BGI Sequencing Data Report

2023/5/15

Table of Contents

1 Project Information

Project code: F20FTSEUHT0946-01_MUSvwagR_poolA
Sample number: 38

2 Data Statistics

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

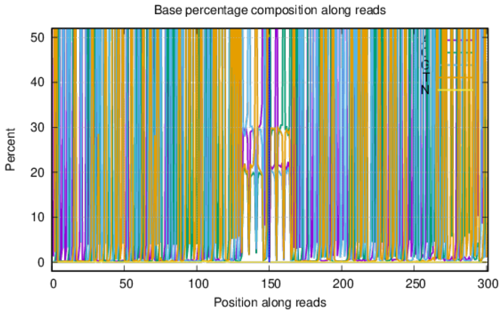
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
LS2N	150;150	95.72;93.99	89.61;84.32	52.03;51.68	14,331	4,299,300
Lop	150;150	98.67;96.58	95.78;91.49	49.59;49.58	40,340,896	12,102,268,800
MoPh11	150;150	97.65;94.24	92.75;85.73	48.61;48.64	34,505,326	10,351,597,800
MoPh14	150;150	97.42;94.09	91.96;85.34	45.57;45.67	32,094,214	9,628,264,200
MoPh7	150;150	97.86;94.79	93.22;86.76	46.50;46.48	28,003,802	8,401,140,600
P125	150;150	98.29;95.87	94.49;89.43	48.65;48.60	12,110,004	3,633,001,200
P126	150;150	98.15;95.73	94.13;89.12	50.00;49.93	74,996,870	22,499,061,000
P131	150;150	98.17;96.00	94.11;89.71	46.91;46.88	10,950,971	3,285,291,300
P132	150;150	98.34;95.84	94.62;89.36	48.09;48.06	13,107,149	3,932,144,700
P133	150;150	98.37;96.08	94.73;89.91	48.07;48.03	9,363,041	2,808,912,300
P135	150;150	98.30;96.43	94.55;90.71	47.31;47.26	10,354,049	3,106,214,700
P136	150;150	98.45;96.65	94.91;91.20	45.88;45.88	9,080,400	2,724,120,000
P137	150;150	98.33;96.05	94.63;89.88	48.18;48.14	16,221,891	4,866,567,300
P139	150;150	98.34;95.98	94.61;89.64	47.91;47.88	12,853,247	3,855,974,100
PICO1_1_cells	150;150	99.67;90.67	97.33;77.33	32.00;31.33	2	600
PICO21	150;150	97.60;91.67	94.13;84.53	29.27;28.13	10	3,000
e1	150;150	97.86;96.92	93.90;92.17	58.79;58.78	6,534	1,960,200
e2	150;150	97.85;96.69	93.87;91.70	58.84;58.81	11,013	3,303,900
e3	150;150	97.82;96.81	93.78;91.95	58.50;58.48	4,849	1,454,700
e5	150;150	97.88;96.84	93.93;92.11	58.74;58.75	7,165	2,149,500
e6	150;150	98.04;97.08	94.41;92.61	58.77;58.78	11,719	3,515,700
mix10	150;150	98.28;95.29	94.47;88.08	49.70;49.65	6,991,515	2,097,454,500
mix25	150;150	98.33;95.49	94.61;88.56	49.90;49.84	6,535,298	1,960,589,400
mix50	150;150	98.20;95.52	94.22;88.56	49.39;49.34	4,895,171	1,468,551,300
mix75	150;150	98.16;95.69	94.07;88.93	48.17;48.16	8,352,114	2,505,634,200
sum_rear1	150;150	96.62;95.32	90.03;87.16	33.61;33.31	1,548,822	464,646,600
sum_rear2	150;150	96.97;96.96	90.82;91.13	34.35;34.15	1,660,111	498,033,300
sum_rear3	150;150	97.34;96.33	91.64;89.78	35.73;34.26	1,377,662	413,298,600
sum_wt1	150;150	97.35;96.65	92.03;91.04	36.34;36.86	1,877,539	563,261,700
sum_wt2	150;150	97.24;96.13	91.36;89.10	35.49;36.60	1,611,923	483,576,900
u1	150;150	97.90;96.42	93.76;90.97	58.85;58.81	15,090	4,527,000
u2	150;150	97.92;96.40	93.83;90.94	58.91;58.90	17,189	5,156,700
u3	150;150	97.99;96.57	94.02;91.29	58.77;58.74	14,392	4,317,600
u4	150;150	98.04;96.80	94.20;91.88	58.96;58.95	6,691	2,007,300
u5	150;150	98.23;96.68	94.72;91.67	58.95;58.94	10,910	3,273,000
u6	150;150	98.19;96.20	94.55;90.55	58.92;58.89	17,143	5,142,900
u7	150;150	98.06;96.44	94.20;91.05	58.89;58.88	14,434	4,330,200
u8	150;150	98.19;96.29	94.65;90.77	58.91;58.90	14,000	4,200,000

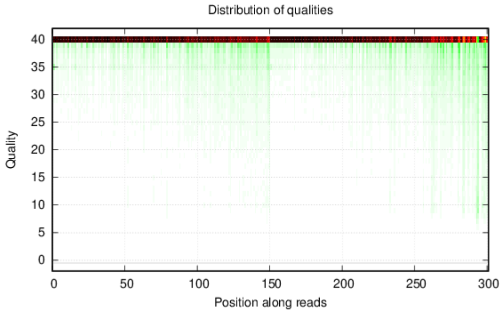
Table Format:
1. Sample: The name of sample
2. Length: The Length of reads
3. Q20 (%): The proportion of nucleotides with quality value larger than 20
4. Q30 (%): The proportion of nucleotides with quality value larger than 30
4. GC Content(%): The proportion of bases G and C
5. Total Reads: The total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

3 Data Quality Control

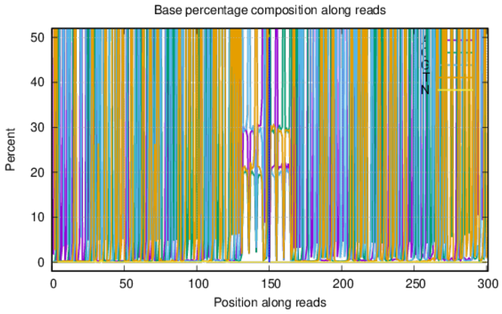
The distribution of base percentage and qualities along reads in data filtering are shown as following(If a sample has multiple lanes, only one of them will be displayed). The left picture is base percentage distribution along reads the sample, the right picture is distribution of qualities along reads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

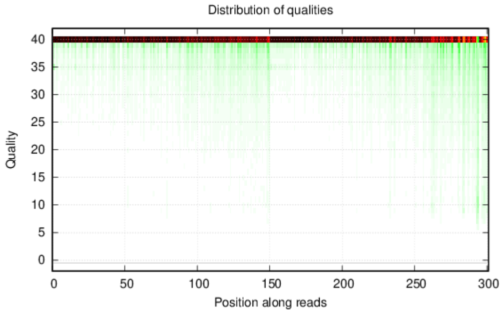
Quality control of sample e1



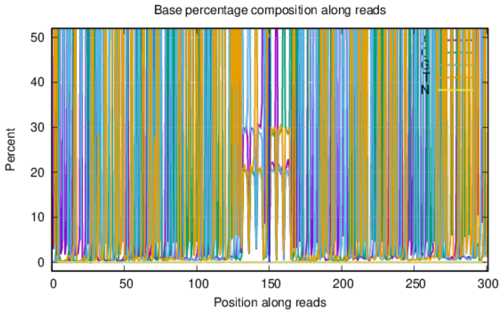


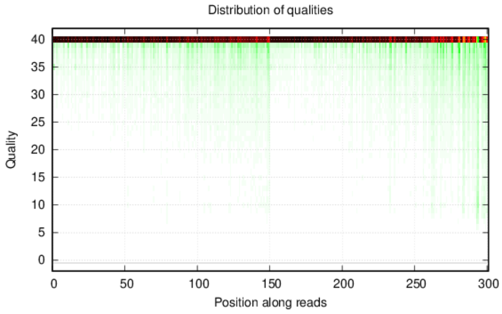
Quality control of sample e2



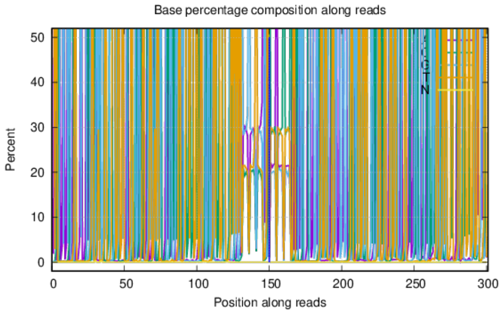


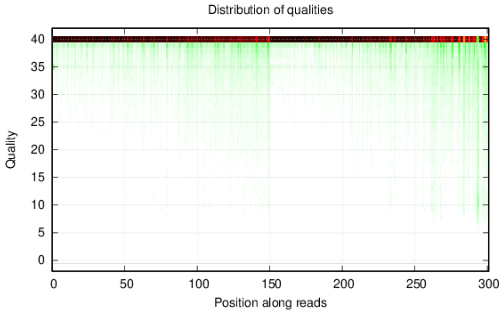
Quality control of sample e3



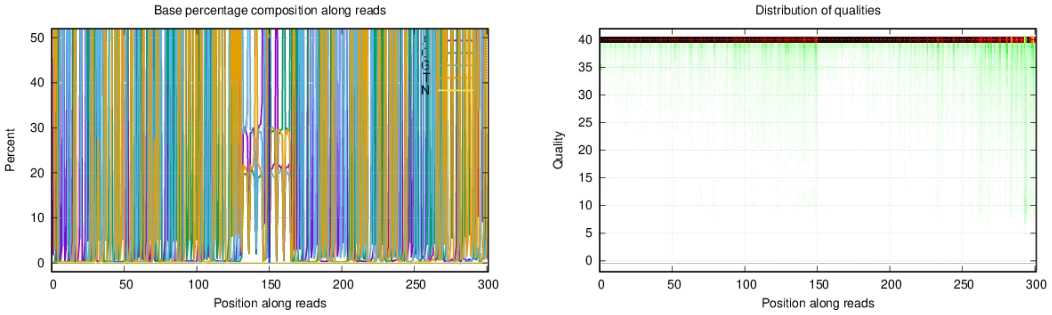


Quality control of sample e5

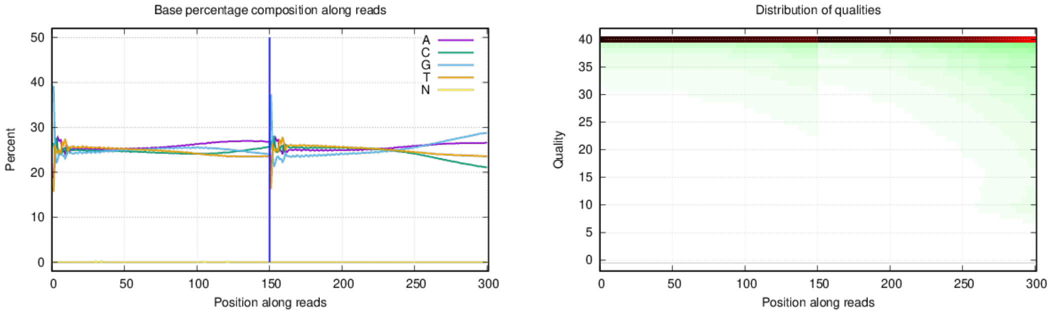




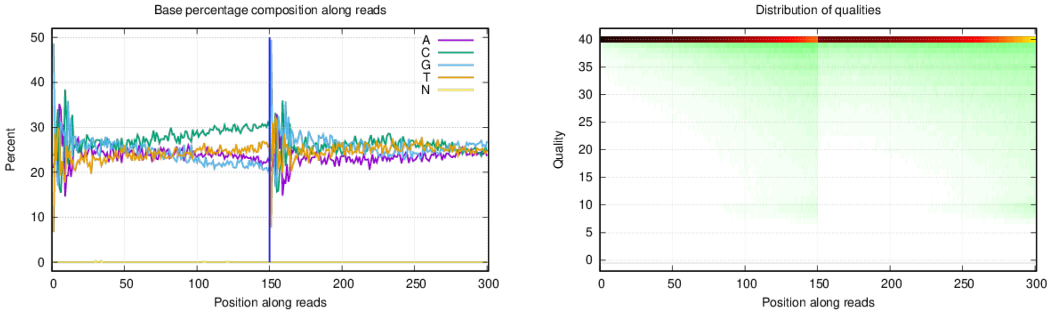
Quality control of sample e6



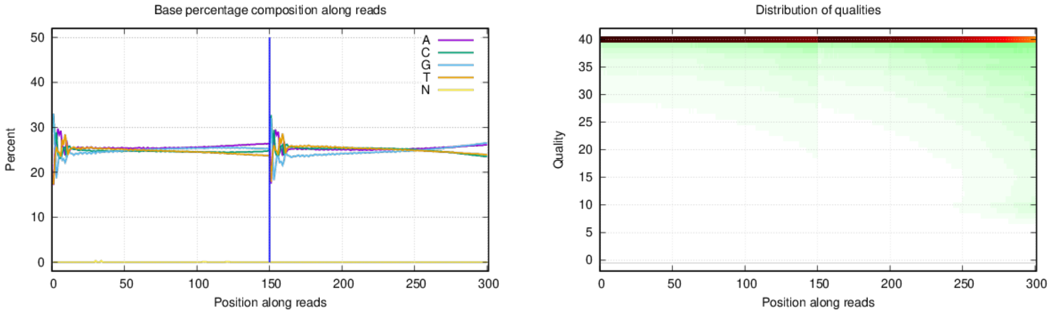
Quality control of sample Lop



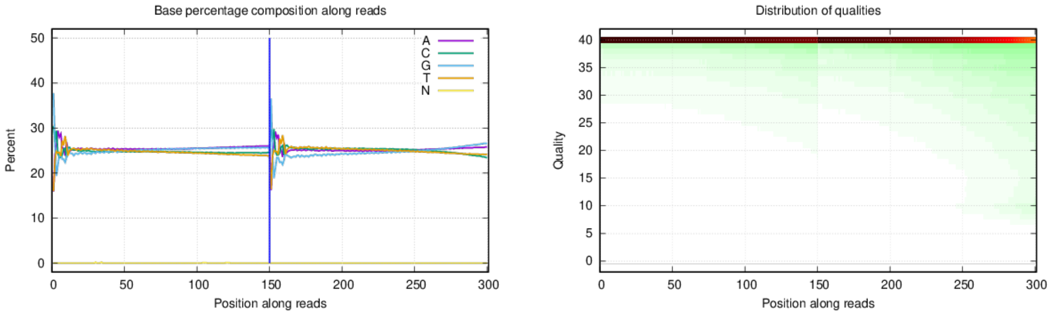
Quality control of sample LS2N



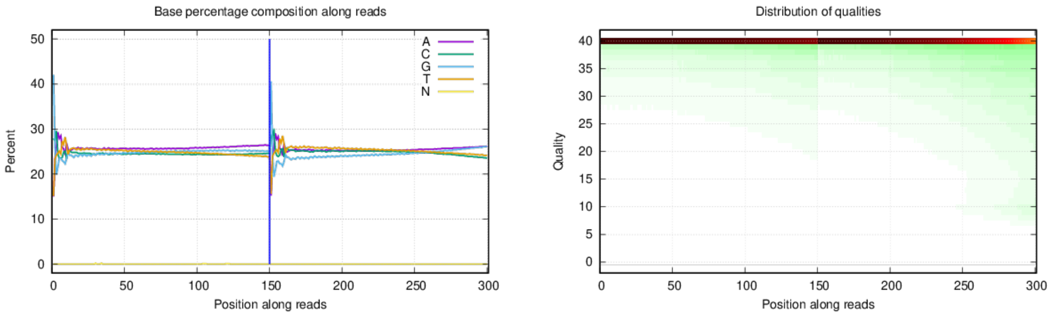
Quality control of sample mix10



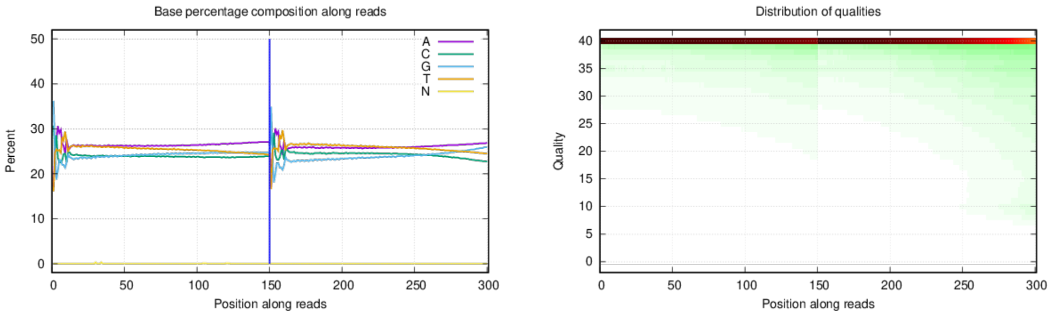
Quality control of sample mix25



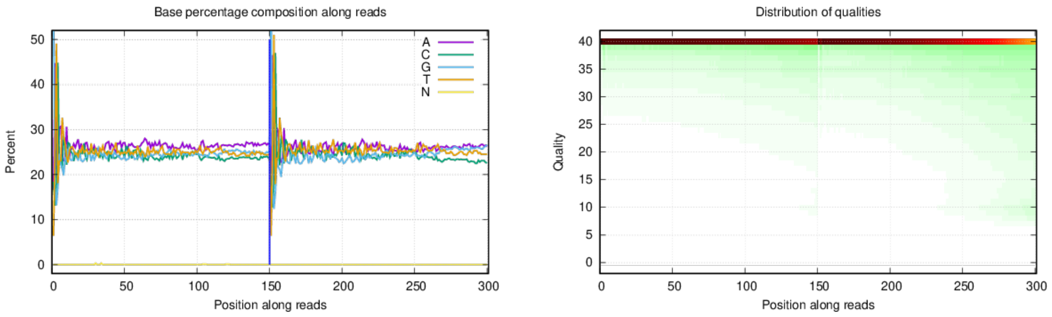
Quality control of sample mix50



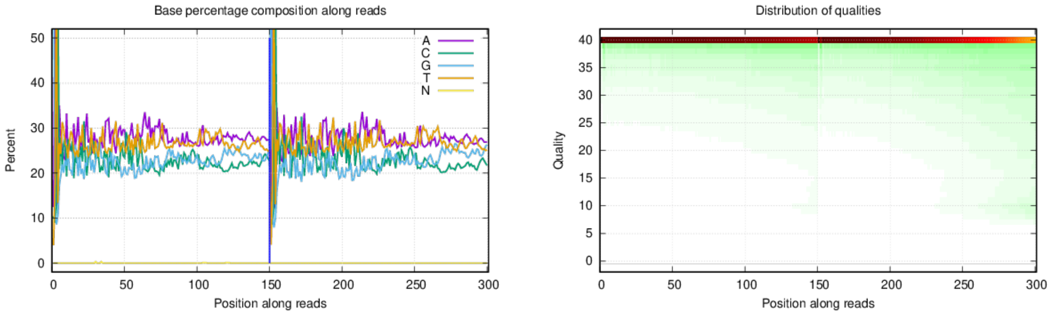
Quality control of sample mix75



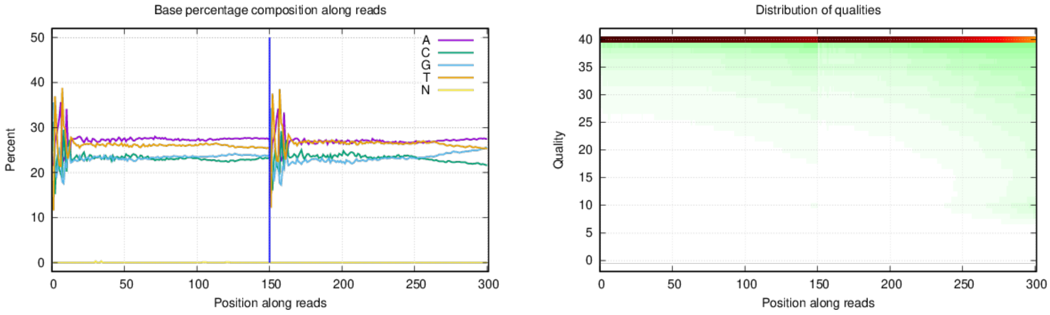
Quality control of sample MoPh11



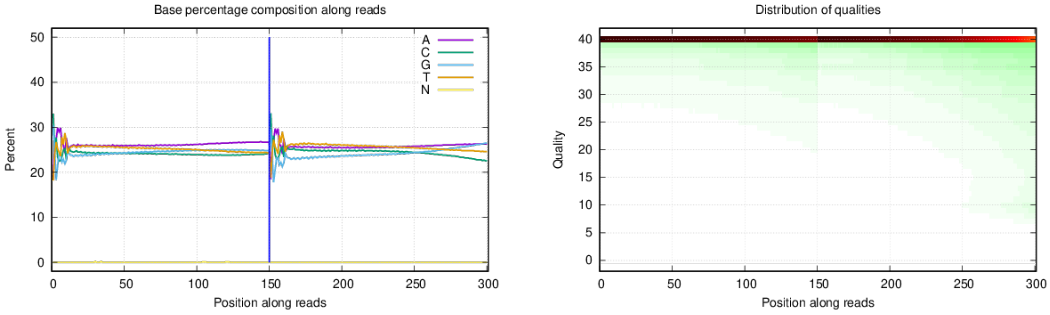
Quality control of sample MoPh14



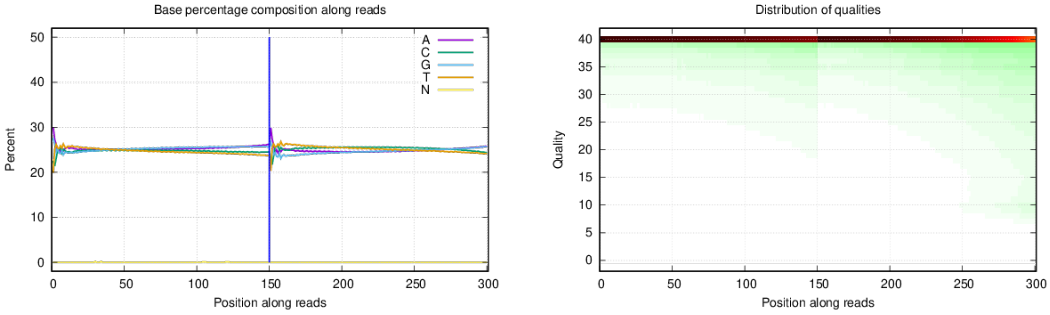
Quality control of sample MoPh7



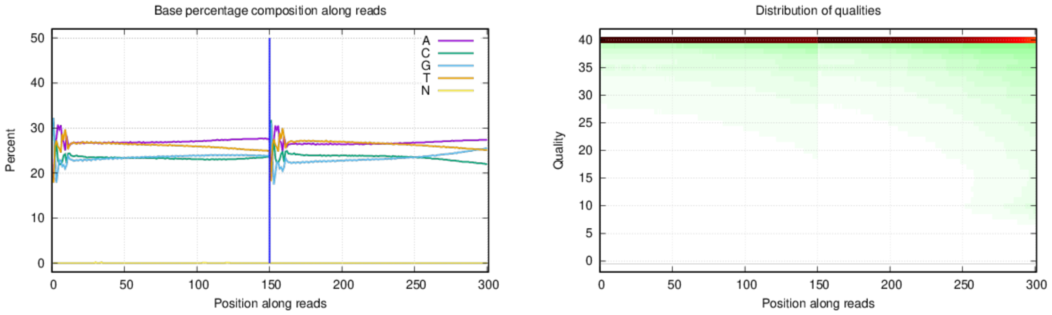
Quality control of sample P125



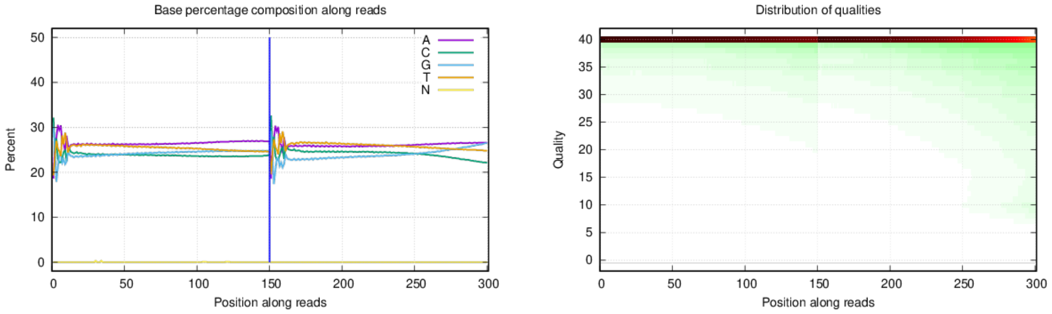
Quality control of sample P126



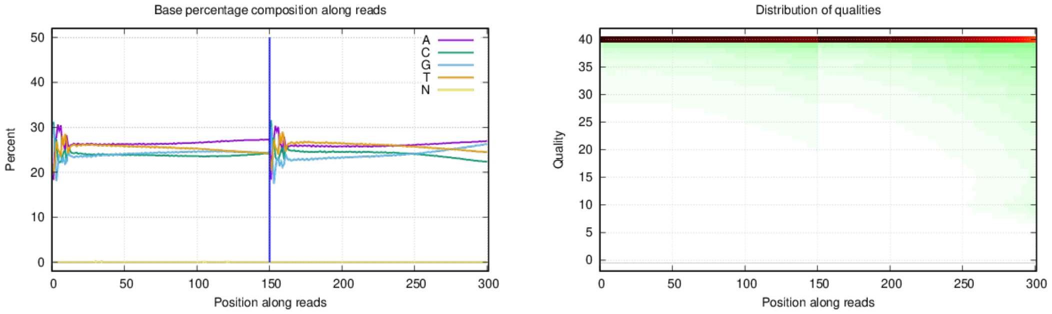
Quality control of sample P131



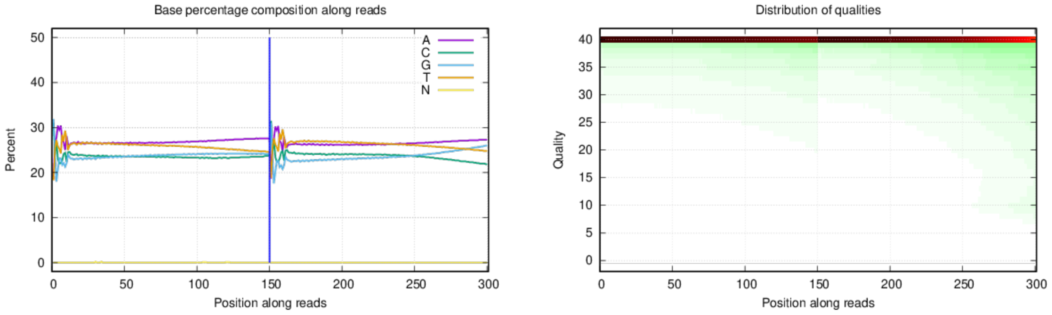
Quality control of sample P132



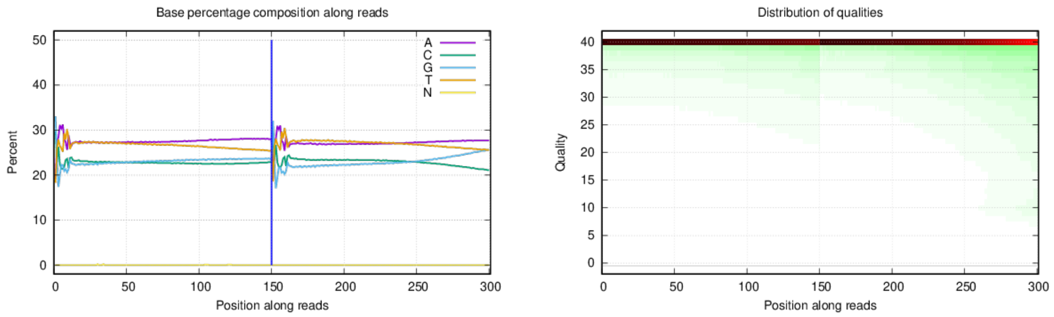
Quality control of sample P133



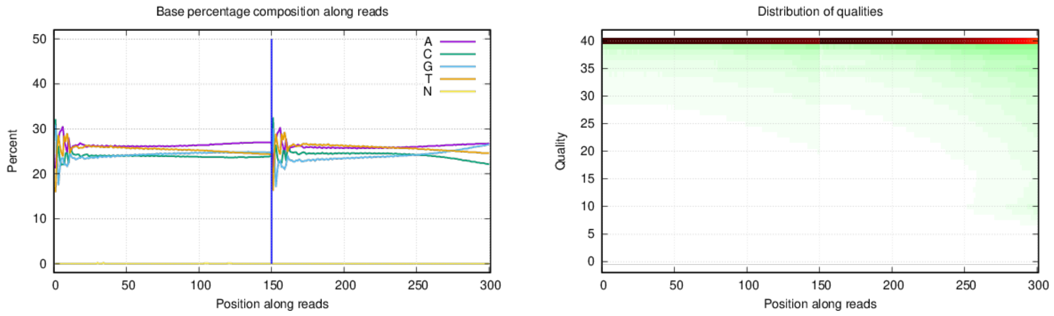
Quality control of sample P135



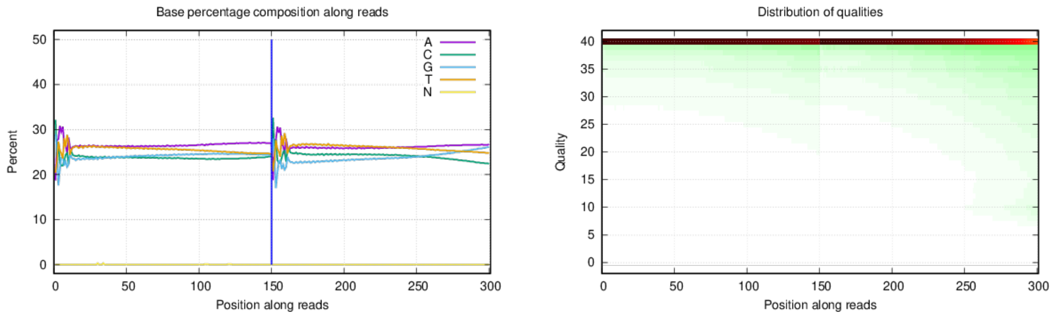
Quality control of sample P136



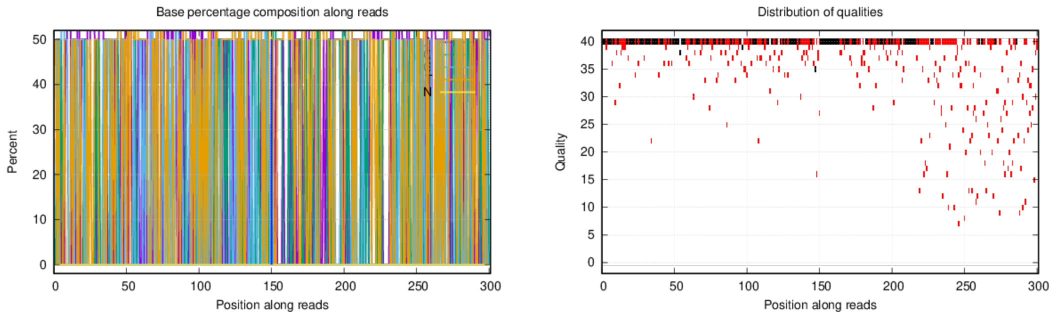
Quality control of sample P137



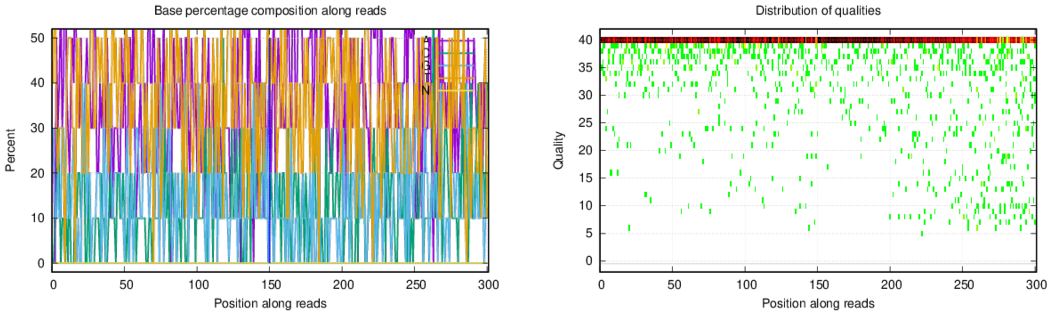
Quality control of sample P139



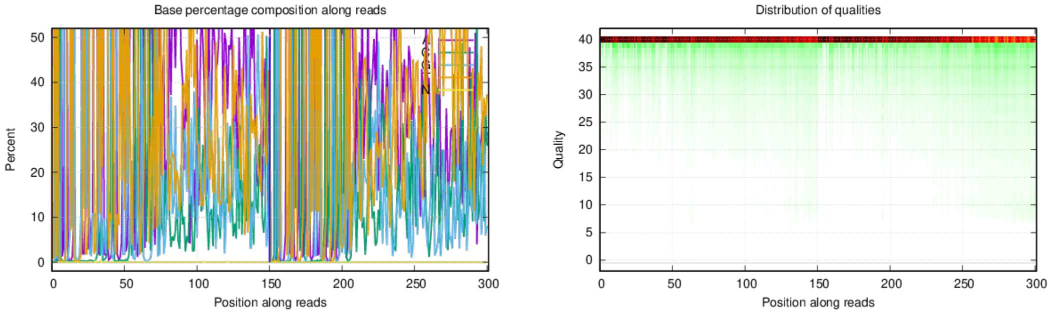
Quality control of sample PICO1_1_cells



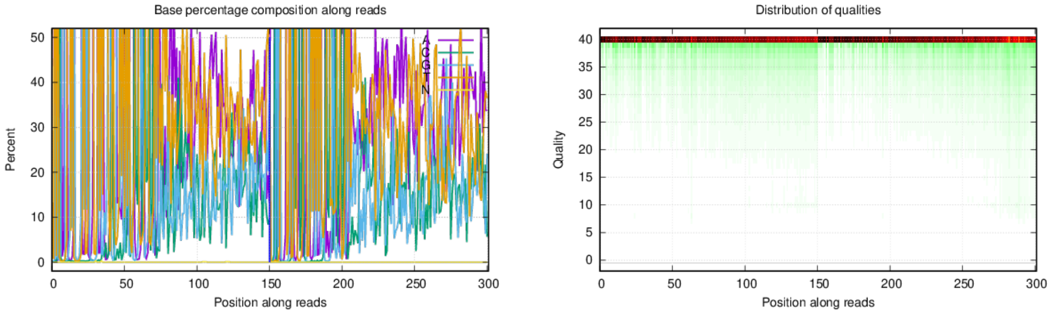
Quality control of sample PICO21



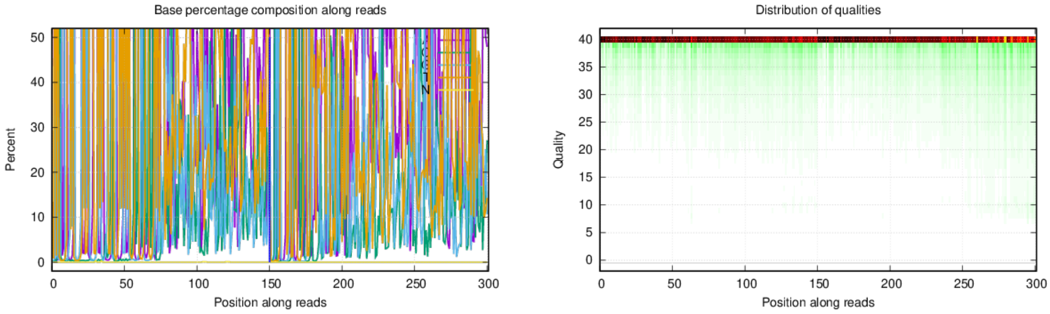
Quality control of sample sum_rear1



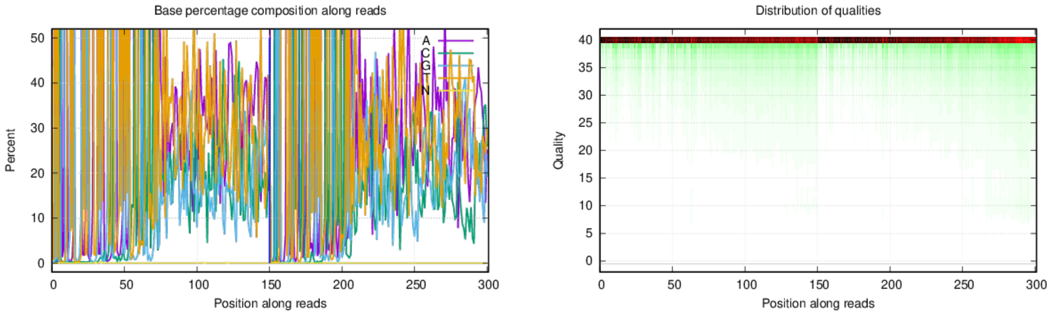
Quality control of sample sum_rear2



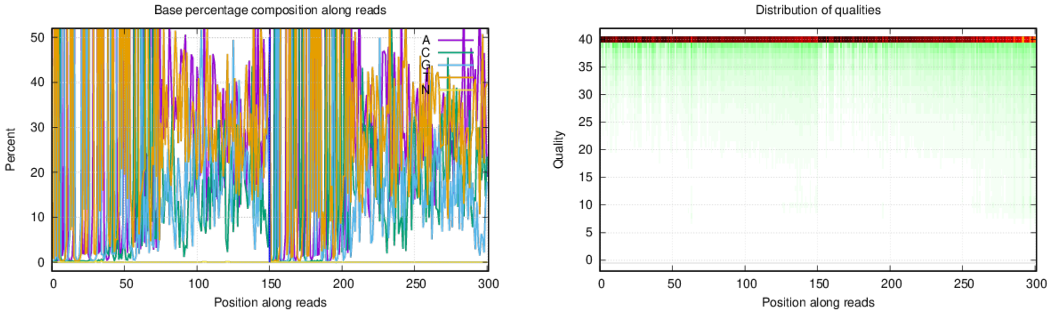
Quality control of sample sum_rear3



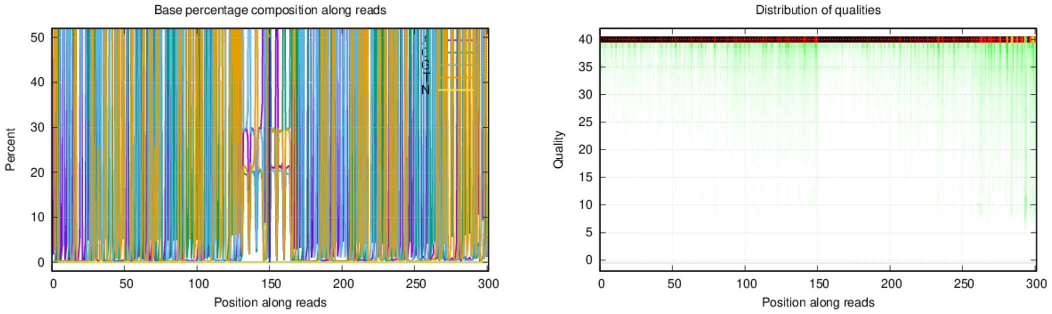
Quality control of sample sum_wt1



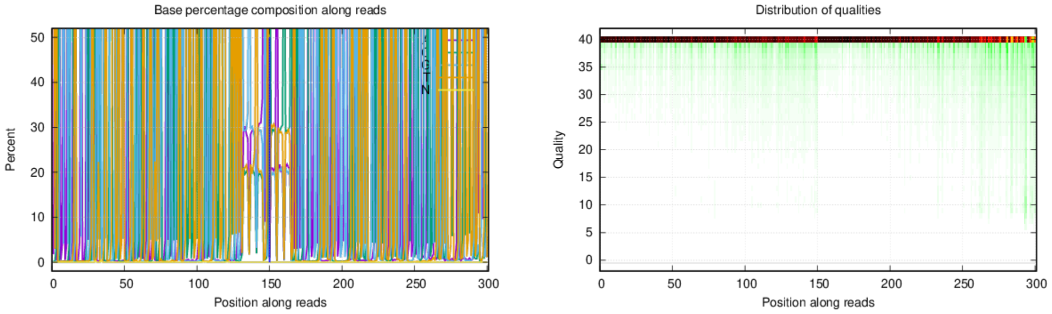
Quality control of sample sum_wt2



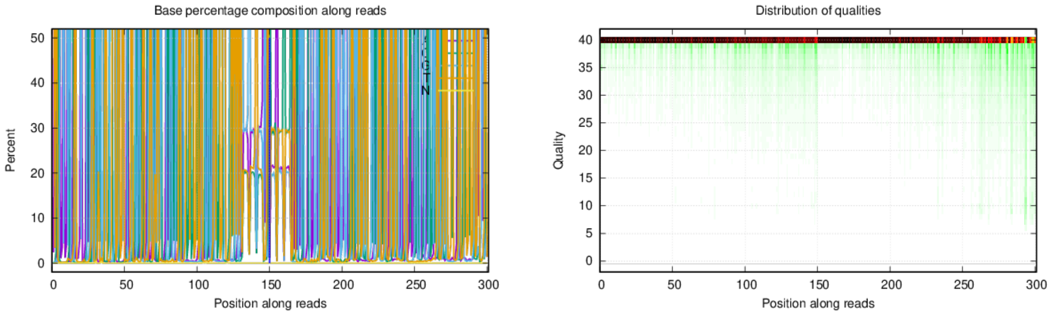
Quality control of sample u1



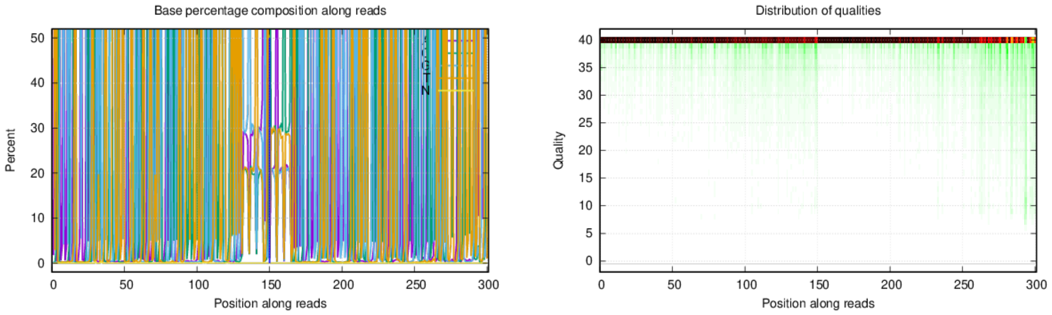
Quality control of sample u2



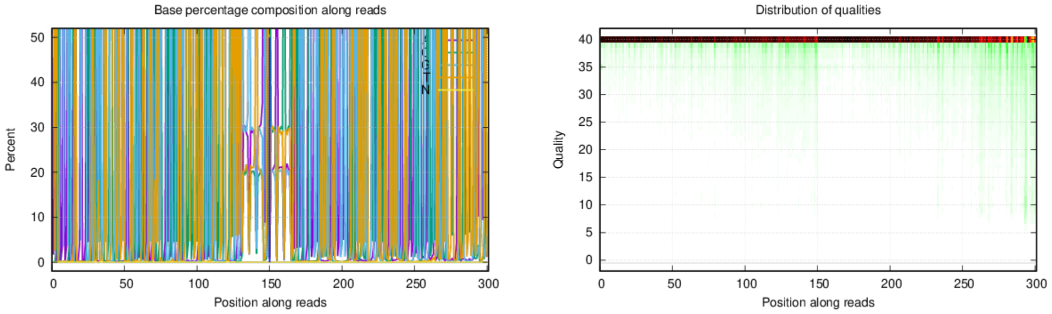
Quality control of sample u3



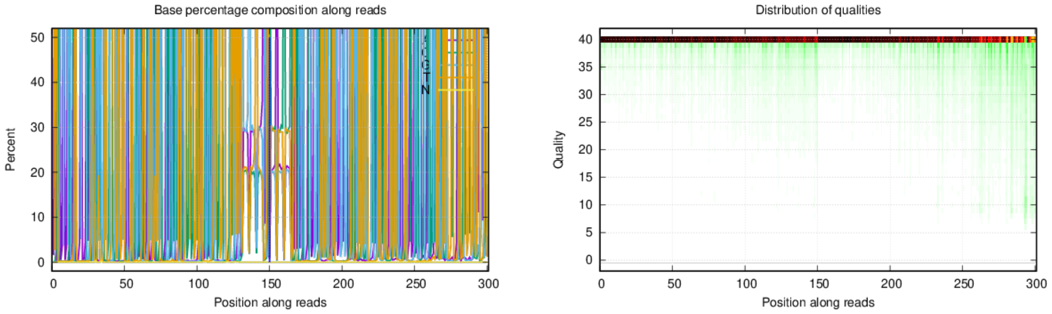
Quality control of sample u4



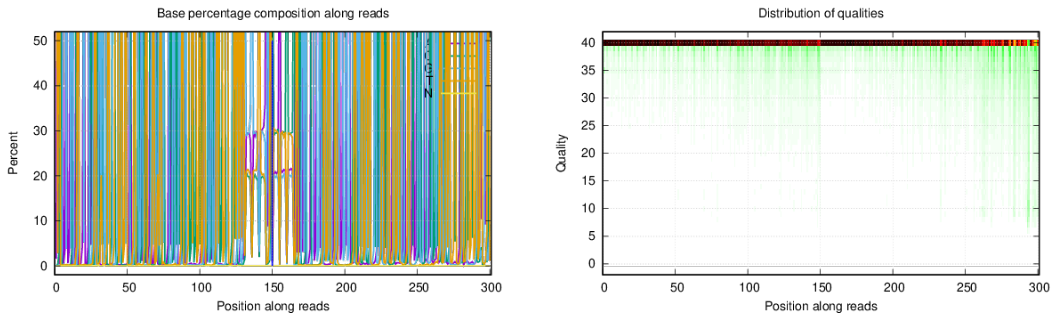
Quality control of sample u5



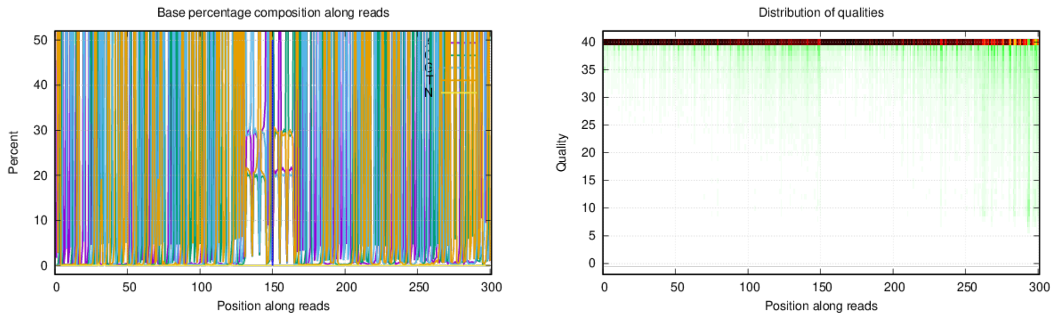
Quality control of sample u6



Quality control of sample u7



Quality control of sample u8



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
+
FFFCFFFFFFFFFFDFEFFFFEFEF0FFFFEFFFFFFFFEFFFFFECGFFFF
```

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	?