BGI Sequencing Data Report

2023/11/16

Table of Contents

1 Project Information

Project code: F20FTSEUHT0946-02_HOMukzaR

Sample number: 37

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
3	150;150	98.48;98.51	95.49;95.72	45.71;45.81	45,813,376	13,744,012,800
Boc_K4	150;150	97.37;96.32	91.85;88.43	42.66;42.64	520,701	156,210,300
Boc_e4	150;150	97.31;95.38	91.48;85.42	42.36;42.40	472,672	141,801,600
N10	150;150	93.89;93.86	84.19;84.82	44.30;44.27	110	33,000
N11	150;150	92.96;92.41	82.03;80.61	48.07;48.22	46	13,800
N14	150;150	97.49;97.60	92.92;92.87	48.93;48.95	810	243,000
N15	150;150	98.19;98.43	95.29;95.43	44.43;43.90	14	4,200
N16	150;150	94.58;96.08	85.92;88.67	47.08;48.67	8	2,400
N35	150;150	98.86;98.86	96.38;97.33	47.62;47.52	7	2,100
N38	150;150	93.66;92.32	82.67;80.22	44.97;44.53	113	33,900
N9	150;150	97.93;97.71	94.17;92.94	49.28;48.58	451	135,300
Ng019	150;150	94.89;96.43	85.44;88.79	44.28;44.20	63,199	18,959,700
Ng025	150;150	97.73;97.21	92.95;91.26	40.73;40.59	82,524	24,757,200
P137	150;150	97.49;95.99	92.03;87.33	42.13;42.05	24,700,944	7,410,283,200
P140	150;150	97.70;97.62	93.00;92.70	46.25;46.08	16,836,366	5,050,909,800
P141	150;150	97.69;97.35	93.02;92.02	49.48;49.27	18,022,641	5,406,792,300
P144	150;150	98.18;97.81	94.43;93.38	47.45;47.48	23,122,866	6,936,859,800
P148	150;150	98.34;97.87	94.94;93.66	48.04;48.09	33,942,659	10,182,797,700
P150	150;150	98.25;97.80	94.66;93.39	48.29;48.34	21,516,721	6,455,016,300
P165	150;150	97.73;97.57	93.15;92.64	48.58;48.39	19,540,457	5,862,137,100
P166	150;150	98.34;97.75	94.97;93.46	50.31;50.33	38,839,568	11,651,870,400
P168	150;150	97.98;97.58	93.81;92.71	49.44;49.43	25,156,544	7,546,963,200
Shur1_e	150;150	95.73;95.26	88.68;87.35	44.74;44.91	39	11,700
Ton1_K	150;150	97.73;95.68	93.09;87.16	54.47;54.46	5,299,721	1,589,916,300
Zap_K2	150;150	97.16;95.14	91.02;84.65	42.99;42.96	670,502	201,150,600
Zap_e2	150;150	97.37;95.39	91.63;85.44	42.37;42.40	487,460	146,238,000
Zap_e3	150;150	97.24;95.40	91.32;85.54	42.29;42.31	375,462	112,638,600
mel-kit30-CTCF-rep1	150;150	97.52;97.42	92.55;92.13	47.11;47.08	21,866,641	6,559,992,300
mel-kit30-CTCF-rep2	150;150	97.37;97.33	92.08;91.79	46.43;46.45	20,445,650	6,133,695,000
mel-kit30-input-rep1	150;150	97.19;97.70	91.63;92.88	42.20;42.19	25,515,220	7,654,566,000
mel-kit30-input-rep2	150;150	97.16;97.72	91.55;92.92	42.21;42.21	19,728,808	5,918,642,400
mel-wt-CTCF-rep1	150;150	95.67;96.57	87.38;89.12	43.32;43.26	7,554,339	2,266,301,700
mel-wt-CTCF-rep2	150;150	97.33;96.98	91.88;90.51	43.75;43.78	8,259,438	2,477,831,400
mel-wt-input-rep1	150;150	96.88;97.43	90.66;91.85	42.20;42.18	16,415,110	4,924,533,000
mel-wt-input-rep2	150;150	96.98;97.21	90.93;91.17	42.62;42.62	10,575,403	3,172,620,900
sar11	150;150	97.50;98.04	92.59;94.09	45.91;45.90	14,461,156	4,338,346,800
sar3	150;150	97.06;97.71	91.35;93.05	45.97;45.90	5,949,851	1,784,955,300

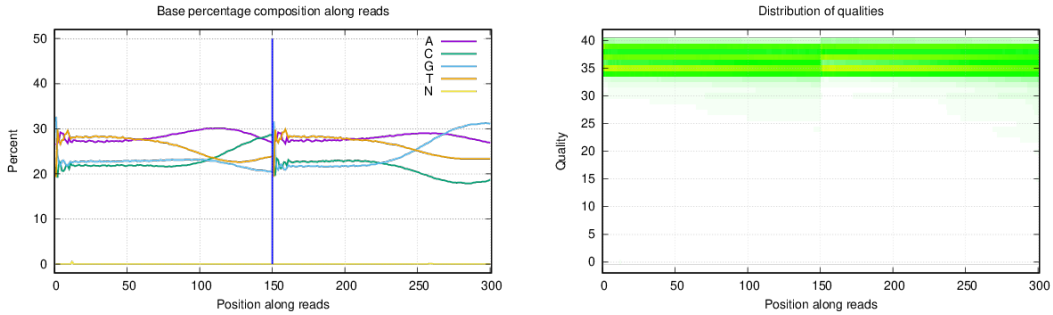
Table Format:

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

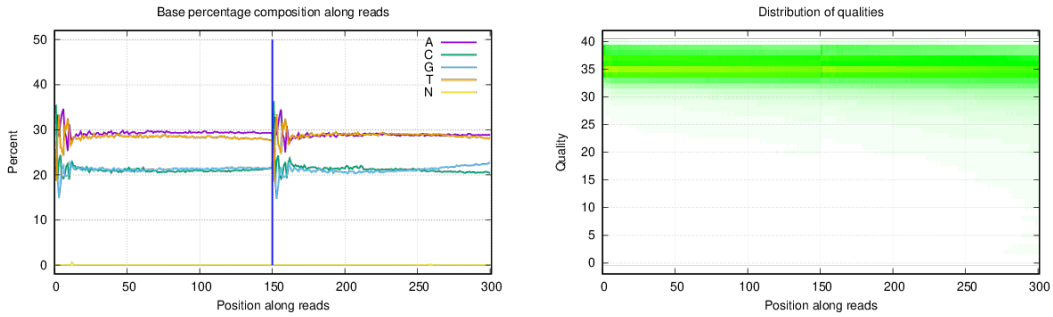
3 Data Quality Control

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.

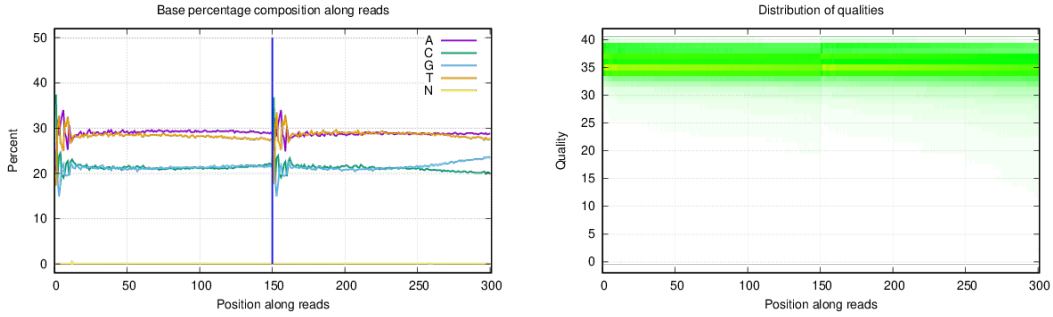
Quality control of sample 3



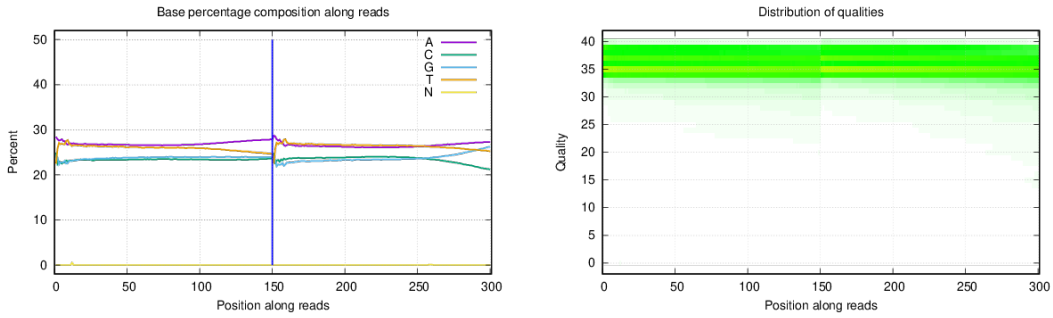
Quality control of sample Boc_e4



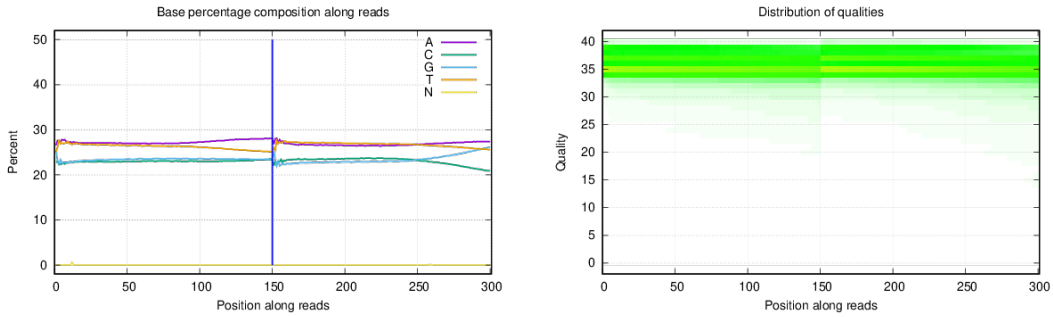
Quality control of sample Boc_K4



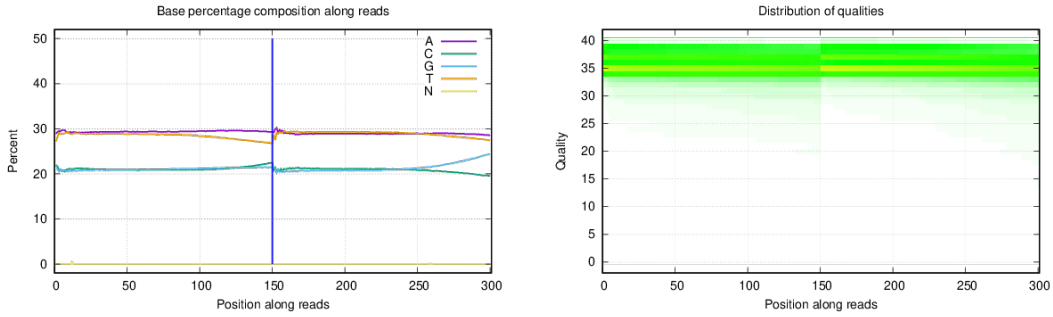
Quality control of sample mel-kit30-CTCF-rep1



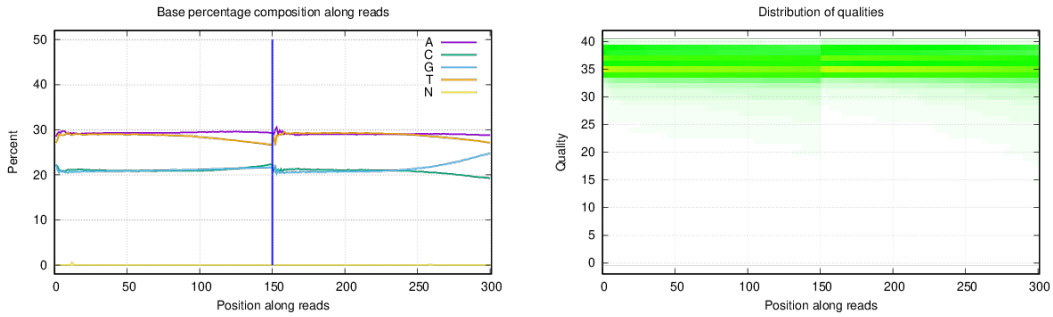
Quality control of sample mel-kit30-CTCF-rep2



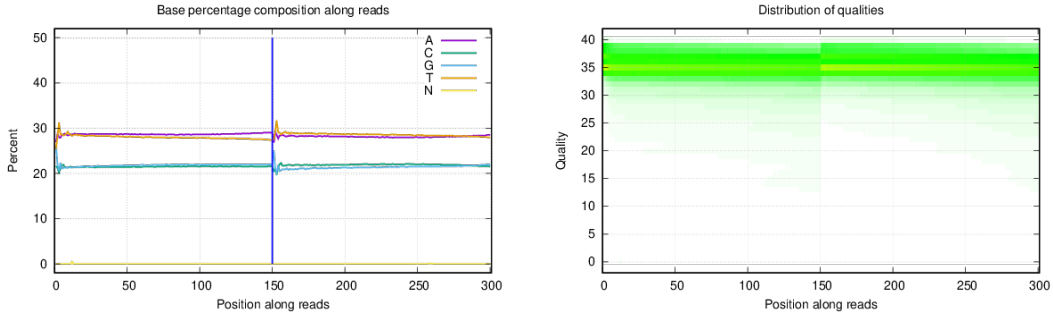
Quality control of sample mel-kit30-input-rep1



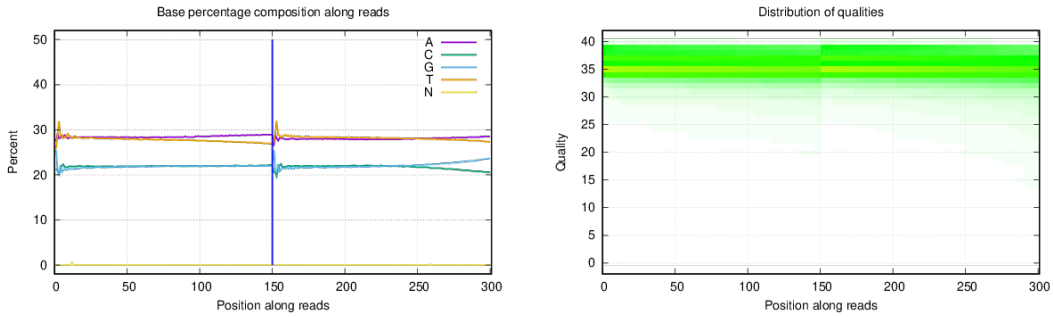
Quality control of sample mel-kit30-input-rep2



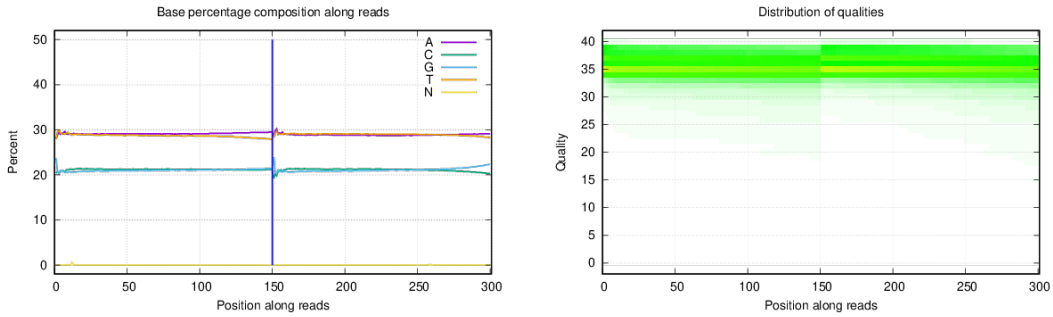
Quality control of sample mel-wt-CTCF-rep1



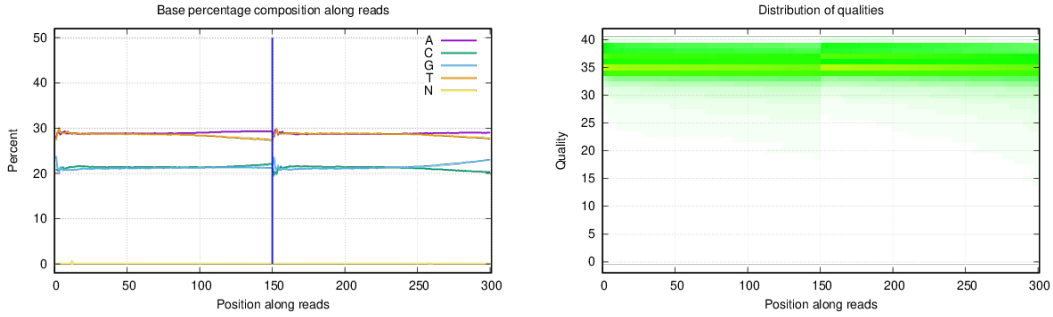
Quality control of sample mel-wt-CTCF-rep2



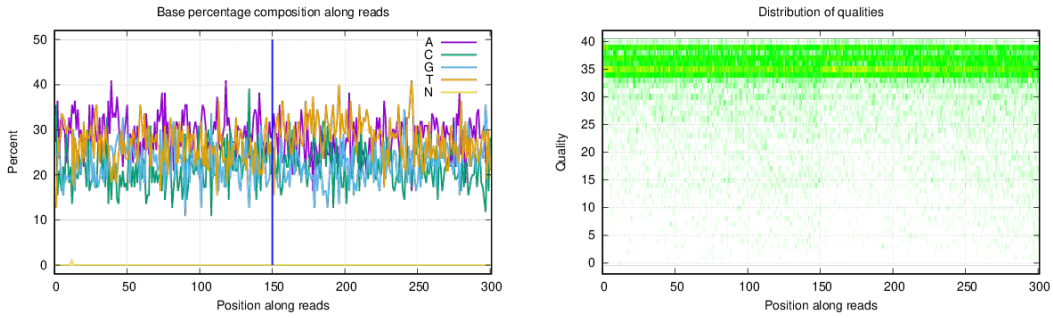
Quality control of sample mel-wt-input-rep1



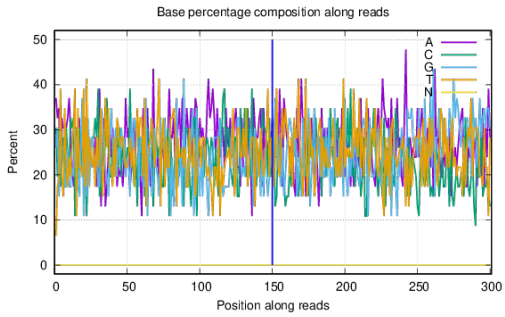
Quality control of sample mel-wt-input-rep2

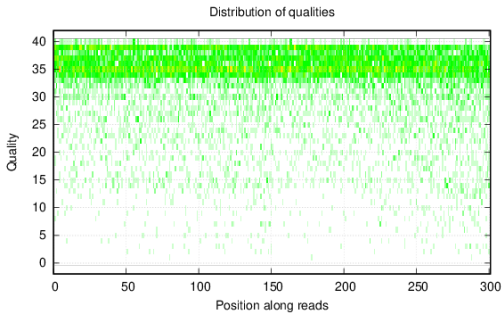


Quality control of sample N10

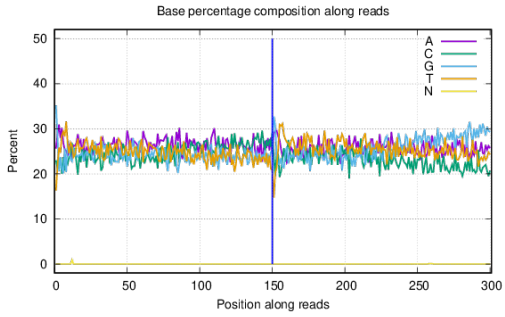


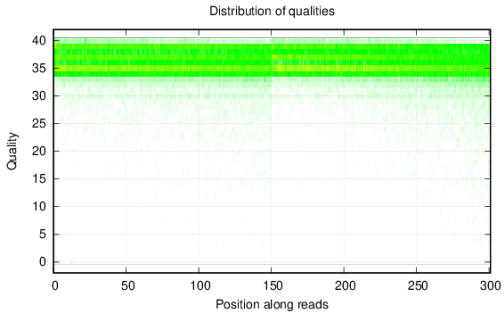
Quality control of sample N11



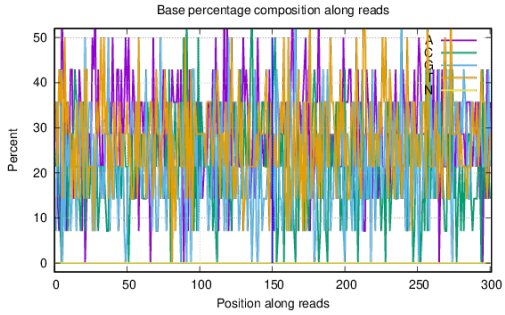


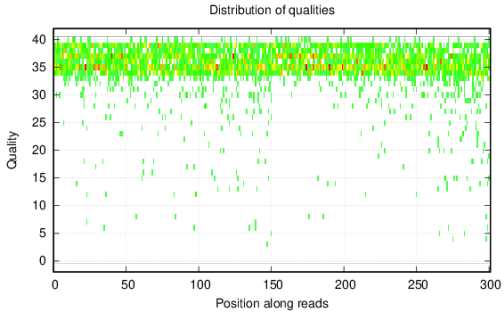
Quality control of sample N14



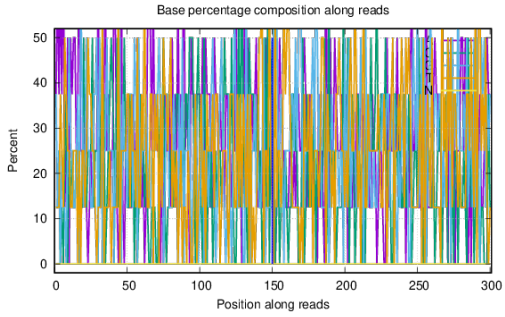


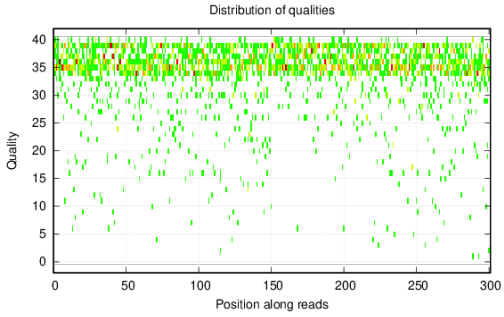
Quality control of sample N15



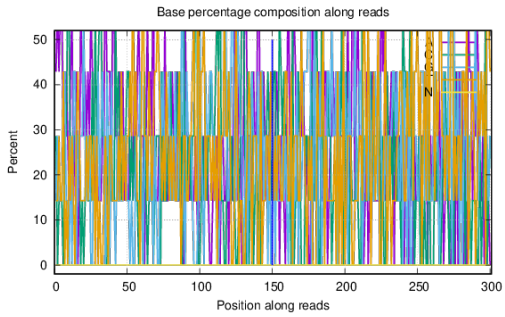


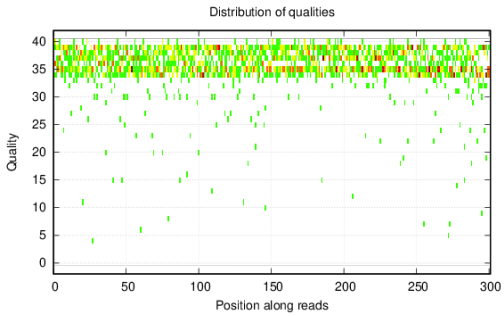
Quality control of sample N16



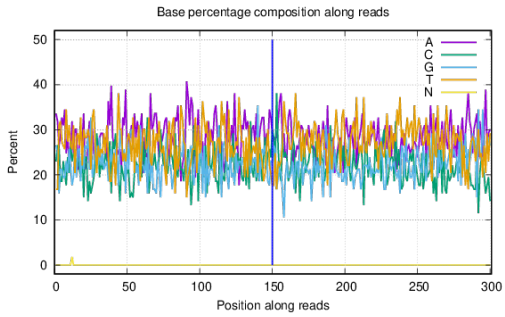


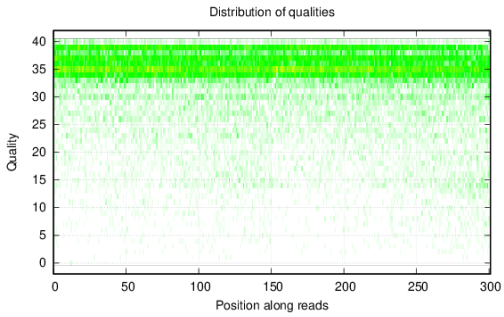
Quality control of sample N35



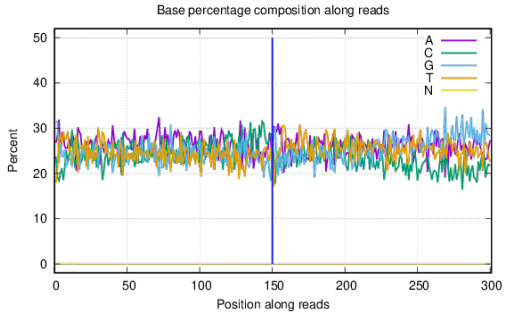


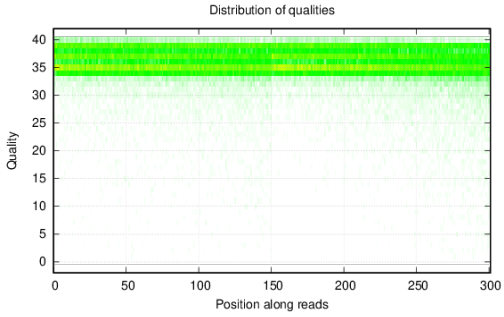
Quality control of sample N38



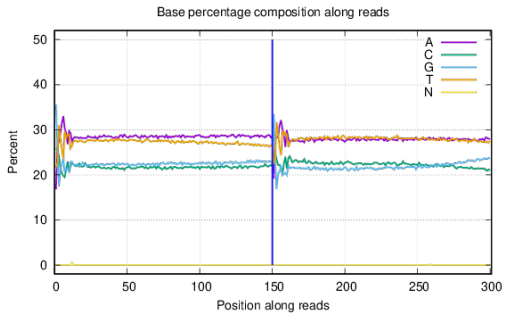


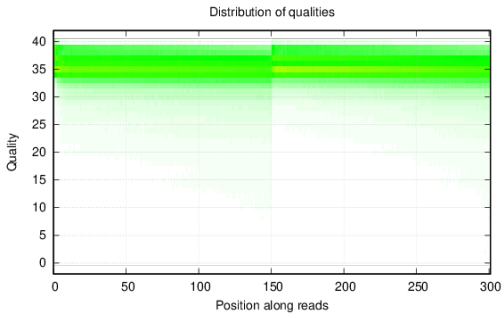
Quality control of sample N9



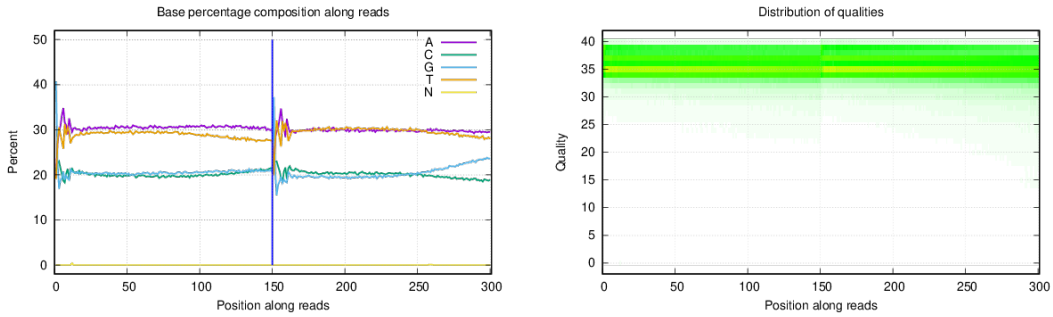


Quality control of sample Ng019

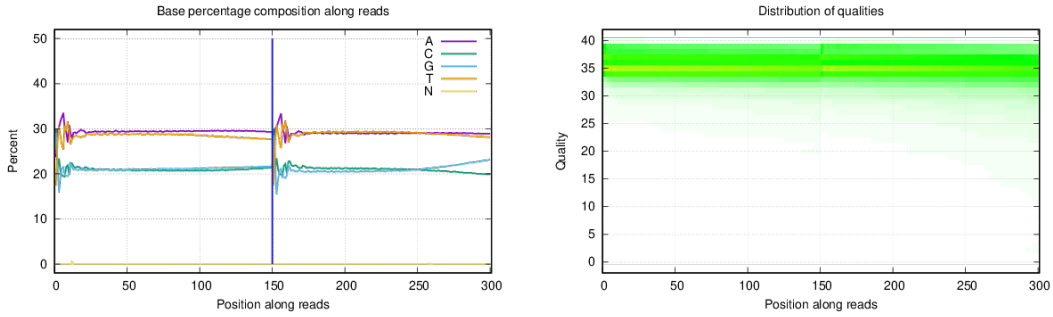




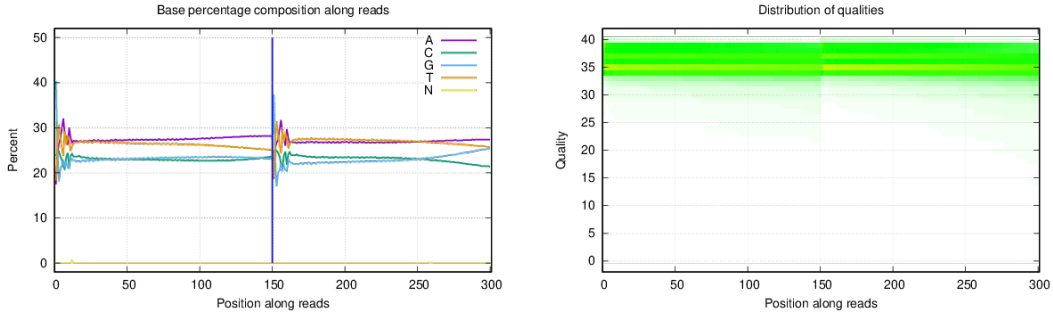
Quality control of sample Ng025



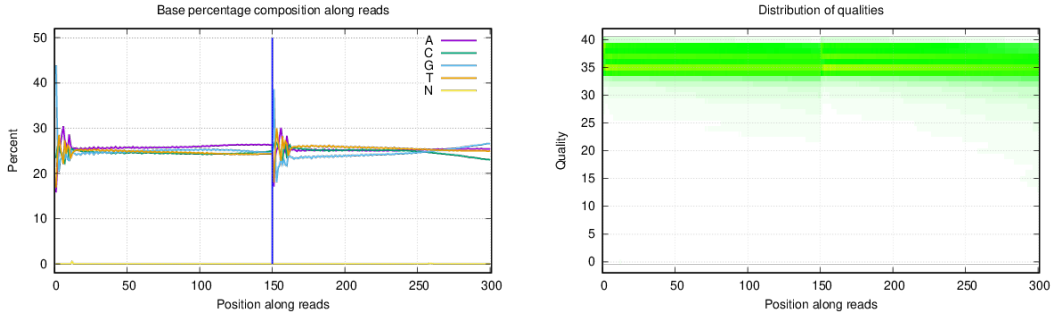
Quality control of sample P137



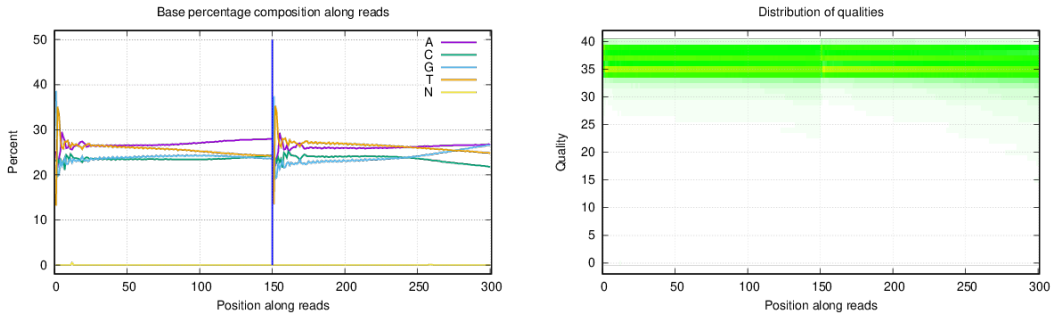
Quality control of sample P140



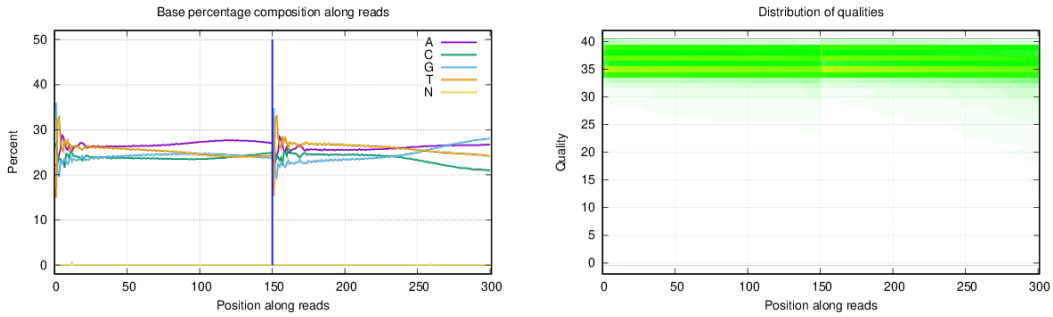
Quality control of sample P141



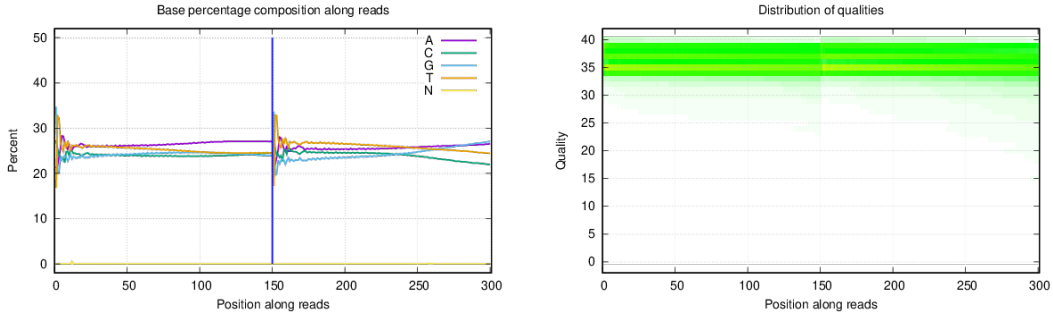
Quality control of sample P144



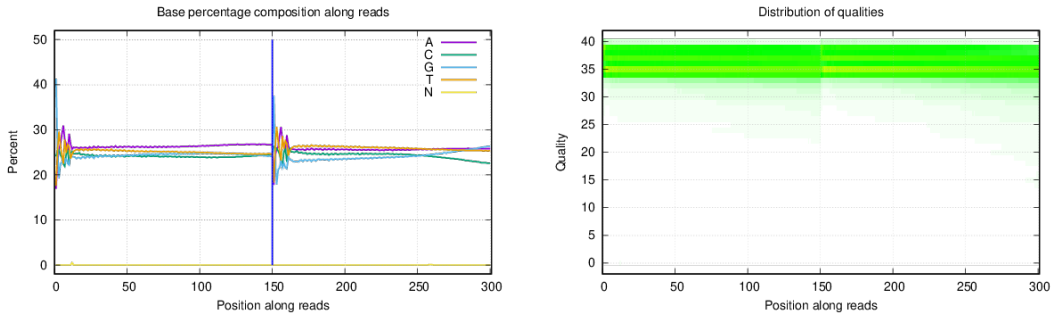
Quality control of sample P148



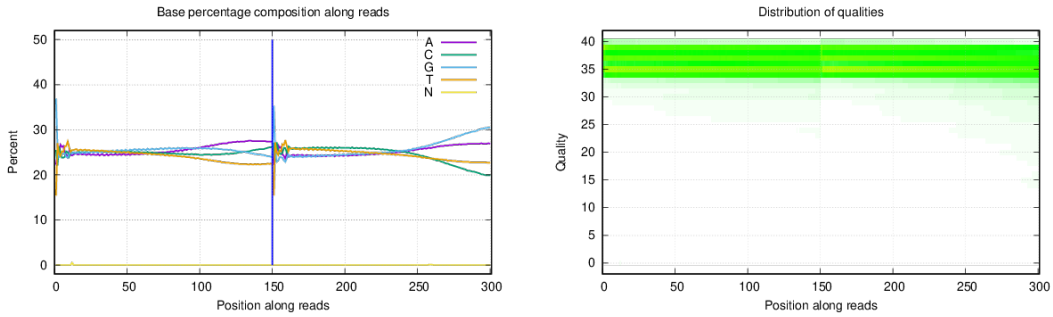
Quality control of sample P150



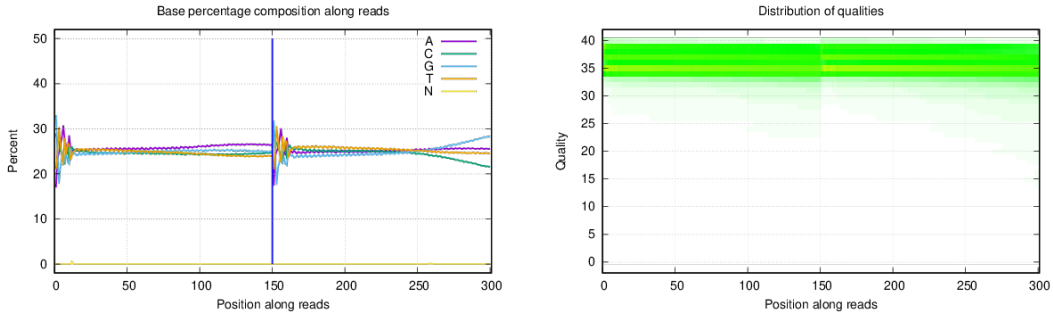
Quality control of sample P165



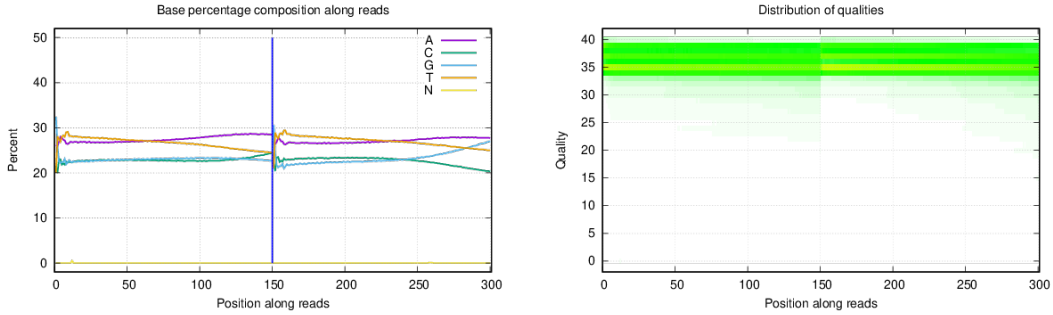
Quality control of sample P166



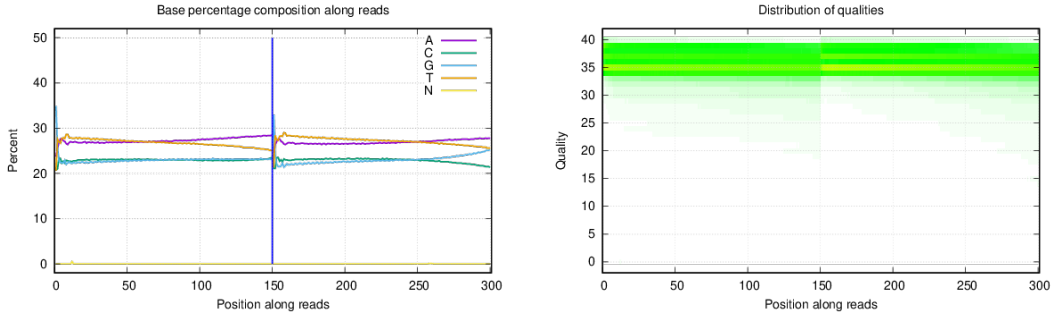
Quality control of sample P168



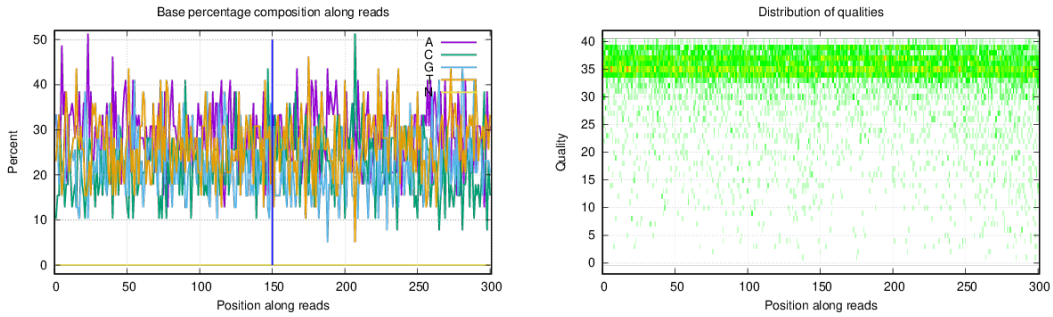
Quality control of sample sar11



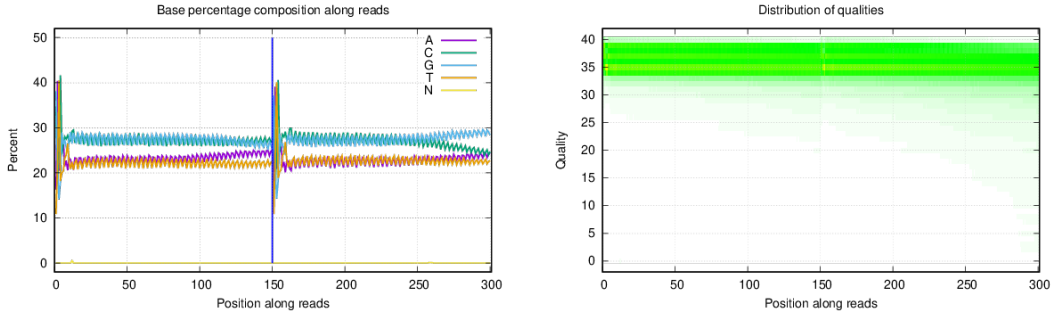
Quality control of sample sar3



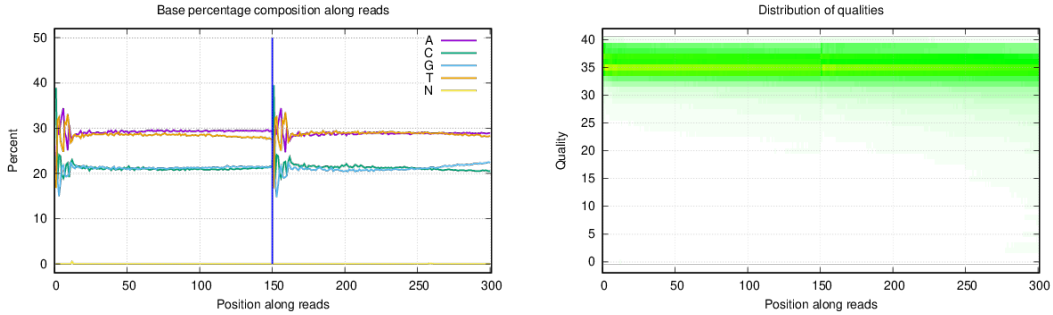
Quality control of sample Shur1_e



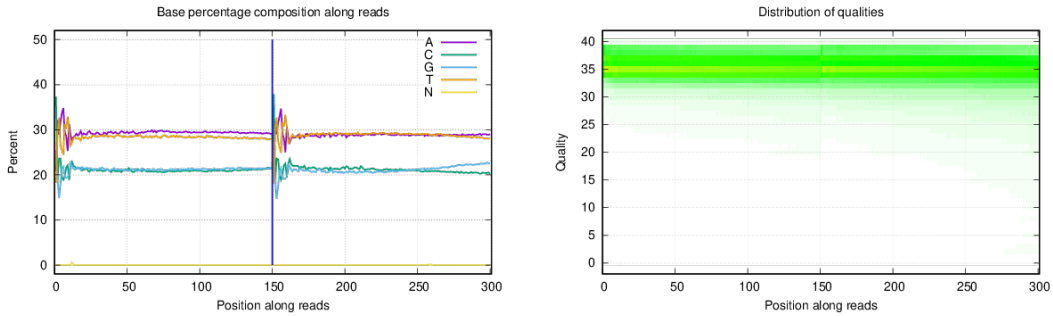
Quality control of sample Ton1_K



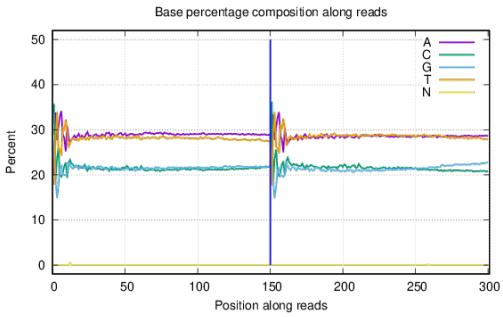
Quality control of sample Zap_e2

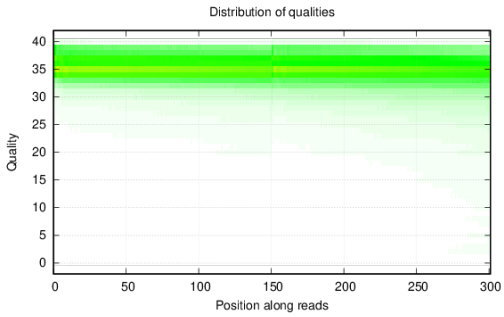


Quality control of sample Zap_e3



Quality control of sample Zap_K2





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	?