BGI Sequencing Data Report

Project: F22FTSEUHT2056_MUSpaxtR

Date: 2023.12.21

Note: For Research Use Only.

BGI Tech, Omics For All

ShenZhen BGI Genomics Co., Ltd.

Table of Contents

Data Statistics

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

There are 20 samples in this project, the statistics of fastq data is shown below.

Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
7407_contr	150;150	98.76;98.93	95.78;96.19	46.01;46.08	60,064	18,019,200
7407_treated	150;150	98.56;98.90	95.09;96.04	44.52;44.50	52,891	15,867,300
CapH2_C_R1	150;150	98.80;99.01	95.90;96.44	45.07;45.10	73,606,455	22,081,936,500
CapH2_C_R2	150;150	98.64;99.07	95.40;96.63	44.64;44.68	61,353,432	18,406,029,600
CapH2_aux_R1	150;150	97.95;98.12	92.86;92.84	44.22;44.10	57,644,195	17,293,258,500
CapH2_aux_R2	150;150	98.34;98.96	94.34;96.17	44.59;44.56	38,439,705	11,531,911,500
CapH_C_R1	150;150	97.81;98.34	92.45;93.51	43.97;43.88	38,975,103	11,692,530,900
CapH_C_R2	150;150	97.93;98.10	92.82;92.73	43.51;43.43	55,794,097	16,738,229,100
CapH_aux_R1	150;150	98.06;98.49	93.37;94.38	43.93;43.88	25,473,764	7,642,129,200
CapH_aux_R2	150;150	98.20;98.60	93.80;94.66	44.13;44.03	50,347,627	15,104,288,100
Clover_1w	150;150	92.50;93.79	77.82;70.95	57.36;57.42	1,776	532,800
Clover_2w	150;150	91.65;93.02	75.54;68.57	57.64;57.74	680	204,000
Clover_ctr	150;150	92.78;93.87	78.95;72.00	57.24;57.69	251	75,300
Red_1w	150;150	93.83;93.73	81.55;72.75	58.44;58.68	1,410	423,000
Red_2w	150;150	94.42;94.20	83.53;74.35	58.64;58.82	1,034	310,200
Red_ctr	150;150	93.15;93.25	79.94;70.48	58.71;59.13	1,349	404,700
U87	150;150	98.36;98.49	94.26;94.25	42.40;42.41	122,708	36,812,400
kit30plus-CTCF-rep1	150;150	97.30;97.84	90.82;91.86	45.45;45.36	21,483,045	6,444,913,500
kit30plus-input-rep1	150;150	96.67;98.31	89.90;94.21	44.53;44.67	13	3,900
kit30plus-input-rep2	150;150	97.05;99.62	89.24;97.52	47.90;44.29	7	2,100

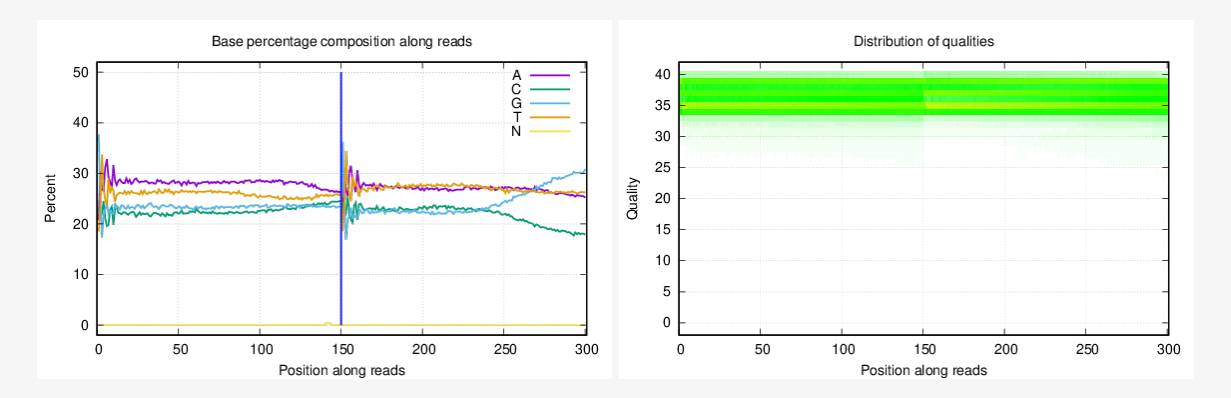
Table Format:

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

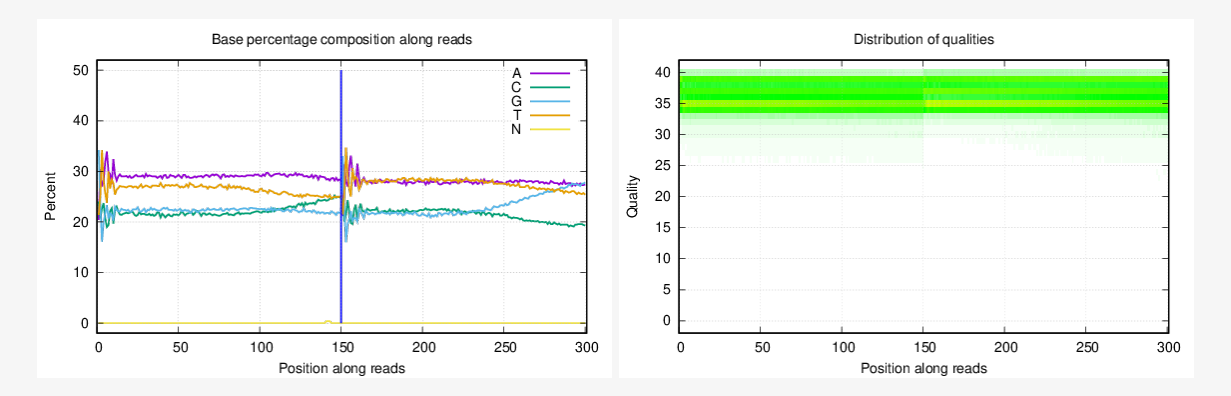
Data Quality Control

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.

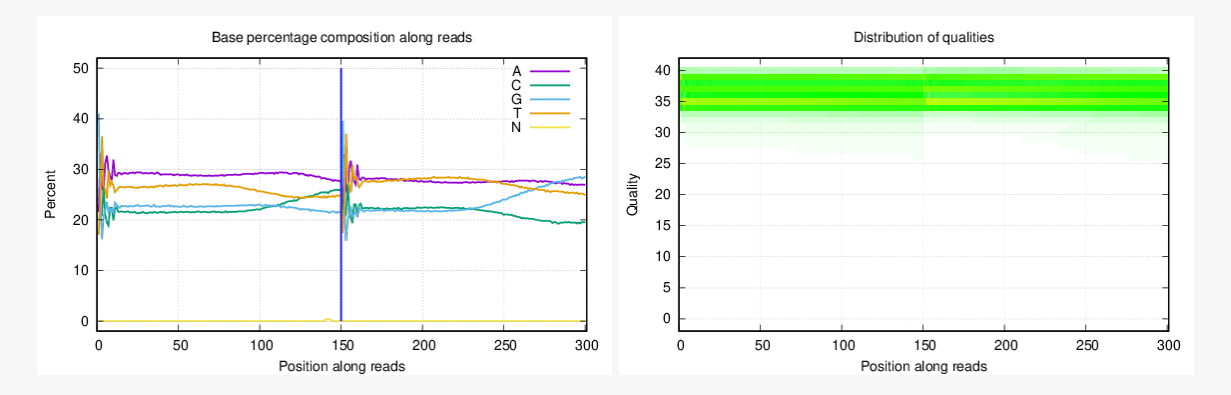
7407_contr



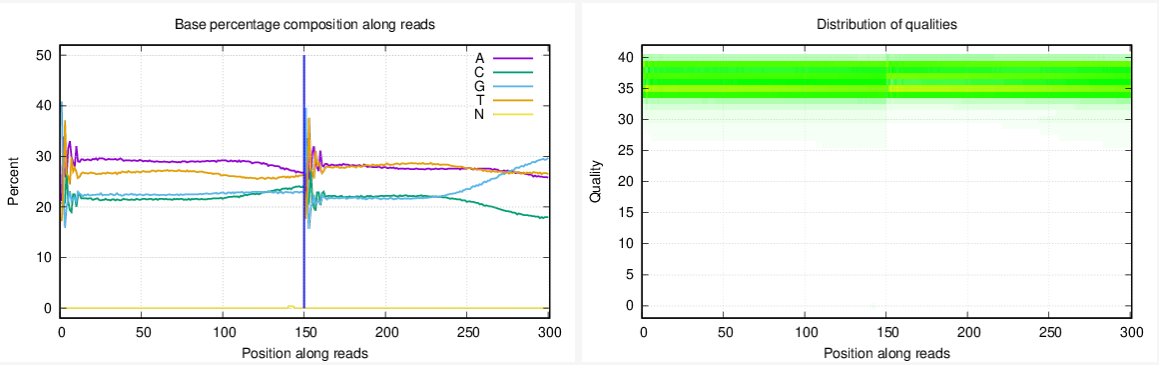
7407_treated



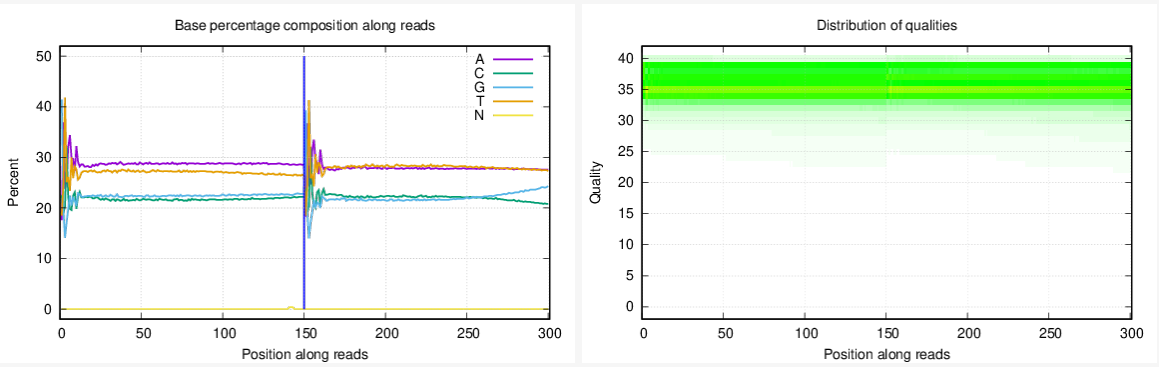
CapH2_C_R1



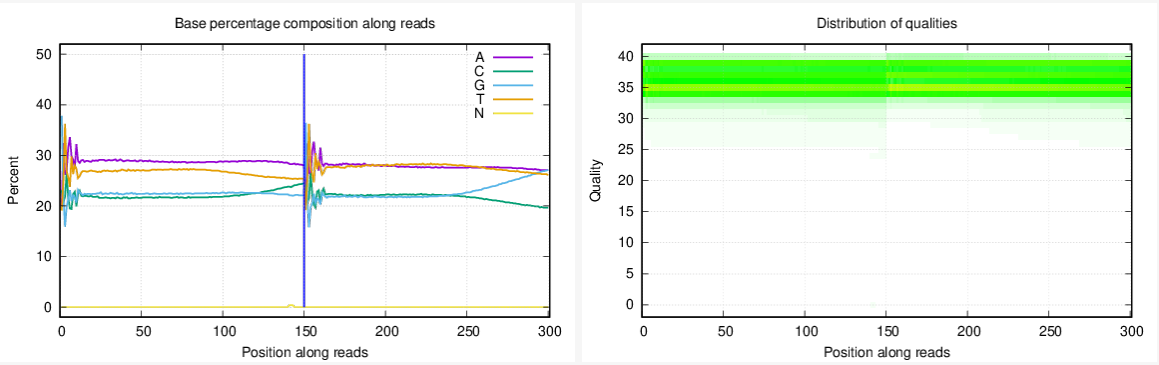
CapH2_C_R2



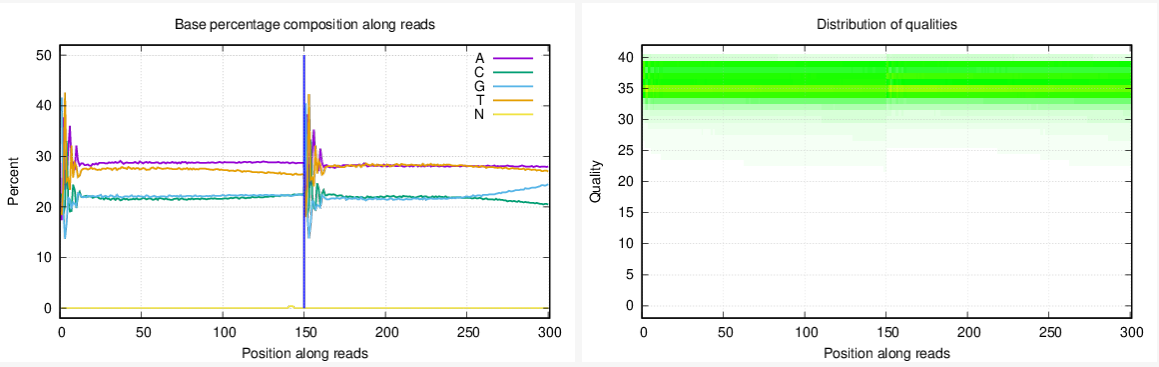
CapH2_aux_R1



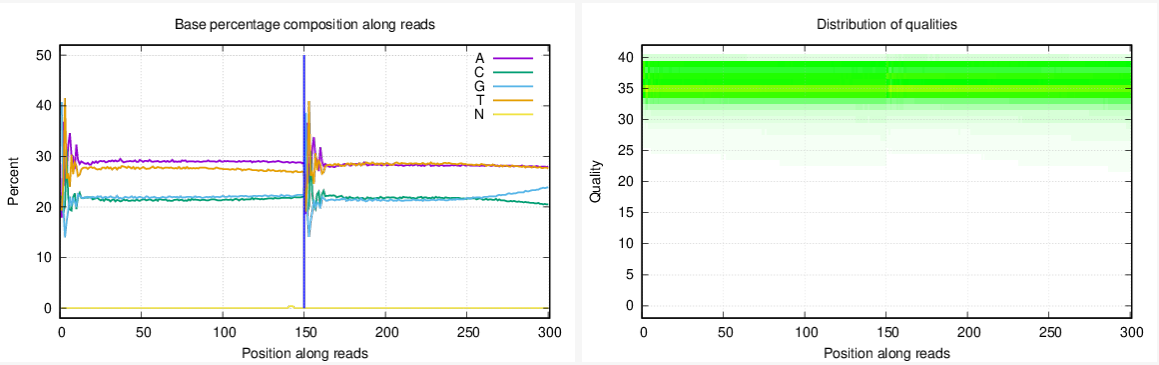
CapH2_aux_R2



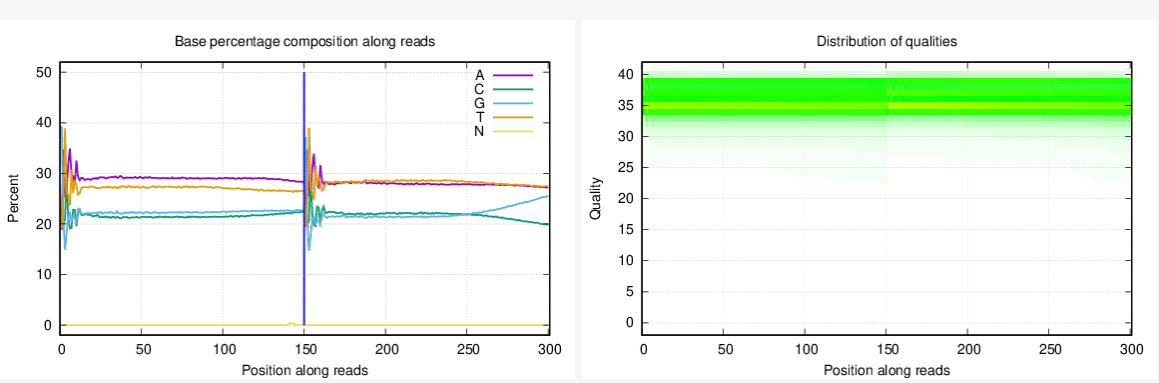
CapH_C_R1



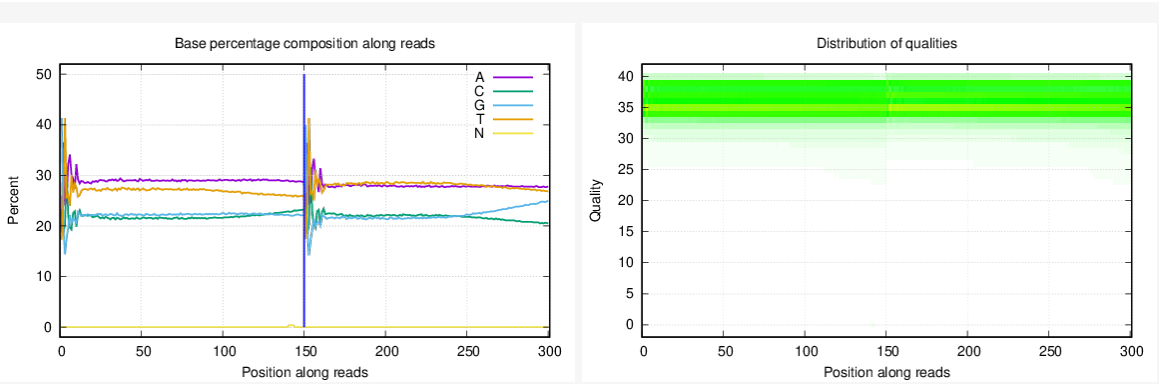
CapH_C_R2



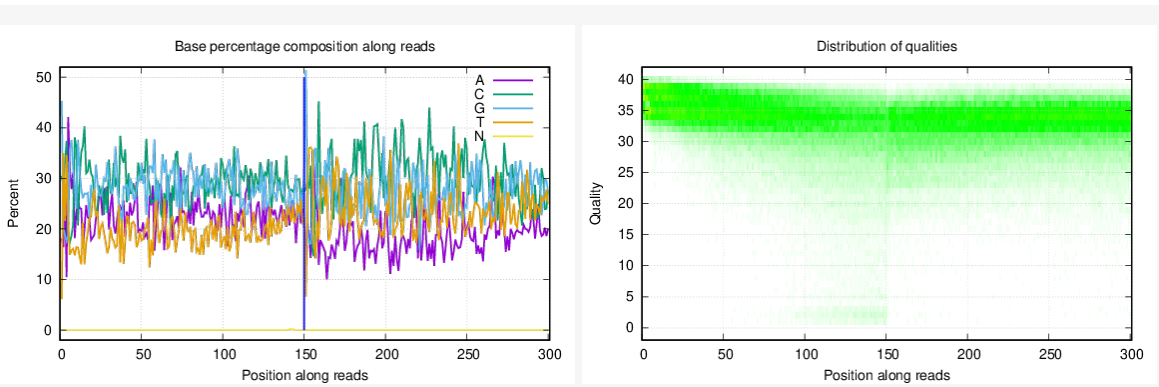
CapH_aux_R1



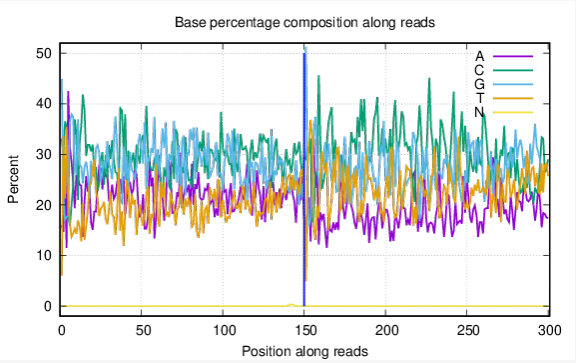
CapH_aux_R2

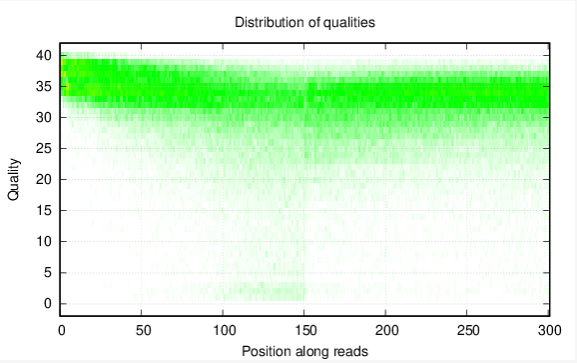


Clover_1w

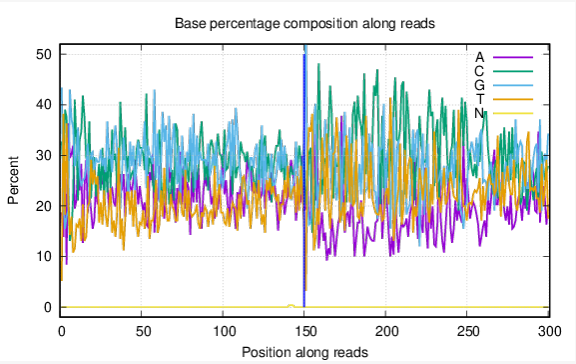


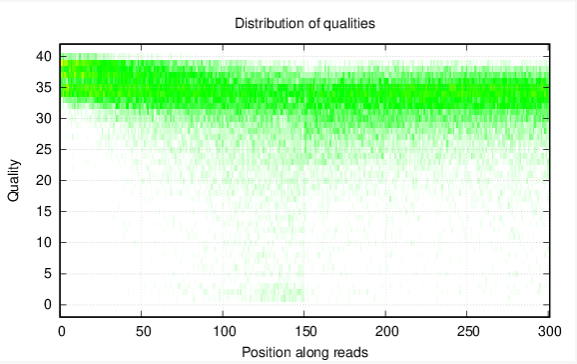
Clover_2w



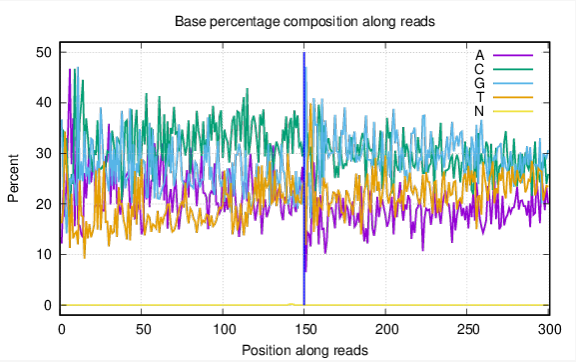


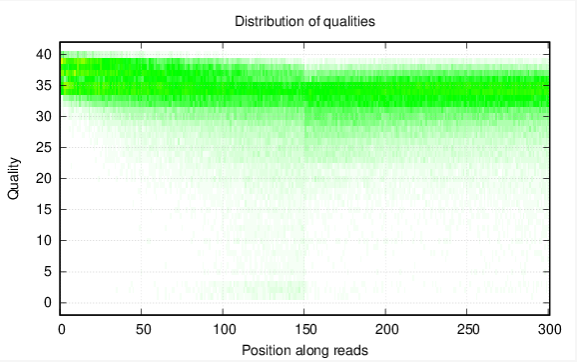
Clover_ctr



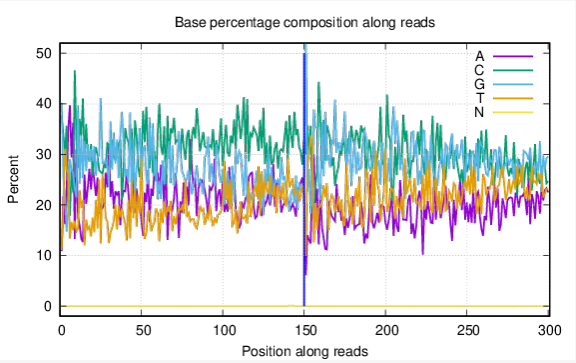


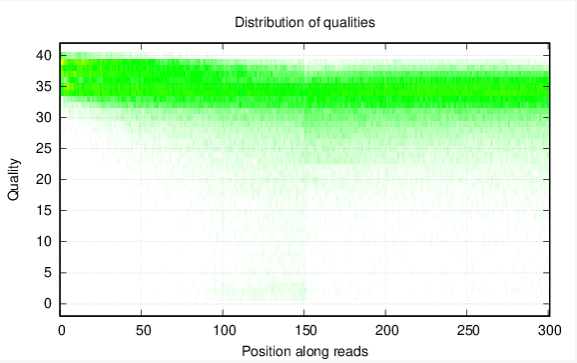
Red_1w



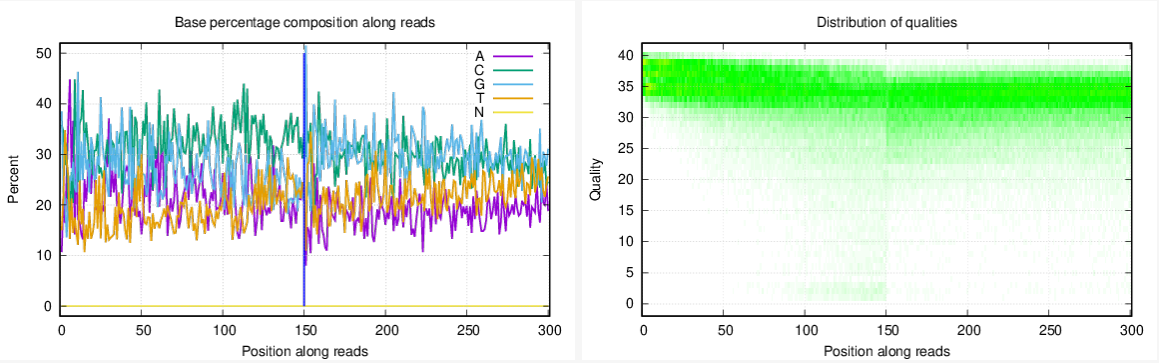


Red_2w

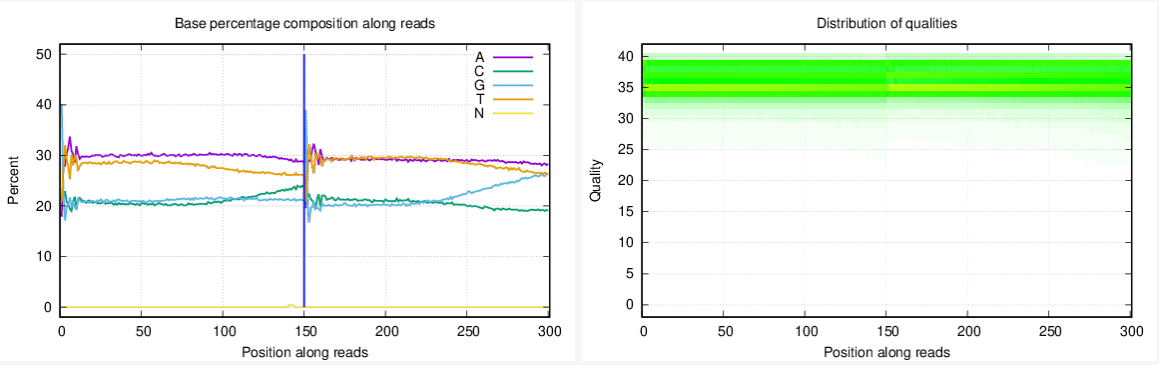




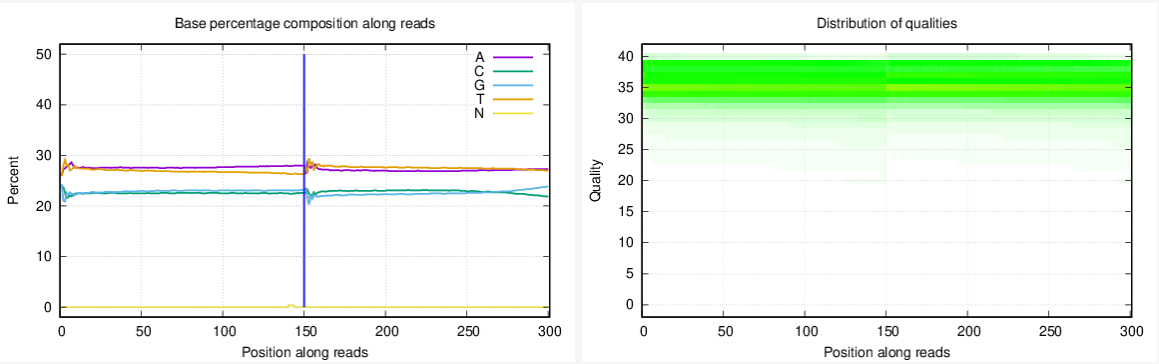
Red_ctr



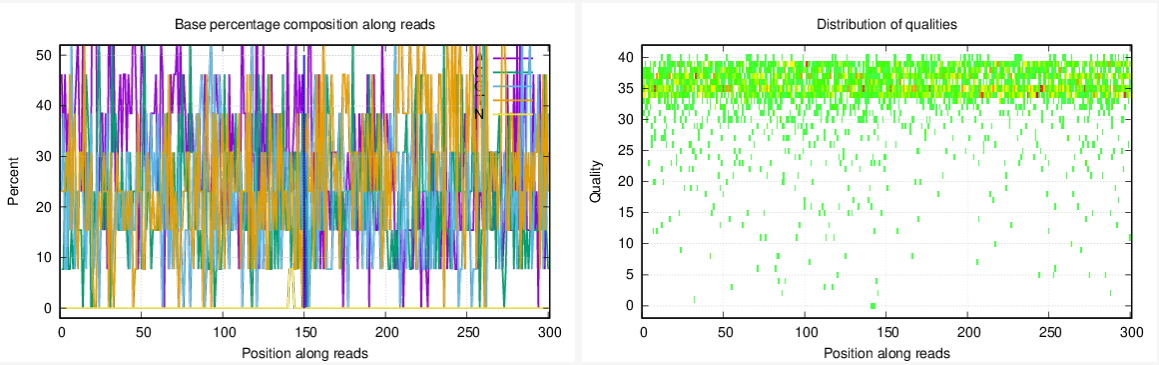
U87



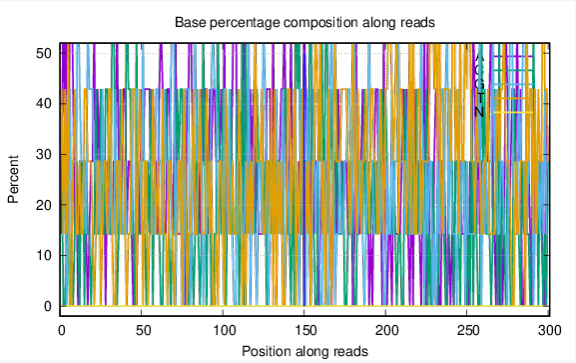
kit30plus-CTCF-rep1

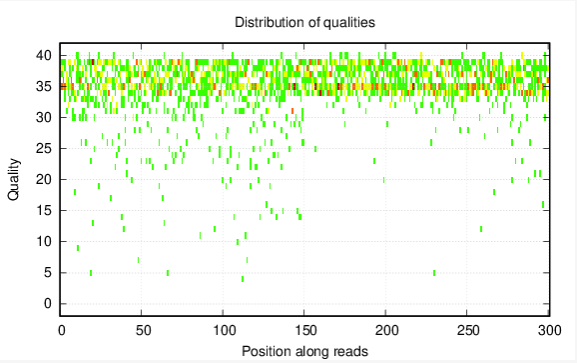


kit30plus-input-rep1



kit30plus-input-rep2





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	?

华大基因
BGI

Omics For All

Shenzhen China (Headquarter)

Address: 9 Yunhua Road, Meisha Street, Yantian District, Shenzhen City, Guangdong Province, China.

Contact: 400-706-6615

E-mail: info@genomics.cn