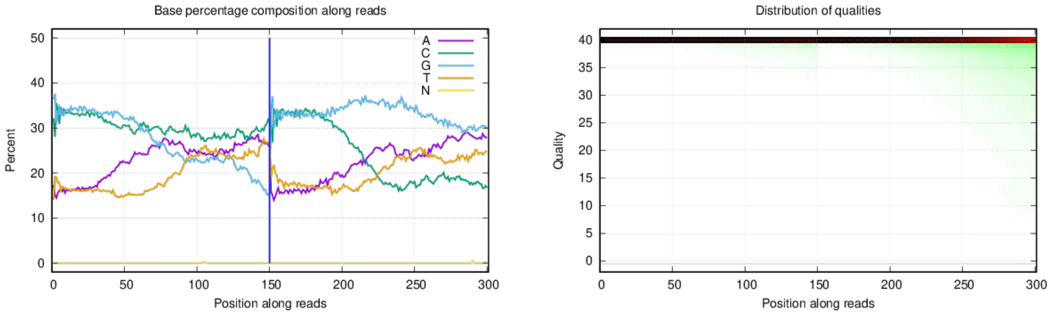
Quality control of sample M-1



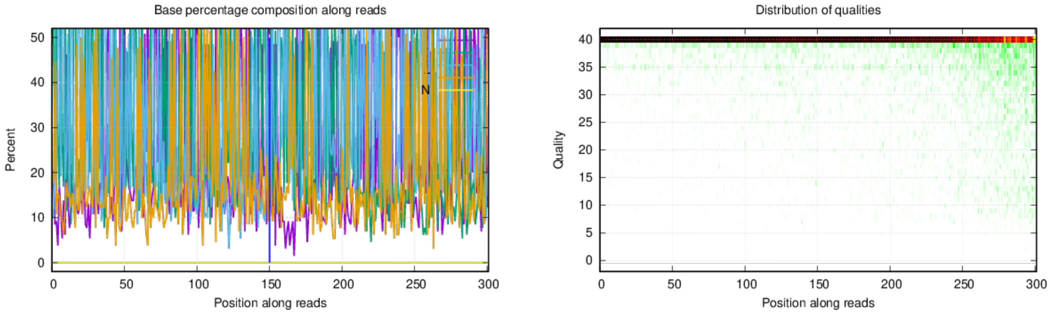
Quality control of sample Raz1-1



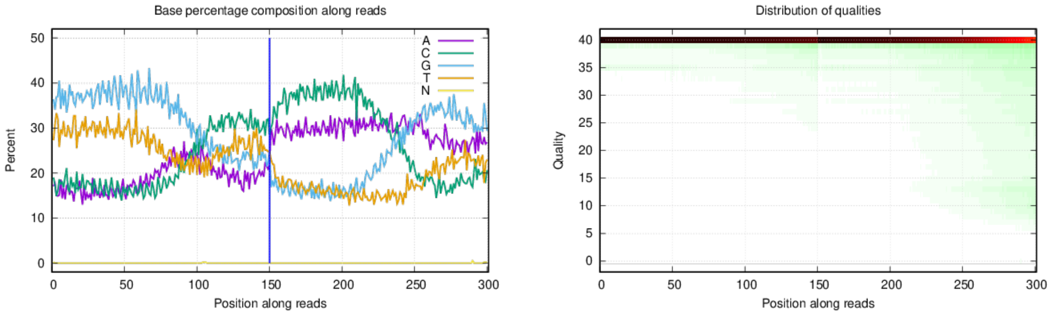
Quality control of sample Rub-1



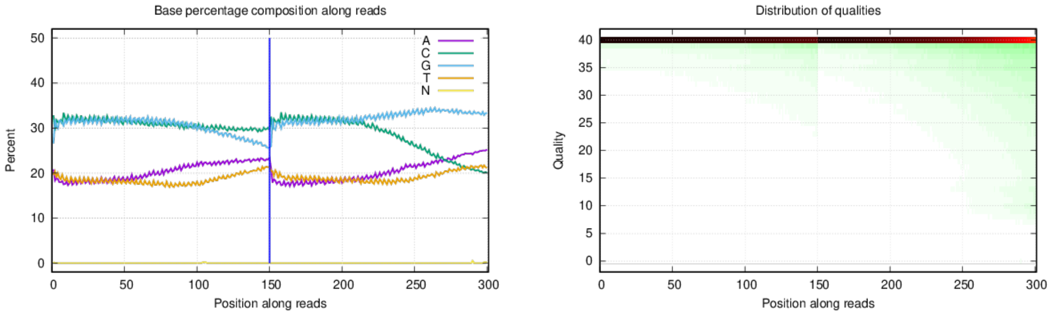
Quality control of sample Rub-2



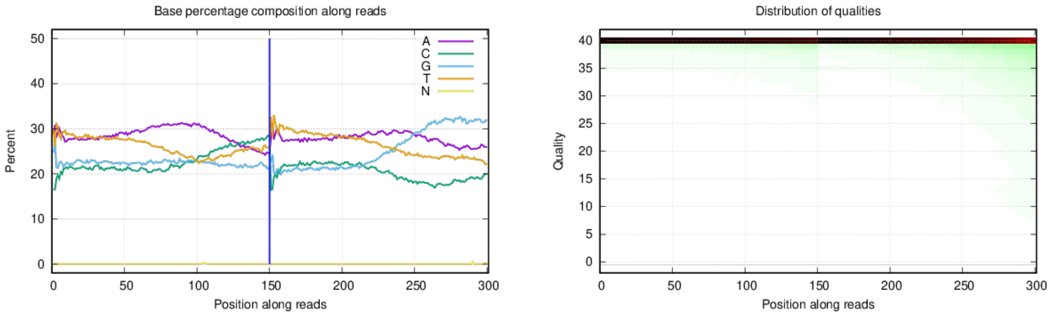
Quality control of sample Shu-2



Quality control of sample Tas-1



Quality control of sample Ust1-1



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
+
FFFCFFFFFFFFFFDFEFFFFFEFEF0FFFFEFFFFFFFFEFFFFFECGFFFF
```

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	?