BGI Sequencing Data Report

2023/2/21

Table of Contents

1 Project Information

Project code: F22FTSEUHT2056_MUSyuazR
Sample number: 16

2 Data Statistics

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

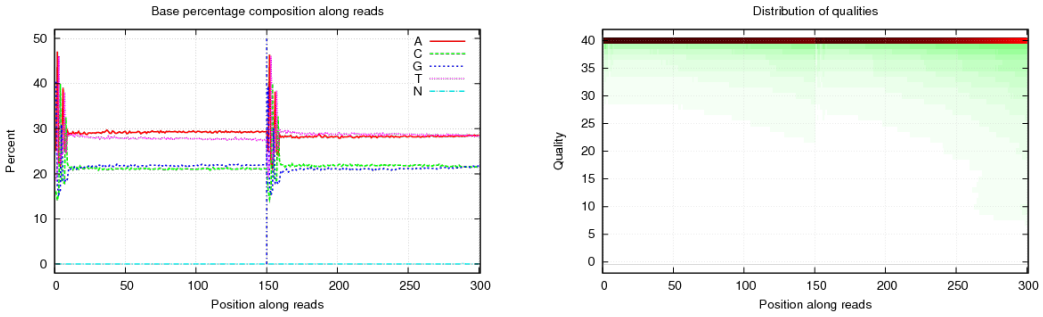
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
B2-15r1	150;150	98.34;96.51	94.79;90.66	42.93;42.97	104,171,171	31,251,351,300
B2-15r2	150;150	98.44;96.54	95.07;90.76	42.65;42.70	131,760,678	39,528,203,400
B2-6r1	150;150	98.39;96.58	94.91;90.82	43.18;43.22	97,903,974	29,371,192,200
B2-6r2	150;150	98.37;96.98	94.83;91.83	43.08;43.12	108,892,952	32,667,885,600
PICO13	150;150	98.07;96.25	94.38;91.58	35.19;35.36	2,164,877	649,463,100
PICO14	150;150	97.49;95.48	92.99;89.95	35.11;35.16	3,655,537	1,096,661,100
PICO15	150;150	97.38;95.81	92.85;90.75	35.86;35.93	4,479,697	1,343,909,100
PICO16	150;150	97.06;94.87	91.90;88.78	35.22;35.45	2,605,232	781,569,600
PICO18	150;150	97.45;95.47	92.91;89.92	33.17;33.26	3,003,832	901,149,600
PICO20	150;150	97.42;95.37	92.99;89.87	33.17;33.22	3,610,752	1,083,225,600
Pico19	150;150	97.13;94.92	92.21;88.96	33.98;34.07	3,218,189	965,456,700
Pico22	150;150	96.40;93.76	90.48;86.50	31.65;31.85	1,747,326	524,197,800
dMR231-A	150;150	98.70;96.99	95.71;91.90	44.21;44.26	101,650,528	30,495,158,400
dMR231-C	150;150	98.66;96.53	95.53;90.59	43.87;43.92	108,355,200	32,506,560,000
dMR70-A	150;150	98.49;96.97	95.18;92.00	44.13;44.18	135,752,021	40,725,606,300
dMR70-C	150;150	98.57;96.87	95.31;91.63	44.33;44.38	110,520,211	33,156,063,300

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

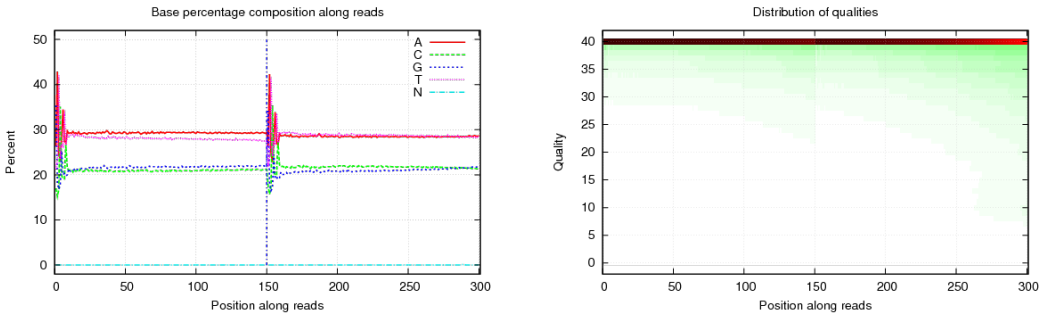
3 Data Quality Control

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.

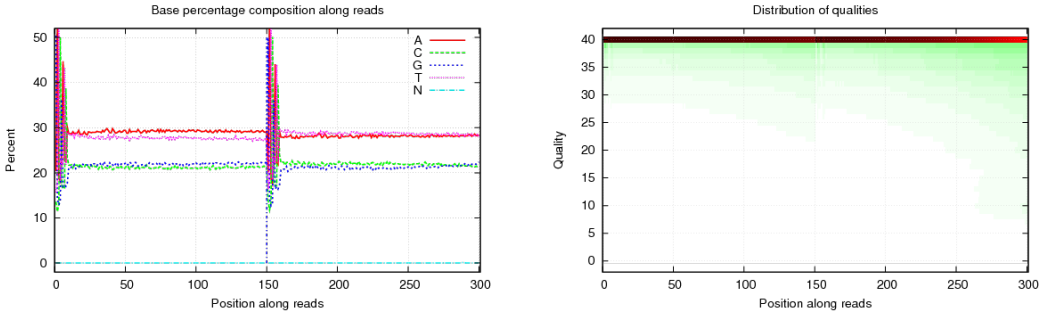
Quality control of sample B2-15r1



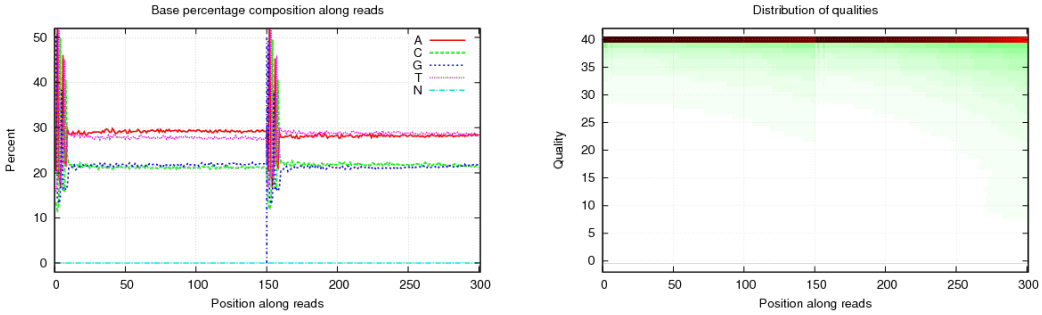
Quality control of sample B2-15r2



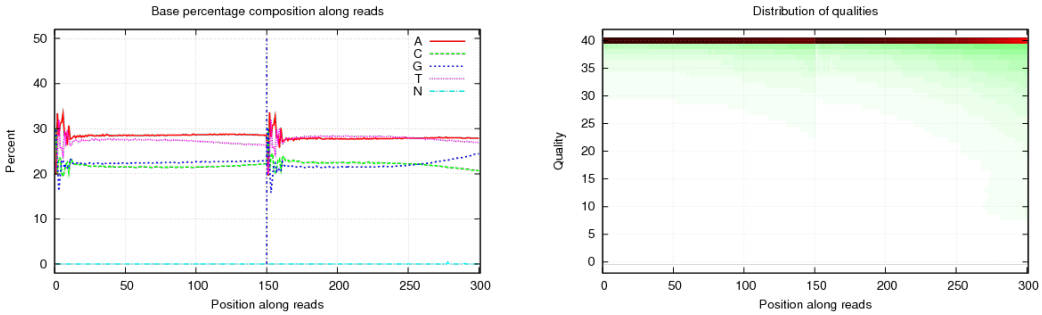
Quality control of sample B2-6r1



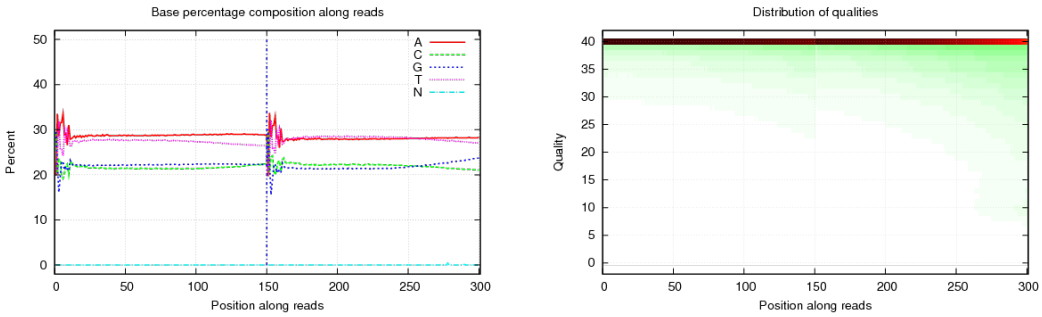
Quality control of sample B2-6r2



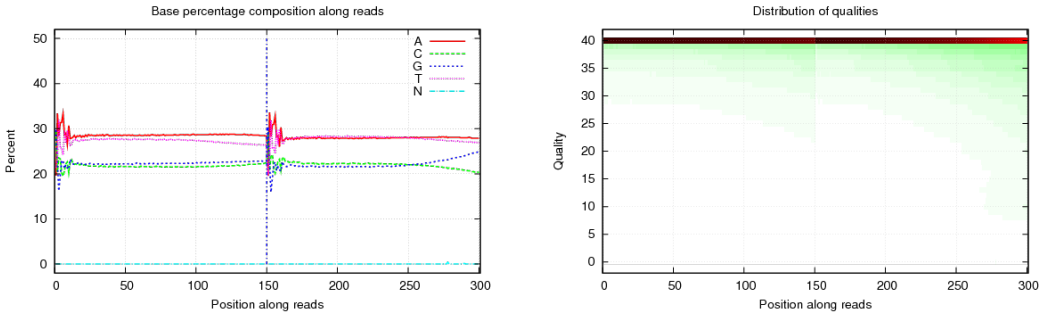
Quality control of sample dMR231-A



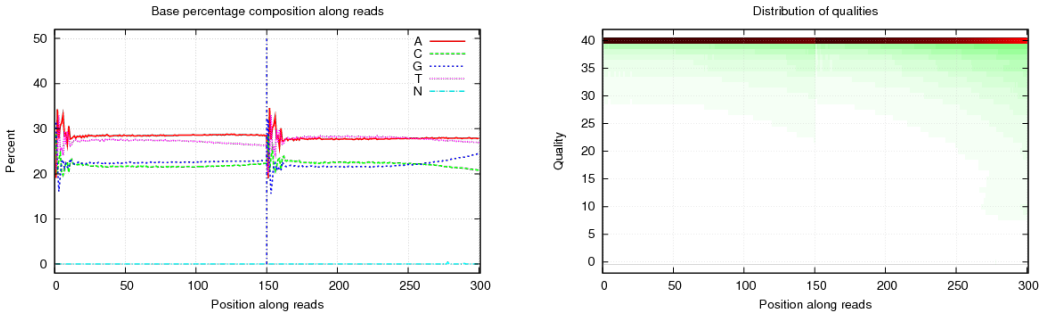
Quality control of sample dMR231-C



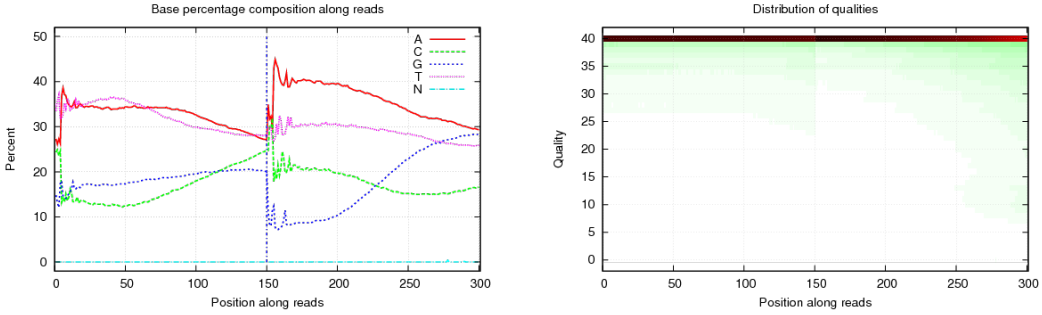
Quality control of sample dMR70-A



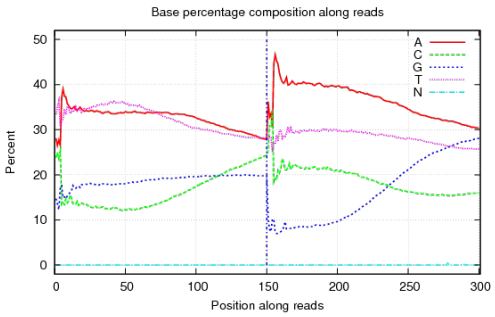
Quality control of sample dMR70-C

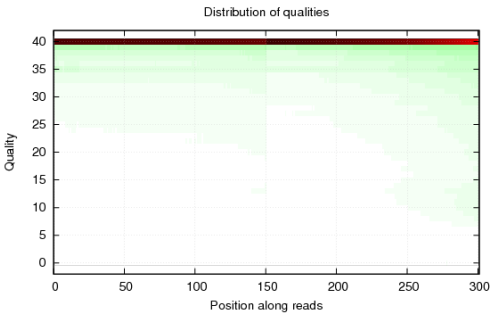


Quality control of sample PICO13

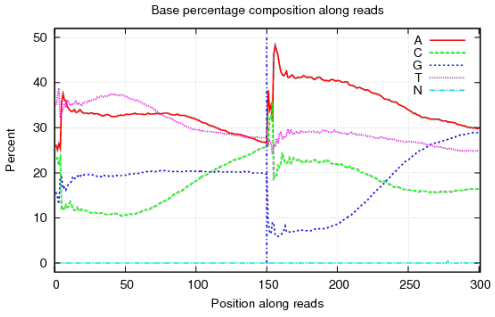


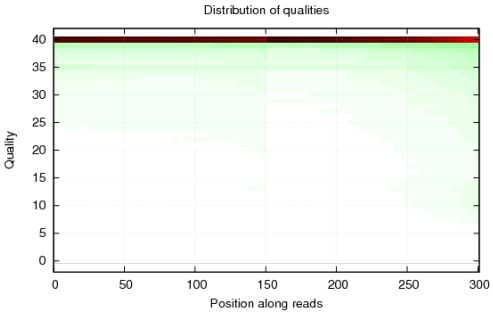
Quality control of sample PICO14



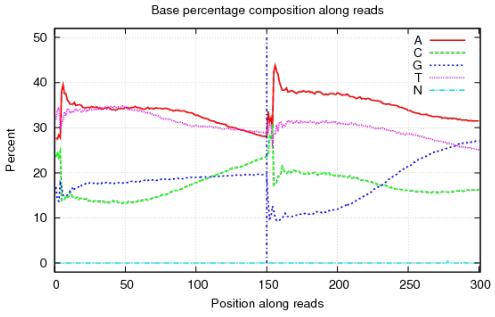


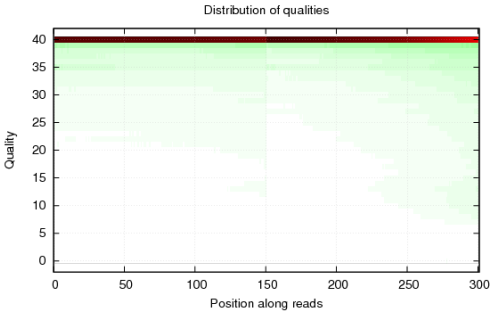
Quality control of sample PICO15



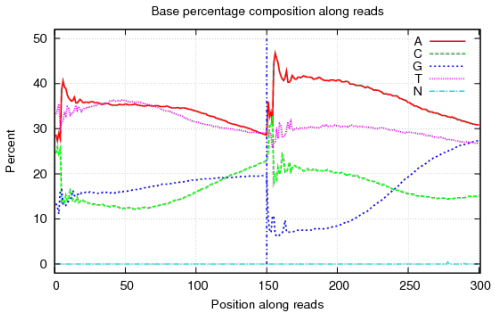


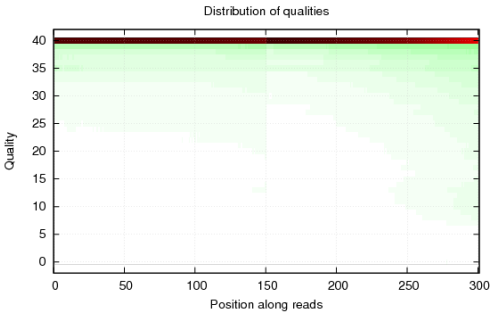
Quality control of sample PICO16



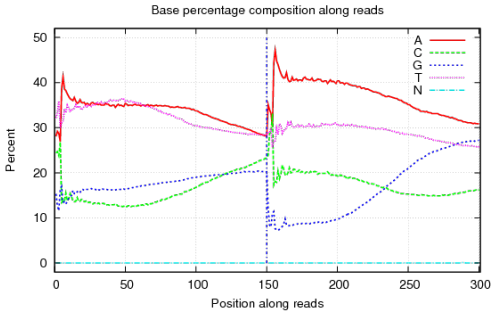


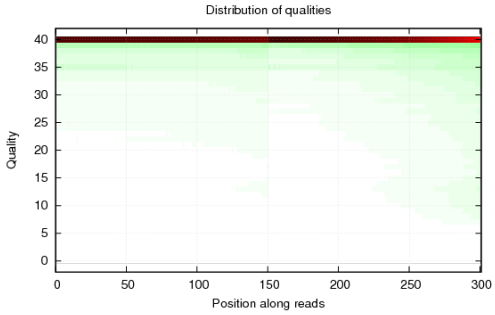
Quality control of sample PICO18



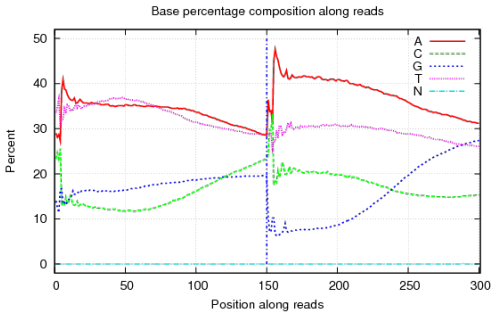


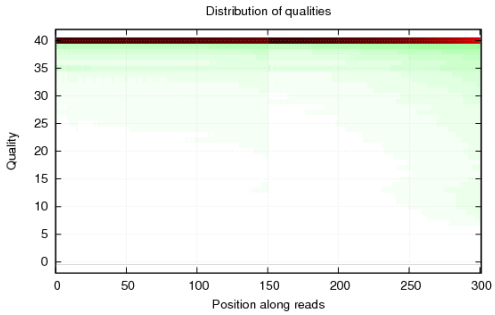
Quality control of sample Pico19



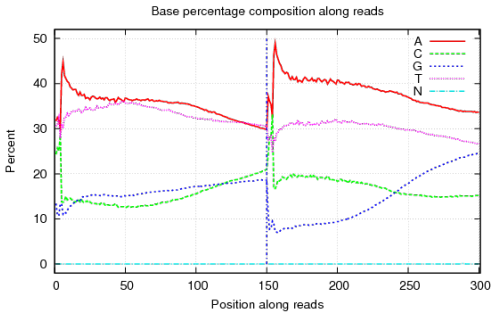


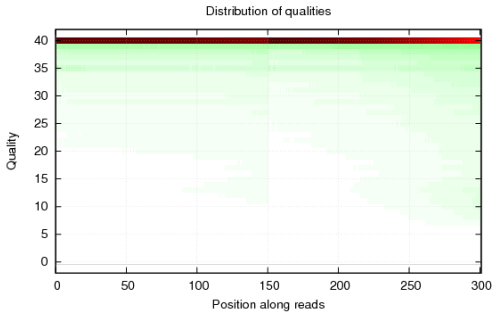
Quality control of sample PICO20





Quality control of sample Pico22





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	?