Sequencing Data Report

Project: F20FTSEUHT0946-02_MUShuavR1

Date: 2024.7.24

Note: For Research Use Only.

Table of Contents

Data Statistics

Raw reads produced from sequencer contain adapters, unknown or low quality bases.

There are 24 samples in this project, the statistics of fastq data is shown below.

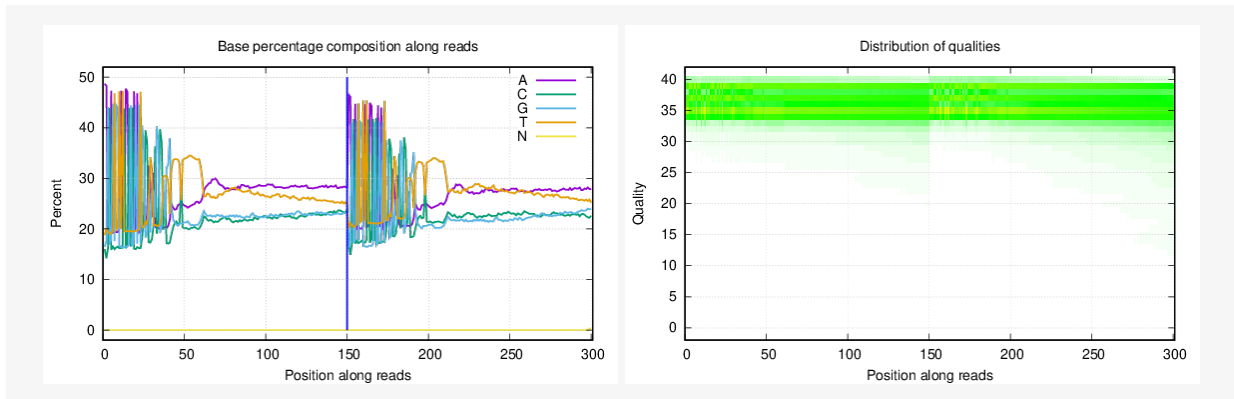| Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
10_1	150;150	98.08;97.30	94.06;92.24	45.86;45.74	36,060,430	10,818,129,000
10_2	150;150	97.97;97.41	93.75;92.42	45.65;45.63	27,081,100	8,124,330,000
14	150;150	97.85;96.90	93.36;91.03	45.95;45.84	22,756,703	6,827,010,900
2	150;150	94.53;93.92	85.70;83.70	45.73;45.45	809	242,700
25	150;150	97.68;97.06	92.93;91.48	46.13;46.08	31,298,740	9,389,622,000
27	150;150	98.09;97.75	94.16;93.49	45.99;45.92	37,913,013	11,373,903,900
28	150;150	97.97;97.55	93.77;92.87	46.46;46.41	20,214,844	6,064,453,200
3	150;150	97.94;97.50	93.69;92.72	45.63;45.57	33,593,670	10,078,101,000
30	150;150	98.07;98.02	94.10;94.09	46.94;46.86	29,088,202	8,726,460,600
31	150;150	97.46;97.50	92.40;92.58	46.62;46.51	28,088,113	8,426,433,900
35	150;150	89.08;91.20	72.60;77.21	43.70;42.57	23,105	6,931,500
37	150;150	94.70;93.59	85.66;82.79	46.11;45.87	4,035	1,210,500
39	150;150	97.84;96.95	93.36;91.09	45.77;45.85	43,140,024	12,942,007,200
5	150;150	97.89;97.73	93.58;93.35	46.12;46.05	27,647,111	8,294,133,300
525_K	150;150	97.41;97.11	92.41;91.78	45.91;45.79	20,167	6,050,100
6	150;150	98.01;97.71	93.91;93.38	45.80;45.75	36,821,358	11,046,407,400
7493_K	150;150	93.41;92.32	81.40;78.50	46.32;46.41	7,905	2,371,500
Aks_K	150;150	93.64;91.06	82.50;76.52	45.89;45.71	140,240	42,072,000
Bag4_K	150;150	96.75;96.19	90.78;89.39	45.69;45.52	22,447	6,734,100
Cher2_K	150;150	88.72;91.12	71.49;77.06	43.90;43.00	6,986	2,095,800
Fuks2_K2	150;150	93.76;91.97	83.57;79.45	45.22;45.02	48,863	14,658,900
Kur3_K	150;150	97.37;97.29	92.33;92.26	46.29;46.19	73,954	22,186,200
Mar3_K	150;150	97.48;96.36	92.39;89.43	43.68;43.60	788,142	236,442,600
T1	150;150	97.34;96.70	92.02;90.38	44.45;44.41	79,050	23,715,000

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

5. GC Content(%): The proportion of bases G and C

6. Total Reads: The total number of read pairs

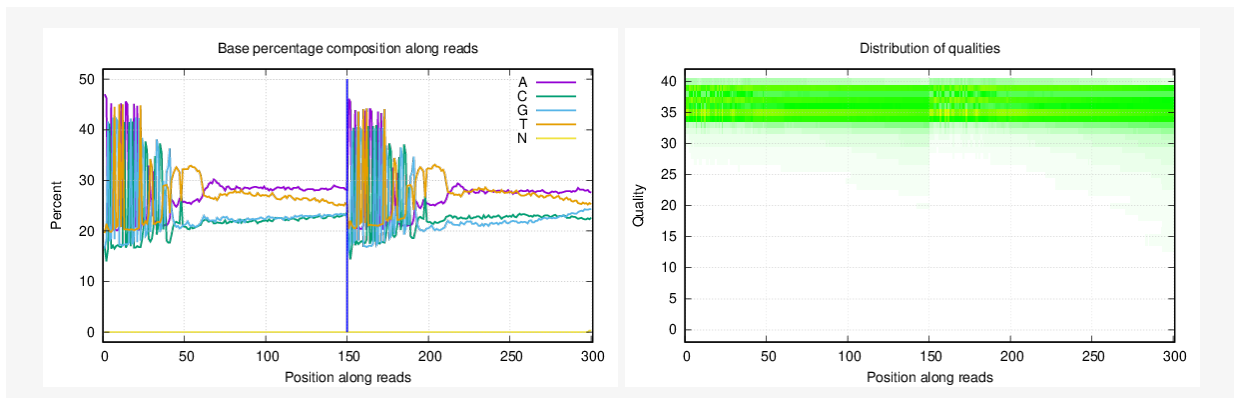7. Total Bases: The total nucleotides number of reads

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.

10_1

10_2

14
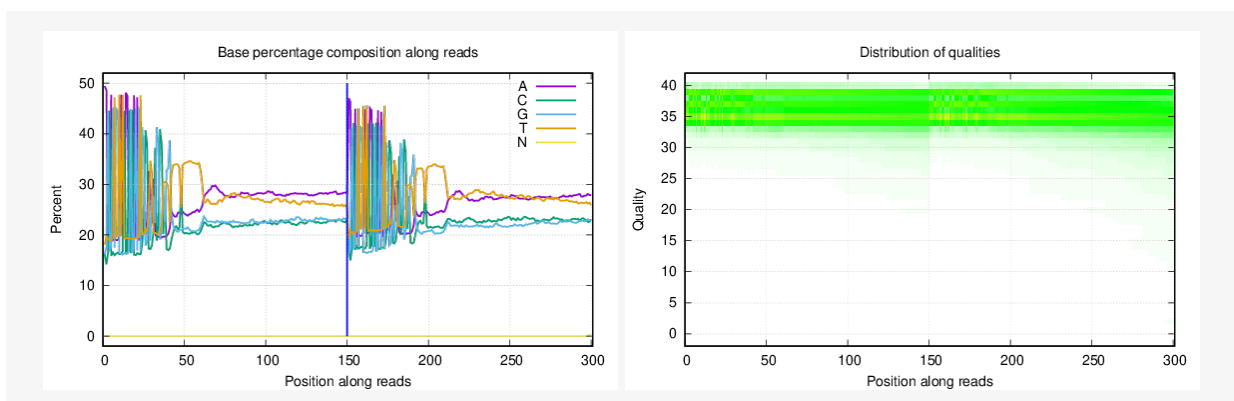
2
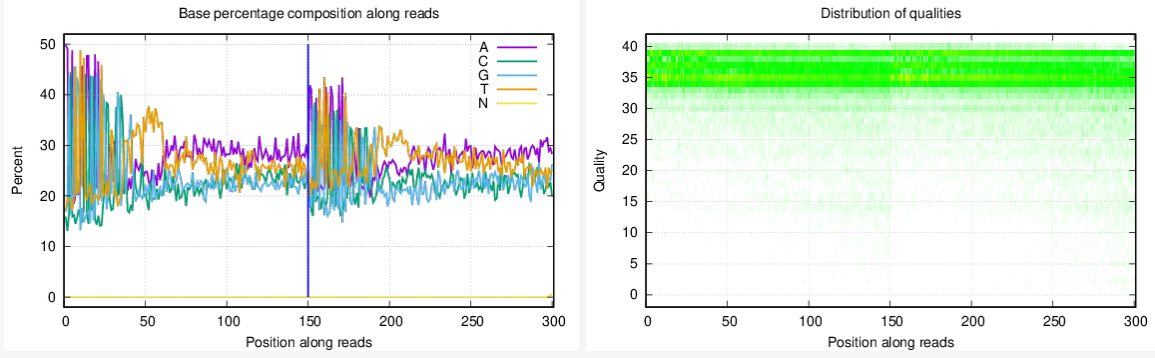
25
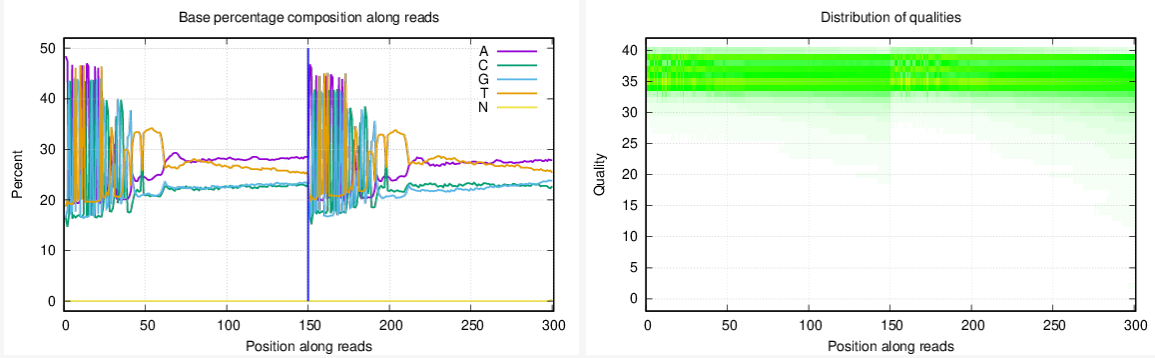
27
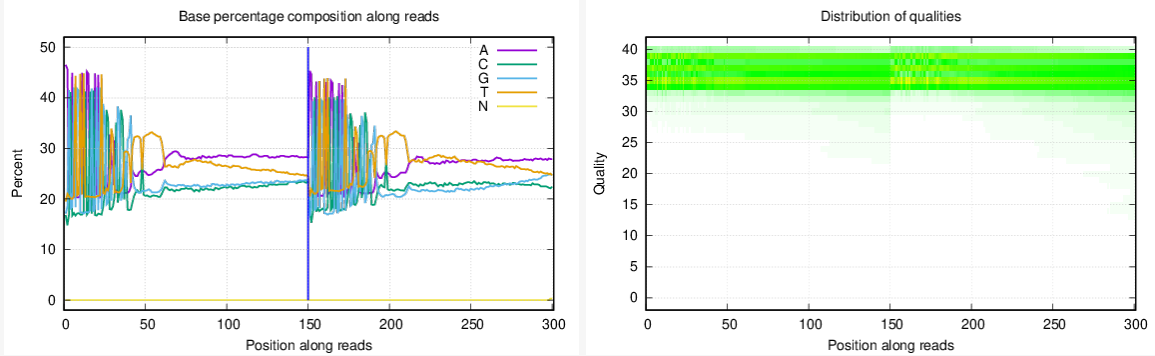
28
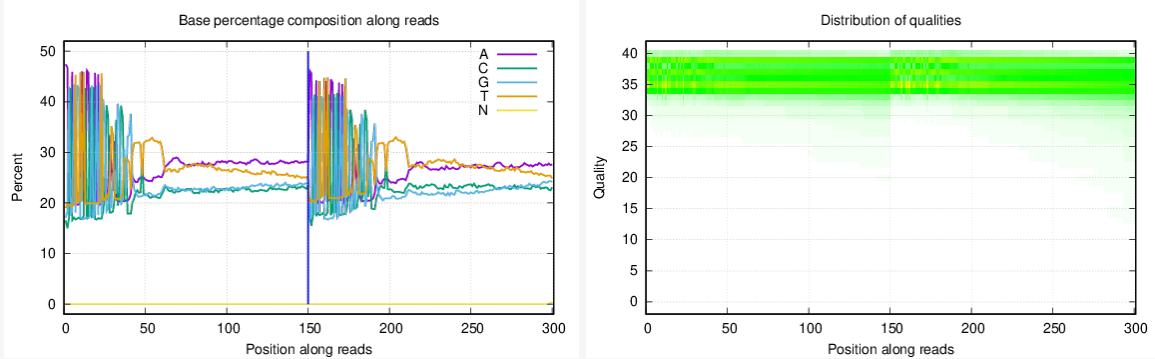
3
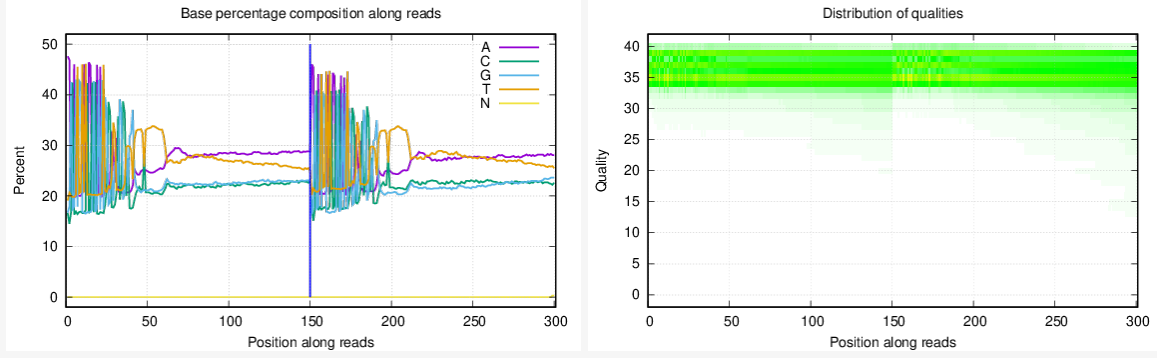
30
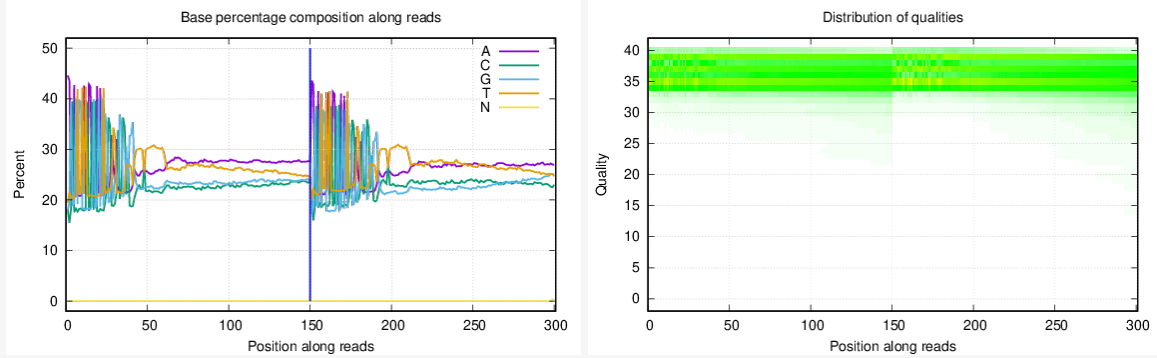
31
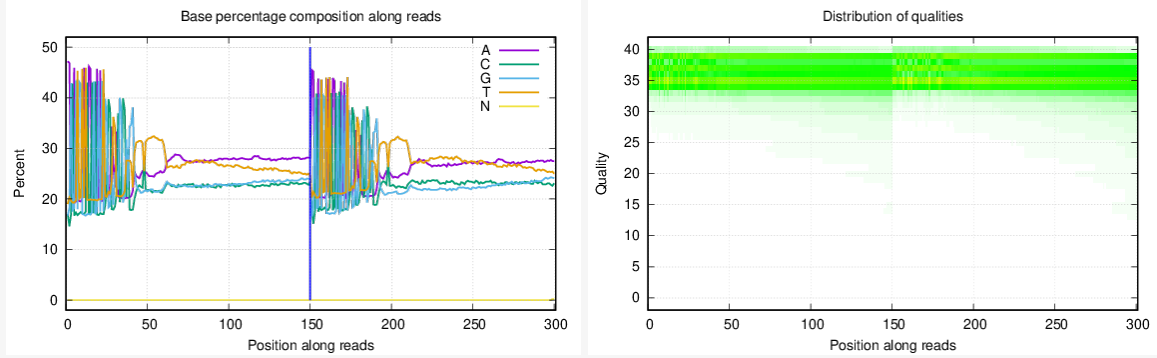
35
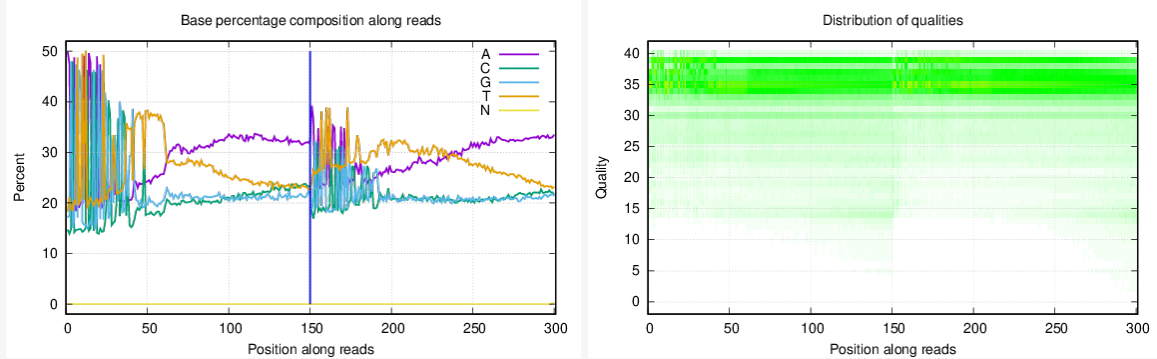
37
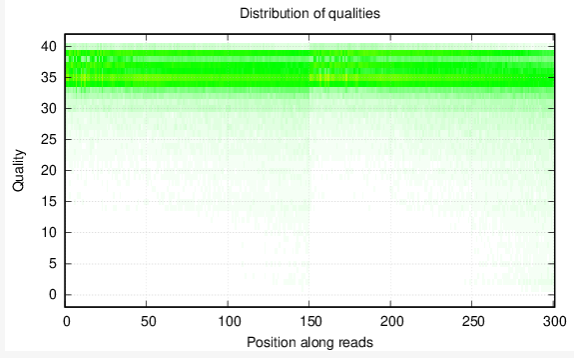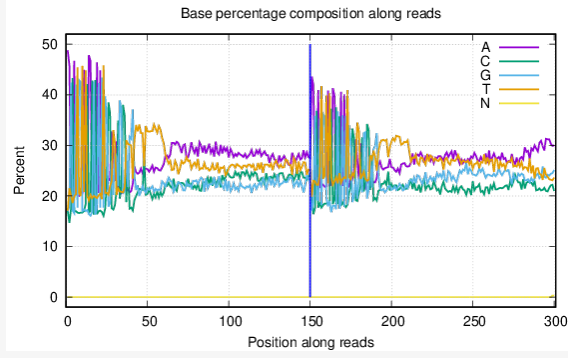
39
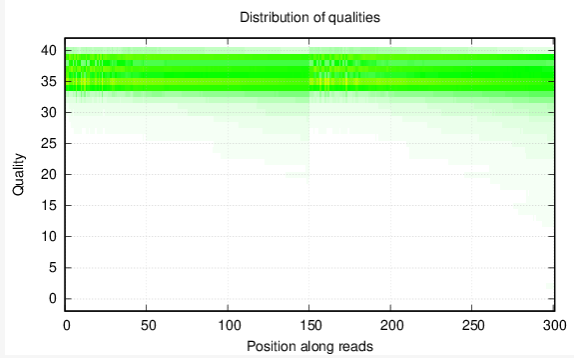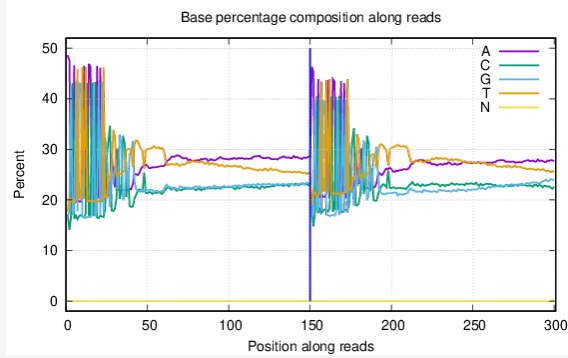
5
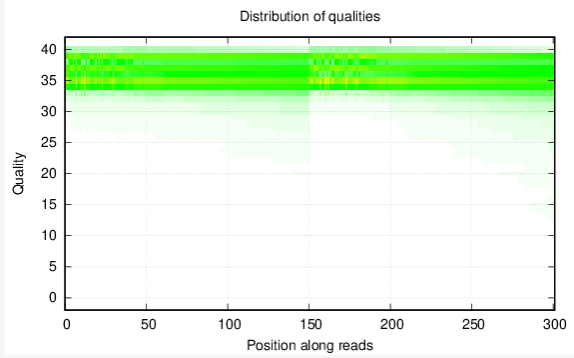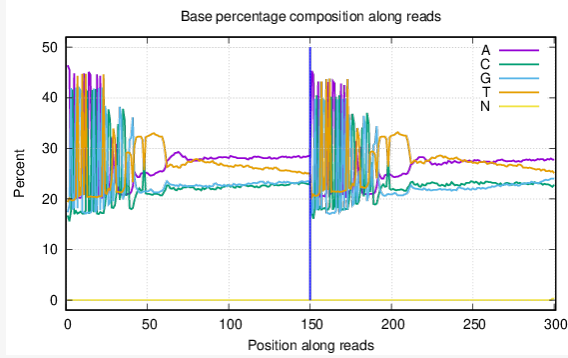
525_K
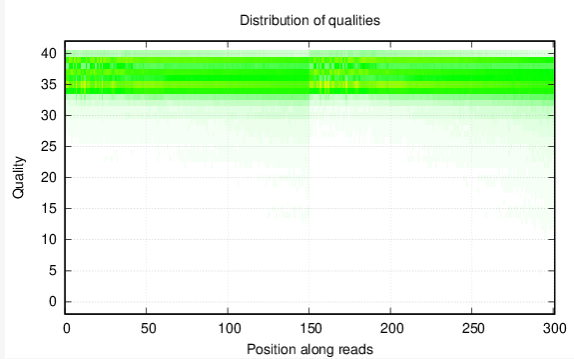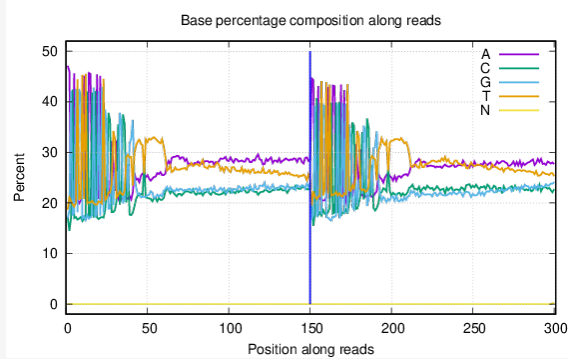
6
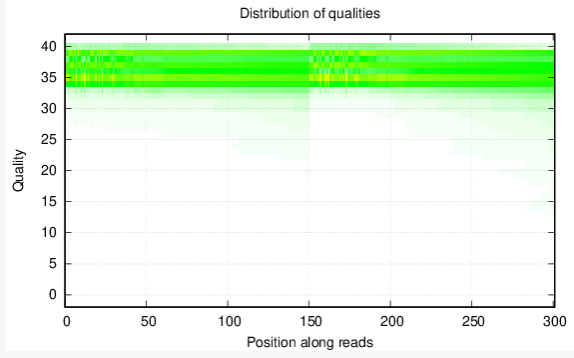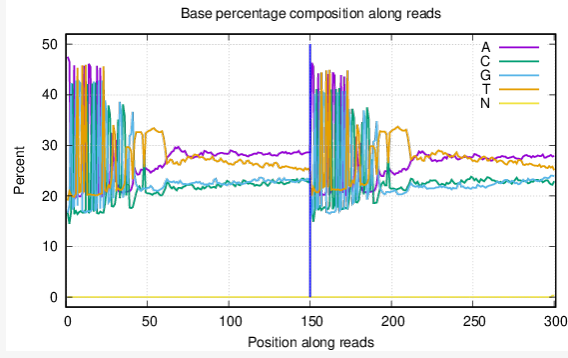
7493_K
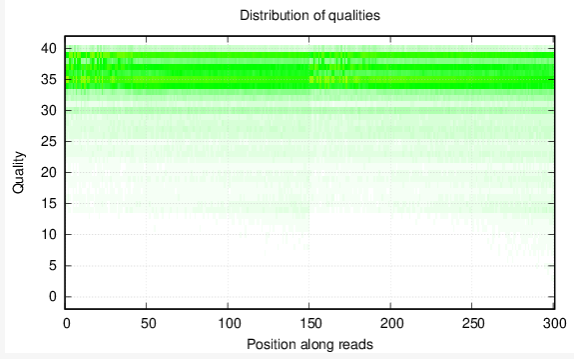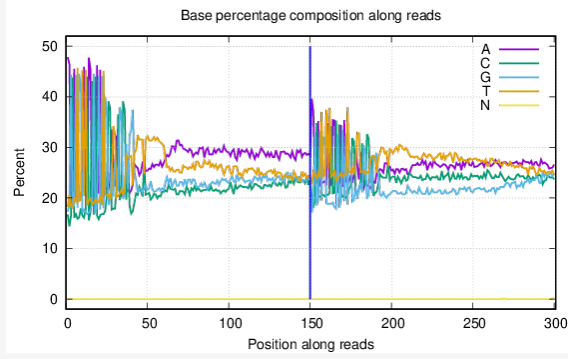
Aks_K
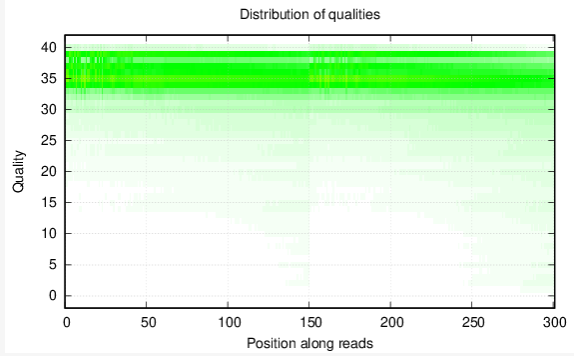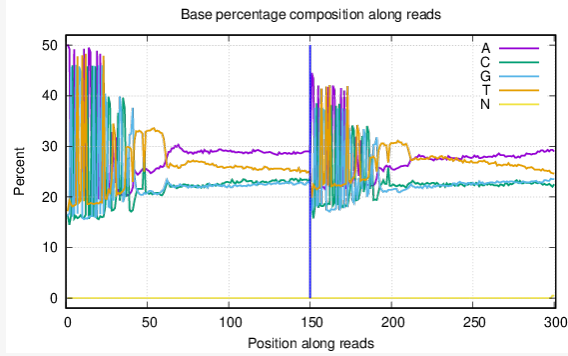
Bag4_K
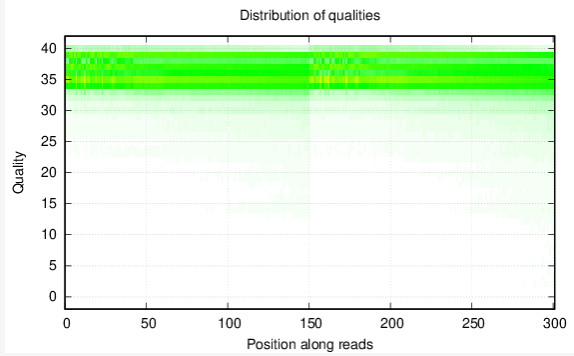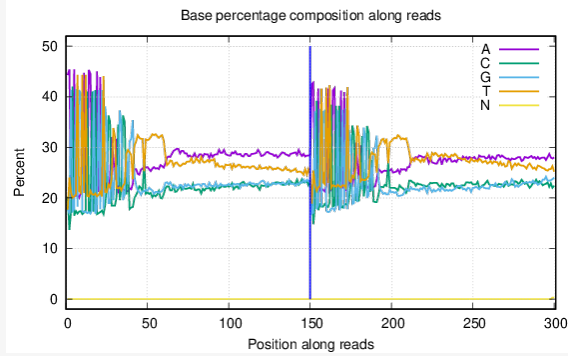
Cher2_K

Fuks2_K2

Kur3_K

Mar3_K

T1

Base percentage composition along reads

Distribution of qualities
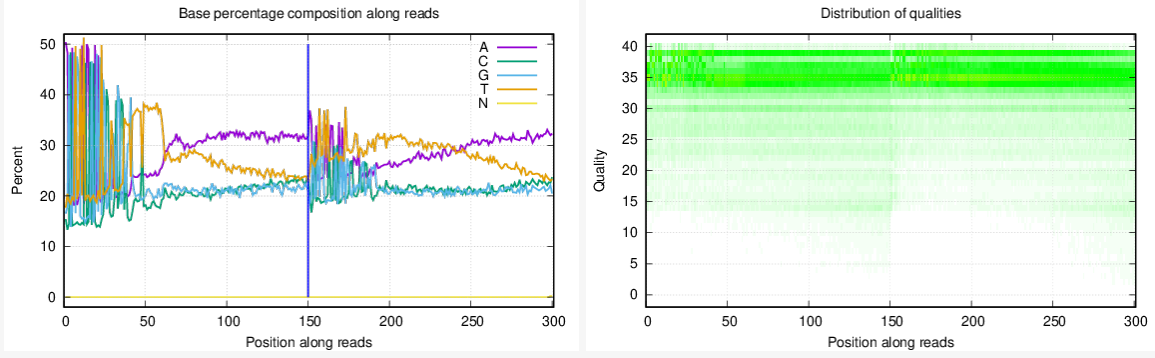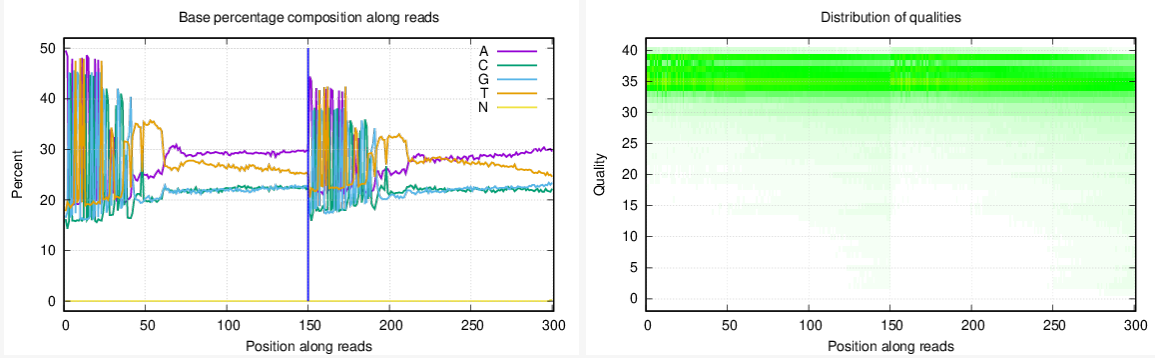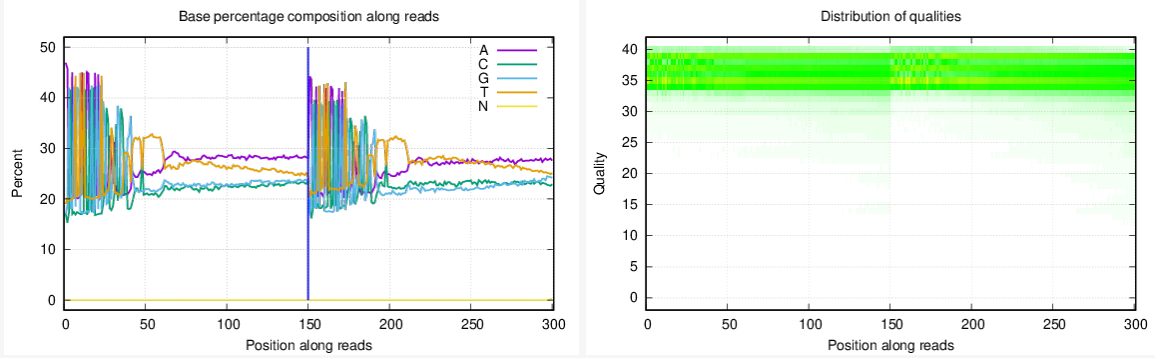
Help Document

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:
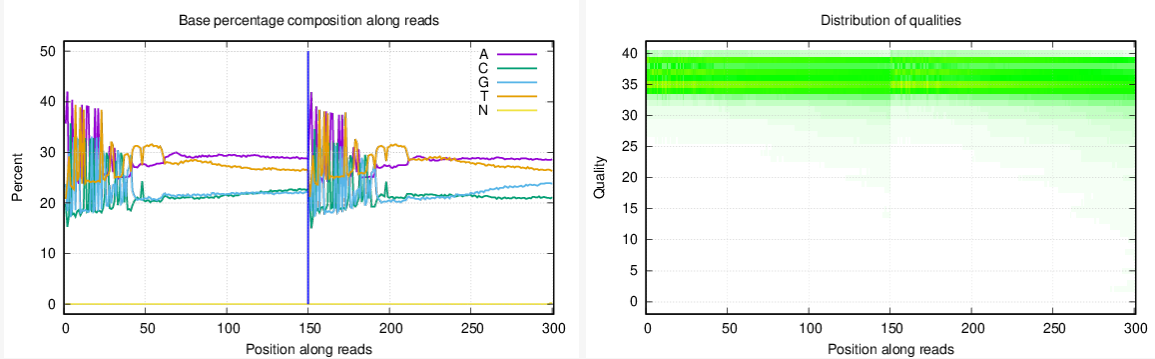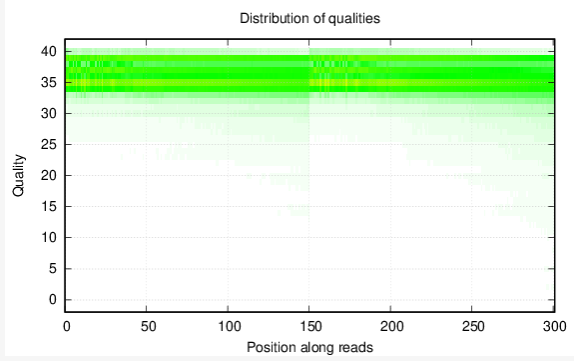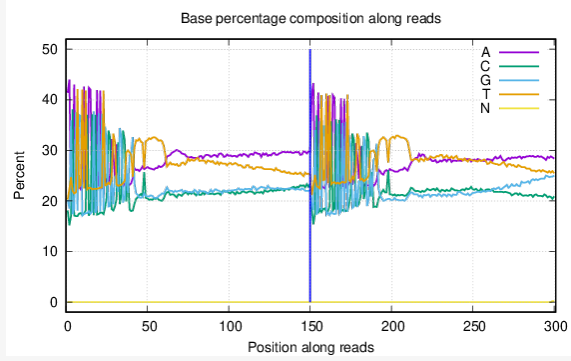
> 1. A sequence identifier with information about the sequencing run and the cluster. The exact contents of this line vary by based on the BCL to FASTQ conversion software used.
> 2. The sequence (the base calls; A, C, T, G and N).
> 3. A separator, which is simply a plus (+) sign.
> 4. The base call 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
+
FFFCFFFFFFFFFDFEFFFFEFEF0FFFFEFFFFFFFFEFFFFFECGFFFF
```

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	?

BGI

Omics For All