BGI Sequencing Data Report

2023/9/19

Table of Contents

1 Project Information

Project code: F20FTSEUHT0946_02_MUSiarzR
Sample number: 39

2 Data Statistics

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

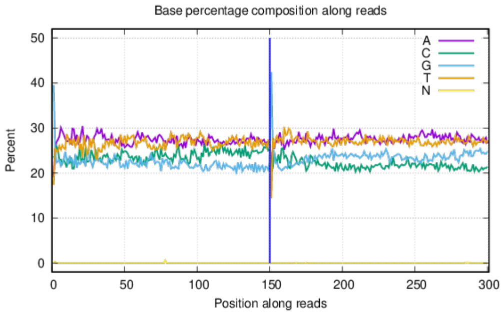
Sample	Length	Q20(%)	Q30(%)	GC Content(%)	Total Reads	Total Bases
DA1	150;150	98.00;95.94	93.42;87.22	44.50;44.37	71,459	21,437,700
DN1	150;150	97.93;96.50	93.28;88.94	44.50;44.40	38,918	11,675,400
Dik_K1	150;150	98.10;96.95	93.76;90.07	42.66;42.65	261,265	78,379,500
Dik_e1	150;150	98.05;97.01	93.65;90.25	42.13;42.14	551,365	165,409,500
EA1	150;150	95.79;92.72	86.84;76.75	42.91;49.22	581	174,300
EA10	150;150	95.64;94.54	86.88;82.60	43.60;50.11	618	185,400
EA11	150;150	93.92;94.21	82.34;81.61	43.42;49.33	557	167,100
EA2	150;150	95.98;94.67	87.73;82.52	44.04;49.82	1,004	301,200
EA3	150;150	94.74;94.49	84.49;81.80	43.27;49.12	860	258,000
EA4	150;150	95.25;93.27	85.37;78.79	42.85;50.26	621	186,300
EA5	150;150	93.81;94.85	82.56;83.00	44.12;49.64	245	73,500
EA6	150;150	93.66;93.53	81.32;78.88	42.84;49.52	594	178,200
EA7	150;150	93.67;94.03	81.69;80.35	42.80;49.62	605	181,500
EA8	150;150	96.87;97.09	90.61;90.65	45.74;48.82	1,115	334,500
EA9	150;150	96.33;95.24	88.04;84.35	43.32;49.68	1,218	365,400
In1	150;150	96.06;96.45	90.42;89.28	54.66;53.71	34,601	10,380,300
In2	150;150	96.10;95.96	90.86;87.99	59.00;57.93	19,411	5,823,300
K562-102	150;150	98.40;98.86	94.91;96.18	42.51;42.49	83,952,922	25,185,876,600
K562-104	150;150	98.12;98.22	93.94;93.97	41.53;41.46	65,666,464	19,699,939,200
K562S1-1000_r1	150;150	97.85;96.63	93.02;89.13	47.66;47.59	24,521,892	7,356,567,600
K562S1-1000_r2	150;150	97.91;96.15	93.17;87.62	46.71;46.59	19,110,823	5,733,246,900
K562S1-10_r1	150;150	97.98;97.68	93.54;92.37	42.53;42.46	4,581,014	1,374,304,200
K562S1-10_r2	150;150	97.89;97.88	93.34;93.06	43.04;42.99	3,702,901	1,110,870,300
K562S1-200_r2	150;150	98.12;97.59	93.89;92.06	42.55;42.48	8,919,287	2,675,786,100
K562S1-500_r1	150;150	98.13;97.65	93.88;92.19	41.95;41.87	14,122,482	4,236,744,600
K562S1-500_r2	150;150	98.24;97.95	94.24;93.15	42.40;42.31	15,976,232	4,792,869,600
LS6N-16	150;150	93.54;89.58	81.61;63.45	53.53;52.56	2,408	722,400
Ore1_K	150;150	98.11;96.87	93.82;89.92	44.56;44.51	8,120,416	2,436,124,800
P114-0	150;150	98.11;97.77	93.84;92.58	44.87;44.81	19,077,680	5,723,304,000
P139-0	150;150	97.83;96.86	92.88;89.75	43.17;43.09	12,881,653	3,864,495,900
Pct_1	150;150	98.18;98.21	94.18;94.06	43.11;43.71	5,031,578	1,509,473,400
Pct_2	150;150	98.17;98.07	94.11;93.66	43.10;43.68	5,662,057	1,698,617,100
S.int6	150;150	94.42;90.15	83.76;65.78	51.19;50.30	1,750	525,000
TAF16	150;150	98.18;97.82	94.04;92.68	41.65;41.59	71,106,892	21,332,067,600
Ton1_K	150;150	97.95;98.05	93.50;93.51	38.56;37.89	666	199,800
ch1	150;150	96.92;96.64	91.69;89.82	46.08;45.40	60,554	18,166,200
ch2	150;150	94.58;96.04	87.23;88.23	45.41;44.05	14,902	4,470,600
dm6-1	150;150	94.45;89.98	84.06;65.26	51.69;50.43	3,449	1,034,700
iP65	150;150	98.17;97.93	94.06;93.01	40.56;40.47	45,835,460	13,750,638,000

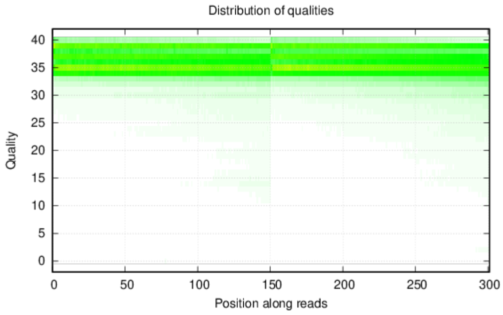
Table Format:
1. Sample: The name of sample
2. Length: The Length of reads
3. Q20 (%): The proportion of nucleotides with quality value larger than 20
4. Q30 (%): The proportion of nucleotides with quality value larger than 30
4. GC Content(%): The proportion of bases G and C
5. Total Reads: The total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

3 Data Quality Control

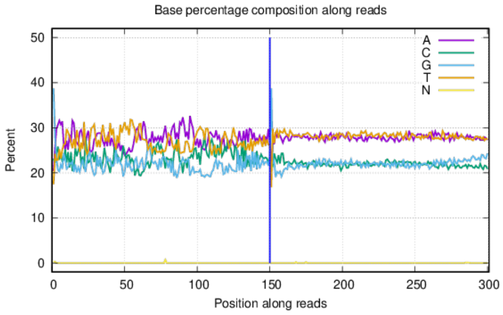
The distribution of base percentage and qualities along reads in data filtering are shown as following(If a sample has multiple lanes, only one of them will be displayed). The left picture is base percentage distribution along reads the sample, the right picture is distribution of qualities along reads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

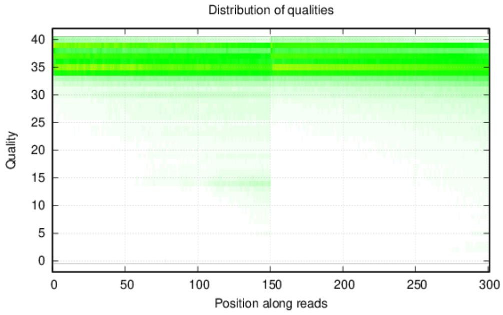
Quality control of sample ch1



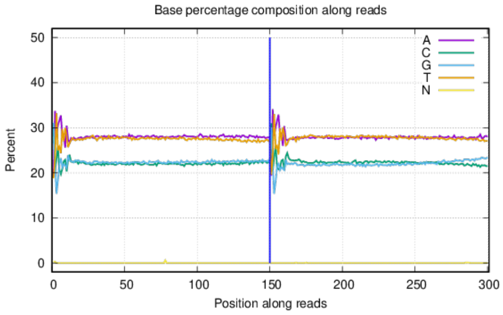


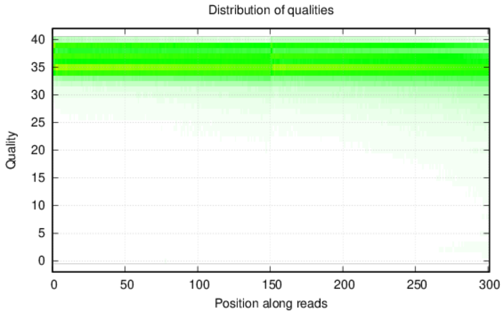
Quality control of sample ch2



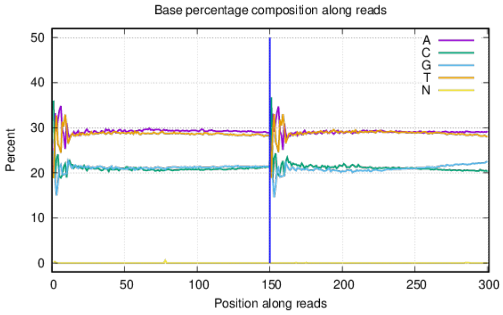


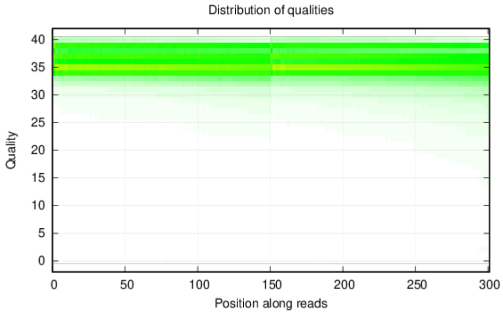
Quality control of sample DA1



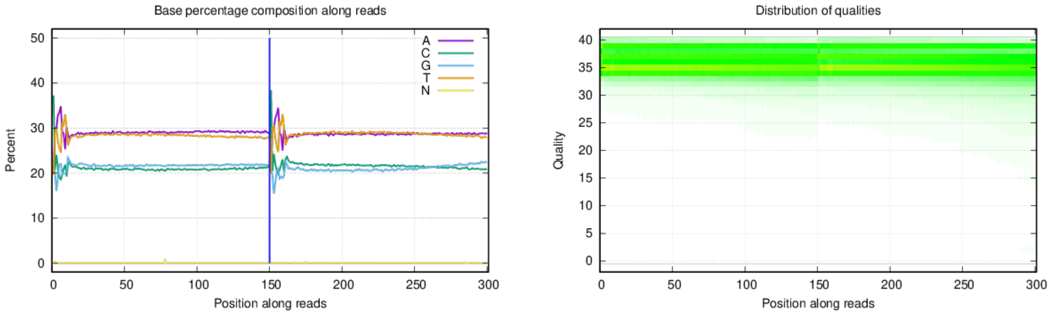


Quality control of sample Dik_e1

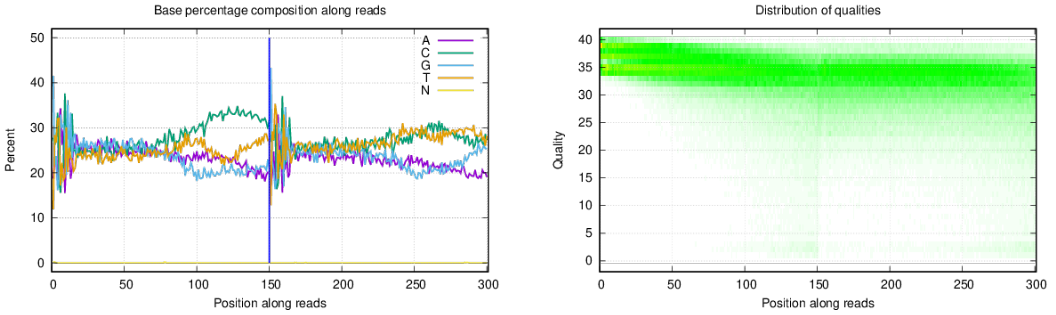




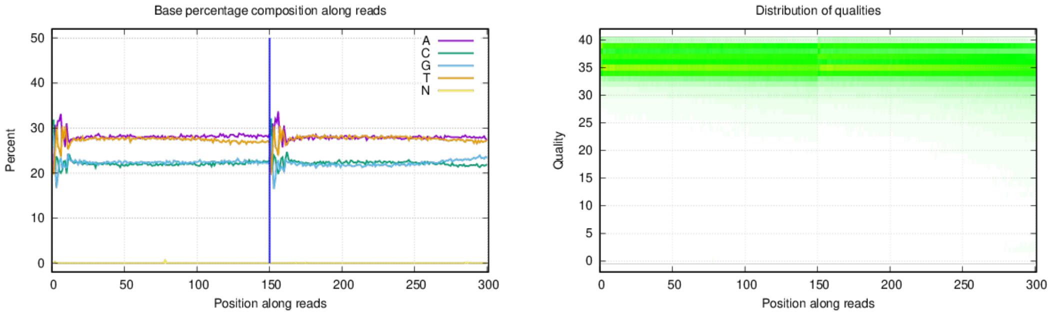
Quality control of sample Dik_K1



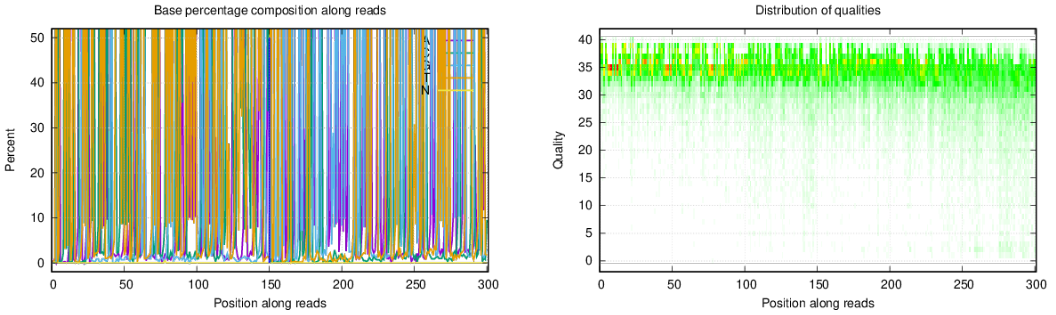
Quality control of sample dm6-1



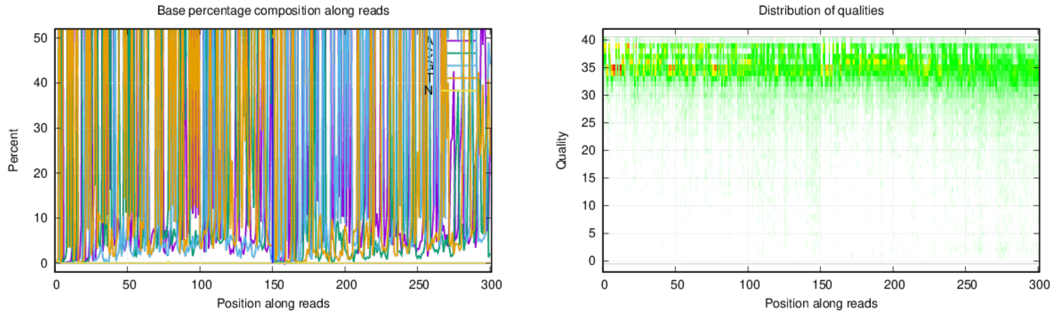
Quality control of sample DN1



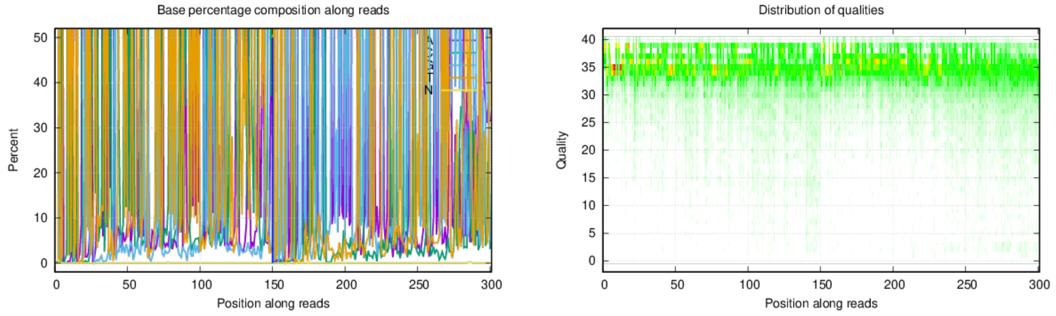
Quality control of sample EA1



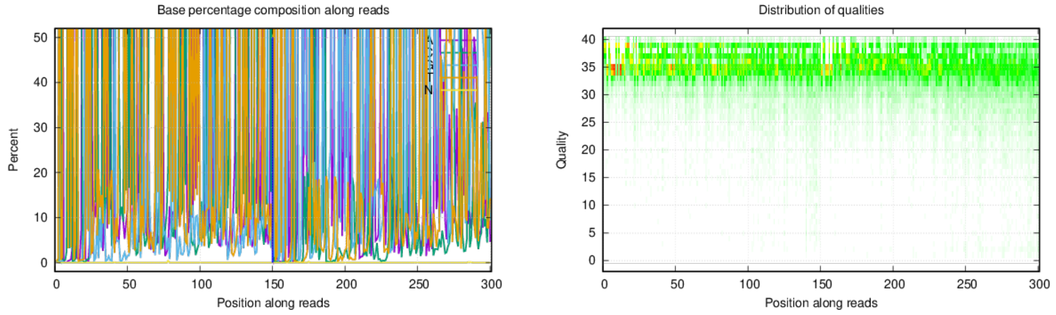
Quality control of sample EA10



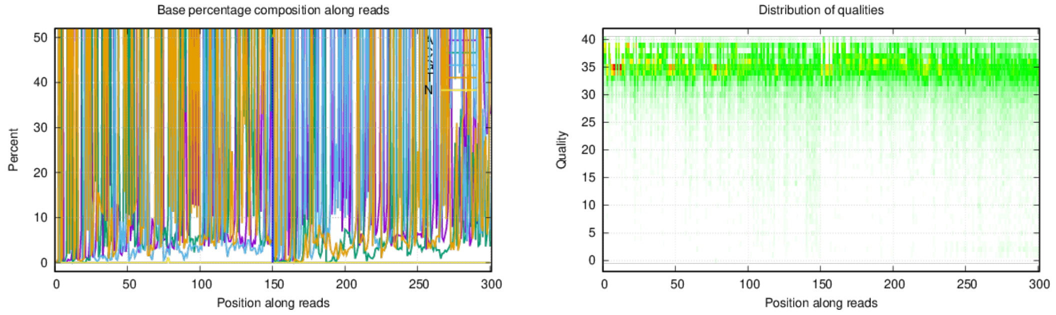
Quality control of sample EA11



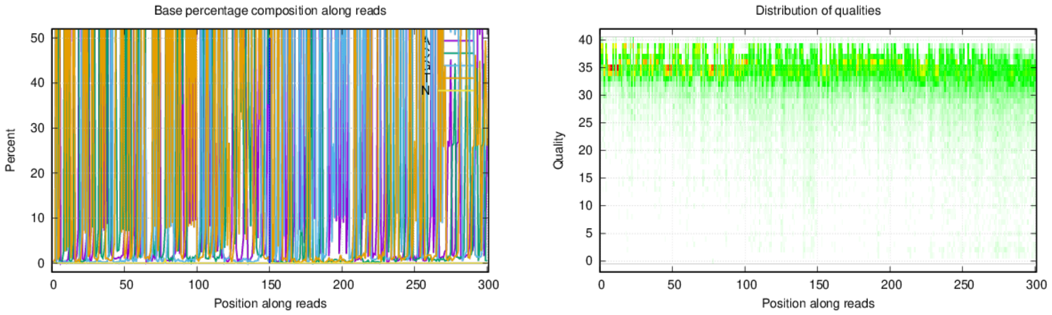
Quality control of sample EA2



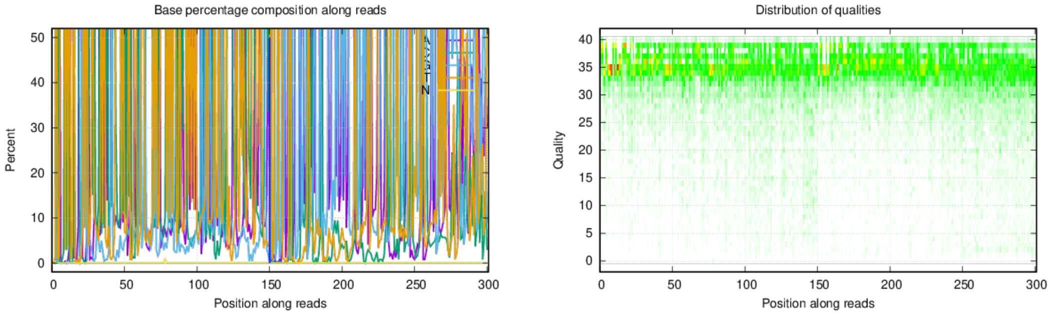
Quality control of sample EA3



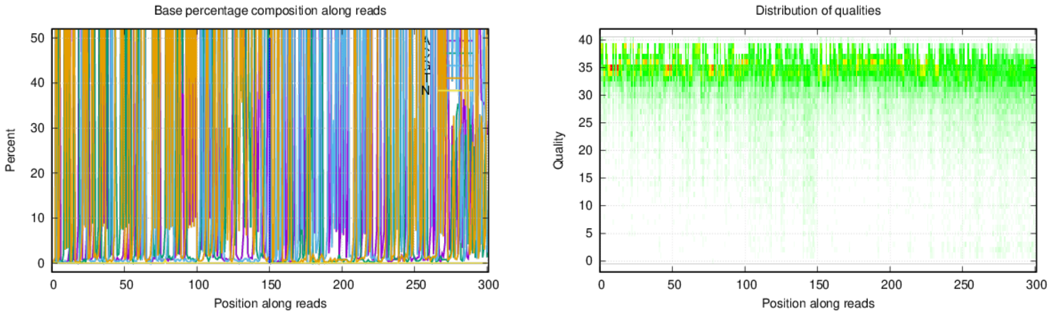
Quality control of sample EA4



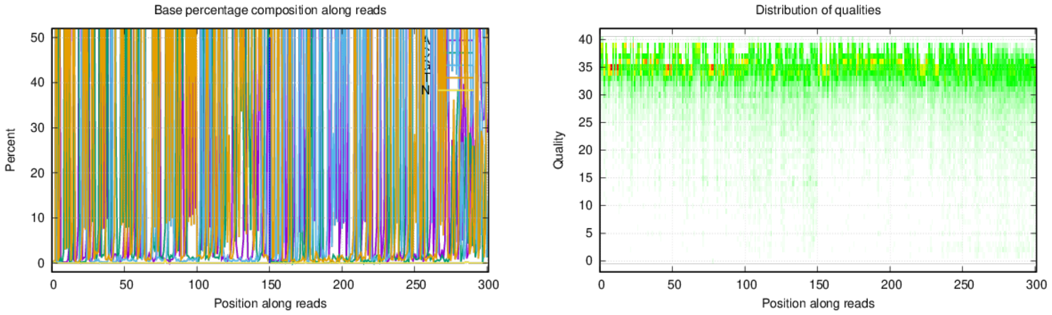
Quality control of sample EA5



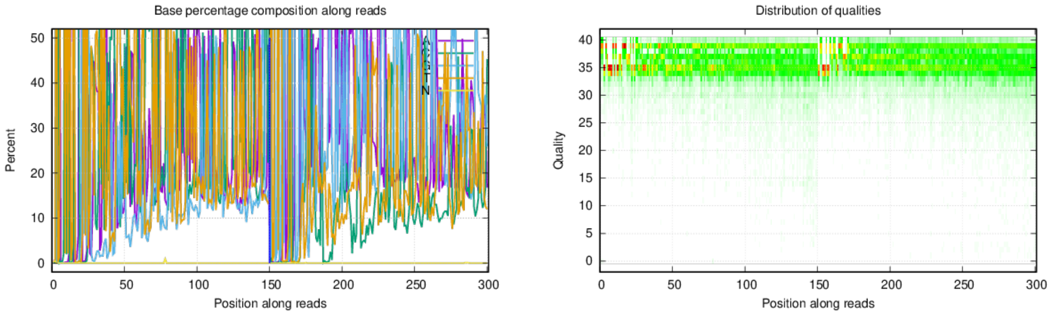
Quality control of sample EA6



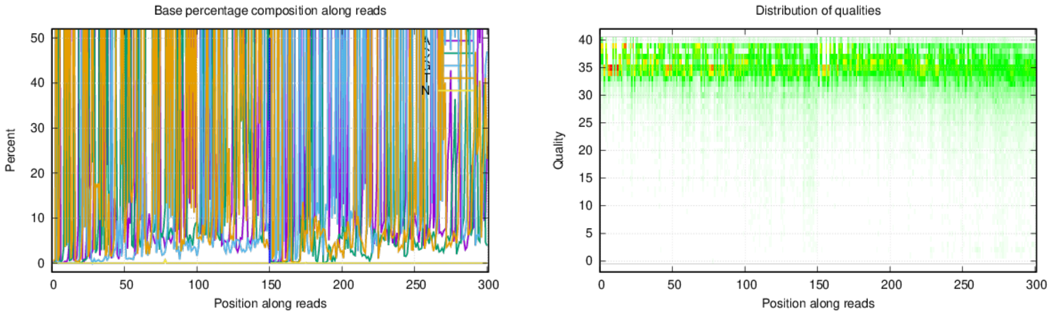
Quality control of sample EA7



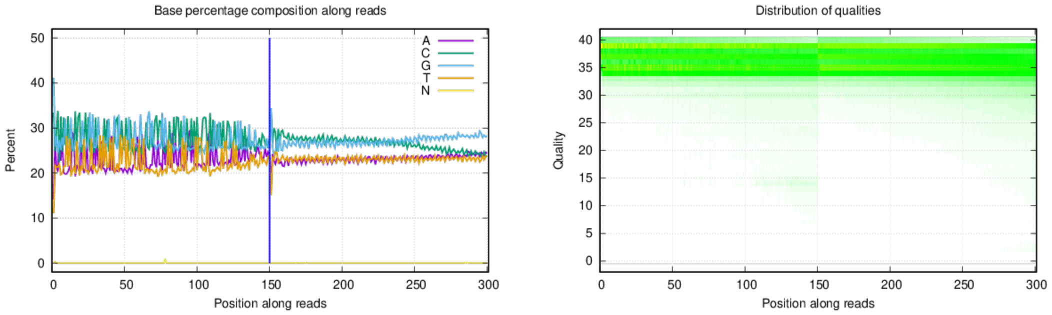
Quality control of sample EA8



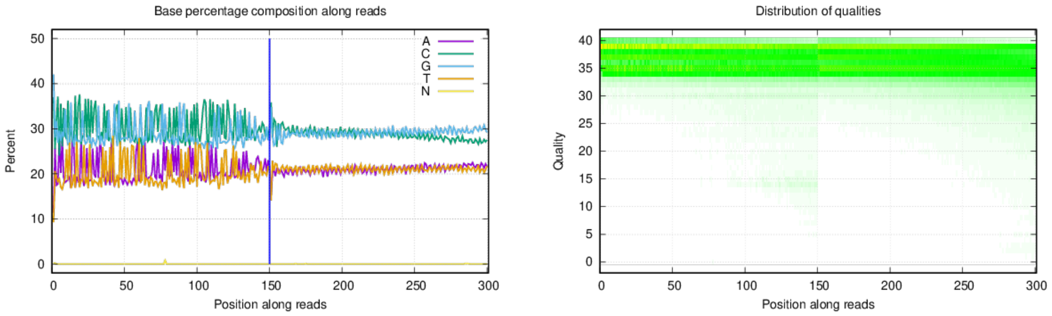
Quality control of sample EA9



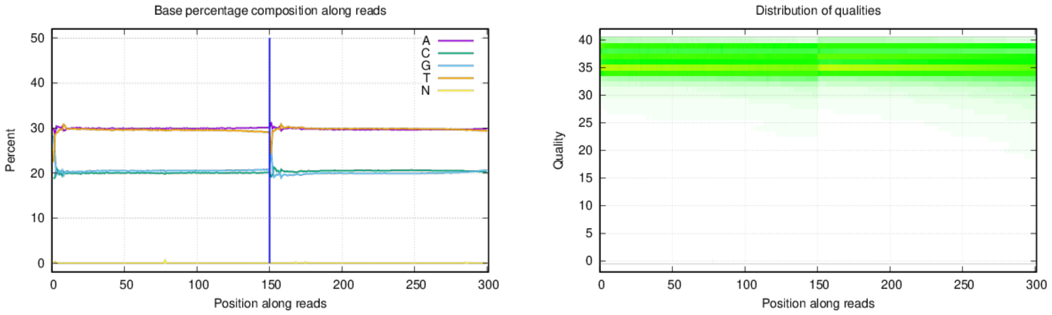
Quality control of sample In1



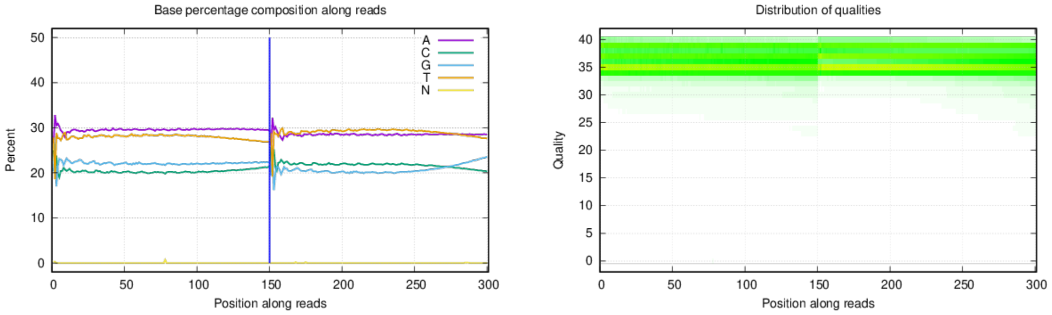
Quality control of sample In2



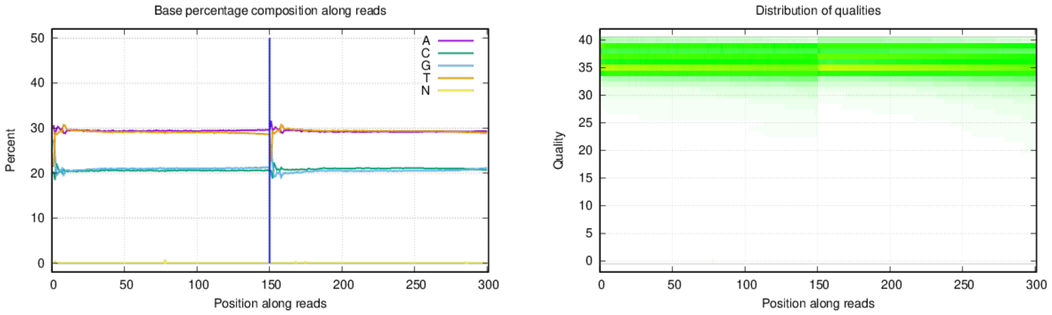
Quality control of sample iP65



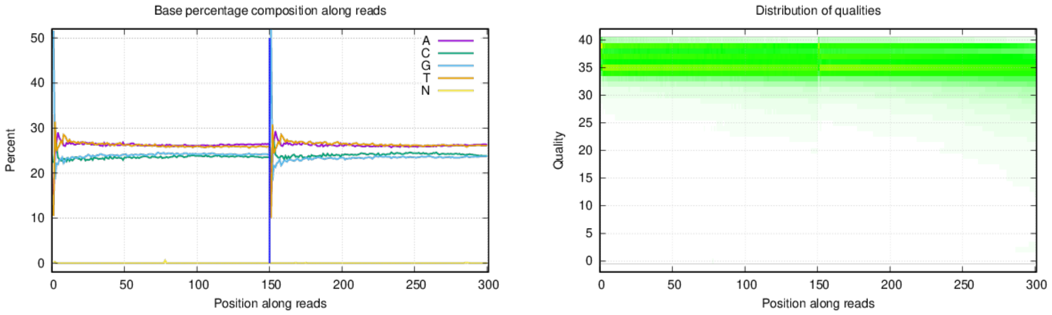
Quality control of sample K562-102



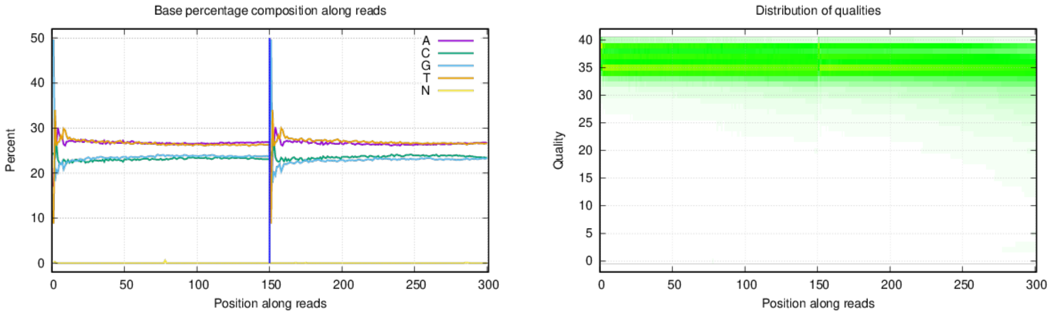
Quality control of sample K562-104



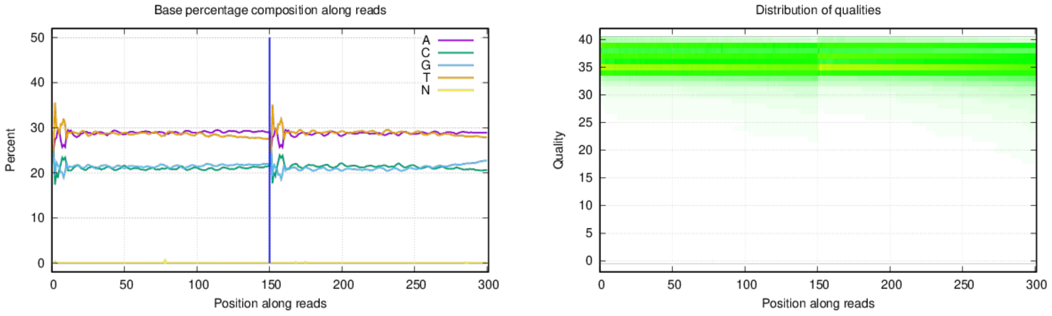
Quality control of sample K562S1-1000_r1



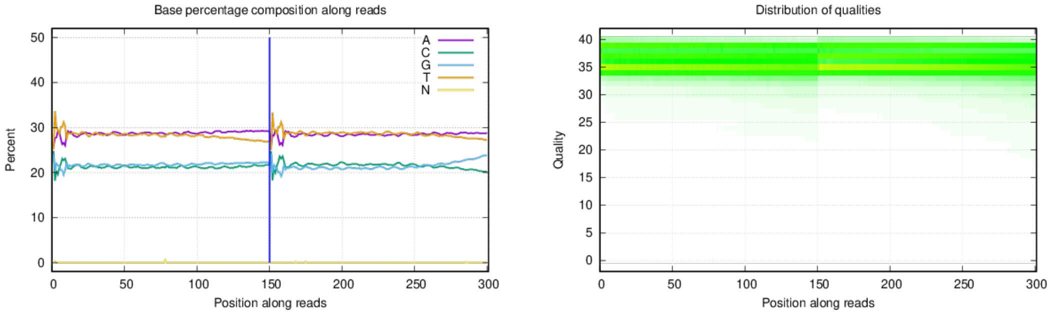
Quality control of sample K562S1-1000_r2



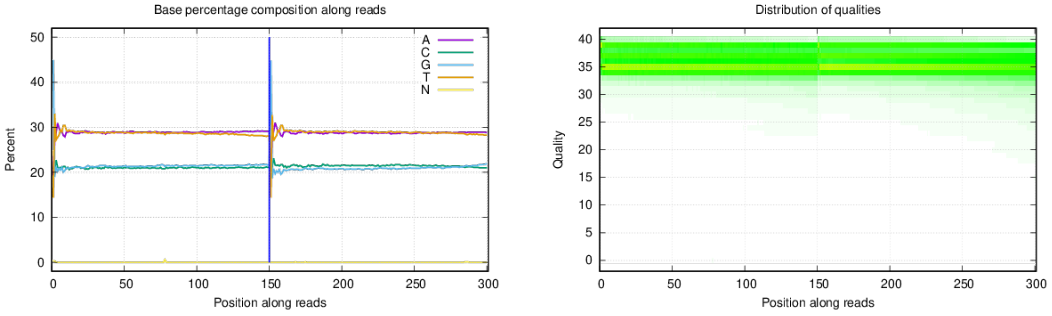
Quality control of sample K562S1-10_r1



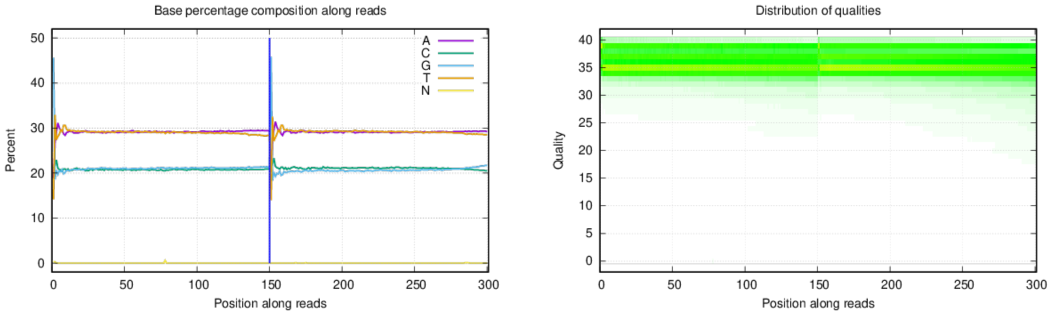
Quality control of sample K562S1-10_r2



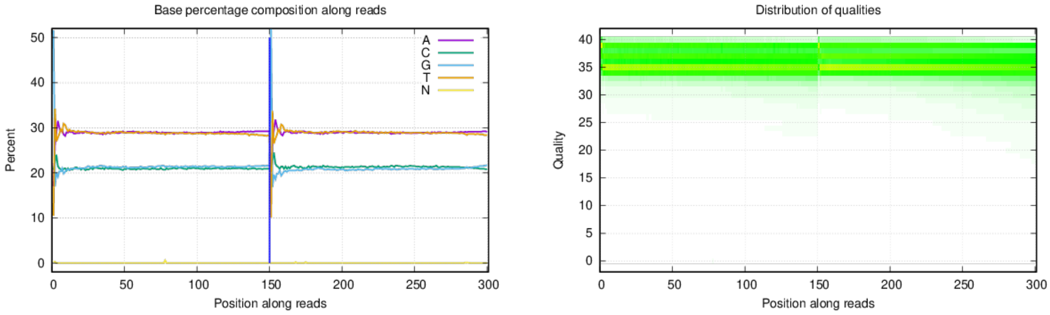
Quality control of sample K562S1-200_r2



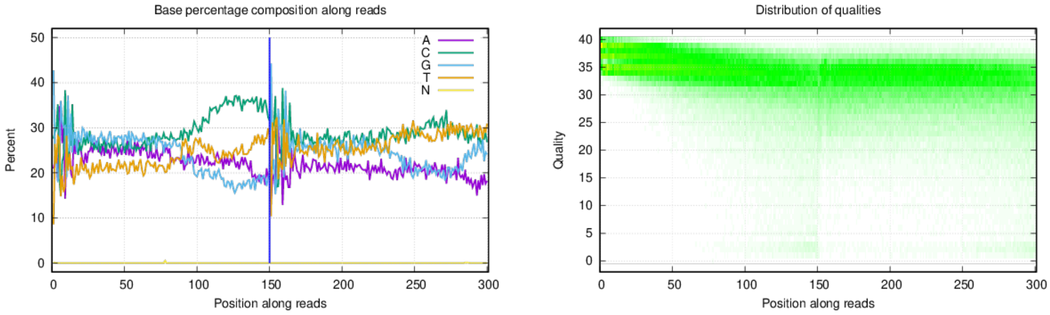
Quality control of sample K562S1-500_r1



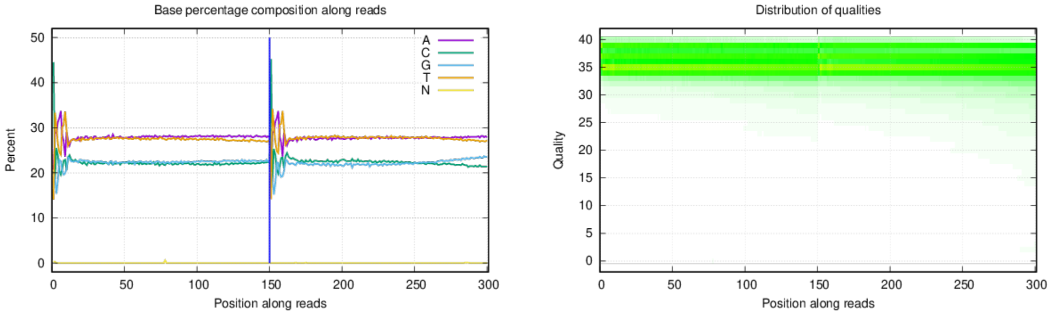
Quality control of sample K562S1-500_r2



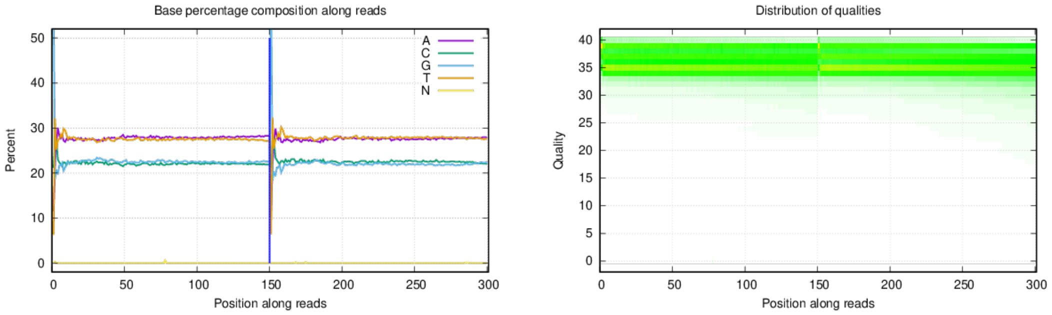
Quality control of sample LS6N-16



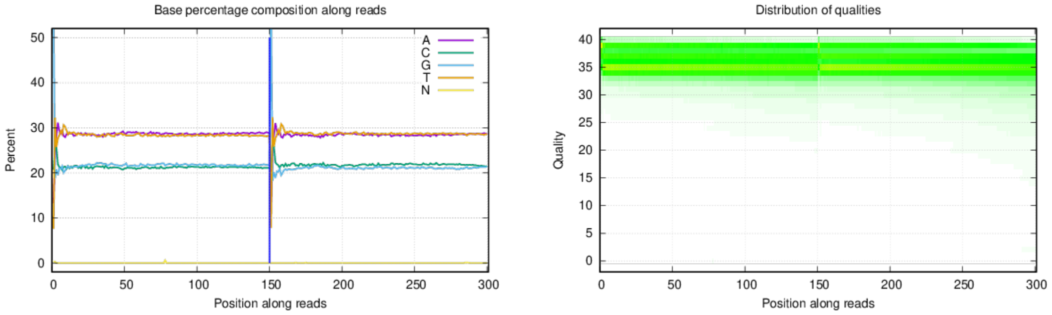
Quality control of sample Ore1_K



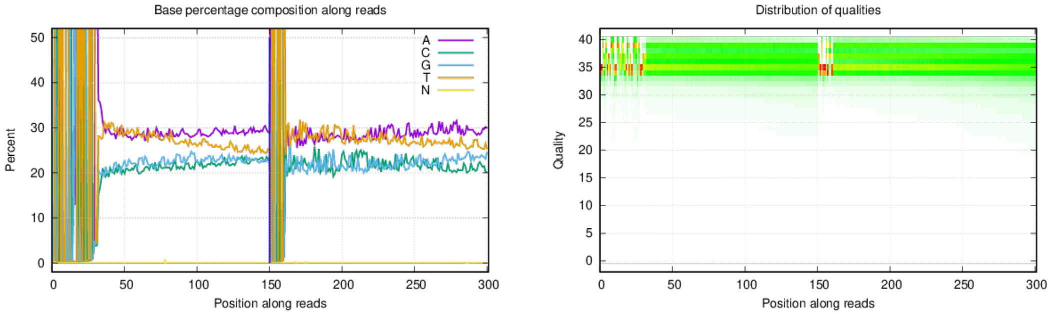
Quality control of sample P114-0



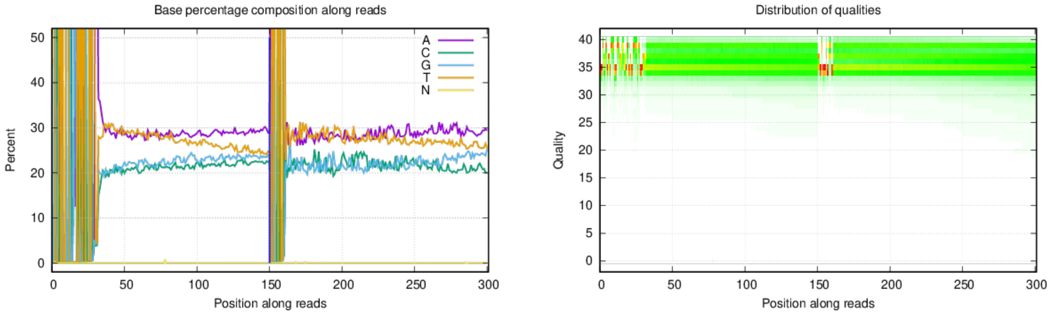
Quality control of sample P139-0



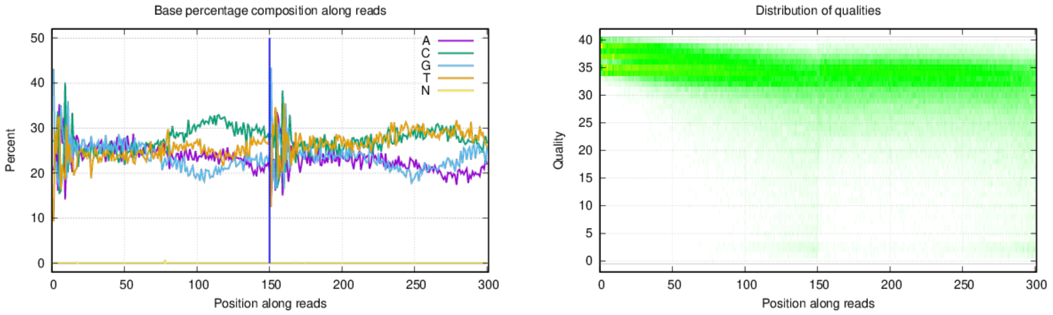
Quality control of sample Pct_1



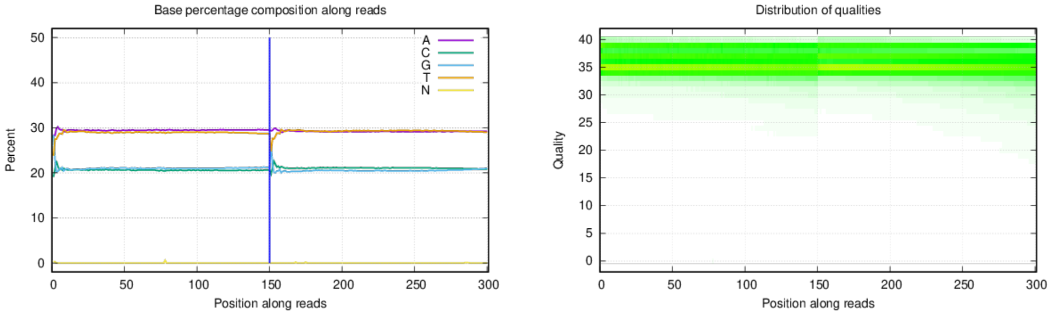
Quality control of sample Pct_2



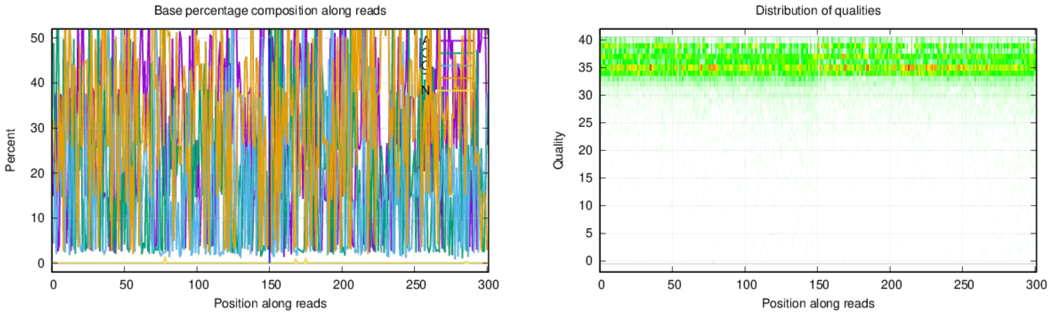
Quality control of sample S.int6



Quality control of sample TAF16



Quality control of sample Ton1_K



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	?