BGI Sequencing Data Report

2023/10/27

Table of Contents

1 Project Information

Project code: F23A430000924_RIPzryuR_6
Sample number: 6

2 Data Statistics

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

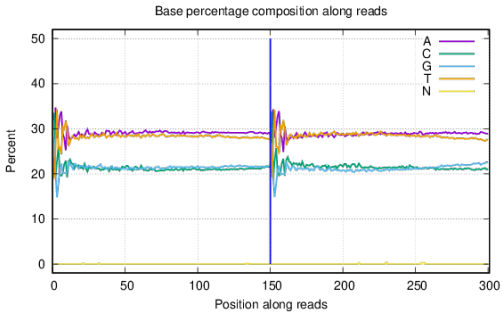
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
BOC_K1	150;150	97.94;96.00	93.55;90.10	43.03;43.18	250,419	75,125,700
BOC_e1	150;150	97.68;95.51	92.75;88.54	42.67;42.77	322,184	96,655,200
CHR_K1	150;150	97.62;95.37	92.55;88.03	43.04;43.14	275,904	82,771,200
CHR_e1	150;150	97.63;95.82	92.65;89.48	42.80;42.94	537,024	161,107,200
XAH_K13	150;150	97.84;95.94	93.25;89.94	42.84;42.97	275,546	82,663,800
XAH_e13	150;150	97.66;95.73	92.70;89.22	42.64;42.74	375,828	112,748,400

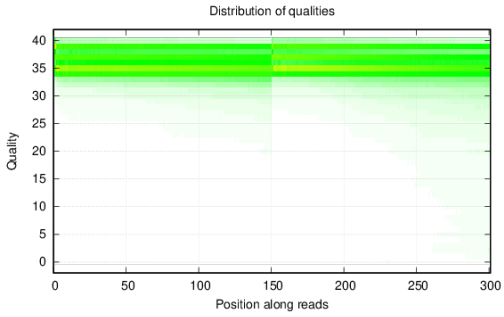
Table Format:
1. Sample: The name of sample
2. Length: The Length of reads
3. Q20 (%): The proportion of nucleotides with quality value larger than 20
4. Q30 (%): The proportion of nucleotides with quality value larger than 30
4. GC Content(%): The proportion of bases G and C
5. Total Reads: The total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

3 Data Quality Control

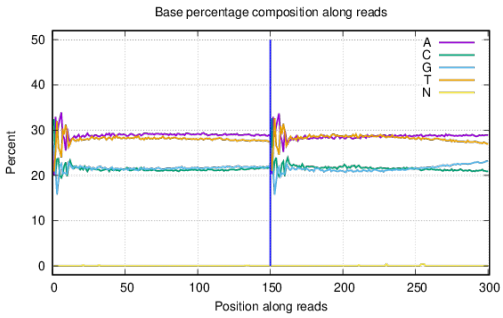
The distribution of base percentage and qualities along reads in data filtering are shown as following(If a sample has multiple lanes, only one of them will be displayed). The left picture is base percentage distribution along reads the sample, the right picture is distribution of qualities along reads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

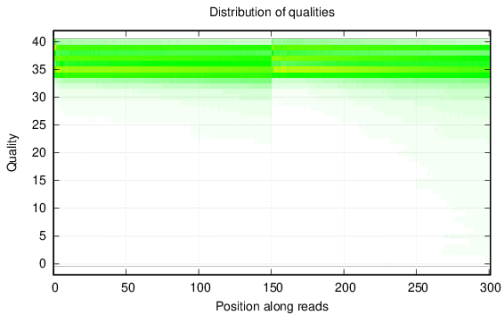
Quality control of sample BOC_e1



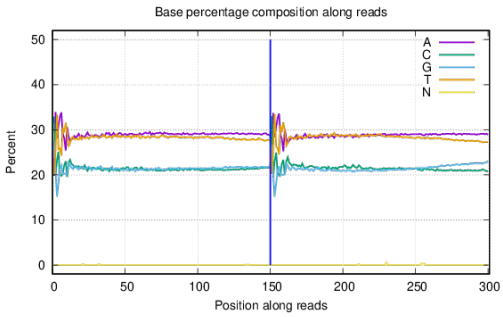


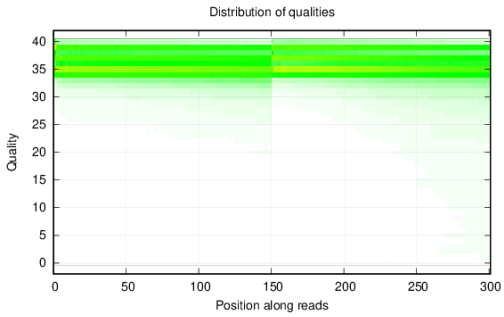
Quality control of sample BOC_K1



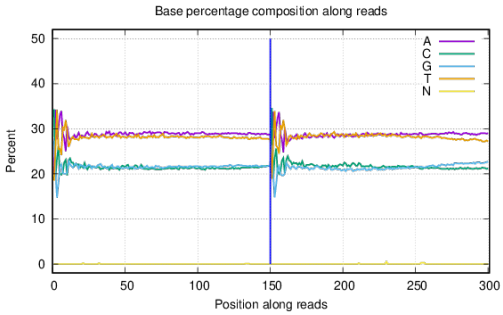


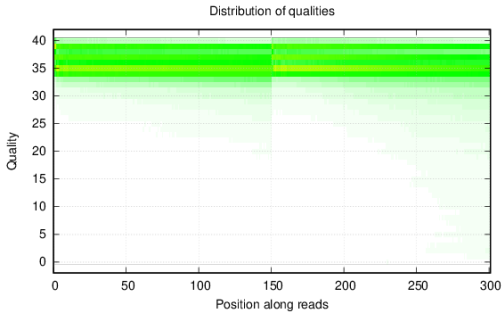
Quality control of sample CHR_e1



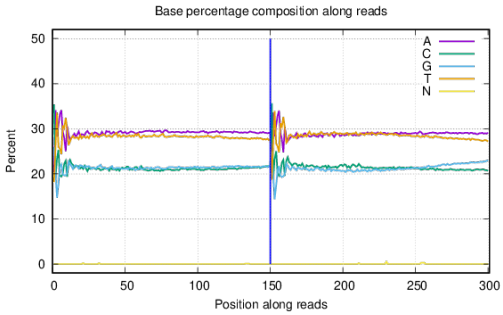


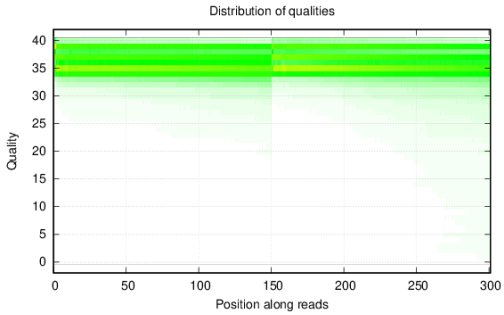
Quality control of sample CHR_K1



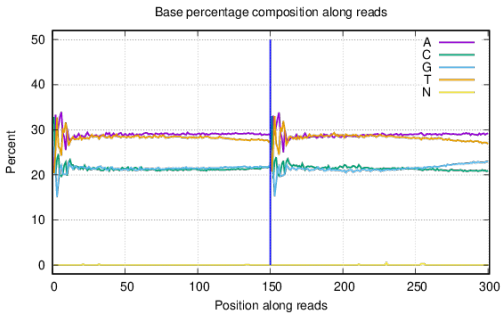


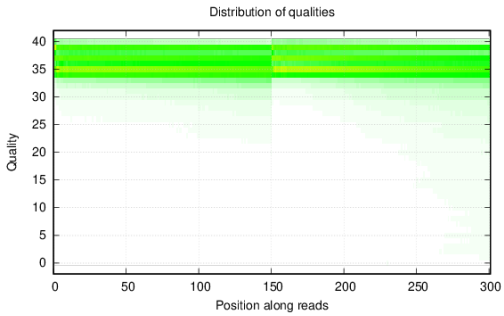
Quality control of sample XAH_e13





Quality control of sample XAH_K13





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	?